UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M11/10
Methodology

Screening Strategies In The Presence Of Interactions

D. Draguljic, D.C. Woods, A.M. Dean, S.M. Lewis, A.E. Vine

Abstract

Product and process improvement can involve a large number of factors which must be varied simultaneously. Understanding how factors interact is a key step in identifying those factors that have a substantial impact on the response. This paper assesses and compares screening strategies for interactions using supersaturated designs, group screening, and a variety of data analysis methods including shrinkage regression and Bayesian methods. Novel methodology is developed to allow application of Bayesian methods in two-stage group screening. Insights on using the strategies are provided through a variety of simulation scenarios and open issues are discussed.

Screening Strategies in the Presence of Interactions

D. Draguljić[1], D. C. Woods[2*], A. M. Dean[3], S. M. Lewis[2] and A. E. Vine[2]

[1]Battelle Memorial Institute, King Avenue, Columbus, OH 43201, USA
[2]Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK
[3]Department of Statistics, The Ohio-State University, Columbis, OH 43201, USA

Product and process improvement can involve a large number of factors which must be varied simultaneously. Understanding how factors interact is a key step in identifying those factors that have a substantial impact on the response. This paper assesses and compares screening strategies for interactions using supersaturated designs, group screening, and a variety of data analysis methods including shrinkage regression and Bayesian methods. Novel methodology is developed to allow application of Bayesian methods in two-stage group screening. Insights on using the strategies are provided through a variety of simulation scenarios and open issues are discussed.

KEY WORDS: Bayesian model selection; Bayesian D-optimality; Gauss-Dantzig Selector; Group screening; Shrinkage regression; Supersaturated designs.

1. SCREENING

In the discovery and development of high quality products and processes, it is increasingly common for *screening experiments* to be run. Screening involves sifting through a large number of potentially important factors to search, as economically and effectively as possible, for the few *active* factors. These are factors whose influence on the measured response is sufficiently large to be of value in improving the system. The active factors are followed up in later studies for building detailed models for prediction and optimization. In complex systems, where generally there are several aspects which must function efficiently together, studies are needed to discover how factors interact. It is then vital that a screening strategy can identify active interactions as well as main effects (see Lewis and Dean, 2001; Phoa, Wong, and Xu, 2009b).

Examples of recent screening studies include: (a) a two-stage group screening experiment at Jaguar Cars to find factors that could be used to improve cold start performance (see Vine, Lewis, Dean, and Brunson, 2008); (b) a 28-run supersaturated design at the specialty chemical company, the Lubrizol Corporation, for determining factors in motor oil that affect the coefficient of friction (Scinto, Wilkinson, and Lin, 2011); and (c) an 18-run experiment on 31 factors to identify those factors that influence the yield from a chemical reaction (Rais, Kamoun, Chaabouni, Claeys-Bruno, Phan-Tan-Luu, and Sergent, 2009). Further applications in analytical chemistry are reviewed by Dejaegher and Vander Heyden (2008).

Traditional experimentation for product and process improvement begins by examining factor main effects only, and uses further experimentation to examine interactions between factors with main effects judged important from the stage 1 results (see, for example, Box, Hunter, and Hunter,

*Corresponding author: D.Woods@southampton.ac.uk

2005, chs. 6 and 7). This approach requires a firm belief in *strong effect heredity* (Hamada and Wu, 1992; Chipman, 1996) which states that interactions occur only between those factors with active main effects. Practical applications provide evidence that strong effect heredity fails to hold quite frequently; see Moore and Epps (1992), Vine et al. (2008), and Scinto et al. (2011). Consequently our preferred method of screening is to include two-factor interactions in the earliest stage of experimentation to screen out interactions of little importance as quickly as possible.

The purpose of this paper is to explore, extend and compare screening strategies that allow investigation of interactions to give insights into how the approaches might work in practice. Strategies that use supersaturated designs and group screening are investigated, together with several methods of shrinkage regression and Bayesian analysis. The novel work presented is an assessment and comparison of these screening strategies as well as the development of prior distributions for a Bayesian analysis of two-stage group screening experiments.

The remainder of this section discusses specific considerations in screening studies. Approaches to design are described in Section 2; regression shrinkage and Bayesian methods for analyzing data are described in Section 3, together with choice of tuning parameters or prior hyperparameters. In Section 4, strategies are compared that use supersaturated designs and group screening with analysis methods from Section 3 through a simulation of the entire screening process, involving design selection, analysis of data, and decisions on the active effects. Issues are raised in Section 5 to stimulate further discussion, development and application of screening methods.

1.1. Choice of factor levels and model

Most experiments for screening a large number of factors examine only two factor levels ('high' and 'low'). We concentrate on this situation, with the intention that a further experiment would use additional factor levels to estimate a more detailed predictive model as needed. We assume that a reasonable approximation, over the region of interest, of the major features of the relationship between a response variable and the main effects of the f independent factors and their $f(f-1)/2$ two-factor interactions is provided by the following linear model

$$\boldsymbol{Y} = \mathbf{1}_n\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ response vector, the corresponding error vector $\boldsymbol{\varepsilon}$ is assumed to follow a $N(\mathbf{0}_n, \boldsymbol{I}_n\sigma^2)$ distribution with $\boldsymbol{I}_n$ the $n \times n$ identity matrix and $\mathbf{0}_n$ the zero n-vector, $\boldsymbol{X}$ is an $n \times p$ matrix with $p = f + f(f-1)/2$, β_0 is the unknown intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ holds p unknown regression parameters. Each column of $\boldsymbol{X}$ corresponds to a factorial effect. Since, in this paper, each factor has two levels, the column corresponding to the main effect of the jth factor has a "-1" in every row for which the jth factor is observed at its low level, and a "$+1$" when the factor is observed at its high level. The column corresponding to the interaction between the kth and lth factors is an elementwise product of the kth and lth main effect columns. Throughout this paper, we refer to the regression parameters as main effects and interactions.

To identify the active or "large" effects in the model (see Section 1.2), we use a p-vector $\boldsymbol{\gamma}$ with first f entries $\gamma_j = 1$ if the jth main effect is large and 0 otherwise ($j = 1, \ldots, f$), and last $p - f$ entries γ_{kl} defined similarly for the interaction between factors k and l ($1 \le k < l \le f$). Then the set $\mathcal{A}_T$ of truly active effects has size $\boldsymbol{\gamma}'\mathbf{1}_p$, where $\mathbf{1}_p$ is the unit p-vector.

1.2. Detection of active factors

In many applications, it is possible to elicit from subject experts the minimum difference, Δ, of substantive interest between two responses. In this paper, we define a factorial effect of a two-level factor as *active*, and hence an element of $\mathcal{A}_T$, if the corresponding β_u $(u = 1, \ldots, p)$ in (1) is larger in absolute value than a threshold t, where the value of t is application-dependent. We set the threshold as though each individual β_u were the only non-zero regression parameter. Suppose, for example, that $\beta_1 \neq 0$ and $\beta_u = 0$ for $u = 2, \ldots, p$. Then the elicited value of Δ represents $E(Y_{HL\ldots L}) - E(Y_{LL\ldots L}) = 2\beta_1$ where, for example, $Y_{HL\ldots L}$ is the response when the first factor is set to its high level (H) and all the remaining factors are set to their low levels (L). It follows that the threshold for β_1 being active is $t = \Delta/2$.

A factor is defined as active if its main effect is active or if it is involved in an active interaction. Thus the analysis of data from a screening experiment can be viewed as using model selection techniques to decide which of the factorial effects (main effects and interactions) satisfy the definition of active. We denote the set of these selected or declared active effects by $\mathcal{A}_S$.

Typically, a screening experiment has many more effects to be examined than observations that can be taken within available resources. The data analysis is then likely to be successful only when there are few active effects and, consequently, few active factors. Even in this situation of *factor sparsity* (Box and Meyer, 1986), it may not be possible to discover all the active factors from among a large set of possibilities without making errors; see, for example, Abraham, Chipman, and Vijayan (1999), Li and Lin (2003), and Marley and Woods (2010) for main effects screening. Thus the goal for a screening strategy is to minimize the probability of making mistakes.

In Section 4, we use the following measures to evaluate and compare screening methods: (i) *Sensitivity*: the proportion of active main effects and interactions that are successfully detected, (ii) *False Discovery Rate* (FDR): the proportion of effects declared active that are actually inactive, (iii) *Type I error rate*: the proportion of inactive main effects and interactions that are incorrectly declared active. If there are no active effects ($\mathcal{A}_T = \emptyset$), then sensitivity is defined as 1; if no effects are selected as active ($\mathcal{A}_S = \emptyset$), then FDR is defined as 0 (see Benjamini and Hochberg, 1995). A further comparison uses the difference, $|\mathcal{A}_S| - |\mathcal{A}_T|$, between the sizes of the selected active and truly active sets of effects. This quantity is called the *Active Set-size Discrepancy* (ASD).

1.3. Elicitation of Prior Information

Prior information is routinely obtained from subject specialists and from pilot runs during the scientific planning of any experiment (Meyer and Booker, 2001; Dupplaw, Brunson, Vine, Please, Lewis, Dean, Keane, and Tindall, 2004; Vine et al., 2008). This includes information on which factors to investigate, their levels, the available budget, and any physical randomization restrictions. In many experiments, for example in engineering and chemistry, experts are often able to provide information on the "direction" of each main effect based on scientific knowledge or previous experience; then the high level of each factor can be set at the level which is most likely to result in the higher response. For group screening (see Section 2.2), a higher rate of detection of active effects can often be achieved when the factor levels are set accordingly. Elicited prior knowledge on the direction and size of effects may be incorporated into the Bayesian design and/or the analysis of an experiment (see, for example, Chipman, Hamada, and Wu, 1997).

2. DESIGNS FOR SCREENING STRATEGIES

Much research has been concerned with designs for screening small or moderate numbers of factorial effects. Recent developments in fractional factorial and non-regular designs have been presented by Wu and Hamada (2009). Other methods include search designs (Srivastava, 1975) which allow a pre-specified set of effects to be estimated, together with a small number of possibly important additional effects (see also DuMouchel and Jones, 1994) and designs that maximize the number of different models that can be fitted, a criterion known as estimation capacity (Cheng, Steinberg, and Sun, 1999). Li (2006) and Jones, Li, Nachtsheim, and Ye (2007) suggested selection of designs with high estimation capacity, followed by application of criteria based on distances between pairs of potential models.

Several authors have developed Bayesian approaches to design for model selection. Box and Hill (1967) proposed a design selection criterion based on the Kullbeck-Liebler distance between the posterior predictive distributions for pairs of models; see also Meyer, Steinberg, and Box (1996). A similar criterion using the Heillinger distance was investigated by Bingham and Chipman (2007). A decision-theoretic approach to this problem was developed by Rose (2008). All of these Bayesian approaches require the specification of a prior probability for each model and a prior distribution for the model parameters, and are more computationally intensive than the frequentist methods.

Two design strategies for screening a large number of factorial effects with far fewer observations are now briefly reviewed and are then evaluated in Section 4.

2.1. Supersaturated designs

We define a supersaturated design as having fewer runs than effects to be estimated. Although the factorial effects cannot all be estimated simultaneously, a variety of submodels will be identifiable. Selection of a models from these submodels is achieved by the methods of analysis discussed in Section 3. For experiments where main effects only models are assumed, the first systematic construction of supersaturated designs was provided by Booth and Cox (1962) via computer search. These authors proposed $E(s^2)$ and r_{max}, the respective average and maximum correlation between columns of $\boldsymbol{X}$, as measures of performance of supersaturated designs. Other measures include the average D-optimality of subdesigns (Wu, 1993), the number of zero correlations (Liu and Dean, 2004), the probability of correct selection of active effects (Allen and Bernshteyn, 2003) and the AM-criterion which combines estimation efficiency with low dependencies within subsets of columns of $\boldsymbol{X}$ (Marley, 2010). In the literature, most of the supersaturated designs selected using the above criteria have been for main effects only models; for example Lin (1993), Nguyen (1996), Li and Wu (1997), Ryan and Bulutoglu (2007), and Georgiou, Draguljić, and Dean (2009).

Very little work has been done on the construction of supersaturated designs for the estimation of factor interactions. As far as we are aware, the only methods that lend themselves to this setting are those of Wu (1993), Liu, Ruan, and Dean (2007), and Jones, Lin, and Nachtsheim (2008). The latter authors used Bayesian D-optimality to find supersaturated designs; we have adopted and extended their method to obtain designs suitable for estimating interactions for the comparison of methodologies in Section 4. Their criterion selects a design that maximises the determinant

$$|\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K}(1/\eta^2)|\,, \tag{2}$$

where η^2 is the variance of the common prior distribution for the regression coefficients, and

$$\boldsymbol{K} = \begin{bmatrix} 0 & \mathbf{0}_p' \\ \mathbf{0_p} & \boldsymbol{I}_p \end{bmatrix}.$$

This criterion provides flexibility in choice of design size and, when combined with a suitable optimization algorithm, is easily incorporated within the framework of a large simulation study.

2.2. Group screening for large numbers of effects

Group screening was introduced by Dorfman (1943) in the context of screening in blood samples and was extended to factor screening by Watson (1961); Morris (2006) has given a review of generalizations and extensions of these ideas.

In two-stage group screening of two-level factors, the f factors are partitioned into g groups at the first stage of experimentation, where the jth group contains $g_j \geq 1$ factors ($j = 1, \ldots, g$). High and low levels for each of the g grouped factors are defined by setting all the individual factors in a group to either their high level or to their low level simultaneously. The first stage of experimentation is performed on the relatively small number of grouped factors. The grouped factors found to have active main effects or to be involved in active interactions are declared active and are carried forward to the second stage, where an experiment is run on the individual factors which constitute the active groups. In the second stage, main effects and interactions between the individual factors within each active group are examined. Where the first stage has identified an active interaction between two grouped factors, the interactions between pairs of individual factors, one from each group, are also investigated. This is the IGS procedure of Lewis and Dean (2001) (see also Vine, Lewis, and Dean, 2005).

3. ANALYSIS STRATEGIES

We now give an overview of various methods for analyzing data arising from experiments that have fewer observations than effects to be estimated: LASSO, SCAD, Gauss-Dantzig Selector and Bayesian model selection and maximum a posteriori (MAP) estimation. These methods will be applied in Section 4 for the supersaturated and group screening designs. In each analysis, to ensure that every model contains the intercept β_0, we use a centered response vector and centered explanatory variables (main effects and interactions) so that $\boldsymbol{y}'\mathbf{1}_n = 0$ and $\boldsymbol{X}'\mathbf{1}_n = \mathbf{0}_p$.

3.1. Frequentist methods

3.1.1. Shrinkage methods These methods (see, for example, Hastie, Tibshirani, and Friedman, 2009, ch. 3) achieve variable selection by biasing, or shrinking, estimated regression coefficients towards zero. The biased estimators typically have lower variance than estimators obtained through ordinary least squares. Most shrinkage techniques can be expressed as a penalized regression problem which seeks estimates of $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{i=1}^{n}\left(y_i - \sum_{u=1}^{p} x_{iu}\beta_u\right)^2 + \lambda \sum_{u=1}^{p} \phi(\beta_u)\,, \tag{3}$$

where $\phi(\cdot)$ is a non-negative penalty function and λ is a constant that controls the relative importance of the penalty term and thus the degree of shrinkage.

In applying shrinkage methods, it is necessary to select values for various different tuning parameters. Selection methods include cross-validation (see Hastie et al., 2009, ch. 7), generalised cross-validation, GCV (Craven and Wahba, 1979) and the Akaike Information Criterion, AIC (for example, Burnham and Anderson, 2002, p. 63). In our study, where the designs are highly structured and the number of possible regression coefficients greatly exceeds the number of observations, we found that cross-validation performed poorly (see also Yuan, Joseph, and Lin, 2007). Following Fan and Li (2001), we used generalized cross-validation with SCAD, and AIC with all other procedures with the effective degrees of freedom approximated by the number of nonzero estimated regression coefficients (as suggested by Zou, Hastie, and Tibshirani, 2007). For the number of parameters, p, close or equal to the number of runs, n, the standard AIC penalty ($2p$) can lead to models that severely overfit the data, see Burnham and Anderson (2002, p. 66). Hence, as suggested by Hurvich and Tsai (1989), we used AIC with a modified penalty,

$$\text{AIC} = n \log\left(\frac{\text{RSS}}{n}\right) + \frac{2pn}{n-p}\,, \quad \text{for } p < n\,, \tag{4}$$

where RSS denotes the residual sum of squares. The modified penalty behaves similarly to the standard penalty for $p << n$, and tends to infinity as $p \to n$ (when RSS$\to 0$).

Bridge regression and the LASSO : Bridge regression (Frank and Friedman, 1993) is a broad class of shrinkage regression methods with penalty in (3) of the form $\phi(\beta_u) = |\beta_u|^\gamma$. The choice $\gamma = 2$ gives ridge regression and $\gamma = 1$ gives the LASSO (Least Absolute Shrinkage and Selection Operator; Tibshirani, 1996). Lin (1995) found that ridge regression performed poorly for main effects models when the number of factors was considerably larger than the number of runs. Unlike ridge regression, LASSO estimates are nonlinear functions of the data and may be found as the solution to a quadratic programming problem. LASSO regression also has the advantage of shrinking some coefficient estimates to zero for suitable choice of λ. In our simulation study, we declared any non-zero effect to be active if its estimated coefficient value exceeded threshold t, where t was set as described in Section 4.3. In practice, the value of t would be elicited from subject experts (see Section 1.2).

SCAD: Fan and Li (2001) developed SCAD (Smoothly Clipped Absolute Deviation) regression in which $\phi(\cdot)$ in (3) is defined through its first derivative

$$\frac{\partial \phi(\beta_u)}{\partial \beta_u} = \theta_1 \left\{ I_{(\beta_u \le \theta_1)} + \frac{(\theta_2\theta_1 - \beta_u)}{(\theta_2 - 1)\theta_1} I_{(\beta_u > \theta_1)} \right\}\,, \quad \text{for } u = 1, \ldots, p\,,$$

where θ_1 and θ_2 are tuning parameters, and $I_{(a>b)}$ is an indicator function taking value 1 if $a > b$ and 0 otherwise. SCAD achieves subset selection in the same way as the LASSO, by allowing estimates $\hat{\beta}_u$, found via iterative fitting of ridge regressions, to be shrunk to zero. Li and Lin (2002, 2003) gave an example of where SCAD regression for the analysis of a supersaturated design performed well compared with stepwise regression and the Bayesian strategy of Beattie,

Fong, and Lin (2002) under a main effects model. In our SCAD implementation, initial values of β_u for the estimation algorithm were obtained using stepwise regression with $\alpha_{in} = \alpha_{out} = 0.10$; $\theta_2 = 3.7$ (Fan and Li, 2001); θ_1 was chosen using GCV; threshold t was applied as in the LASSO.

Dantzig Selector: Phoa, Pan, and Xu (2009a) suggested using the Dantzig Selector (Candes and Tao, 2007) for the analysis of supersaturated designs where $\hat{\boldsymbol{\beta}}$ is chosen to satisfy

$$\min_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} \sum_{u=1}^{p} |\hat{\beta}_u| \quad \text{subject to } ||\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})||_\infty \leq s\,,$$

with s a tuning constant and $||\boldsymbol{a}||_\infty = \max|a_i|$, $\boldsymbol{a}' = (a_1, \ldots, a_p)$. We selected the value of s using (4), since our initial studies showed that the modification of AIC adopted by Phoa et al. (2009a) returned too few active effects.

Candes and Tao (2007) applied the Dantzig Selector to choose a subset of potentially active effects, and then used standard least squares to fit a reduced linear model. The terms in this model whose coefficient estimates exceeded threshold t were declared active. This procedure is known as the Gauss-Dantzig Selector.

3.1.2. Non-shrinkage regression for the first stage of group screening A regular fractional factorial design may be selected for the first stage of group screening so that traditional analysis of variance or linear regression methods apply. As noted by Abraham et al. (1999), Li and Lin (2002), and others, non-shrinkage regression methods such as forward selection and stepwise regression methods may not be successful for supersaturated designs. These methods may also have problems when there are no truly active effects, see for example Draguljić (2010, Section 3.6.4) and Marley and Woods (2010).

In our simulation, we have included a modification of a forward selection procedure for analysis of group screening designs, as follows. We performed individual one-parameter regressions for each main effect and interaction parameter β_u and ordered these by their p-values. For $h = 1, \ldots, n-1$, the hth largest estimated parameter was added into the model if its inclusion increased R^2 by at least $R^2_{inc} = 0.99/\lceil m+1 \rceil$, where m is the expected number of active effects for a given number of factors and given probabilities of main effects and interactions being active, and $\lceil m+1 \rceil$ denotes the smallest integer greater than $m+1$.

We compared two different analysis methods for group screening. The first used the above procedure for selecting the active grouped effects at stage 1, and the active individual effects from the corresponding groups at stage 2. A final threshold t was applied so that effects with estimated effects less than t were screened out. The second analysis method used the Dantzig Selector (Section 3.1.1) at both stages of group screening.

3.2. Bayesian methods

Screening via *model selection* fits naturally within the Bayesian paradigm, where posterior probabilities for individual models of the form (1) can be calculated and compared (see, for example, O'Hagan and Forster, 2004). In this section, we describe Bayesian model selection for screening experiments and a choice of hyperparameters for the prior distributions necessary for implementation. Each possible model may be described by a p-vector $\boldsymbol{\gamma}$ with entries 1 or 0

according to whether or not the corresponding effect is active (see Section 1.1). Given data $\boldsymbol{y}$, the screening problem of identifying which of the p effects should be classified as active and which as inactive can then be viewed as selecting the best choice of $\boldsymbol{\gamma}$.

Bayesian model selection methods can be applied directly to supersaturated designs, as demonstrated by Chipman et al. (1997). For two-stage group screening, the application requires the derivation of first-stage prior distributions for the grouped effects (see Section 3.2.1). Model selection is via interrogation of the (approximated) posterior density. In this paper, we take two approaches, using (i) *model selection* and (ii) *model averaging*:

(i) We identify the subset of models (values of $\boldsymbol{\gamma}$) that have high posterior probabilities, and then declare as active the effects that occur in these models. In Section 4, we select the effects from models whose posterior probabilities exceed one third the probability of the posterior modal model(s), a procedure that gives a more reasonable trade-off between sensitivity and Type I error rate than similar empirical alternatives. More formal methods such as cross-validation or the use of intrinsic Bayes factors (as applied by Beattie et al., 2002) could be employed to identify a subset of effects from the high probability models. However, as these methods subdivide the data into training and test sets, they work best with larger experiments or smaller numbers of truly active effects. Also, the column correlations in the model matrices for training and test subsets may become undesirably large.

(ii) We find, for each β_u, an approximation to the marginal distribution (a model-averaged mixture t-distribution) by using a posterior sample from MCMC and kernel density estimation. We then select as active those effects whose regression coefficients have maximum a posteriori (MAP) estimates greater than threshold t. These MAP estimates correspond to a 0-1 loss function for β_u. Alternatively, a squared or absolute error loss function could be employed with a comparison of the model-averaged posterior mean or median, respectively, with the threshold. MAP estimation is used because it is analogous to the shrinkage methods of Section 3.1.1 but with possibly different shrinkage parameters for each β_u; see Lu and Zhang (2007) for a similar frequentist approach.

In our study, the specification of conjugate prior distributions for $\boldsymbol{\gamma}$, σ^2 and $\boldsymbol{\beta}$ follows that of George and McCulloch (1997). The prior distribution for σ^2 is an inverse gamma, $IG(\nu/2, \nu\lambda/2)$, with $\nu > 0$ and $\lambda > 0$; i.e. $\nu\lambda/\sigma^2 \sim \chi^2_\nu$. The conditional prior distribution for $\boldsymbol{\beta}$ is a mixture of Normal distributions, $\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2 \sim N(\mathbf{0}, DRD\sigma^2)$. Here D is a $p \times p$ diagonal matrix with entries $D_{uu} = a_u\tau_u$, with $a_u = 1$ if the uth entry in $\boldsymbol{\gamma}$ is 0 and $a_u = c_u > 1$ otherwise, and R is the $p \times p$ prior correlation matrix for $\boldsymbol{\beta}$. This choice gives each active effect a more diffuse (larger variance) prior distribution than each inactive effect. Hence, an effect corresponding to a large regression coefficient has a higher prior probability of being active. The choice of τ_u and c_u is ideally informed by threshold t (Section 1.2), see details below and also George and McCulloch (1993). The hyperparameters ν, λ and π are chosen to reflect prior beliefs.

For main effects, we take the prior distribution for γ_j to be Bernoulli with parameter $0 \leq \pi_j \leq 1$, for $j = 1, \ldots, f$. Prior interaction probabilities can be assigned via the effect heredity principle (as described by Chipman, 1996) which allows the probability of an interaction being active to depend upon whether or not each of the two "parent" main effects is active:

$$P(\gamma_{kl}=1|\gamma_k,\gamma_l)=\pi_{kl}=\begin{cases}\pi_{kl}^{(00)} & \text{if } \gamma_k=\gamma_l=0\\ \pi_{kl}^{(10)} & \text{if } \gamma_k=1,\gamma_l=0\\ \pi_{kl}^{(01)} & \text{if } \gamma_k=0,\gamma_l=1\\ \pi_{kl}^{(11)} & \text{if } \gamma_k=\gamma_l=1\,.\end{cases} \tag{5}$$

The prior probability for $\boldsymbol{\gamma}$ is then $p(\boldsymbol{\gamma})=\prod_{j=1}^{f}\pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}\prod_{k<l}\pi_{kl}^{\gamma_{kl}}(1-\pi_{kl})^{1-\gamma_{kl}}$. The posterior densities, updated in light of observed data, are available in closed form, together with unnormalised posterior probabilities for each model. When p and hence the number of competing models is large, a numerical search of the model space is more efficient than complete enumeration. Samples from the posterior distributions can be obtained via a Gibbs sampling algorithm (see George and McCulloch, 1993, 1997).

In this paper, all the regression parameters are assigned the same prior distributions, with $\tau_u=\tau$ and $c_u=c$ $(u=1,\ldots,p)$. As described below, τ and c may be treated as tuning parameters and chosen by graphical investigation of the following conditional distribution for γ_u:

$$P(\gamma_u=1|\boldsymbol{\gamma}_{(u)},\boldsymbol{\beta},\sigma^2)=\frac{P(\gamma_u=1,\boldsymbol{\gamma}_{(u)})}{P(\gamma_u=1,\boldsymbol{\gamma}_{(u)})+P(\gamma_u=0,\boldsymbol{\gamma}_{(u)})r(\boldsymbol{\gamma},\boldsymbol{\beta},\sigma^2)}\,, \tag{6}$$

where $\boldsymbol{\gamma}_{(u)}$ is formed from $\boldsymbol{\gamma}$ by deletion of entry γ_u $(u=1,\ldots,p)$, and

$$r(\boldsymbol{\gamma},\boldsymbol{\beta},\sigma^2)=\frac{f(\boldsymbol{\beta}|\gamma_u=0,\boldsymbol{\gamma}_{(u)},\sigma^2)}{f(\boldsymbol{\beta}|\gamma_u=1,\boldsymbol{\gamma}_{(u)},\sigma^2)}\,.$$

Under the prior distribution for $\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma^2$ with $R=I$, the latter ratio simplifies to

$$r(\boldsymbol{\gamma},\boldsymbol{\beta},\sigma^2)=c\exp\left[-\beta_u^2(1-c^{-2})/(2\tau^2\sigma^2)\right]\,. \tag{7}$$

The graphical procedure for selecting τ and c begins with examination of (6) for $\beta_u=t$ (the threshold) and an initial value of σ^2; values of τ and c are selected to make the conditional posterior probability for an active effect close to 1, and $c\tau\geq 1$. This constraint reduces the influence of shrinkage on the coefficients of the active effects which is particularly important for MAP estimation. The shrinkage arises from the conditional posterior distribution, $\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma^2,\boldsymbol{y}\sim N(A\boldsymbol{X}'\boldsymbol{y},A\sigma^2)$ with $A=[\boldsymbol{X}'\boldsymbol{X}+D^{-2}]^{-1}$. In practice, a variety of σ^2 values should be explored of sufficient size to reflect the inflated residual sum of squares due to the exclusion of inactive effects of moderate size from the model.

The chosen values of τ and c are fine-tuned by investigating (6) as the value of β_u approaches t. As an illustration, suppose that $t=17$, the prior probability of an effect being active is 0.1, and $\sigma^2=1$. Inspection of the contours in Figure 1(a) suggests a value of $\tau=3$, where the posterior probability of the uth effect being declared active when $\beta_u=t$ exceeds 0.9. A wide range of possible values of c achieves high conditional posterior probability and $c\tau\geq 1$. Plots such as that shown in Figure 1(b) enable a final choice of c to be made. Selection of a higher value of c will result in a steeper curve and hence a lower Type I error rate but may also reduce sensitivity. The choice of $c=10$, as shown, gives a high probability (>0.9) of declaring the uth effect to be active when the value of β_u is close to the threshold of 17.

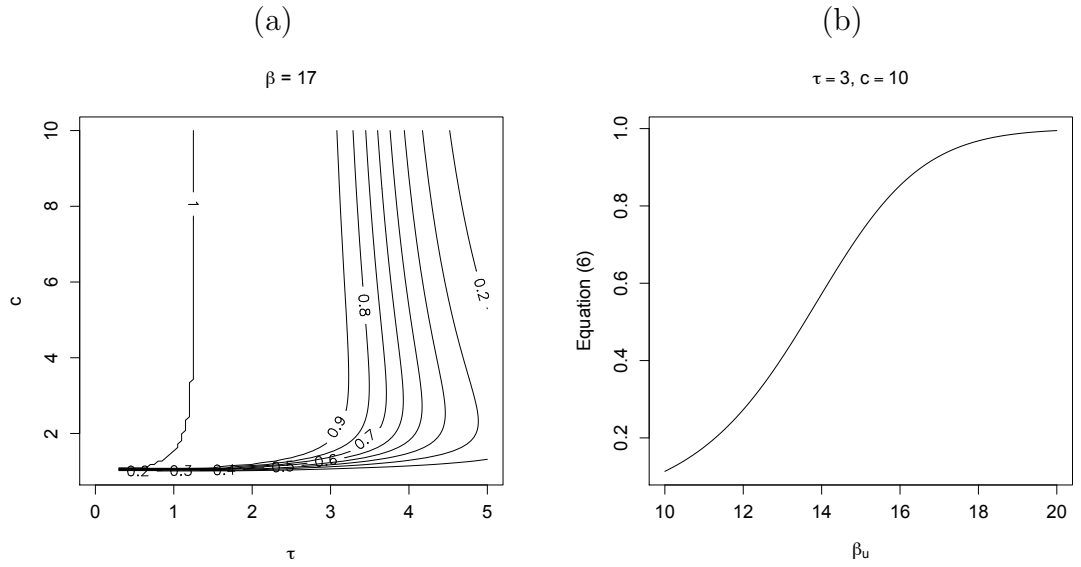


Figure 1: Investigation of choices for τ and c: (a) contour plot of conditional posterior probability (6) obtained using (7), as a function of c and τ; (b) probability (6) as a function of β_u.

3.2.1. Bayesian analysis of a two-stage group screening experiment This section describes the key steps in our Bayesian analysis of the first stage of group screening; particularly, the construction of prior distributions for the grouped factor effects from those of the individual effects.

Suppose that, at stage 1, the individual factors are labelled so that the first g_1 factors are in group 1, the second g_2 factors are in group 2, and so on, where $f = \sum_{j=1}^{g} g_j$. Suppose also that factor levels $x_{1i}, \ldots, x_{fi}$ are applied in the ith run of the first stage experiment. Define $s_1 = 0$, $s_h = s_{h-1} + g_{h-1}$ $(h = 2, \ldots, g+1)$. Then the level of the jth grouped factor in the ith run is

$$x_{ji}^G = x_{(s_j+1)i} = \ldots = x_{(s_j+g_j)i}\,, \quad i = 1, \ldots, n;\; j = 1, \ldots, g\,.$$

A linear model for the response Y_i at stage 1 may be expressed as follows:

$$\begin{aligned} Y_i &= \beta_0 + \sum_{j=1}^{f} x_{ji}\beta_j + \sum_{k=1}^{f-1}\sum_{l=k+1}^{f} x_{ki}x_{li}\beta_{kl} + \varepsilon_i \\ &= \left[\beta_0 + \sum_{j=1}^{g}\sum_{q=s_j+1}^{s_{j+1}-1}\sum_{r=q+1}^{s_{j+1}} \beta_{qr}\right] + \sum_{j=1}^{g} x_{ji}^G \left[\sum_{t=s_j+1}^{s_{j+1}} \beta_t\right] + \sum_{k=1}^{g-1}\sum_{l=k+1}^{g} x_{ki}^G x_{li}^G \left[\sum_{q=s_k+1}^{s_{k+1}}\sum_{r=s_l+1}^{s_{l+1}} \beta_{qr}\right] \\ &= \beta_0^G + \sum_{j=1}^{g} x_{ji}^G\beta_j^G + \sum_{k=1}^{g-1}\sum_{l=k+1}^{g} x_{ki}^G x_{li}^G\beta_{kl}^G + \varepsilon_i\,, \end{aligned} \tag{8}$$

which shows the deliberate aliasing of regression coefficients resulting from factor grouping. The vector of $p^G = (g + g^2)/2$ grouped regression coefficients is $\boldsymbol{\beta}^G = (\beta_1^G, \ldots, \beta_g^G, \beta_{12}^G, \ldots, \beta_{(g-1)g}^G)'$. Each possible model is defined by $\boldsymbol{\gamma}^G = (\gamma_1^G, \ldots, \gamma_g^G, \gamma_{12}^G, \ldots, \gamma_{(g-1)g}^G)$ whose entries are p^G indicator

variables with uth entry equal to 1 if and only if the uth grouped effect is active. Prior probabilities for each γ_j^G $(j = 1, \ldots, g)$ are obtained from the relationships between individual and grouped probabilities, see Vine et al. (2005, Section 2). Prior probabilities for γ_{kl}^G $(1 \leq k < l \leq g)$ are obtained using the heredity principle as in (5) and, in Section 4.3, we assume the same conditional probabilities as for the individual interactions.

Setting $\gamma_{s_j+1} = \ldots = \gamma_{s_j+g_j} = \gamma_j^G$ and $\gamma_{(s_k+1)(s_l+1)} = \ldots = \gamma_{(s_k+g_k)(s_l+g_l)} = \gamma_{kl}^G$, to ensure that individual factors involved in the same active grouped effect are all brought forward to the second stage, and setting $R = I_p$ in the prior density for $\boldsymbol{\beta}$, results in the following conditional prior distributions derived using (8): $\beta_j^G|\boldsymbol{\gamma}^G, \sigma^2 \sim N\left(0, g_j\tau^2\sigma^2[\gamma_j^G c^2 + (1-\gamma_j^G)]\right)$ and $\beta_{kl}^G|\boldsymbol{\gamma}^G, \sigma^2 \sim N\left(0, g_k g_l\tau^2\sigma^2[\gamma_{kl}^G c^2 + (1-\gamma_{kl}^G)]\right)$. Gibbs sampling can be used to generate a posterior sample from the joint distribution for $\boldsymbol{\gamma}$ (George and McCulloch, 1997) and hence to approximate the posterior model probabilities. Marginal posterior probabilities for each factorial effect can be approximated by the proportion of visited models that include the effect. To decide which individual effects are investigated at the second stage, we declare active those grouped effects contained in a subset of models with high posterior probability, see (i) above. Individual factors are carried forward to stage 2 if they are in groups having a declared active grouped main effect or involved in a declared active grouped interaction.

The outcome of the first-stage experiment is a realization of the random vector $\boldsymbol{\gamma}$, which is represented by the p-vector $\tilde{\boldsymbol{\gamma}}$ with p_2 entries equal to 1 (corresponding to each of the p_2 individual effects selected as active) and $p - p_2$ entries 0. The p_2 selected individual effects are then investigated in the stage 2 experiment. The analysis uses the stage 1 prior distributions to calculate the posterior densities using Gibbs sampling or, for small numbers of effects, explicit calculations. The closed-form stage 1 posterior distributions for $\boldsymbol{\beta}$ and σ^2, conditional on $\tilde{\boldsymbol{\gamma}}$, are not used to construct second stage prior distributions. This is because complete aliasing of individual effects at stage 1 creates ambiguities in the data analysis and interpretation due to high correlations in the prior distribution for $\boldsymbol{\beta}$.

4. EMPIRICAL COMPARISON OF STRATEGIES

We compared the performance of screening strategies using single-stage supersaturated designs and two-stage group screening procedures together with the analysis methods described in Section 3. The smallest number of factors investigated was $f = 10$, leading to 10 main effects and 45 two-factor interactions to be screened (a total of 55 factorial effects). We also investigated screening with $f = 15$ factors (120 factorial effects) and $f = 20$ factors (210 factorial effects). Interactions between three or more factors were set to zero.

4.1. Designs used in the simulation

For the first stage of the two-stage group screening procedure, the f factors were divided into five equal-sized groups and a 2_V^{5-1} fraction was selected with 16 runs and defining relation $I = G_1G_2G_3G_4G_5$, where G_i is the label of the ith group; the IGS procedure of Lewis and Dean (2001) was used, as described in Section 2.2. The required total number of observations in the stage 2 design was set equal to the number of effects to be estimated plus five extra observations

to avoid a saturated design. The decision to use five groups for the stage 1 design was based on minimizing the probabilities of missing active effects, as calculated through the *GiSEL* software (Dupplaw et al., 2004).

To allow dynamic construction of designs of various different sizes for estimating particular sets of effects at stage 2, the algorithm of Jones et al. (2008) was used to generate Bayesian D-optimal designs. This method was used regardless of the type of analysis undertaken due to its flexibility in generating designs within a large simulation. For the frequentist analysis, $\eta^2 = 0$ in (2) was used to generate a standard D-optimal design. For the Bayesian analysis, the prior distribution suggested by Jones et al. (2008) was employed, with $\eta^2 = 5$.

For the one-stage supersaturated design with f factors, the number of runs was chosen to be similar to the number required by group screening. This number was found by calculating the median number of runs used by group screening in the simulation study (see below). In this study, all main effects are assumed to have the same probability, q_{me}, of being active, with interaction probabilities calculated using effect heredity (5). A compromise was made in the run sizes resulting from the four different values of q_{me} used in the simulation and this resulted in 32, 58, and 94 runs for $f = 10$, 15 and 20 factors respectively. Bayesian D-optimal supersaturated designs for these sizes were found using $\eta^2 = 5$. An $E(s^2)$-optimal supersaturated design (Section 2.1) could have been used instead, and the findings of Marley and Woods (2010) suggest that similar results would have been obtained. For each f, a single generation of the design was used for all the simulations. The pairwise main effect and interaction column correlations for each of the $f = 10, 15, 20$ designs were small, with 50% of correlations below 0.071, 0.069, 0.048, and 95% below 0.31, 0.24, 0.17, respectively, and with maximum pairwise column correlations of 0.45, 0.41 and 0.40, respectively. Our view is that it is not necessary to have zero correlations for effective screening, provided the correlations are sufficiently small (c.f. Chen and Lin, 1998; Liu et al., 2007).

4.2. Data generation

At the start of each "batch" of 1,000 runs of the simulation, the value for the probability of each main effect being active was set equal to the common value q_{me}, selected from $\{0.0, 0.05, 0.1, 0.15, 0.2\}$ and held constant throughout the batch of runs. The probability of a two-factor interaction being active was determined in two ways:

(i) given the activity status of the constituent main effects, the interaction probability was calculated using relaxed weak heredity (Chipman, 1996) as in (5) with values

$$\pi^{(00)} = 0.005, \quad \pi^{(01)} = \pi^{(10)} = 0.125, \quad \pi^{(11)} = 0.25\,; \tag{9}$$

(ii) marginal interaction probabilities q_{int} were calculated usingthe conditional probabilities in (9). Each interaction was then set active with probability q_{int} independently of the status of the main effects.

Results obtained using the first method are included in this paper.

The success of identifying active factors correctly is *highly dependent* upon which columns of the supersaturated design are assigned to these factors. Thus, in a departure from many other

Table 1: Distributions for data generation and choice of prior distribution hyperparameters for the analysis of supersaturated designs and stage 2 group screening

	Distributions			Prior	
Setting	Active Effect	Inactive Effect	Error	τ	c
1	N(6, 1)	N(0, 1)	N(0, 1)	0.4	10
2	N(12, 4)	N(0, 1)	N(0, 1)	0.7	20
3	N(24, 4)	N(0, 1)	N(0, 1)	3	10
4	N(24, 4)	N(0, 16)	N(0, 1)	3	10

papers (for example, Phoa et al., 2009a; Li and Lin, 2002), the designation of active effects in our simulation is not fixed throughout each batch of runs, nor are the effect values. In addition, the non-active main effects and two-factor interactions are not set to zero, but are selected from the distributions listed in Table 1. This means that the success rates for detecting active effects in our simulations tend to be lower than those in other published studies.

For a batch of 1000 runs, values of f, q_{me} and one of the four sets of effect and error distributions in Table 1 were selected. For each run in a batch, a binary vector $\boldsymbol{\delta}$ was created to indicate the activity or non-activity of each effect. For the uth effect, δ_u was set equal to 1.0 with probability q_{me} or via heredity as appropriate, and set equal to zero otherwise. Values, $\beta_u^\star$, for each main effect and interaction parameter were then generated from the selected effect distributions to give a vector of true regression coefficients $\boldsymbol{\beta}^\star$. For main effects, the direction of the effects was assumed known and hence each $\beta_u^\star$ was generated as the absolute value of a draw from $N(\mu_{act}, \sigma_{act}^2)$; for the interactions, each parameter value was drawn from $N(\mu_{inact}, \sigma_{inact}^2)$ or from $N(-\mu_{inact}, \sigma_{inact}^2)$ with equal probabiliity. There was a very small probability that, on any given run, the generated value $|\beta_u^\star|$ of an active (inactive) effect would be less than (greater than) the chosen threshold t, and hence violate the definition of an active (inactive) effect; in such cases, $\beta_u^\star$ was regenerated.

A vector of observations $\boldsymbol{y}$ was then generated from the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, 1)\,,$$

where $\boldsymbol{X}$ is a model matrix (excluding an intercept column) corresponding to the design used. For single-stage supersaturated designs and at the first stage of group screening, this matrix has columns corresponding to each of the individual main effects and interactions (including completely aliased effects for the group screening designs). At stage 2 of group screening, $\boldsymbol{X}$ corresponds to a Bayesian D-optimal design in the individual effects brought forward to the second stage; for any grouped factor declared non-active at the end of stage 1, we set the constituent individual factors to their nominal levels, labeled 0, for the second-stage experiment and include the corresponding constant columns in $\boldsymbol{X}$.

4.3. Choice of tuning parameters for the analyses

For the shrinkage analysis methods, the tuning parameters were chosen as described in Section 3.1.1. The threshold, throughout, was set as $t = \mu_{act} - 3.5\sigma_{act}$, so that it was linked to the active effect distribution on each run of the simulation. For the Bayesian analysis of supersaturated designs and stage 2 group screening, the values of τ and c, given in Table 1, were chosen for

each of the four settings of active and inactive effects using the graphical methods of Section 3.2 to compromise between sensitivity and Type I error rate. For each of these choices, the marginal probability of declaring the uth effect active when $\beta_u > t$ is equal to, or very close to, one. Note that this does not take account of how the correlations between the columns of $\boldsymbol{X}$, and other aspects of the design, affects the posterior distribution of $\boldsymbol{\beta}$. For stage 1 of group screening, we recommend a choice of τ and c that gives less conservative results to avoid excessive screening of effects; we applied the default values of $\tau = 1/6$ and $c = 10$ (see Chipman, 2006).

For the inverse gamma prior distribution for σ^2, we chose $\nu = 5$ and, following an empirical investigation and in the same spirit as Chipman et al. (1997), set $\lambda = s/5$, where s^2 is the sample variance from the data. In an investigation not reported here, we found that the simulation results were fairly robust to the values of ν and λ used.

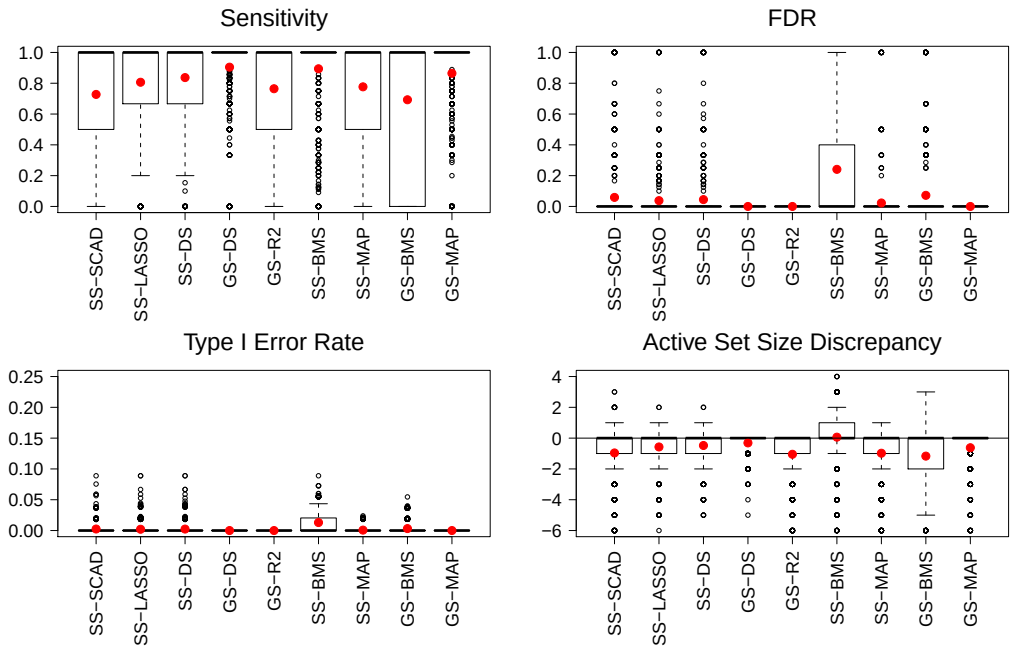
4.4. Results from the study

We used the four measures sensitivity, false discovery rate (FDR), Type I error rate, and active set size discrepancy (ASD), defined in Section 1.2, to compare the various strategies. Empirical distributions of these measures were obtained for each procedure using the 1000 simulations of each setting. Figures 2–4 show boxplots of results, mainly for more challenging cases where there is less separation between the active and inactive effect distributions. For clarity in the ASD plots, the range has been chosen so that up to 1.5% of values in the tails of the distribution are excluded. The occurrence of any grossly outlying values is discussed in the text. The nine screening strategies are five frequentist methods labelled SS-SCAD, SS-LASSO, and SS-DS (Gauss-Dantzig Selector) as in Section 3.1 for the supersaturated designs, GS-DS and GS-R2 (group screening with 5 equal-sized groups at stage 1 using the Gauss-Dantzig Selector or using R^2 analysis, respectively) as in Section 2.2, and four Bayesian methods SS-BMS, SS-MAP, GS-BMS, and GS-MAP (using, respectively, a supersaturated design with model selection and with MAP estimation, group screening with 5 groups and model selection, and group screening with MAP estimation; see Section 3.2).

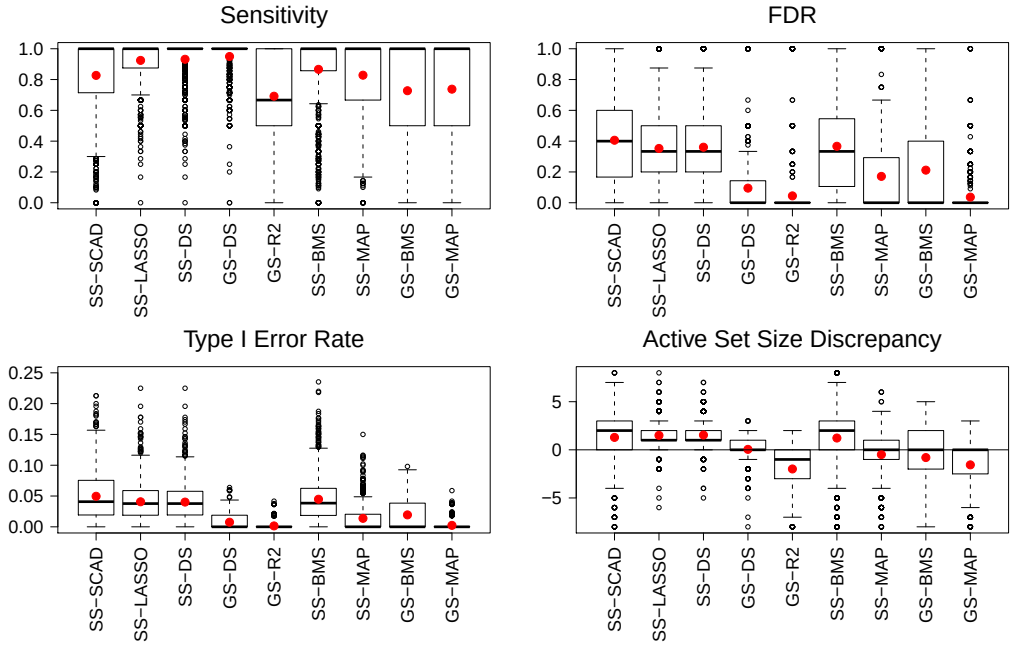
We first discuss results from two relatively easy settings: active and inactive effect distribution pairs N(12, 4), N(0, 1) and N(24, 4), N(0, 1). Our results show that SS-SCAD, SS-LASSO, SS-DS, GS-DS, SS-BMS and SS-MAP all tend to perform well, regardless of the number of factors and values of other simulation parameters. The SS-SCAD and SS-BMS methods tend to have somewhat larger FDR values resulting from overestimation of the true model size. Group screening without the Gauss-Dantzig Selector (GS-R2, GS-BMS and GS-MAP) tends to have lower sensitivity than the other procedures. For these settings, GS-DS has slightly lower sensitivity than SS-DS and SS-LASSO for 15 and 20 factors, see Figures 3(a) and 4(a), which is possibly due to the larger group sizes. In contrast, for the harder settings discussed below, GS-DS is consistently the most sensitive method.

The difficult settings in the study were the N(6, 1), N(0, 1) and N(24, 4), N(0, 16) cases (for example, Figures 2(a), 2(b), 3(b) and 4(b)) where there is least separation between active and inactive effects. Generally, and not surprisingly, higher sensitivity is often accompanied by an over-fitting of the model, so that FDR and Type I error rate are non-zero. However, most screening procedures keep the Type I error rate under control (< 0.25 for all procedures, and < 0.1 for GS-DS) regardless of the setting. The exception is for SS-BMS where, for $f = 15$ and all

Figure 2: Performance measures for screening strategies for $f = 10$ factors; for strategy labels, see Section 4.4

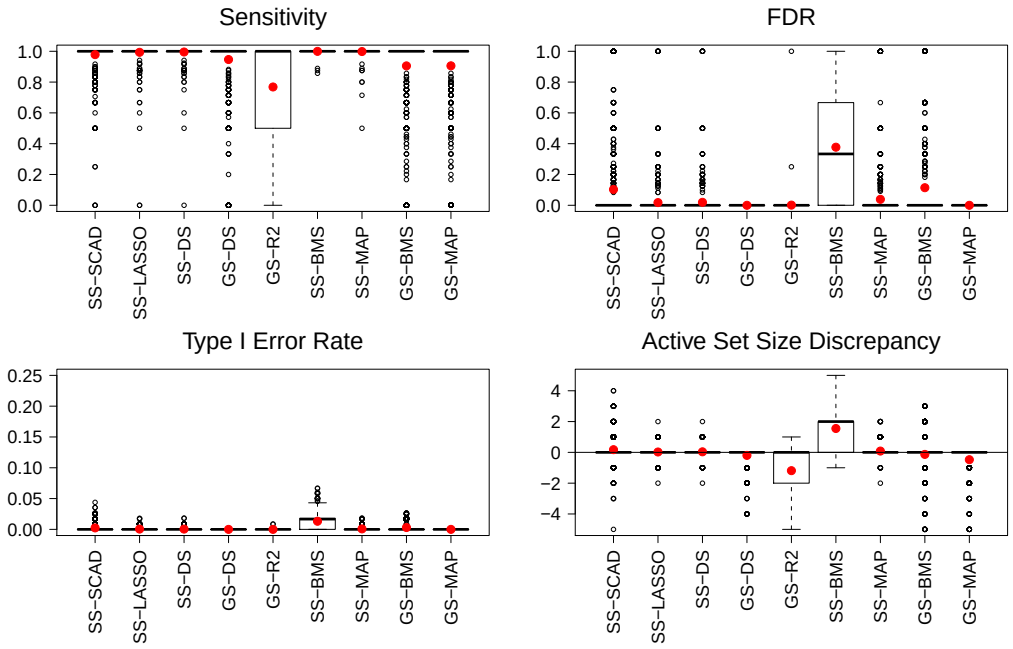


(a) $q_{me} = 0.10$, Active $N(24, 4)$, Inactive $N(0, 16)$

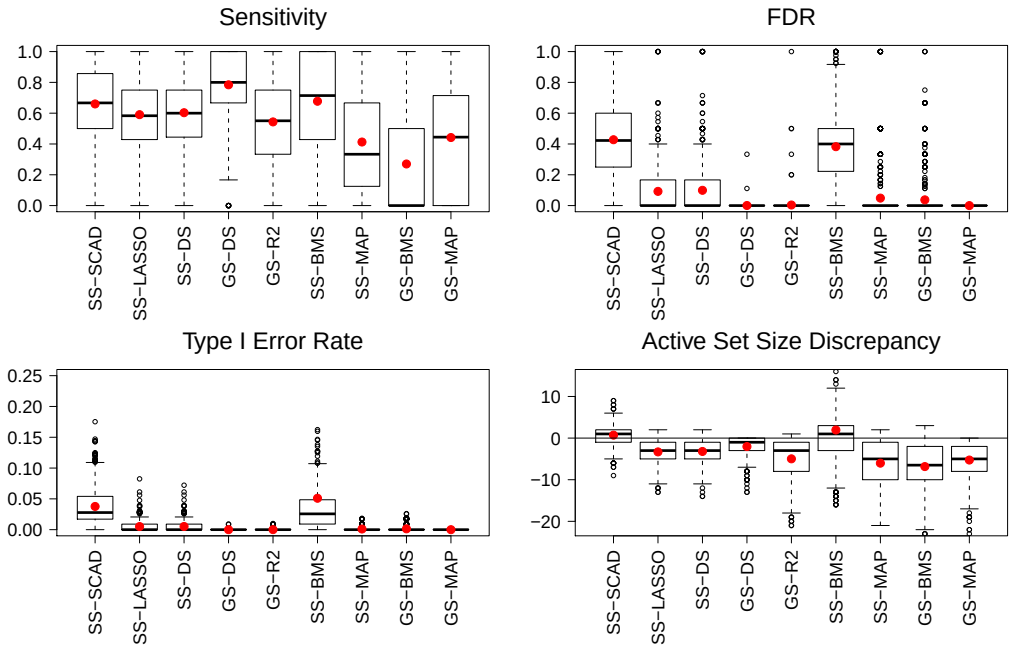


(b) $q_{me} = 0.20$, Active $N(6, 1)$, Inactive $N(0, 1)$

Figure 3: Performance measures for screening strategies for $f = 15$ factors

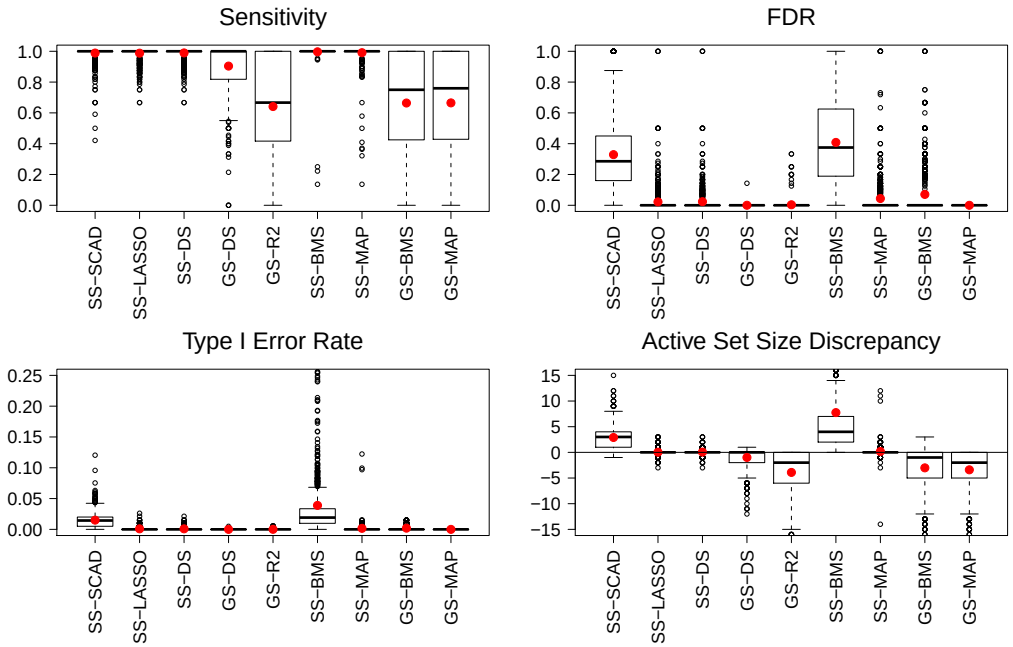


(a) $q_{me} = 0.05$, Active $N(12, 4)$, Inactive $N(0, 1)$

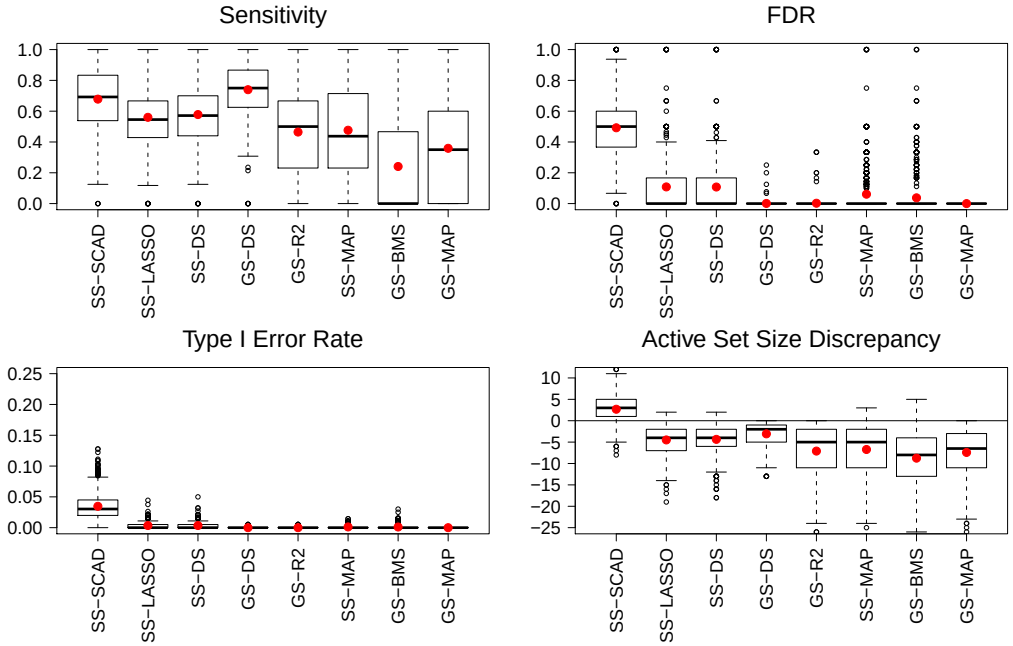


(b) $q_{me} = 0.2$, Active $N(24, 4)$, Inactive $N(0, 16)$

Figure 4: Performance measures for screening strategies for $f = 20$ factors

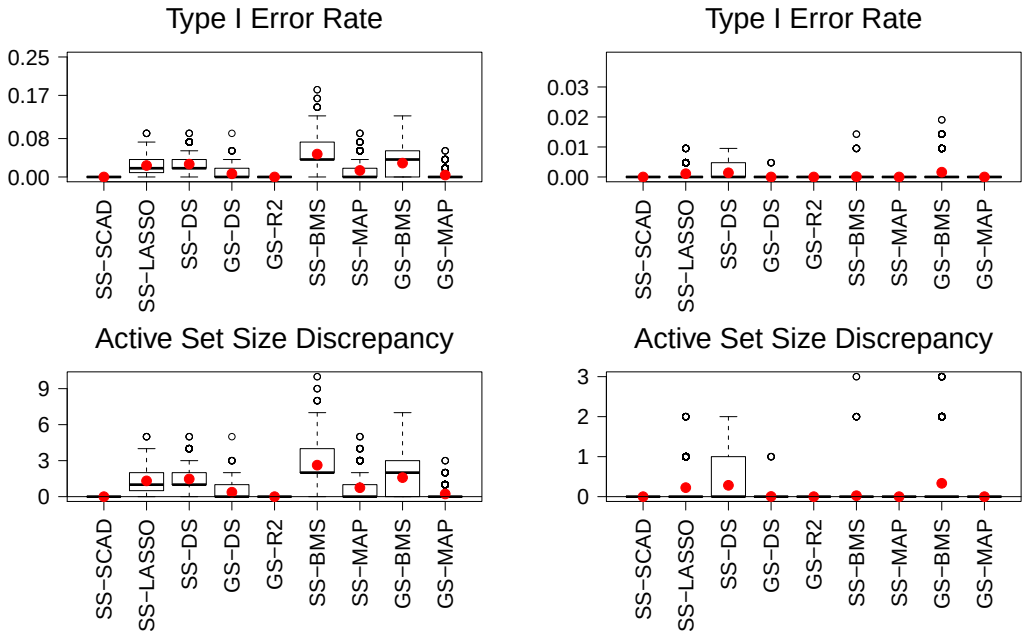


(a) $q_{me} = 0.1$, Active $N(12, 4)$, Inactive $N(0, 1)$



(b) $q_{me} = 0.15$, Active $N(24, 4)$, Inactive $N(0, 16)$

Figure 5: Performance measures for screening strategies with no active effects ($q_{me} = 0.0$)



(a) $f = 10$, Inactive $N(0, 1)$, $t = 2.5$

(b) $f = 20$, Inactive $N(0, 16)$, $t = 17$

values of q_{me}, the procedure declares all, or nearly all, effects active for the N(6,1), N(0,1) setting; for $f = 20$, SS-BMS declares most effects active for both the difficult settings. Hence this method has been removed from Figure 4(b). This tendency occurs to a far lesser extent for other settings, for example, in less than 1.5% of the simulated experiments summarised in Figures 3(b) and 4(a). The choice of scale for Type I error rate and ASD in these plots excludes the corresponding points. In general, for these difficult cases, the Bayesian methods with either supersaturated designs or group screening experience a greater drop in performance than the frequentist methods.

Overall, GS-DS seems to be the best performing procedure, followed by SS-LASSO and SS-DS which performed similarly to each other. Generally, SS-SCAD and GS-R2 tend to perform less well than the other procedures, with the latter method controlling the Type I error rate but having poor sensitivity. In general, Type I error rate is highest for N(6, 1), N(0, 1), where there is least distinction between the active and inactive effects; sensitivity is lowest for N(24, 4), N(0, 16), possibly due to the large inactive effects making it hard to detect the truly active effects. For sensitivity and Type I error, the results are poorer for larger numbers of factors.

We further evaluated the screening strategies for the separate detection of active main effects and active interactions. We found that all procedures had higher sensitivity and lower FDR for screening main effects than for interactions, or for screening main effects and interactions together. An explanation for this finding is that the direction of each active main effect was assumed known and the corresponding $\beta_u^\star$ was set positive whereas, for an active interaction, the sign of $\beta_u^\star$ was chosen at random.

Results obtained from the analysis procedures when there are no truly active case are shown in Figure 5. The methods perform similarly well except that SS-BMS and, to a lesser extent, GS-BMS have slightly poorer performance for $f = 10$ factors in Figure 5(a). Notice that the SS-BMS and GS-BMS methods do not use the threshold as a hard cut-off on the estimated parameter values and this may explain their slightly poorer performance for $f = 10$.

5. DISCUSSION AND OPEN ISSUES

5.1. Threshold for active effects

The threshold, t, defines the minimum absolute value for an effect to be classified as active. Its use can improve the performance of screening strategies as it allows overfitting in model selection to be followed by "model-pruning" to obtain low Type I error rate and FDR. In a given simulation run, we viewed t as a common threshold that applies to each regression parameter regardless of whether the effect is a main effect or interaction. In practice, however, one may wish to set the threshold differently for these different types of effects.

At the first stage of group screening, we chose not to use a threshold in our frequentist analysis. If a threshold were to be used, however, we recommend that it be adjusted for the fact that a group effect is a sum of individual effects (see (8)). The distribution of active grouped effect values has a larger variance than that of individual effect values and, under effect sparsity, this would argue for a smaller threshold for this first stage.

5.2. Design issues

In group screening, main effects of individual factors in the same group are completely aliased at stage 1, as are all two-factor interactions between factors from each of two specified groups. As in all fractional factorial experiments, there is some danger that aliased small effects may amalgamate, resulting in a non-active grouped factorial effect appearing to be active at stage 1. However, a high FDR can be corrected at stage 2 through screening out individual effects that have spuriously come forward from stage 1. There is also a small chance in group screening that aliased active effects may cancel each other and not be taken through to stage 2. This possibility can be minimized for main effects by matching the high and low levels of the factors (Section 1.3).

For an experiment with a moderate number of factors, an alternative may be to use a regular resolution III or IV fractional factorial design. However, such designs often have aliasing relationships that are too complicated for the screening setting involving two-factor interactions. For example, the 32-run regular 2^{10-5} Resolution IV fraction listed in Table 5A.3 of Wu and Hamada (2009) links eight of the ten factors in a single alias string, and the design listed by Montgomery (2009) links all ten. Although these designs may be preferable for other settings, they are not ideal for screening interactions. In contrast, for the same number of runs, group screening in conjunction with a higher resolution fraction for the grouped factors at stage 1 links at most g^2 factors together (where g is the maximum group size), no matter how many groups are present. Our simulations showed that a strategy of two-stage group screening and the Gauss-Dantzig Selector tended to produce slightly better results than the single-stage supersaturated design procedures. We believe

that its success was most likely due to the fact that sufficient unimportant factor groups were removed at stage 1 to allow the second stage to sort through many fewer correlated effects than the one-stage procedure.

At stage 2 of group screening, a regular fraction or, as in our simulations, a non-regular design can be used. For a larger number of factors, it would be possible to use supersaturated designs at both stages. Alternatively, two-stage group-screening can be extended to multiple-stage group screening. In the extreme, one could start with only two groups, and continue subdividing the active groups; a procedure called *sequential bifurcation*, (see, for example, Kleijnen, Bettonvil, and Persson (2006), and Wan, Ankenman, and Nelson (2005)). The performances of such strategies are topics for future study.

In practice, some effects may be of more interest than others. For example, if the experiment involves noise factors in addition to control factors, then usually control×noise interactions are of primary interest. Similarly, it may be possible to classify effects into classes such as "very likely" or "less likely" to be active". In group screening, it is advantageous to place the "very likely active" effects into the same group and allow the observations to shed more light on the other effects. Similarly, the groups can be formed so that effects of most interest can be estimated at the first stage; for example, by keeping control and noise factors in different groups (c.f. Vine et al. (2008)).

For supersaturated designs, our simulations confirmed the well-known fact that the particular columns which happen to be assigned to active effects have a bearing on how easily the active effects can be detected. For example, one of the situations considered by Phoa et al. (2009a) and Li and Lin (2003) involved a supersaturated design with 14 runs and 23 factors under a main effects model, where columns 1, 5, 9, 13 and 17 were assigned to active factors. We found that if these five active factors are associated, instead, with columns 1-5 of the same design, then an average of only 1.24 of the active effects were detected by the Gauss-Dantzig Selector as compared with 4.95 for the original choice of columns.

Marley and Woods (2010) showed that active effects assigned to columns having low average correlations with all other columns have a greater probability of being detected. Consequently, effects of most interest should be assigned to such columns in the design. Alternatively, if a Bayesian D-optimal supersaturated design is generated, then (more) diffuse prior distributions with $\tau \to 0$ can be assigned to effects of greater interest to force the construction of designs that provide more information on these effects.

A drawback of a two-level screening design is that factors with non-linear effects may not be detected (Laycock (2001), Torsney (2001)). One remedy is to set factor levels on the same side of an anticipated turning point (as in Vine et al., 2008), which allows an active effect with curvature to be detected via a linear component using only two levels. Another possibility is to add a centre point to a 2-level design which is common practice in response surface methodology. Although a third level for such a factor could be used, this leads to an increased number of effects to be estimated which can complicate the screening process especially in the presence of interactions. For a rapidly evolving literature on multi-level designs that could be used for moderate numbers of factors, see for example Cheng and Wu (2001), Jones and Nachtsheim (2010), Chen and Liu (2008), Liu and Lin (2009).

5.3. Simulation issues

All the methods studied in this paper are likely to be under-performing in comparison with an expert analysis of a single data set due to the need to use, for example, automatic tuning and decision rules. For instance, in the Bayesian model selection procedure, the choice of active factors would be made by inspecting the posterior model probabilities, and the posterior distribution for $\boldsymbol{\beta}$. Similarly, at the end of stage 1 of group screening, a decision is made whether or not to send each group of factors through to stage 2 (see, for example, Vine et al. (2008), where not all "active groups" were investigated at stage 2). However, in a simulation study, this type of control is not possible.

In comparison with other published simulations on some of the procedures studied in this paper, our inactive effects are quite large, being selected from a $N(0, 1)$ or $N(0, 16)$ distribution, rather than being set to zero. This reduces the success rate for all of the procedures. When compared with the previous group screening simulation of Dean and Lewis (2002), the mean sizes of the active effects are much smaller, which again reduces success rate.

Most other simulation studies in the literature fix the number of active effects, rather than the proportion of the main effects and interactions that are active, as in our study. Consequently, we explored a wider range of true models and our results are more variable. In particular, there will be a number of much harder scenarios (such as more factors or smaller active effects) "hidden" in our results, and it is for these cases that many of the methods struggle to identify the active effects. Many published studies also fix the columns of the supersaturated design assigned to the active effects. As illustrated in Section 5.2, the column choice affects the results greatly, and hence we assigned columns at random to active effects in our simulations, and computed performance measures over a large number of such column assignments.

In our study, the number of runs for the supersaturated design was set approximately equal to the median number of runs required by the group screening procedure with the same number of factors and five equal-sized groups. These decisions came from practical considerations. For a single experiment experiment, the expected size of a group screening design can be calculated theoretically (Vine et al., 2005) and a comparison made with a similar sized supersaturated design. A future study might try to match the sizes more closely and compare the performances of the various strategies on each individual run of the simulation.

There are various other extensions that could be made in order to encompass a wider range of situations. For example, on any given run of the simulation, the active main effects and the active interaction effects could be drawn from different distributions. Similarly, to mimic a scenario where effects can be categorized in advance by their likelihood of being active, the means and variances of the active effect distributions could be set differently for the different categories.

The Bayesian approaches of Section 3.2 are more computationally intensive than the frequentist methods of Section 3.1. An open problem for future study is to refine the procedures so that larger numbers of factors can be handled within the Bayesian framework.

ACKNOWLEDGEMENTS

The work of Dean and Draguljic was partly supported by grants SES-0437251 and DMS-0806134 from the National Science Foundation. Part of the work was undertaken while Dean was

visiting the Southampton Statistical Sciences Research Institute. The authors thank Dr Sarah Carnaby for help in performing the simulation studies. We note with deep regret that Dr Anna Vine passed away before the completion of this work.

REFERENCES

Abraham, B., Chipman, H., and Vijayan, K. (1999), "Some risks in the construction and analysis of supersaturated designs," *Technometrics*, 41, 135–141.

Allen, T. T. and Bernshteyn, M. (2003), "Supersaturated designs that maximize the probability of identifying active factors," *Technometrics*, 45, 90–97.

Beattie, S. D., Fong, D. K. H., and Lin, D. K. J. (2002), "A two-stage Bayesian model selection strategy for supersaturated designs," *Technometrics*, 44, 55–63.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, 57, 289–300.

Bingham, D. R. and Chipman, H. A. (2007), "Incorporating prior information in optimal design for model selection," *Technometrics*, 49, 155–163.

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

Box, G. E. P. and Hill, W. J. (1967), "Discrimination among mechanistic models," *Technometrics*, 9, 57–71.

Box, G. E. P., Hunter, J. S., and Hunter, W. G. (2005), *Statistics for Experimenters: Design, Innovation, and Discovery*, Hoboken, NJ: Wiley, 2nd ed.

Box, G. E. P. and Meyer, R. D. (1986), "An analysis for unreplicated fractional factorials," *Technometrics*, 28, 11–18.

Burnham, K. P. and Anderson, D. R. (2002), *Model Selection and Multimodel Inference*, New York: Springer, 2nd ed.

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

Chen, J. and Lin, D. K. J. (1998), "On the identifiability of a supersaturated design," *Journal of Statistical Planning and Inference*, 72, 99–107.

Chen, J. and Liu, M.-Q. (2008), "Optimal mixed-level k-circulant supersaturated designs," *Journal of Statistical Planning and Inference*, 138, 4151–4157.

Cheng, C.-S., Steinberg, D. M., and Sun, D. X. (1999), "Minimum aberration and model robustness for two-level fractional factorial designs," *Journal of the Royal Statistical Society B*, 61, 85–93.

Cheng, S.-W. and Wu, C. F. J. (2001), "Factor screening and response surface exploration (with discussion)," *Statistica Sinica*, 11, 553–604.

Chipman, H. A. (1996), "Bayesian variable selection with related predictors," *The Canadian Journal of Statistics*, 24, 17–36.

— (2006), "Prior distributions for Bayesian analysis of screening experiments," in *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*, eds. Dean, A. M. and Lewis, S. M., New York: Springer, pp. 235–267.

Chipman, H. A., Hamada, M. S., and Wu, C. F. J. (1997), "A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing," *Technometrics*, 39, 372–381.

Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions," *Numerische Mathematik*, 31, 377–403.

Dean, A. M. and Lewis, S. M. (2002), "Comparison of Group Screening Strategies for Factorial Experiments," *Computational Statistics and Data Analysis*, 39, 287–297.

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

Dorfman, R. (1943), "The detection of defective members of large populations," *Annals of Mathematical Science*, 14, 436–440.

Draguljić, D. (2010), "Design and Analysis of Computer Experiments for Screening Input Variables," Ph.D. thesis, Department of Statistics, The Ohio State University, Columbus, Ohio USA.

DuMouchel, W. and Jones, B. (1994), "A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model," *Technometrics*, 36, 37–47.

Dupplaw, D. P., Brunson, D., Vine, A. E., Please, C. P. P., Lewis, S. M., Dean, A. M., Keane, A. J., and Tindall, M. J. (2004), "A web-based knowledge elicitation system (GISEL) for planning and assessing group screening experiments for product development," *Journal of Computing and Information Science in Engineering*, 4, 218–225.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Frank, I. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools (with discussion)," *Technometrics*, 35, 109–148.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

— (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.

Georgiou, S. D., Draguljić, D., and Dean, A. M. (2009), "An overview of two-level supersaturated designs with cyclic structure," *Journal of Statistical Theory and Practice*, 3, 489–504.

Hamada, M. and Wu, C. F. J. (1992), "Analysis of designed experiments with complex aliasing," *Journal of Quality Technology*, 24, 130–137.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer, 2nd ed.

Hurvich, C. M. and Tsai, C. L. (1989), "Regression and time series model selection in small samples," *Biometrika*, 76, 297–307.

Jones, B. A., Li, W., Nachtsheim, C. J., and Ye, K. Q. (2007), "Model discrimination—another perspective on model-robust designs," *Journal of Statistical Planning and Inference*, 137, 1576–1583.

Jones, B. A., 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. A. and Nachtsheim, C. J. (2010), "A class of screening designs robust to active second-order effects," in *mODa9 – Advances in Model-Oriented design and Analysis*, ed. A. Giovagnoli, et al., New York: Springer.

Kleijnen, J. P. C., Bettonvil, B., and Persson, F. (2006), "Screening for the important factors in large discrete-event simulation models: sequential bifurcation and its applications," in *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics.*, eds. Dean, A. M. and Lewis, S. M., New York: Springer, pp. 287–307.

Laycock, P. J. (2001), "Discussion of *Detection of Interactions in Experiments on Large Numbers of Factors*," *Journal of the Royal Statistical Society B*, 63, 664.

Lewis, S. M. and Dean, A. M. (2001), "Detection of interactions in experiments on large numbers of factors (with discussion)," *Journal of the Royal Statistical Society B*, 63, 633–672.

Li, R. and Lin, D. K. J. (2002), "Data analysis of supersaturated designs," *Statistics and Probability Letters*, 59, 135–144.

— (2003), "Analysis methods for supersaturated designs: some comparisons," *Journal of Data Science*, 1, 249–260.

Li, W. (2006), "Screening designs for model selection," in *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, eds. Dean, A. M. and Lewis, S. M., New York: Springer, pp. 205–234.

Li, W. and Wu, C. F. J. (1997), "Columnwise-pairwise algorithms with applications to the construction of supersaturated designs," *Technometrics*, 39, 171–179.

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

— (1995), "Generating systematic supersaturated designs," *Technometrics*, 37, 213–225.

Liu, M.-Q. and Lin, D. K. J. (2009), "Construction of optimal mixed-level supersaturated designs," *Statistica Sinica*, 19, 197–211.

Liu, Y. and Dean, A. (2004), "K-circulant supersaturated designs," *Technometrics*, 46, 32–43.

Liu, Y., Ruan, S., and Dean, A. (2007), "Construction and analysis of $E(s^2)$ efficient supersaturated designs," *Journal of Statistical Planning and Inference*, 137, 1516–1529.

Lu, W. and Zhang, H. H. (2007), "Variable selection for the proportional odds model," *Statistics in Medicine*, 26, 3771–3781.

Marley, C. J. (2010), "Screening experiments using supersaturated designs with application to industry," Ph.D. thesis, University of Southampton.

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.

Meyer, M. A. and Booker, J. M. (2001), *Eliciting and Analyzing Expert Judgment: a Practical Guide*, Philadelphia: SIAM.

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

Montgomery, D. C. (2009), *Design and Analysis of Experiments*, New York: Wiley, 7th ed.

Moore, M. A. and Epps, H. H. (1992), "Accelerated weathering of marine fabrics," *Journal of Testing and Evaluation*, 20, 139–143.

Morris, M. D. (2006), "An overview of group factor screening," in *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, eds. Dean, A. M. and Lewis, S. M., New York: Springer, pp. 191–206.

Nguyen, N.-K. (1996), "An algorithmic approach to constructing supersaturated designs," *Technometrics*, 38, 69–73.

O'Hagan, A. and Forster, J. J. (2004), *Kendall's Advanced Theory of Statistics 2B: Bayesian Inference*, London: Arnold, 2nd ed.

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

Phoa, F. K. H., Wong, W. K., and Xu, H. (2009b), "The need of considering the interactions in the analysis of screening designs," *Journal of Chemometrics*, 23, 545–553.

Rais, F., Kamoun, A., Chaabouni, M., Claeys-Bruno, M., Phan-Tan-Luu, R., and Sergent, M. (2009), "Supersaturated design for screening factors influencing the preparation of sulfated amides of olive pomace oil fatty acids," *Chemometrics and Intelligent Laboratory Systems*, 99, 71–78.

Rose, A. D. (2008), "Bayesian Experimental Design for Model Discrimination," Ph.D. thesis, University of Southampton.

Ryan, K. J. and Bulutoglu, D. A. (2007), "$E(s^2)$-optimal supersaturated designs with good minimax properties," *Journal of Statistical Planning and Inference*, 137, 2250–2262.

Scinto, P. R., Wilkinson, R. G., and Lin, D. K. J. (2011), "Screening for fuel economy: a case study of supersaturated designs in practice," *Quality Engineering*, 23, 15–25.

Srivastava, J. N. (1975), "Designs for searching non-negligible effects," in *Statistical Design and Linear Models*, New York: Elsevier, pp. 507–520.

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

Torsney, B. (2001), "Discussion of *Detection of Interactions in Experiments on Large Numbers of Factors*," *Journal of the Royal Statistical Society B*, 63, 665.

Vine, A. E., Lewis, S. M., and Dean, A. M. (2005), "Two-stage group screening in the presence of noise factors and unequal probabilities of active effects," *Statistica Sinica*, 15, 871–888.

Vine, A. E., Lewis, S. M., Dean, A. M., and Brunson, D. (2008), "A critical assessment of two-stage group screening through industrial experimentation," *Technometrics*, 50, 15–25.

Wan, H., Ankenman, B., and Nelson, B. L. (2005), "Controlled sequential bifurcation: a new factor-screening method for discrete-event simulation," *Operations Research*, 54, 743–755.

Watson, G. S. (1961), "A study of the group screening method," *Technometrics*, 3, 371–388.

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

Wu, C. F. J. and Hamada, M. (2009), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley, 2nd ed.

Yuan, M., Joseph, V. R., and Lin, Y. (2007), "An efficient variable selection approach for analyzing designed experiments," *Technometrics*, 49, 430–439.

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the "degrees of freedom" of the lasso," *Annals of Statistics*, 35, 2173–2192.