Stop Treating Supersaturated Designs like Other Screening Designs

Maria L. Weese\textsuperscript{1}, Jonathan W. Stallrich\textsuperscript{2}, Byran J. Smucker\textsuperscript{3}, and David J. Edwards\textsuperscript{4}

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

Abstract

The history of supersaturated designs is full of unrealized promise. Despite the vast amount of literature on supersaturated designs, there is a scant record of their use in practice. We argue this imbalance is due to the designs’ inability to meet practitioners’ analysis expectations and the existing literature’s lack of clearly stated expectations. To resolve this issue, we first summarize practitioner concerns and expectations of supersaturated designs gathered from a survey. Next we tie exploration and optimization, two different goals of experimentation, to tradeoffs between power and type 1 error. The best strategy, which integrates a design’s analysis with its construction, is shown to depend on the goal. 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 and so are recommended for more exploratory studies. Var(s+) designs (Weese et al. 2017), when paired with the Dantzig selector, are recommended as the initial experiment for optimization studies, as they can identify many active factors with a low type 1 error. The construction of both designs is less intuitive than traditional supersaturated designs and forces reflection on current best practices.

Keywords: Dantzig selector, GO-SSD, Orthogonality, Power, Sparsity, Var(s+)

1 Introduction

“I think it is perfectly natural and wise to do some supersaturated experiments.”–John Tukey
Screening experiments increase understanding of complex, expensive systems by efficiently identifying the few factors that most influence those systems. Supersaturated designs (SSDs) claim to effectively screen even when the number of runs is less than the number of factors under study. 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 (for instance, 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). This paper proposes a reassessment of the construction and analysis of SSDs, arguing that the best pairing depends on the ultimate purpose of screening. This perspective will help practitioners evaluate whether an SSD is appropriate, and to choose the best construction/analysis approach.

For small-run screening to be feasible, the experimenter 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 estimating the parameters from the collected data. If all the estimated parameters associated with a factor are negligible, that factor is declared inactive. Otherwise, it is declared active.

The most basic model used in screening experiments includes just the linear main effects:

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

where \( n \) is the number of experimental runs, \( \epsilon_i \sim N(0, \sigma^2) \) and independent, and \( \beta_j \) is the effect a factor has on the mean of \( Y \). The model can be written equivalently as \( Y = X\beta + \epsilon \), where \( X \) is the \( n \times (k+1) \) model matrix, \( \beta \) is a \( (k+1) \)-vector of model parameters, and \( Y \) and \( \epsilon \) are \( n \)-vectors, with \( \epsilon \sim N(0, \sigma^2 I) \). The \( j \)-th factor is then considered active if \( |\beta_j| > t \) for some threshold \( t \geq 0 \). Model (1) is simplistic because it assumes each factor’s effect on the response is independent of the levels of the other factors, i.e. there are no interaction effects. Even though the model is likely inaccurate, it is reasonable to believe this simple model 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). Such designs tend to require more runs than those targeting estimation of the main effect model, especially if $k$ is large.

The goal of a screening analysis is to correctly classify factors as active or inactive. For the classification to be done correctly, the decision rule must minimize the probability of a false positive (type 1 error) and maximize the probability of a true positive (power). Simultaneously achieving both becomes more difficult as $n$ decreases so a trade-off between the two becomes necessary. Screening is the first part of a sequential experimental procedure. If the ultimate goal of the entire procedure is understanding of the entire system, the experimenter may be more concerned with maximizing power and willing to accept a higher type 1 error rate. On the other hand, if the ultimate goal is to optimize the system as quickly as possible, the experimenter may be satisfied with identifying only a handful of the active factors with smaller type 1 error. We argue here that the best choice of SSD construction and analysis depends on which of these two trade-offs the experimenter is most comfortable with.

The fundamental principle of experimental design is that the data collection procedure strongly influences the properties of your parameter estimates. Knowledge of this relationship should be leveraged whenever possible to compare and rank designs; this is well understood for least-squares estimation of the main effect model. The main effect model is said to be estimable if the least-squares estimator is unique, which holds when the model matrix $X$ has full column rank. The estimator, $\hat{\beta}_{LS}$, is unbiased (when the main effect model is correct) and has covariance matrix $\sigma^2 (X^T X)^{-1}$. The type 1 error and power for hypothesis tests in least-squares regression are straightforward to calculate so an experimenter has reasonable expectations for what the analysis can accomplish. In particular, $\hat{\beta}_{LS}$ has good screening properties for a given design if $\sigma^2 (X^T X)^{-1}$ is “small.” The optimal $X^T X$ matrix comes from regular and nonregular fractional factorial designs having a model matrix, $X$, with orthogonal columns. By estimating the $\beta_j$’s with maximal precision, we maximize power while minimizing the type I error rate. For SSDs, an orthogonal design cannot exist and,
even worse, the main effect model is not least-squares estimable.

Penalized regression estimation (e.g. LASSO (Tibshirani 1996) and Dantzig selector (Candes and Tao 2007)) is an alternative to the standard least squares analysis. Such estimators penalize the magnitude of the $\beta_j$ estimates and shrink them toward 0, often producing unique estimates even when $X$ does not have full column rank. Indeed, some estimates are forced to be exactly 0. These estimates are biased but tend to have smaller variance than least-squares estimators. Hence, they may be better at performing screening (i.e., classifying factors as active or inactive) than least-squares estimators. For example, suppose we have five factors and the first three are active, with $\beta_1 = \beta_2 = \beta_3 = 5$ and $\beta_4 = \beta_5 = 0 = t$. A penalized estimator may give the estimates $\hat{\beta}_j = 1$ for $j = 1, 2, 3$ and $\hat{\beta}_j = 0$ for $j = 4, 5$. The screening results would be perfect, but the estimators would be poor. Unfortunately, it is more difficult to calculate type I error and power under such estimators so simulations are often used.

The lack of least-squares estimators with SSDs and no clear connections between $X$ and penalized estimators makes it difficult to justify a criterion for constructing a good SSD. Instead, designs have been compared by heuristic measures of orthogonality based on the off-diagonals of $X^TX = (s_{ij})$. For example, the $E(s^2)$-criterion forces $s_{0j} = 0$ and minimizes $E(s^2) = (k/2)^{-1}\sum_{1\leq i<j\leq k}s_{ij}^2$. Such criteria are based on least-squares intuition, even though other estimation methods such as stepwise selection or penalized regression must instead be used. While connections between the orthogonality-based design criteria and the estimation procedures used for SSDs are believable, they are not rigorous as they are for traditional screening designs. 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 classes of SSDs have been developed that challenge the current SSD near-orthogonality paradigm and treat design and analysis as a package. The SSDs constructed in Weese et al. (2017) are based on the $Var(s+)$ criterion that forces the average off-diagonals of $X^TX$ to be positive, but not too large, while minimizing the variability. Through an extensive simulation study, the authors found the $Var(s+)$-optimal designs had higher power and small type I error rates compared to traditional SSDs when effect directions were known and when analysis
was performed with the Dantzig selector. When effect directions were unknown, their designs had equivalent performance to traditional SSDs. The superior performance is then tied to the analysis method and additional assumptions about $\beta$. The group orthogonal SSDs (GO-SSDs) by Jones et al. (2019) create SSDs with a group factor structure along with extra fake factors that 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 $\beta$ and ideal partitions of active factors across the groups.

This article aims to convince practitioners of the potential value of certain pairings of SSD and analysis strategies, and equip them with the knowledge and tools to appropriately use them. In Section 2 we summarize survey results from screening practitioners about SSD awareness and implementation to better understand their concerns and/or misconceptions that have suppressed their use. In Section 3, we review traditional SSD construction and analysis recommendations, and recent SSD research of the $\text{Var}(s+)$-criterion and GO-SSDs. Section 4 presents an improved analysis method for the GO-SSDs of Jones et al. (2019) that maximizes power and is able to provide information regarding sparsity. Section 5 shows results from several simulation studies to more fully investigate $\text{Var}(s+)$-optimal designs paired with the Dantzig selector and the GO-SSDs paired with the new analysis method. Finally, in Section 6 we conclude the paper with a discussion of our results, how these results can help allay concerns raised in the survey, and practical advice for both the design and analysis of SSDs.

2 The Current State of Supersaturated Designs in Practice

To assess the use of SSDs in practice, we conducted an informal survey to reach the greater design of experiments (DOE) community, using the authors’ networks and social media. We received 63 responses to a twenty question survey gathering information on current and past experience with DOE and SSDs (the survey is included in the Supplementary Material). Note that these results cannot be reasonably taken as representative of any more general, identifiable population. Instead, it serves as a snapshot of the thinking of the population defined by the survey respondents. We report the results of this survey here and discuss them in light of current literature.
2.1 Practitioner Survey Results

Of the 63 survey respondents, 55% work in the field of Engineering, Manufacturing and Technology and 26% in Academia or Education; 80% have a Master’s degree or a PhD in statistics (see Figure 1). Figure 2 shows that most respondents use Response Surface Methodology, Fractional factorials/Plackett Burman designs and/or Split plot designs. Regression/ANOVA is unsurprisingly the most common analysis technique with reported use by 95% of the respondents. Interestingly, the next most common analysis is LASSO and/or other penalized regression techniques, followed closely with Bayesian methods and Gaussian Process Models. Only six respondents indicated SSDs as a design technique they use on a regular basis. Most respondents reported choosing a design technique with which they are comfortable and/or which is associated with a straightforward analysis method. Several self-reported reasons for design choice include the particular experimental situation and the efficiency of the design.

With regard to SSDs, 68% said they were familiar with them, and only 13 respondents reported using a SSD in practice. Table 1 provides examples of descriptions of the use of SSDs in practice with some successes and a failed experiment. Nineteen respondents voluntarily provided detailed explanations of concerns they have regarding SSDs; these are categorized into groups and shown in Figure 3. The most common concern with using a SSD is the sparsity assumptions and/or power (Figure 3). For example, one respondent’s concern regarding SSDs was stated as “Not identifying significant factors that are appearing insignificant”; this response was categorized both as “Spar-
“100+ factors 64 runs; failed experiment”

“Bayesian D-optimal design with many terms that were unable to be estimated by the design, but were able to be estimated after unimportant factors were removed.”

“Analytical Method Robustness testing. Successful.”

“Testing to characterize drill bit effectiveness as a function of many input parameters. Experiment was successful due to engineering expertise for interpretation.”

“Confounded effects that were not believed to be important or could be estimated in aggregate. Without knowing the truth, I think the design seemed successful.”

Table 1: Summary of SSD use in practice from survey responses.

Comments such as “Lack of [Degrees of Freedom]” and “You gamble. There is no guarantee...”
that you find the important few [factors]; you can find a few but not necessarily the important ones” gives the impression that some practitioners consider SSDs as a one-shot experiment. Practitioners also believe that sparsity is a necessary requirement for SSDs to be successful and that the analysis is uninformative. We will show this need not be true if one is willing to accept larger type 1 errors, and to focus their attention on factor classification rather than estimation. While this survey is by no means representative of the entire community of DOE users, the lack of application of SSDs in practice suggests that this viewpoint is not isolated to our survey sample.

2.2 Reflections from the Literature

The trepidation surrounding the use of SSDs is perhaps due to very conflicting recommendations of their use in the literature. For example, in the article that develops the $E(s^2)$ criterion, Booth and Cox (1962) 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 types of conflicting statements continue into the recent SSD literature. For example, Georgiou (2014) states

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) 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.

Perhaps the most prominent paper to date giving practical advice on SSD use is Marley and Woods (2010). They provide practical recommendations on SSD size and true model sparsity based on simulation power. In general they state that “the number of runs should be at least three times the number of active factors.” For instance, if an experimenter has allocated 12 runs for screening, the user should not expect to be able to find more than four important factors. They also assert that the level of saturation of a SSD—the number of factors relative to the number of runs, \(k/n\)—should be less than 2. Their statements were for traditionally constructed SSDs and analysis is defined as keeping type 1 errors as low as possible. We revisit these assertions in the Appendix.

3 Supersaturated Design Construction and Analysis

We first review SSD construction criteria developed under the standard assumptions of near-orthogonality. These designs are based on least-squares intuition, but were not constructed with a particular analysis method in mind. We then review the analysis methods that have been recommended for these designs. Lastly we consider two recent SSD approaches that closely link the design to the analysis.

3.1 Classical Supersaturated Designs and Separate Analysis Methods

Porting least-squares-based intuition to the SSD setting, early work in the area (Lin 1993; Wu 1993) focused on \(E(s^2)\) designs with small pairwise column correlations. Let \(S = X^TX = (s_{ij})\) where \(i, j = 0, 1, \ldots, k\). A design’s proximity to orthogonality is measured by a summary of the
The $E(s^2)$-criterion minimizes (2) over all balanced designs with $n$ runs and we call such a design $E(s^2)$-optimal. The unconditional $E(s^2)$-criterion, or $UE(s^2)$-criterion, studied by Jones and Majumdar (2014) and Weese et al. (2015), is

$$UE(s^2) = \frac{2}{k(k+1)} \sum_{0\leq i<j\leq k} s_{ij}^2,$$

which is similar to the $E(s^2)$-measure but drops the balance constraint and includes the $s_{0j}^2$ in its evaluation. Cheng et al. (2018) point out the large number of available $UE(s^2)$-optimal designs and propose a secondary discriminating criterion based on average projection properties of subset models of (1).

Marley and Woods (2010) compared $E(s^2)$-optimal and Bayesian $D$-optimal (Jones et al. 2008) designs and found little difference between them, though they suggested a preference for Bayesian $D$-optimal designs because they had slightly better power. Weese et al. (2015) compared a number of supersaturated design types using simulation, including $E(s^2)$-optimal, Bayesian $D$-optimal, model-robust (Jones et al. 2009; Smucker and Drew 2015), and $UE(s^2)$-optimal design (Jones and Majumdar 2014). No clear winner emerged among the criteria.

Table 2 summarizes analysis methods that have been used to analyze SSDs in the literature. 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, tend to have poor model selection properties, while regularization methods, such as the Dantzig selector, have higher power and lower type 1 error, see Table 3.

The Dantzig selector (Candes and Tao 2007; Phoa et al. 2009) is a regularization method which
constrains an \( \ell_1 \)-estimator as follows:

\[
\hat{\beta}_{DS} = \arg\min_{\hat{\beta}} ||\hat{\beta}||_1 \text{ subject to } ||X^T(y - X\hat{\beta})||_\infty \leq \delta ,
\]

(4)

where \( || \cdot ||_\infty \) denotes the largest element of the vector argument. Note that when \( \delta = 0 \) the resulting \( \hat{\beta}_{DS} \) will be an \( \ell_1 \)-sparse, least-squares estimator since it satisfies the normal equations \( X^TX\hat{\beta}_{DS} = X^Ty \). When analyzing a design using this procedure, the use of profile plots is recommended (see Section 6), but when using this method in simulation studies, a non-graphical approach must be taken. We follow Phoa et al. (2009) for automated model selection:

1. Center the response vector, \( y \) and center and scale the columns of \( X \) to have unit variance. Drop the intercept column from \( X \).

2. Solve (4) for many values of \( \delta \) up to \( \delta_0 = \max |x_j^Ty| \) where \( x_j \) is the \( j \)-th column of \( X \);

3. For each Dantzig solution, set all \( \hat{\beta}_j < \gamma \) to 0. Call this the reduced Dantzig solution;

4. For each reduced Dantzig solution, refit via least squares, using only the predictors retained from Step 3;

5. Compute BIC for each least squares fit, and choose the \( \delta \) which minimizes the BIC = 
\[
\ln(SSE/n) + \ln(n).
\]

For the chosen \( \delta \), choose as active effects those in the reduced Dantzig solution from Step 3.

Step 1 is important when \( X \) is unbalanced since the columns will be correlated with the intercept. The specification of \( \gamma \) is important in this automated procedure. In our simulation study (see Section 5), we consider two different values of \( \gamma \). An idealized case is \( \gamma = \sigma \), the true error variance, was used in Marley and Woods (2010) and Weese et al. (2017). A more realistic, data-driven approach suggested by Phoa et al. (2009) is to choose \( \gamma = 0.1 \times \max |\hat{\beta}_j| \) where \( \hat{\beta}_j \) is estimated 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 \( \gamma \) is smaller than \( \sigma \). In Section 5, we also consider a third set of Dantzig estimates: the solution at \( \delta = 0 \), which we refer to as a “sparse LS” solution.
LS-based Methods | Regularization Methods | Other Methods
---|---|---

Table 2: Literature review of analysis methods used to analyze SSDs.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forward Selection</th>
<th>Dantzig</th>
<th>Bayesian</th>
<th>LASSO</th>
<th>SCAD</th>
<th>SA</th>
<th>PLSVS</th>
<th>SVDPR</th>
<th>MA</th>
<th>SRRS</th>
<th>SIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoa et al. (2009)</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marley and Woods (2010)</td>
<td>x</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen et al. (2013)</td>
<td>1</td>
<td>2 (CGS)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draguljić et al. (2014)</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weese et al. (2015)</td>
<td>x</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weese et al. (2017)</td>
<td>x</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drosou and Koukouvinos (2018)</td>
<td>2</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jones et al. (2019)</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: A comparison of methods: “x” indicates the method was included in the study; “1” indicates best performer; “2” indicates the method out performed “1” under certain conditions.

Based on this review of the current literature, one might conclude that the Dantzig selector should be used to analyze SSDs and among traditional SSDs, such as $E(s^2)$-optimal and Bayesian D-optimal, there is no clearly superior design type. These recommendations were largely developed within an existing paradigm that assumes that less correlation between design columns, as measured by the $s_{ij}^2$, is desirable, that the analysis and design can be developed separately, and smaller type I error is best. In the next section, we will reexamine existing wisdom by studying two recent SSD methodologies that challenge these standard assumptions.
3.2 Pairing SSDs and Analysis

3.2.1 \( \text{Var}(s+) \) Designs and the Dantzig Selector

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

\[
\text{Var}(s+) = UE(s^2) - UE(s)^2 \quad \text{s.t.} \quad \frac{UE^*(s^2)}{UE(s^2)} > c \text{ and } UE(s) > 0,
\]

where \( UE^*(s^2) \) is the optimal value, \( UE(s) \) is the average of the \( s_{ij} \), and \( c \) is a specified efficiency that determines how near to \( UE(s^2) \)-optimal the design is required to be. This criterion allows the off-diagonal elements to be, on average, more positive than those in the \( UE(s^2) \)-optimal design, but with less variability.

Weese et al. (2017) found that \( \text{Var}(s+) \) designs are superior to \( UE(s^2) \)-optimal and Bayesian \( D \)-optimal designs when effect directions were correctly specified in advance. That is, they were more powerful to detect active factors, without elevated type I error rates. Furthermore, even when the effect directions were misspecified, the \( \text{Var}(s+) \) designs fared no worse than the classical designs. The superiority of these designs was most clear when the Dantzig selector was employed for analysis. \( \text{Var}(s+) \) designs are not yet well-understood, but their effectiveness appears to be connected to the average positive off-diagonals so that when the design is used to estimate active main effects (certain to be positive if effect directions are assumed known), the estimates are exaggerated making them easier to detect (Weese et al. 2017). The Dantzig selector appears to further aid in this exaggeration of estimates. Research is currently underway that seeks to more fully understand the efficacy of pairing \( \text{Var}(s+) \) SSDs with the Dantzig selector.

3.2.2 GO-SSDs

Jones et al. (2019) constructed SSDs that can estimate \( \sigma^2 \) by introducing a group of fake factor columns that are orthogonal to the true factor columns. The true factor columns are further partitioned into equally-sized groups that are mutually orthogonal, and so the designs are called group orthogonal SSDs (GO-SSDs). Jones et al. (2019) recommend 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) generate designs through a Kronecker product of a Hadamard matrix, $H_m$
and a small generating SSD, $T_{w \times p}$. 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, $r < p$. The first
group includes the intercept column, and $p - 1$ fake factors. The remaining columns are assigned
to the actual factors. Let $X_g$ denote the $g$-th group of actual factor columns, $g = 1, \ldots, p - 1$.

Jones et al. (2019) provide recommendations when constructing a GO-SSD. First, $w > p/2$ to
prevent complete confounding between two columns. They also recommend that $m$ be as large
as possible, producing more factor groups with fewer factors in each. Finally, 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.

Figure 4 shows the pairwise column correlations of SSDs constructed by the $UE(s^2)$, $Var(s+)$
and GO-SSD criteria for $n = 12$ and $k = 24$. Notice the $Var(s+)$ criterion tends to produce columns
with mostly positive correlation whereas the $UE(s^2)$-optimal design has fairly equal amounts of
positive and negative correlation between the columns. The bottom plot in Figure 4 shows the
group orthogonal structure of the GO-SSD. The GO-SSD has $m = 4$, $w = 3$, and $p = 6$ which
violates the recommendation $w > p/2$. leading to complete confounding between columns within a
group.

The distinguishing feature of the GO-SSD approach is its pre-variable selection estimate of $\sigma^2$
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. Group screening starts by
calculating the mean sum-of-squares for the $g$-th factor group, denoted $MS_g$ and has $r$ degrees of
freedom. The test statistic $F_g = MS_g/MSE$ follows an $F$-distribution with $r$ and $r - 1$ numerator
Figure 4: Correlation color plots of three $n = 12$, $k = 24$ SSDs: (a) $UE(s^2)$-optimal; (b) $Var(s+)$; and (c) GO-SSD. Blue represents positive correlation; red represents negative correlation.

and denominator degrees-of-freedom, respectively, under the null hypothesis $H_0 : \mathbf{X}_g \beta_g = 0$, where $\beta_g$ denotes the coefficients for the factors in $\mathbf{X}_g$. Hence the null hypothesis is satisfied whenever $\beta_g$ is in the null space of $\mathbf{X}_g$, which includes more vectors than just $\beta_g = 0$, the desired null hypothesis. If the dimension of the null space is low, we can accept this as an approximate test for $\beta_g = 0$.

Jones et al. (2019) recommend group screening be done with a backwards elimination procedure, starting with the group that has the smallest $MS_g$. If this group is declared inactive, then pool the corresponding $MS_g$ with $MSE$ to improve estimation of $\sigma^2$. Denote this potential estimate by $MSE^*$. To maximize power, they only recommend replacing $MSE$ with $MSE^*$ when

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

where $F(\cdot)$ denotes the critical value under significance level $\alpha$, $r$ numerator degrees-of-freedom, and $df_d$ degrees-of-freedom for the current $MSE$. When a group is declared inactive, the group with the next smallest $MS_g$ is tested. The process continues until all remaining groups are deemed active.

For each active group, Jones et al. (2019) perform factor screening with a modified forward or sequential all-subsets procedure, starting with the active group that has the smallest $MS_g$. Let $g_1$ and $g_2$ denote partitioning subgroups of $g$ where the factors in $g_1$ comprise the current model of interest. The goal is to identify the sparsest $g_1$ that does not exhibit lack of fit, measured by the
ratio of a model’s mean residual sum-of-squares, denoted $MSE_{g_1}$, to $MSE$.

First, all $p$ one-factor models are considered so every $MSE_{g_1}$ has $r - 1$ degrees-of-freedom and $F_{g_1} = MSE_{g_1}/MSE$ is compared to the critical value $F(1 - \alpha, r - 1, df_d)$. Any $g_1$ model with $F_{g_1} < F(1 - \alpha, r - 1, df_d)$ 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 $F_{g_1}$’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 this eventually happens for all models of size $r - 1$, then all $p$ factors are included in the model because all models with $r$ factors would have $MSE_{g_1} = 0$. If the best model for a group has less than $r$ factors, then $MSE_{g_1}$ may be pooled with $MSE$ following the previous pooling rule. This process is performed on the active group having the second smallest $MS_g$ and continues until all active groups have been analyzed. We propose two modifications to this analysis strategy to improve its power.

4 Augmented Two-Stage Analysis for GO-SSDs

There are potential issues with the two-stage analysis recommended by Jones et al. (2019), particularly for within-group factor testing. In this section we propose an augmented approach to maximize power. Let $X_{g_1}$ be the submatrix of $X_g$ corresponding to the subgroup $g_1$ and similarly let $\beta_{g_1}$ be the corresponding elements of $\beta_g$. Let $P_g$ be the orthogonal projector onto the column space of $X_g$. Then we have

$$MSE_{g_1} = \frac{y^T(P_g - P_{g_1})y}{r - r_1},$$

where $r_1$ is the number of factors in $g_1$. For $F_{g_1}$ to be a valid test statistic for lack of fit, it must follow a central $F$-distribution if and only if $\beta_{g_2} = 0$. Just like with the group screening test, which has null hypothesis $X_g\beta_g = 0$, this turns out not to be true. To see this, consider the expected value of $MSE_{g_1}$:

$$E(MSE_{g_1}) = \sigma^2 + \frac{\beta_{g_2}^T X_{g_2}^T (P_g - P_{g_1}) X_{g_2} \beta_{g_2}}{r - r_1}.$$
The correct null hypothesis is then \((P_g - P_{g1})X_{g2}\beta_{g2} = 0\). As \(r_1\) increases, so does the null space dimension for \((P_g - P_{g1})X_{g2}\) so there are more \(\beta_{g2} \neq 0\) that satisfy the null hypothesis.

The above result gives us cause for concern about the power of the factor screening procedure. First, the recommended approach ignores the possibility that many models of size \(r_1\) could have \(F_{g1} < F(1 - \alpha, r - r_1, df_d)\). Ignoring this possibility and simply choosing the model with the smallest \(MSE_{g1}\) could be a poor strategy. Even if only one model failed to reject, there is still 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. As the dimension of the null space of \((P_g - P_{g1})X_{g2}\) increases, so does the probability that a \(\beta_{g2} \neq 0\) is in or near that space, meaning the test is a poor approximation.

We propose a modification to the Jones et al. (2019) method to address the possibility that one or more columns of \(X_{g2}\) lie in the column space of \(X_{g1}\). If true, the corresponding column(s) in \((P_g - P_{g1})X_{g2}\) would be 0, automatically eliminating those factors’ corresponding coefficients in \(\beta_{g2}\) from contributing to lack of fit. There is a simple fix for this: after choosing the best model for a given size, add to the model all factors in \(g_2\) whose columns lie in the column space of the best model’s \(X_{g1}\). This modification is necessary if analyzing designs with completely confounded columns, although this was not recommended in Jones et al. (2019). Otherwise, the original procedure would randomly select one factor from the set of correlated columns to include in the model. Henceforth, we incorporate this modification as part of the method by Jones et al. (2019).

The above issues were not discussed in Jones et al. (2019) nor were they observed in their simulation study; recall that they allowed at most two active factors in a group. We now report results from a short simulation study that explored the factor power and type 1 error of the Jones et al. (2019) analysis method for groups with more than two active factors. We examined the most powerful scenario for within-group factor screening by only allowing one group’s factors to be active, allowing the remaining groups’ sums-of-squares to be pooled with the \(MSE\). The number of active factors in the chosen group started with 2 (the maximum setting for Jones et al. (2019)) and increased to the group size. The active factors in the group were randomly assigned to the group’s columns. The magnitude of the significant effects were randomly generated from \(Exp(1) + SN\),
where $SN$, meaning signal-to-noise ratio, was set to either 1 or 3, with the error variance being $\sigma^2 = 1$. These coefficients either remained positive or their signs were randomly set to +1 or −1 with probability 0.5, corresponding to known and unknown effect directions, respectively.

Figure 5 shows factor power and type 1 error rates based on the number of active factors in a group for three GO-SSDs, denoted by $(n, k)$: (12, 24), (24, 28), and (40, 56). The total number of truly active factors in the chosen group are shown on the x-axis. Group power is at or near 100% for all scenarios. Factor power is near 100% for 2 active factors but dips below 80% for more active factors. Except with extremely sparse systems, it is reasonable to expect a group to have more than two active factors.

![Figure 5: Power and Type 1 error for different group sizes where only one group is assigned as active.](image)

To overcome the deficiencies with the lack of fit test, we propose the following. First, for a given model size, we collect all models whose $F_{g_1} < F(1-\alpha, r - r_1, df_d)$ and then classify all the factors in these models as active. If $F_{g_1} > F(1-\alpha, r - r_1, df_d)$ for all $g_1$ models, increase the model size by
one and repeat. However, instead of increasing the model size up to \( r - 1 \), 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 \( F_{g_i} > F(1 - \alpha, r - r_1, df_d) \), we designate all factors in the group as active.

We expect this augmented, two-stage strategy to have similar power and type 1 error as seen in the above simulation study for sparsely active groups. For groups with more than two active factors, the strategy significantly reduces the probability of missing an active factor, improving the power although at a cost of higher type 1 error. As previously discussed, this may be a reasonable trade-off for some screening experiments. Moreover, this strategy’s classification of many factors as active is indicative of low sparsity, which is itself informative. In Section 5.3 we explore the power and type 1 error trade-off for the strategy of Jones et al. (2019), with the modification we detailed earlier in this section, and our modification (denoted as MaxPower). We also compare them to existing methods in Section 5.4.

5 Empirical Studies to Investigate Promising SSD/Analysis Combinations

In this section, we study the new SSD approaches described in Section 3.2 that challenge the standard assumptions, and compare them to a classical SSD approach. First, \( Var(s^+) \) designs are analyzed with the three Dantzig selector strategies described in section 3.1. Then we compare the two analysis strategies described in Section 4 for GO-SSDs. Finally, we compare these two designs with \( UE(s^2) \) designs analyzed with the Dantzig selector. We measure the power and type I error under various conditions of complexity, sparsity, and knowledge of the true effect directions.

5.1 Simulation Protocol

Five different design sizes were considered, denoted by \((n, k)\): \((8, 12)\), \((12, 24)\), \((16, 28)\), \((24, 28)\) and \((40, 56)\). Considering the effect sizes, sparsity levels, and \((n, k)\) combinations, we attempted to consider both settings that are relatively easy (large effect sizes; high sparsity; favorable \( n/k \) ratios)
as well as scenarios that were so difficult that they in some sense would “break” the methods. For instance, three of the \((n, k)\) combinations, \((8, 12), (12, 24)\) and \((16, 28)\), are not recommended GO-SSD sizes due to full confounding of factor columns within each group.

Model sparsity, i.e., the number of active factors, was varied according to \(0.25n, 0.5n,\) and \(0.75n\). (Note that “high sparsity” refers to the setting with the fewest active effects, whereas “low sparsity” is when the number of active effects is larger.) We generated each active effect \(\beta_a \sim \text{Exp}(1) + SN\) where \(SN\), the signal to noise ratio, was set to either 1 or 3. Inactive factors had coefficients set to 0.

Because both the \(\text{Var}(s+)\) and the GO-SSDs benefit from correctly prespecifying effect directions, we considered effect directions known (all coefficients are positive) and all effect directions unknown (all coefficients are assigned a positive coefficient with 0.5 probability). In all, there were 12 scenarios simulated for each of five design sizes and each of the three design/analysis combinations.

A total of 5000 iterations were performed for each simulation scenario, design size, and design/analysis combination. For each iteration, the \(a\) active factors were randomly assigned to the factor columns of \(X\) and a new set of \(\beta_a\) were generated. If sign directions were unknown, we also randomly generated positive and negative signs for each iteration. The responses were then generated according to model (1) where \(\epsilon \sim N(0, 1)\).

### 5.2 \(\text{Var}(s+)\) Simulation Results

Figures 6 and 7 show the power and type I error for the \(\text{Var}(s+)\) SSDs for the three versions of the Dantzig selector \((\gamma = 1, \gamma = 0.1 \times \max|\hat{\beta}_i|, \text{ and } \delta = 0;\) see Section 3.2). In general as the true models became less sparse, the power declined, but for the \((24,28)\) and \((40,56)\) design sizes this decline was less pronounced. As expected, there is a noticeable increase in power when the effect directions were known compared to when they were unknown. The sparse LS version of the Dantzig selector has the highest power but also has the largest type I error rate. The data-driven method of Phoa et al. (2009) has slightly smaller power, but with smaller type I errors than sparse LS. Using \(\gamma = 1\) gives the lowest type I error and power but has low power for \(a = 0.25n\) and \(0.75n\).
Figure 6: Power vs. Sparsity level for the $\text{Var}(s+)$ designs by sign specification (known, unknown) and model complexity (SN) for the different values of $\gamma$ and the sparse LS solution.

5.3 GO-SSD Simulation Results

Figures 8 and 9 compare the power and type I error rate, respectively, for GO-SSDs using MaxPower and the analysis of Jones et al. (2019) with the slight modification described in section 4. Figure 8 illustrates the significant improvement in power that MaxPower has over Jones et al. (2019). Figure 9 shows that the type I error for the MaxPower method is higher than the method of Jones et al. (2019), though for low sparsity cases the (12, 24), (24, 28) and (40, 56) SSDs have a reasonable type I error rate at approximately 25%. In both the (8, 12) and (16, 28) cases the power (see Figure 8) is nearly 1, but type 1 error (Figure 9) hovers around 50%, indicating that often the entire group is being selected. This is because these design sizes have groups with two completely confounded factors, which is not recommended by Jones et al. (2019).
Figure 7: Type I Error vs. Sparsity for $\text{Var}(s^+)$ designs by sign specification (known, unknown) and model complexity (SN) for the different values of $\gamma$ and the sparse LS solution.

5.4 Comparing $\text{Var}(s^+)$, GO-SSD, and $\text{UE}(s^2)$

Figures 10 and 11 compare $\text{Var}(s^+)$-optimal and $\text{UE}(s^2)$-optimal designs, both analyzed with the Dantzig selector using the data-driven threshold, and the GO-SSD/MaxPower approach. GO-SSD/MaxPower has the most consistently high power, especially in the non-sparse scenarios. $\text{Var}(s^+)$ performs better than $\text{UE}(s^2)$ when the effect directions are known except for $(24, 28)$ where the level of saturation is small. The power for the $\text{Var}(s^+)$ and $\text{UE}(s^2)$ designs were generally comparable to GO-SSD for high sparsity settings, but with lower type 1 error. For $(24, 28)$ and $(40, 56)$, the designs have similar power even for medium sparsity.

From the simulation results, we can make the following statements regarding the performance of the three design/analysis combinations studied in this section. $\text{Var}(s^+)/\text{Dantzig}$ with the data-driven threshold produce results that dominate the $\text{UE}(s^2)/\text{Dantzig}$ designs, in terms of power and type I error, when the effect directions are known. When the effect directions are unknown, the per-
Figure 8: Power vs. Sparsity level for GO-SSDs by size, sign specification (known, unknown) and model complexity (SN) for each analysis method.

Performance of these two design/analysis combinations are indistinguishable. Thus it is advantageous to use a $Var(s^+)$ SSD over a $UE(s^2)$-optimal SSD and attempt to specify the effect directions ahead of time. 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 approach introduced in Jones et al. (2019), though it does show increased type I error over the approach of Jones et al. (2019). However, the rate is reasonable for situations with higher sparsity, where SSDs are recommended for use, and the number of factors declared active provides a rough index of the difficulty of the experimental scenario.

We consider the GO-SSD/MaxPower approach to be the conservative choice compared to $Var(s^+)/$Dantzig with a threshold. In high sparsity situations (e.g. $a = 0.25n$) the methods will give you relatively equivalent power, but the GO-SSD/MaxPower will choose larger models. However, those larger models maintain high power for GO-SSD/MaxPower across medium and low...
levels of sparsity, whereas $Var(s+)/$Dantzig does not.

6 Discussion and Recommendations

We argue that a supersaturated screening design should effectively classify factors as active and inactive, and ultimately provide information to guide follow-up experiments. Supersaturated experiments will best achieve these goals if the design and analysis method are synergistic, and such SSDs will address many practitioner concerns (Section 2.1). In particular, such approaches will increase confidence that the design is either (1) giving high-quality information about the factors, or (2) producing a signal indicating that more extensive experimentation is required. Either way, an SSD is a reasonable first experiment when there are many factors that have not previously been screened.
In this article, we have studied two recent approaches to SSDs, both of which pair a design construction method with a particular analysis method. In both cases, the designs stray from the traditional near-orthogonal-based paradigm that has dominated thinking in supersaturated experiments since they were introduced into the literature. Traditional screening experiments, such as strength 2 or 3 orthogonal arrays, have structure that gives maximum advantage to least squares analysis. Similarly, SSDs should provide a structure that gives maximum advantage to factor identification for some analysis method. It is clear from the simulation results presented in Section 5 that SSDs which exploit the structure of the design matrix in the analysis outperform traditional SSDs in terms of factor classification and/or informing future experimentation.

$Var(s+)$ designs introduce a fairly small amount of extra correlation among the factor columns, but in a way that exaggerates the estimates of important effects while reasonably controlling the Type I error (Weese et al. 2017). For the simulations we’ve explored, the $Var(s+)$ designs, analyzed using the Dantzig selector, dominate the unbalanced $E(s^2)$-optimal design. That is, when the effect

![Figure 10: 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).](image-url)
Figure 11: Type I 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).

directions are known, the $Var(s^+)$ designs have higher power and lower Type I error rates. When the effect directions are unknown, the $Var(s^+)$ and $UE(s^2)$ have very similar performance. So, in terms of classification, the $Var(s^+)$ design is preferred to the classical design.

We have also explored and improved the GO-SSD design and analysis approach of Jones et al. (2019). These designs have a group orthogonal structure that facilitate a model-independent estimate of the error variance, which can be used to perform both group and factor screening. The improved analysis method presented in Section 3.1 is a principled way to achieve high power while screening inactive factors out to the extent allowed by the effect sizes and sparsity. The results in Sections 5.3 and 5.4 demonstrate that this approach will reliably detect active effects, even under challenging sparsity and effect size conditions, though in return, the type I error eventually becomes large. On the other hand, the type I error indicates the complexity of the experimental setting. Either there is reasonable sparsity, in which case the type I 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 active most of the factors in the design.

Neither the \( Var(s+) \) nor the GO-SSD approaches 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 approach is preferred. It is relatively conservative in its screening, and provides information regarding the certainty or ambiguity of each identified factor. The designs are easy to construct and are analyzed using a method reminiscent of ANOVA (see Figure 2, which suggests that regression/ANOVA methods are the preferred analysis methods of the survey respondents). If possible, active factors should be spread across groups to maximize factor power. One disadvantage of the GO-SSD 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. Practitioners interested in this method can access these designs and analysis via JMP software (SAS Institute, Inc. 2019).

Alternatively, if the goal is to correctly classify both active and inactive effects, then experimenters should use the \( Var(s+) \) design. Any available domain knowledge should be used to specify effect directions, and the designs should be analyzed with the Dantzig selector, centered and scaled as described in Section 3.2.1. This strategy will screen fairly aggressively, producing relatively high power and low type I error rate in settings with reasonably high sparsity. (We provide results in Appendix 6 which study guidelines on required 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 I error increases somewhat. Information on the ambiguity of the analysis is gained by viewing the profile plot of the Dantzig selector, which is fairly popular among practitioners (at least, among those we surveyed; see Figure 2). Figure 12 shows two examples of Dantzig selector profile plots for a \((12, 24) Var(s+)\) design for two different simulated response vectors. The plot on the left shows three clearly dominant factors and perhaps
one more worth investigating. Follow-up experimentation should focus on the identified factors (e.g. augmenting to investigate interaction effects). The right-hand plot is more ambiguous and a case where follow up-experimentation is necessary in order to simply fine-tune which factors actually impact the response. Because these designs are algorithmically generated, there is great flexibility in \((n,k)\) combinations. We have provided a catalog of \(\text{Var}(s+)\) 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 material.

![Scaled Dantzig Coefficient Estimates](image)

Figure 12: A clear and ambiguous Dantzig Selector profile plot of the Dantzig estimates vs. the shrinkage parameter, \(\delta\) for the \(n = 12, k = 24 \text{Var}(s+)\) SSD.

We believe a paradigm shift for SSDs has implications for both practitioners as well as researchers. From the survey described in Section 2.1, prominent concerns included “unreasonable assumptions”, “ambiguous analysis”, and “sparsity/power”. The work we’ve highlighted and studied, with paired design and analysis, helps to address these three issues which are strongly interrelated. Both the \(\text{Var}(s+)\)/Dantzig and GO-SSD approaches have mechanisms to detect whether the sparsity assumptions (see the Appendix) demanded of the designs are too strict for a particular experiment. The feedback the user obtains can be used to follow up with appropriate experimentation. We have also shown that under reasonable sparsity conditions, we can reliably detect active factors and screen out inactive factors.

Both the \(\text{Var}(s+)\) and GO-SSD procedures are ripe for additional research. Better theoret-
ical understanding of the $\text{Var}(s+)$ designs needs to be developed, and this may suggest further improvements to this approach. Since these designs have been successfully analyzed using the Dantzig selector, a natural approach would be to construct designs to exploit the Dantzig structure more directly; we are actively investigating this approach. As shown in this paper, there is room to improve the GO-SSD analysis method and there are many opportunities to use the basic idea of generating a model-independent estimate of $\sigma^2$ in other screening settings. 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. 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, just as they are in 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**

1. PDF copy of the survey questions presented in Section 2.1.

2. Catalog of $\text{Var}(s+)$ designs for $5 \leq n \leq 50$ with $n + 1 \leq k \leq 2n$.

3. R code for Dantzig selector function with data-driven $\gamma$ and profile plot.

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

**Appendix: Design Size and Factor Sparsity**

Recall the insights of Marley and Woods (2010) regarding SSD sample size requirements as a function of number of factors and number of active factors. In general, their simulations suggested that the run size should be at least three times the number of active factors for successful screening.
with SSDs. They also assert that the level of saturation of a SSD—the number of factors relative to the number of runs, $k/n$—should be less than 2.

To further investigate the recommendations of Marley and Woods (2010), we briefly present simulation results that include a larger variety of design sizes than previously studied. We consider 20 different $(n, k)$ combinations (Table A.1) ranging from (6, 10) to (40, 100), constructed with both the $\text{Var}(s^+)$ and $\text{UE}(s^2)$ criteria (giving a total of 40 SSDs). Two levels of sparsity are considered ($a = n/4$ and $a = n/2$) and two levels of complexity ($SN=1$ and $SN=3$). We analyze the designs using the Dantzig procedure outlined in Section 3.

Figure A.1 shows power as a function of $n$ for each of the 40 SSDs. Each design is indicated by the ratio of $n$ to the number of active factors, $a$, in the simulation. It is clear that the scenarios where the number of runs was at least three times the number of active factors ($n/3 \geq a$) produce higher power than situations where $n/3 < a$. This pattern is present regardless of the complexity of the simulation scenario (compare $SN=1$ to $SN=3$ in Figure A.1). We note that for designs with fewer than 25 or so runs, violation of this rule of thumb seems to be particularly problematic for analysis.

Figure A.2 examines how the level of saturation, $k/n$, is associated with power. Average power over both model complexities ($SN = 1$ and $SN = 3$) and design types versus the $k/n$ ratio is shown. The marker size represents the number of runs in the SSD. There is a clear degradation in performance as the level of supersaturation increases. The designs with larger $n$ generally have higher power for a similar level of $k/n$.

Thus, we confirm the recommendation from Marley and Woods (2010) that $n/a$ be larger than 3, but find no clear evidence that $k/n = 2$ is a changepoint. Rather, there is a steady degradation of performance as the level of supersaturation grows larger. We also note that if the run size is small, there is an additional possibility of reduced effectiveness.

**References**


30
Table A.1: SSD sizes used in simulations.

<table>
<thead>
<tr>
<th>n</th>
<th>k</th>
<th>n</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>10</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>20</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
<td>31</td>
<td>33</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>40</td>
<td>56</td>
</tr>
<tr>
<td>14</td>
<td>24</td>
<td>40</td>
<td>85</td>
</tr>
<tr>
<td>15</td>
<td>35</td>
<td>40</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure A.1: Power vs. run size (40 SSDs shown).
Figure A.2: Average Power vs. $k/n$


