DEPARTMENT OF ENGINEERING MANAGEMENT

A Mixed Integer Optimization Approach for Model Selection in Screening Experiments

Alan Vázquez-Alcocer, Eric D. Schoen & Peter Goos

UNIVERSITY OF ANTWERP
Faculty of Applied Economics
City Campus
Prinsstraat 13, B.226
B-2000 Antwerp
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
www.uantwerpen.be

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF ENGINEERING MANAGEMENT

A Mixed Integer Optimization Approach for Model Selection in Screening Experiments

Alan Vázquez-Alcocer, Eric D. Schoen & Peter Goos

RESEARCH PAPER 2018-007
MAY 2018

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@uantwerpen.be

D/2018/1169/007

A Mixed Integer Optimization Approach for Model Selection in Screening Experiments

Alan Vázquez-Alcocer[1], Eric D. Schoen[1, 2, 3], and Peter Goos[1, 3]

[1]University of Antwerp, Belgium
[2]TNO, Zeist, Netherlands
[3]University of Leuven, Belgium

May 7, 2018

Abstract

After completing the experimental runs of a screening design, the responses under study are analyzed by statistical methods to detect the active effects. To increase the chances of correctly identifying these effects, a good analysis method should: (1) provide alternative interpretations of the data, (2) reveal the aliasing present in the design, and (3) search only meaningful sets of effects as defined by user-specified restrictions such as effect heredity or constraints that include all the contrasts of a multi-level factor in the model. Methods like forward selection, the Dantzig selector or LASSO do not posses all these properties. Simulated annealing model search cannot handle other constraints than effect heredity. This paper presents a novel strategy to analyze data from screening designs that posses properties (1)–(3) in full. It uses modern mixed integer optimization methods that returns the results in a few minutes. We illustrate our method by analyzing data from real and synthetic experiments involving two-level and mixed-level screening designs. Using simulations, we show the capability of our method to automatically select the set of active effects and compare it to the benchmark methods.

Keywords: Dantzig selector, definitive screening design, LASSO, sparsity, two-factor interaction.

1 Introduction

Screening experiments permit the study of many factors using a small number runs. Successful screening requires the assumption that only a small proportion of the factors' effects on the responses of interest matter. This assumption is known as effect sparsity. Screening experiments are commonly carried out using two-level orthogonal screening designs (Mee et al., 2017; Schoen et al., 2017), three-level orthogonal screening designs (Cheng and Wu, 2001; Xu et al., 2004), mixed-level orthogonal screening designs (Wu and Hamada, 2009, ch. 7), or definitive screening designs (Jones and Nachtsheim, 2011). These designs have economical run sizes but posses complex aliasing structures that make the identification of the influential or *active* effects challenging.

The analysis of data from screening designs involves a model selection problem characterized by a small number of observations and a large number of effects. The goal is to find the smallest model that includes the active effects. Abraham et al. (1999) and Mee (2013) provide guidelines on the use of model selection methods to analyze data from screening designs. According to these authors' findings, a method suitable for the analysis has several specific characteristics. We formulate these characteristics as desirable properties of a good model selection method:

Property 1. A good model selection method creates a list of models that are compatible with the data.

Often, there is more than one model that explains the observed data well. Therefore, having a list of good models rather than a single overall-best model provides more information on the potentially active effects. Moreover, if the screening design used is not satisfactory, such a list of good models suggests which effects should be considered for further investigation in a follow-up experiment.

Property 2. A good model selection method reveals the aliasing present in the design.

When using a screening design to study many factors, the number of possible models is large, and, due to the limited number of runs in a screening design, many effects are aliased. In that case, some of these models will be highly aliased and therefore have a similar fit to the data. The active effects can then not be identified unambiguously. A good model selection method should reveal any aliasing present in the data and thereby recognize the limitations of the screening design used and refrain the experimenter from drawing unwarranted conclusions.

Property 3. A good model selection method allows the user to specify restrictions on the model search.

Model selection methods that possess Property 3 can restrict the search space to specific sets of meaningful models such as those obeying weak or strong effect heredity (Hamada and Wu, 1992), where it is assumed that an interaction or a quadratic effect can be active only if its corresponding main effect(s) are also active. Based on a meta-analysis of a large number of two-level factorial experiments and response surface experiments, Li et al. (2006) and Ockuly et al. (2017) showed that effect heredity generally holds in practice. It therefore makes sense to incorporate heredity restrictions in model selection. Other situations in which user-specified restrictions are useful in the context of model selection include experiments involving multi-level categorical factors. In the analysis of data from such experiments, it is generally desirable to select all contrast vectors corresponding to the categorical factor simultaneously. Finally, restrictions on the model search can also be used to set bounds on the model size or the number of factors in the model, or to avoid models that are known in advance to be misleading.

Abraham et al. (1999) and Mee (2013) demonstrated that Properties 1, 2 and 3 lead to a more informed decision on the set of active effects. Another attractive attribute of a model selection method is its capacity to analyze data from any screening design, regardless of whether it is a two-level design, a three-level design with continuous or categorical factors, or a mixed-level design. So, the following property is also desirable for a good model selection method:

Property 4. A good model selection method is applicable to any kind of design.

Model selection methods available in the literature can be categorized into shrinkage and nonshrinkage methods. Shrinkage methods (Hastie et al., 2009, ch. 3) perform model selection by biasing, or shrinking, some of the effect estimates toward zero. They include the Dantzig selector (Candes and Tao, 2007) as well as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1999), and its extensions to impose effect heredity (Yuan et al., 2007; Choi et al., 2010; Bien et al., 2013). Although the primary focus of these methods is to build good predictive models, several authors use them for analyzing data from screening designs and thus for identifying active effects (Phoa et al., 2009; Marley and Woods, 2010; Draguljić et al., 2014; Weese et al., 2015; Errore et al., 2017). Nonshrinkage methods perform model selection without biasing the effect estimates. They include forward selection (Westfall et al., 1998) and simulated annealing model search (SAMS; Wolters and

Bingham, 2011). None of these shrinkage and nonshrinkage methods possess all the four desired properties listed above.

In this paper, we present a model selection method to analyze data from screening designs that does posses Properties 1–4. Our method is based on best-subset selection (Miller, 2002) and uses modern mixed integer optimization methods to find high-quality models of any size. We introduce our method, called MIO because of its use of Mixed Integer Optimization, and discuss its strengths in Section 2. In Section 3, we illustrate the efficiency of MIO by analyzing data from synthetic and real experiments. We discuss advantages and disadvantages of the aforementioned shrinkage and nonshrinkage methods in Section 4. In Section 5, we present a simulation study to compare the performances of MIO and the benchmark methods when it comes to correctly identifying active effects. We conclude the paper and mention avenues for future research in Section 6.

Throughout the paper, we consider a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is an $n \times 1$ vector of responses and n the number of observations, $\boldsymbol{\beta}$ is a $p \times 1$ vector of p unknown parameters, $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of independent and normally distributed random errors with zero mean and variance σ^2, and $\mathbf{X}$ is an $n \times p$ model matrix. Note that $\mathbf{X}$ can include contrast vectors associated with the main effects (MEs), two-factor interactions (2FIs) and quadratic effects (QEs) of continuous factors as well as contrast vectors associated to the effects of multi-level categorical factors. We assume that the columns of $\mathbf{X}$ have been standardized to have zero means and to have the same length, and that $\mathbf{y}$ is centered around zero to exclude the intercept from the model. We denote the element in the ith row and uth column of $\mathbf{X}$ by x_{iu}. Similarly, we denote the ith element of $\mathbf{y}$ by y_i.

2 MIO in full

For any given model size up to a user-specified maximum, $k_{\max}$, MIO lists the best models in terms of the residual sum of squares (RSS). For a given size, the list contains the top M models that satisfy any user-specified model search restrictions. Due to the fact that it produces a list of models, MIO possesses Properties 1 and 2. Because of its ability to include model search restrictions, MIO also possesses Properties 3 and 4. A key feature of MIO is that it also visualizes the list of models using raster plots, which allows for detecting patterns in the selected effects. The most important factor effects are those that appear consistently in the best feasible models. These can then be declared active.

We first describe the core integer optimization procedure used by MIO to find the best-fitting models. Next, we show how to incorporate user-specified restrictions in the model search and then we introduce MIO's sequential algorithm to list the best M feasible models for any given model size. Finally, we introduce the raster plots and discuss the implementation of our MIO approach.

2.1 MIO's optimization problem

2.1.1 Basic idea

For a given model size k, MIO searches for the model that minimizes the RSS value. In other words, it seeks the model that has the best least-squares fit to the data and thus performs a best-subset selection (Miller, 2002). Essentially, best-subset selection solves the following problem:

$$\min_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \sum_{u=1}^{p} x_{iu} \hat{\beta}_u \right)^2 \text{ subject to } \sum_{u=1}^{p} I(\hat{\beta}_u \neq 0) \leq k, \tag{1}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$ is the vector of parameter estimates and $I(c)$ is an indicator function that takes the value 1 if condition c is satisfied and 0 otherwise. The nonzero $\hat{\beta}_u$ values correspond to the ordinary least squares (OLS) estimates of the parameters corresponding to the selected model terms. The constraint in the optimization problem ensures that at most k terms are selected for inclusion in the regression model. More specifically, it ensures that at most k model parameters are nonzero. The parameter k thus has a simple interpretation, and the parameter estimates produced are plain OLS estimates. So, MIO does not use any shrinkage.

The best-subset selection problem in (1) is an NP-hard problem (Natarajan, 1995), which means that it cannot be solved in polynomial time. Indeed, current state-of-the-art algorithms for best-subset selection, as implemented in SAS 9.4 or JMP 13, or in the 'leaps' package in R, do not allow to solve the problem when it involves more than 30 effects (i.e., when $p > 30$). Literally fitting all possible models for a given value of k obviously quickly becomes infeasible as the number of potential terms p increases. The 'leaps' package in R avoids a complete enumeration of all possible models by using the leaps and bounds algorithm of Furnival and Wilson (1974), which combines computationally efficient matrix operators with a branch and bound procedure. Despite these attractive features, the leaps and bounds algorithm remains fairly slow, even for moderate model sizes. Therefore, in spite of its intuitive appeal, by many, best-subset selection is considered infeasible for screening experiments involving many factors. However, due to its use of modern mixed integer optimization methods in modern state-of-the-art solvers such as Gurobi or CPLEX, our newly proposed MIO approach makes best-subset selection feasible for a broad range of screening designs.

2.1.2 Problem formulation

Mixed integer optimization is an optimization method to determine the values of a set of decision variables, which can be discrete or continuous, so as to maximize or minimize a particular linear or quadratic objective function, while satisfying a set of linear constraints (Bertsimas and Weismantel, 2005). Solvers such as Gurobi and CPLEX can be used to tackle mixed integer optimization problems. The solvers provide both feasible solutions and bounds for the objective function's optimal value. As the solver progresses toward the optimal solution, the bounds improve and provide an increasingly better guarantee of optimality, which is especially useful if the solver is stopped before it converges to the global optimum. In contrast, heuristic algorithms do not provide such a certificate of optimality.

Bertsimas et al. (2016) formulated the best-subset selection problem in (1) as a mixed integer optimization problem to solve large instances with associated certificates of optimality, using a computational time that is acceptable for practical applications. Their approach is a dramatic improvement over the leaps and bounds algorithm of Furnival and Wilson (1974) due to recent developments in computer hardware and to both theoretical and practical advances in mixed integer optimization such as cutting plane theory, disjunctive programming for branching rules, and improved heuristic and linear optimization methods (Bixby, 2012). Bertsimas et al. (2016) present two versions of the MIO problem. One is intended for problems in which the number of parameters p is smaller than the number of runs n, while the other is for problems in which the number of runs n is smaller than the number of parameters p. The two versions differ in the number of quadratic variables they use. We adopt the latter version because of the fact that $n < p$ is a basic characteristic of screening experiments, especially when considering 2FIs in addition to MEs. Therefore, our MIO problem is the following:

$$\min_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}},\mathbf{z}} \hat{\boldsymbol{\eta}}^T\hat{\boldsymbol{\eta}} - 2(\mathbf{X}^T\mathbf{y})^T\hat{\boldsymbol{\beta}} + \mathbf{y}^T\mathbf{y} \tag{2}$$

subject to

Model constraints:

$$\mathrm{SOS}_1(\hat{\beta}_u, 1 - z_u), \quad u = 1, \ldots, p, \tag{3}$$

$$\sum_{u=1}^{p} z_u \leq k, \tag{4}$$

Technical constraints:

$$\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{5}$$

$$z_u \in \{0, 1\}, \quad u = 1, \ldots, p, \tag{6}$$

Boosting constraints (optional):

$$-B \leq \hat{\beta}_u \leq B, \quad u = 1, \ldots, p, \tag{7}$$

$$\sum_{u=1}^{p} |\hat{\beta}_u| \leq B^L, \tag{8}$$

$$-E \leq \hat{\eta}_i \leq E, \quad i = 1, \ldots, n, \tag{9}$$

$$\sum_{i=1}^{n} |\hat{\eta}_i| \leq E^L. \tag{10}$$

In this problem formulation, $\hat{\beta}_u$ represents the uth element of $\hat{\boldsymbol{\beta}}$, z_u is a binary variable associated with it, $\hat{\eta}_i$ is the ith element of the $n \times 1$ vector $\hat{\boldsymbol{\eta}}$, $\mathbf{z} = (z_1, z_2, \ldots, z_p)^T$, and B, B^L, E and E^L are auxiliary constants larger than zero. This formulation involves $p + n$ continuous decision variables contained within $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$, p binary decision variables z_u and $3p + 2n + 3$ constraints. The binary variable z_u is zero when the corresponding parameter $\hat{\beta}_u$ is set to zero. It takes the value 1 when the corresponding model parameter is nonzero. A z_u variable taking the value one indicates that the term corresponding to the uth column of $\mathbf{X}$ belongs to the best subset, with an effect estimate equal to $\hat{\beta}_u$.

The objective function (2) expresses the RSS value in the objective function in problem (1) in terms of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$. In the new objective function, the vector $\mathbf{X}\hat{\boldsymbol{\beta}}$ is replaced by $\hat{\boldsymbol{\eta}}$, so that the number of quadratic terms in the problem formulation is n rather than p. This feature improves the computing time required by the solver when p is much larger than n, which is typical for screening experiments.

The problem formulation involves three kinds of constraints: model constraints, technical constraints and boosting constraints. There are two types of model constraints. The first type involves specially ordered sets of type 1 (SOS_1; Beale and Forrest, 1976). A specially ordered set of type 1 is a set of decision variables at most one of which can be different from zero. Using such sets generally speeds up branch and bound algorithms. The specially ordered set in (3) implies that $\hat{\beta}_u$ or $1 - z_u$ is zero. In other words, it implies that either the uth model term is selected and can have a nonzero parameter estimate, or that the uth term is not selected and has a zero parameter estimate. The second type of model constraint in (4) ensures that at most k model terms are selected and have a nonzero parameter estimate.

The technical constraint in (5) takes care of the substitution of vector $\mathbf{X}\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\eta}}$ in the objective function. The constraints in (6) ensure that the variables z_u are binary.

The boosting constraints in (7)–(10) are optional: we do not need these constraints for the MIO approach to provide optimal solutions. However, the boosting constraints avoid the need to explore all possible real values for the elements of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$, which significantly improves the computing time required by the solver to certify optimality. The constants B and E bound the absolute $\hat{\beta}_u$ and $\hat{\eta}_u$ values, respectively, while B^L and E^L bound the sums of the absolute $\hat{\beta}_u$ and $\hat{\eta}_u$ values. Bertsimas et al. (2016) showed that the elements of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\eta}}$ satisfy constraints (7)–(10) when the following bounds are used:

$$B = \tau\beta^\star,$$

$$B^L = kB,$$

$$E = \left(\max_{i=1,\ldots,n} \|\mathbf{x}_i\|_{1:k}\right) B,$$

and

$$E^L = \min\left\{B^L \sum_{i=1}^{n} \|\mathbf{x}_i\|_\infty, \sqrt{n}\|\mathbf{y}\|_2\right\}.$$

In these expressions, $\beta^\star$ is the smallest absolute $\hat{\beta}_u$ value known to be infeasible, $\tau \geq 1$ is a constant to safeguard against a misspecification of $\beta^\star$, $\mathbf{x}_i$ represents the ith row of matrix $\mathbf{X}$, $\|\mathbf{x}_i\|_r$ is the l_r-norm of vector $\mathbf{x}_i$, and $\|\mathbf{x}_i\|_{1:k}$ is the sum of the k largest absolute entries of $\mathbf{x}_i$. Note that our expression for E^L differs from that in Equation (2.12) in Theorem 2.1 of Bertsimas et al. (2016): $\sqrt{k}$ in their Equation (2.12) needs to be replaced by $\sqrt{n}$ (see their Supplementary Section 8.3). Borrowing ideas from projected gradient descent methods in first-order convex optimization problems (Nesterov, 2004, 2013), Bertsimas et al. (2016) proposed a data-driven algorithm to specify the value of $\beta^\star$.

After solving the MIO problem to optimality, the output is twofold. First, the nonzero z_u values indicate the best subset, i.e., the set of k model terms that minimize the RSS value. Second, the corresponding $\hat{\beta}_u$ values are the OLS estimates of the parameters corresponding to the selected model terms. Note that an optimal solution to the MIO problem in (2)–(10) always involves a model with the maximum number of effects, k, specified by constraint (4). This is because adding a term to a model will always result in a RSS value that is at least as good as the previous one.

2.2 Search restrictions

Bertsimas and King (2016) explained that the MIO problem can include linear constraints to incorporate subject-matter expertise in the model selection. These constraints have the form $\mathbf{a}^T\mathbf{z} \leq b$, where $\mathbf{a}$ is a $p \times 1$ vector, $\mathbf{z}$ is a $p \times 1$ vector of binary decision variables, and b is a constant. In this section, we show that, due to the possibility to add such constraints, MIO possesses Properties 1, 3 and 4. More specifically, we show that the constraints allow us to impose weak and strong effect heredity, to limit the number of factors in a model, to properly deal with multi-level categorical factors and to exclude specific models from the model search.

2.2.1 Effect heredity

Restricting the model search to the class of models obeying effect heredity is a well-established practice among practitioners searching for active second-order effects, i.e., 2FIs and QEs. The meta-analyses of Li et al. (2006) and Ockuly et al. (2017) of large numbers of published data sets from two-level factorial experiments and from response surface experiments, respectively, provide empirical support for the use of effect heredity in model selection. More specifically, both studies revealed that second-order effects are much more likely to be active when their corresponding MEs are active. This implies that, when searching for good models, it is generally a waste of effort to study models including certain 2FIs and QEs, but not the corresponding MEs.

Two types of effect heredity are commonly used: strong and weak heredity. Under strong heredity, a 2FI can be included in the model only if both of the corresponding MEs are considered active and therefore included in the model as well. Under weak heredity, a 2FI can be included in the model when at least one of the corresponding MEs is considered active and included in the model too. Under both strong and weak heredity, the QE of a factor can be considered for inclusion in the model provided its ME has already been included because it is considered active. So, for QEs, weak and strong heredity are equivalent. Note that effect heredity should not be confused with functional marginality (Nelder, 1977, 1994). Functional marginality states that, if a second-order effect is included in a model, the corresponding MEs must enter the model too, regardless of whether these MEs are active.

To see how heredity can be imposed in the MIO approach, denote the number of factors in the data set under investigation by m, the binary variable associated with the 2FI between the factors u and v by z_{uv}, the binary variable associated with the quadratic effect of factor u by z_{uu}, and the binary variables associated with the MEs of the factors u and v by z_u and z_v. To impose strong heredity for the 2FIs, we should either add the constraints

$$2z_{uv} \leq z_u + z_v, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m, \tag{11}$$

to the MIO formulation in (2)–(10), or, equivalently, the constraints

$$z_{uv} \leq z_u, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m,$$

and

$$z_{uv} \leq z_v, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m.$$

We prefer the former option because it requires the addition of fewer constraints ($m(m-1)/2$ instead of $m(m-1)$). To impose weak heredity for the 2FIs, we should add the constraints

$$z_{uv} \leq z_u + z_v, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m. \tag{12}$$

To impose (weak or strong) heredity for the QEs, we need to add the constraints

$$z_{uu} \leq z_u, \quad u = 1, \ldots, m. \tag{13}$$

The meta-analysis of data from response surface experiments by Ockuly et al. (2017) indicated that 2FIs are also more likely to be active when the QEs of the factors involved are active. This motivated these authors to introduce the concepts of strong and weak quadratic/interaction heredity. Under strong quadratic/interaction heredity, a 2FI can be included in the model only if the QEs of both factors involved are considered active and therefore included in the model. Under weak quadratic/interaction heredity, a 2FI can be included in the model when the QE of at least one of the factors is considered active and therefore included in the model. By adding the constraints

$$2z_{uv} \leq z_{uu} + z_{vv}, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m, \tag{14}$$

to the MIO formulation, we can enforce strong quadratic/interaction heredity in the search for good models, while, by adding the constraints

$$z_{uv} \leq z_{uu} + z_{vv}, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m, \tag{15}$$

we can enforce weak quadratic/interaction heredity.

2.2.2 Factor sparsity

The results of screening experiments can be analysed effectively only if the assumption of effect sparsity holds. According to this assumption, just a small proportion of the factor's effects on the responses of interest matter. Sometimes, it is assumed instead, or in addition, that only a limited number of factors drive the responses. This assumption is known as factor sparsity (Box and Meyer, 1986).

We can embed factor sparsity in the MIO approach by imposing an upper bound on the number of factors that can be included in the selected model. This requires additional constraints, involving a new kind of binary decision variable w_u, which takes the value 1 if the uth factor is included in the model and the value 0 if it is not. For instance, for a continuous factor u, w_u should take the value 1 as soon as the ME, one of the 2FIs or the QE of that factor enter the model. To this end, we can add the following four types of constraints to a MIO formulation involving only continuous factors:

$$\begin{aligned} z_u &\leq w_u, \quad u = 1, \ldots, m, && (16)\\ 2z_{uv} &\leq w_u + w_v, \quad u = 1, \ldots, m-1;\ v = u+1, \ldots, m, && (17)\\ z_{uu} &\leq w_u, \quad u = 1, \ldots, m, && (18)\\ w_u &\in \{0, 1\}, \quad u = 1, \ldots, m. && (19) \end{aligned}$$

The key constraint imposing factor sparsity is given by

$$\sum_{u=1}^{m} w_u \leq f, \tag{20}$$

where f is the maximum number of factors that is allowed to enter the model.

It is possible to simultaneously impose effect sparsity and factor sparsity in the MIO approach, by incorporating the constraint in Equation (4) (which imposes effect sparsity) as well as the constraints in Equations (16)–(20) (which impose factor sparsity) simultaneously.

2.2.3 Categorical factors

The levels of an l-level categorical factor are coded using a set of $l-1$ contrast vectors. So, adding the ME of an l-level categorical factor to a model implies the addition of $l-1$ terms and the simultaneous estimation of $l-1$ additional parameters. To ensure that the MIO model selection procedure either enters all $l-1$ terms in the model or none of them, we need to add the following constraint to the MIO problem:

$$z_{j_1} = \cdots = z_{j_{l-1}}, \tag{21}$$

where $j_1, \ldots, j_{l-1}$ identify the columns of $\mathbf{X}$ containing the $l-1$ contrast vectors associated with the l-level categorical factor f, and j_r denotes the rth contrast vector of factor f. We refer to (21) as a grouping constraint. A grouping constraint ensures that the estimates of the $l-1$ parameters corresponding to the categorical factor's ME are either all zero at the same time or not.

Grouping constraints can be used together with heredity constraints to impose strong or weak effect heredity for categorical factors. For instance, let G_a denote the set of columns of $\mathbf{X}$ containing the contrast vectors associated with an l_1-level categorical factor a, G_b denote the set of columns of $\mathbf{X}$ containing the contrast vectors associated with an l_2-level categorical factor b, and $G_{a\times b}$ denote the set of columns of $\mathbf{X}$ containing the $(l_1-1)(l_2-1)$ contrast vectors associated with the 2FI involving a and b. Strong effect heredity can then be imposed by adding one constraint of the type

$$2z_q \leq z_i + z_j, \tag{22}$$

where $q \in G_{a\times b}$, $i \in G_a$ and $j \in G_b$, to the MIO problem, together with the grouping constraints for G_a, G_b and $G_{a\times b}$. Which decision variables z_q, z_i, z_j are used in this constraint does not impact the final solution, provided they correspond to elements of $G_{a\times b}$, G_a and G_b, respectively. Constraint (22) implies that the $(l_1-1)(l_2-1)$ interaction contrast vectors can only be entered into the model when the MEs of the two factors involved are considered active and all $(l_1-1)+(l_2-1)$ ME contrast vectors are therefore included in the model. To impose weak heredity rather than strong heredity, we need to replace the coefficient of z_q in the left-hand side of constraint (22) by 1.

We can also develop expressions similar to (16) and (17) which, together with grouping constraints, impose factor sparsity for categorical factors. For a categorical factor a, w_a should take the value 1 as soon as the l_1-1 parameters corresponding to its ME or the $(l_1-1)(l_2-1)$ parameters corresponding to one of its 2FIs with another categorical factor b having l_2 levels, enter the model. This can be achieved by including the grouping constraints for G_a, G_b and $G_{a \times b}$ in the MIO formulation, together with constraint (20) and the following three types of constraints:

$$\begin{aligned} z_i &\leq w_a, \quad i \in G_a, \\ 2z_q &\leq w_a + w_b, \quad q \in G_{a \times b}, \ \forall b \neq a, \\ w_a &\in \{0, 1\}. \end{aligned}$$

2.2.4 Subset constraints

In some cases, we may want to exclude certain models from the space of models MIO is exploring. For instance, we may want to avoid models or combinations of effects that are known in advance to be misleading. Also, after having identified the overall best model, we may want to identify the second best model. This can be done by solving the MIO problem with an additional constraint that states that the overall best model can no longer be selected.

To achieve this, we have to define a subset of terms that cannot enter into the model simultaneously. If we denote that subset by S and its cardinality by $|S|$, then the constraint

$$\sum_{u \in S} z_u \leq |S| - 1 \tag{23}$$

ensures that not all terms in S can be included in the model simultaneously. We refer to (23) as a subset constraint. In Section 2.3, subset constraints play a key role in our algorithmic approach to create lists of best-fitting models.

In certain cases, it may be useful to consider only models that contain a specific set of terms. For example, when analyzing data from a blocked experiment, it is recommended to include the main effect(s) of the blocking factor(s) in all models under investigation. In such cases, we need to define another set, say S', of terms that must be included in the model, and add the following constraint to the MIO formulation:

$$\sum_{u \in S'} z_u = |S'|. \tag{24}$$

2.3 A sequential algorithm to list the best models

When looking for suitable models, we often desire a list of the best fitting models. The MIO approach is ideal for creating that list without enumerating and evaluating all possible models. To create lists of the best M fitting models of given sizes k, we need to embed the MIO approach in a sequential algorithm. The basic idea of the sequential algorithm is to add subset constraints sequentially to the MIO formulation so that previously selected models are removed from the model search. The outline of our sequential algorithm is shown in Algorithm 1.

Besides the model matrix and the response vector, the input to the algorithm consists of a MIO formulation (which may include user-specified search restrictions, for instance to impose some form of heredity), the maximum model size $k_{\max}$, and the number of models, M, to be generated for each model size. The algorithm begins by initializing the list of models L and the value of k, which indicates the current model size (see lines 1 and 2 in Algorithm 1). Next, the algorithm starts creating the list of the M best models with one term. To this end, the algorithm first solves the original MIO problem with $k = 1$, and adds the optimal subset of terms to the list L (see lines 6 and 7 in Algorithm 1). The algorithm then adds a subset constraint of the type (23) to the MIO formulation to exclude the best fitting model from the search space (see line 8), and solves the resulting MIO problem. This leads to the second best model, which is also added to the list L and for which a new subset constraint is added to the MIO formulation. Solving that new modified MIO problem produces the third best fitting model. This procedure continues until the M best models with one term have been identified (see lines 9–13 in Algorithm 1). The list with M models for $k = 1$ is thus obtained by solving M different MIO formulations, each with one extra subset constraint.

As soon as the exploration of models with one term is finished, the algorithm shifts its attention to models with two terms and creates the list of M best fitting models with two terms, again by solving M different MIO formulations with an increasing number of subset constraints, this time with $k = 2$. The whole procedure is repeated until the M best models with $k_{\max}$ model terms have been found. The output of the algorithm is the list L containing $Mk_{\max}$ models.

For most data sets and MIO formulations used in practice, the number of terms in the optimal solution of a MIO problem will have exactly k terms. In some cases however, the optimal solution will involve fewer than k terms. For example, consider a model selection problem involving the MEs of two three-level categorical factors and assume that grouping constraints have been added to the MIO formulation, to ensure that the two contrast vectors corresponding to each ME are either included in the model together or not. In that case, it is only possible to construct models of sizes two and four. Solving the MIO formulation

Algorithm 1: Sequential algorithm to generate lists of optimal models.

```
Input: $\mathbf{X}$, $\mathbf{y}$, MIO problem, $k_{\max}$ and $M$
Set $L \leftarrow \varnothing$
Set $k \leftarrow 0$
while $k < k_{\max}$ do
    $k \leftarrow k + 1$
    Set $i \leftarrow 1$
    $S_{k,i} \leftarrow$ optimal subset obtained by solving the MIO problem with at most $k$ effects
    $L \leftarrow L \cup \{S_{k,i}\}$
    Add the constraint $\sum_{u \in S_{k,i}} z_u \leq k - 1$ to the initial MIO formulation
    while $i < M$ do
        $i \leftarrow i + 1$
        $S_{k,i} \leftarrow$ optimal subset obtained by solving the current MIO problem
        $L \leftarrow L \cup \{S_{k,i}\}$
        Add the constraint $\sum_{u \in S_{k,i}} z_u \leq k - 1$ to the current MIO formulation
Output: List $L$ of $M$ best subsets for each $k$
```

with a k value of 3 will then produce a model that involves only two terms.

Two strengths of our sequential algorithm are that it avoids the need for a complete enumeration of all possible models and guarantees that the best subset is found for each value of k. Compared to existing algorithms for best-subset selection, our algorithm has the advantage that the list of models satisfies any user-specified search restrictions. The sequential algorithm ensures that the MIO approach possesses Properties 1 and 2, because the lists of the M best models for the various k values provide alternative interpretations of the data and highlight the aliasing of the effects. More specifically, if two effects frequently alternate between the models in the list, this may indicate that these effects are highly aliased.

Ideally, the value of $k_{\max}$ is obtained from subject-matter experts based on their interpretation of effect sparsity. However, in the absence of subject-matter expertise, several guidelines are available in the literature. Based on a simulation study involving two-level screening designs, Marley and Woods (2010) argue that the number of observations should be at least three times the number of possibly active effects. Following the advice from these authors, we can set $k_{\max} = \lfloor n/3 \rfloor$ when analyzing data from an n-run screening design, as we should not be expect to identify more than $\lfloor n/3 \rfloor$ effects with only n observations. Wolters and Bingham (2011) suggest adding 2, 3 or 4 units to that $k_{\max}$ value, so that overfitted models are also explored. Miller and Sitter (2004) demonstrated that a two-level screening design that permits the estimation of all models including the MEs and up to h 2FIs, can correctly identify up to $h/2$ 2FIs. So, for these designs, we can include all MEs in the model by adding the constraint (24) on these effects in the MIO formulation, and set $k_{\max}$ to $m + h/2$ to detect 2FIs.

2.4 Raster plots

After obtaining the list L of the best fitting models for the different model sizes k, we recommend visualizing the models by means of a raster plot. Raster plots were introduced in the literature on the analysis of screening experiments by Wolters and Bingham (2011), and they complement our sequential MIO algorithm very well. For this reason, we view raster plots as a key component of our MIO approach. A raster plot shows the effects considered on the horizontal axis and the models from the list L on the vertical axis. The models on the vertical axis are ranked according to the RSS value. The best-fitting models are located at the bottom of the raster plot. Additionally, the largest effect estimates in absolute value are visualized by the darkest cells in the plot.

2.5 MIO implementation

We implemented our MIO approach in Gurobi v.6.5.2 and Python. Our implementation includes the MIO formulation in Equations (2)–(10), Algorithm 1 and raster plots. Moreover, by default, it uses the boosting constraints of the MIO problem and specifies the value of $\beta^\star$ using the discrete first-order algorithm in Bertsimas et al. (2016, sec. 3). Since the discrete first-order algorithm provides promising solutions to the standard MIO problem only, we use $\tau = 2$ to safeguard against a misspecification of $\beta^\star$ in the boosting constraints of MIO problems including user-specified constraints. The implementation of our MIO approach is available upon request.

3 Proofs of concept

In this section, we demonstrate the power of the MIO approach by analyzing data from one synthetic experiment and two real-life experiments. The analysis of the data from the synthetic experiment, involving a three-level design, illustrates how Properties 1, 2 and 3 of the MIO approach help to identify the active effects. The next demonstration, involving data from a real-life experiment with two-level factors studied by Mee (2013), shows the computational power of the MIO approach. Finally, using a second real-life experiment, involving a design with seven two-level factors and two four-level categorical factors, we demonstrate that the MIO approach can deal with multi-level categorical factors. The variety in the three examples considered in this section demonstrates that MIO possesses Property 4.

Example 1. We simulated data for a definitive screening design (DSD; Jones and Nachtsheim, 2011) involving ten quantitative factors and 21 runs. Table 1 shows the treatment combinations of this design, along with the simulated responses. The DSD has orthogonal columns for the MEs, and these columns are orthogonal to all columns of $\mathbf{X}$ that correspond to second-order effects. The maximum absolute correlation between two second-order effect columns of $\mathbf{X}$ is 3/4, implying that some of the second-order effects are strongly aliased. We simulated the responses in Table 1 using the following model:

$$Y_i = 2\mathrm{A} + 2\mathrm{C} + 2\mathrm{BC} + \mathrm{CD} + 4\mathrm{C}^2 + 4\mathrm{D}^2 + \epsilon_i, \tag{25}$$

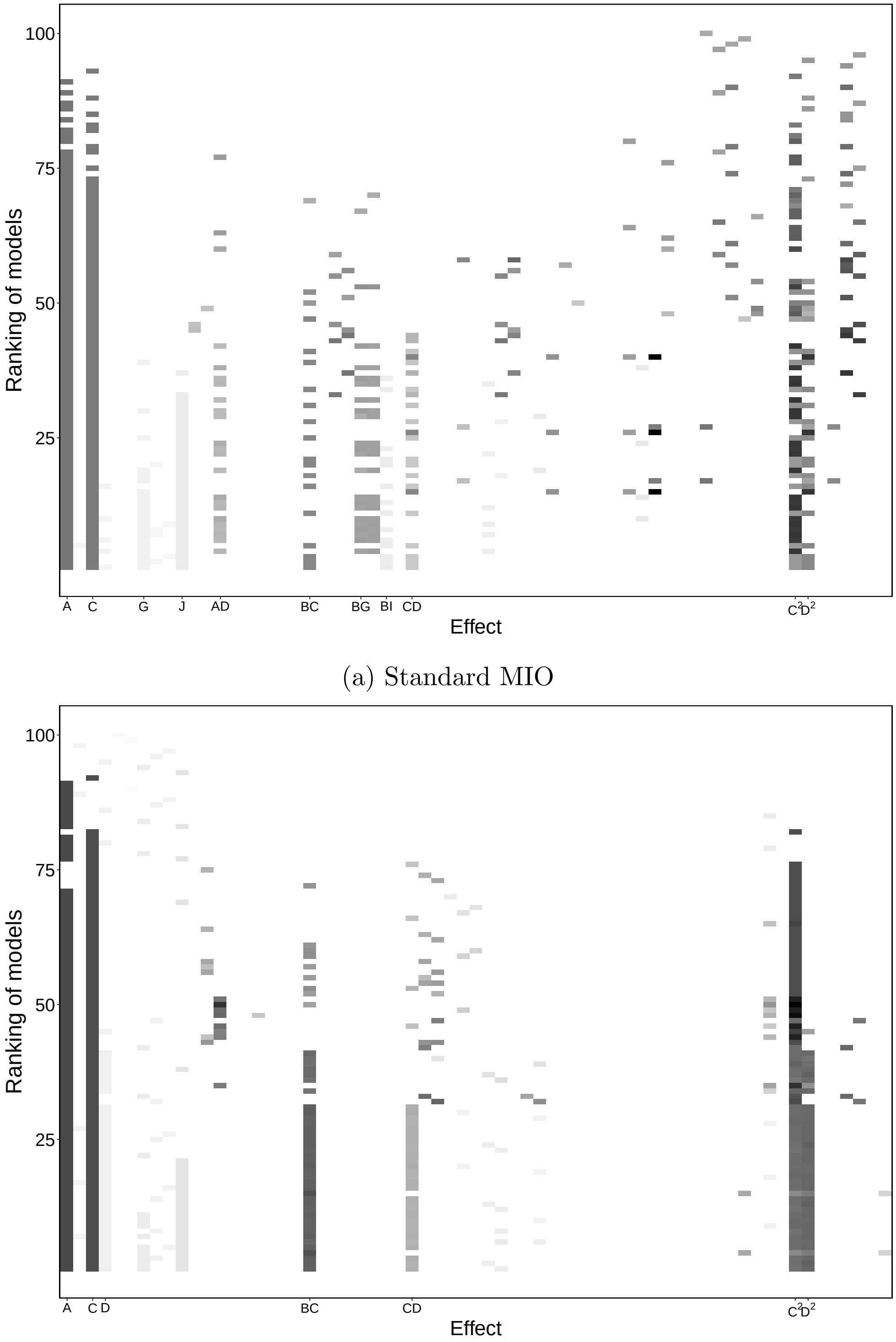
Table 1: Design and response vectors for the synthetic experiment.

Row	A	B	C	D	E	F	G	H	I	J	Y
1	0	1	1	1	1	1	1	1	1	1	13.61
2	0	−1	−1	−1	−1	−1	−1	−1	−1	−1	8.89
3	1	0	−1	−1	−1	−1	1	1	1	1	9.27
4	−1	0	1	1	1	1	−1	−1	−1	−1	8.86
5	1	−1	0	−1	1	1	−1	−1	1	1	5.59
6	−1	1	0	1	−1	−1	1	1	−1	−1	2.07
7	1	−1	−1	0	1	1	1	1	−1	−1	5.28
8	−1	1	1	0	−1	−1	−1	−1	1	1	5.78
9	1	−1	1	1	0	−1	−1	1	−1	1	11.62
10	−1	1	−1	−1	0	1	1	−1	1	−1	3.29
11	1	−1	1	1	−1	0	1	−1	1	−1	10.84
12	−1	1	−1	−1	1	0	−1	1	−1	1	2.81
13	1	1	−1	1	−1	1	0	−1	−1	1	6.1
14	−1	−1	1	−1	1	−1	0	1	1	−1	4.32
15	1	1	−1	1	1	−1	−1	0	1	−1	5.53
16	−1	−1	1	−1	−1	1	1	0	−1	1	6.03
17	1	1	1	−1	−1	1	−1	1	0	−1	12.56
18	−1	−1	−1	1	1	−1	1	−1	0	1	5.8
19	1	1	1	−1	1	−1	1	−1	−1	0	13.52
20	−1	−1	−1	1	−1	1	−1	1	1	0	4.59
21	0	0	0	0	0	0	0	0	0	0	−0.33

where $\epsilon_i \sim N(0, 0.5^2)$. This model has two large QEs, namely C^2 and D^2, the former of which obeys effect heredity. Both 2FIs in the model exhibit weak effect heredity, and the interaction involving C and D obeys strong quadratic/interaction heredity. The detection of the active effects in model (25) is challenging because of the aliasing among the second-order effects in the DSD and the large number of effects typically considered when analyzing data from DSDs: 10 MEs, 45 2FIs and 10 QEs.

We applied the sequential algorithm from Section 2.3 to the data in Table 1 with $k_{\max} = \lfloor n/3 \rfloor + 3$, following the advice of Wolters and Bingham (2011), and generated the $M = 10$ best models for each value of $k \in \{1, \ldots, 10\}$. Initially, we utilize the original MIO formulation in (2)–(10), without adding any extra constraints. So, initially, we did not impose heredity. Figure 1a shows the raster plot for the effect estimates in each of the 100 models created in this way.

The raster plot in Figure 1a shows that the estimates of the MEs of A and C are



(a) Standard MIO

(b) MIO with weak heredity constraints

Figure 1: Raster plots for the synthetic experiment in Example 1 obtained from standard MIO and MIO with weak heredity constraints for the second-order effects.

consistently large in the best models. Therefore, we can declare these MEs active. The successful detection of these effects is, in part, due to the good aliasing properties for the MEs in the DSD. Other effects that are frequently included in the best models are the truly active second-order effects (BC, CD, C^2 and D^2) in addition to the 2FIs AD, BG and BH. All other effects either have small estimates or appear in very few of the good models. Interestingly, Figure 1a shows that the best models either include the set of second-order effects $\{$BC, CD, $D^2\}$, or the set $\{$AD, BG, BH$\}$. A close inspection of these sets reveals that their effects are highly aliased when using the DSD.

To prevent non-hereditary models from entering the list of good models, we also applied the MIO approach with the heredity constraints for the 2FIs and the QEs in (12) and (13), respectively, and the weak quadratic/interaction constraints in (15) for the 2FIs. The resulting raster plot in Figure 1b then clearly identifies the truly active effects. The plot greatly benefits from the search restrictions as it more consistently identifies the true model. Note that, due to the heredity constraints, the models including D^2 must also include the (inactive) ME of D. So, imposing heredity forces the MIO approach to incorporate an inactive effect in the model. We do not consider this problematic because the light color corresponding to the ME estimates for factor D in the raster plot suggests that the ME is close to zero and should therefore not be declared active. This shows that using MIO with heredity constraints is able to identify true models that violate effect heredity.

This example showed the potential of MIO to generate a list of models compatible with the data (Property 1), elucidate the aliasing in the effects (Property 2) and use model search restrictions (Property 3) when identifying active effects.

Example 2. Schoen and Mee (2012) described a 48-run experiment carried out at TNO Science and Industry in the Netherlands to identify the best diamond-turning process for a mirror. The goal of the 13-factor experiment was to detect the active MEs and 2FI effects on the mean surface roughness of the diamonds. The experimenters used a two-level orthogonal design in which the 13 ME contrast vectors are orthogonal to all 78 2FI contrast vectors and the 2FI contrast vector pairs have absolute correlations of at most 1/3. Mee (2013) analyzed the data for this experiment using the method of Lenth (1989), forward selection and SAMS, and concluded that the MEs of factors A, B, E, G and I and the 2FIs AD, BE and GI are active.

We re-analyzed the data using MIO with weak heredity constraints for the 2FIs. For illustrative purposes, we set the value of $k_{\max}$ to the number of active effects (according to Mee (2013)) plus 3 and we generated the 10 best models for each $k \in \{1, \ldots, 11\}$. Even though this example requires searching a space of 5.46×10^{13} models, the sequential MIO algorithm takes less than 45 minutes to find the list of optimal models on a standard CPU (Intel(R) Core(TM) i7 processor, 2.8Ghz, 8 GB). We also searched for the best models using the best-subset selection algorithms in JMP 13 and the 'leaps' package in R. JMP 13 could not scale up to models of size 8, even when imposing strong heredity. The best-subset selection algorithm in the 'leaps' package did not finish the search for the best models within three hours.

Figure 2 shows the raster plot for the 110 models we found. The figure suggests the presence of nine active effects. More specifically, the raster plot shows that the MEs of A, B, E, G and I, and the interactions AD, BE, EH and GI have large estimates (in absolute value) and frequently appear in the best-fitting models. The remaining effects do not show any pattern. Therefore, MIO supports the findings of Mee (2013) and suggests that also the interaction EH is active.

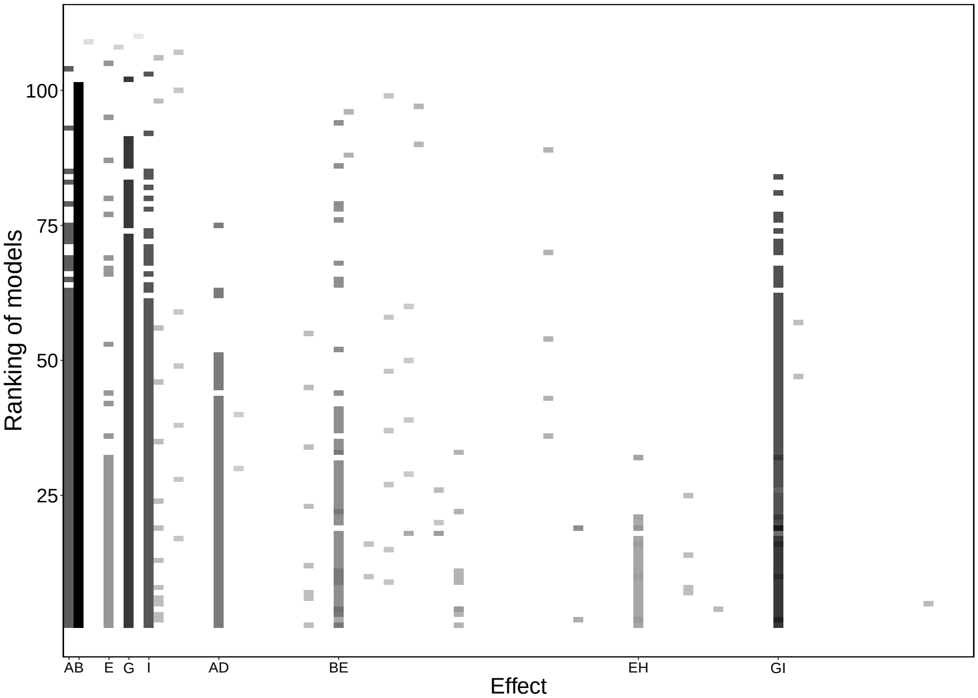


Figure 2: Raster plot obtained from MIO for the mirror-polishing experiment in Example 2. All models follow weak effect heredity.

Example 3. Phadke (1986) discussed a router bit experiment involving seven two-level factors and two four-level categorical factors. The goal of the experiment was to find a model that explains the router-bit life. The data from the experiment was analyzed in Wu and Hamada (2009, ch. 7). The two-level factors were labeled using the letters A–C and F–J, whereas the four-level categorical factors were labeled using the letters D and E. In the analysis of Wu and Hamada (2009, ch. 7), factor D was replaced by three two-level contrast vectors labeled D_1, D_2 and D_3. Similarly, factor E was replaced by three two-level contrast vectors labeled E_1, E_2 and E_3. The design used for this experiment was a regular resolution-III design. This means that some MEs are fully aliased with 2FIs; details regarding the aliasing structure of this design are discussed by Wu and Hamada (2009, ch. 7).

Following Phadke (1986), we considered the MEs of all factors and only the 2FIs involving the two-level factors. We use MIO with the strong heredity constraints for the 2FIs in (11) and the grouping constraints in (21) for the MEs of the four-level factors D and E. We generated the 10 best models for each $k \in \{1, \ldots, 14\}$, where $k_{\max}$ was set to $\lfloor n/3 \rfloor + 4$, following the advice of Wolters and Bingham (2011). The corresponding raster plot is shown in Figure 3. The figure shows that the MEs of the factors F, G and J, and the 2FIs GJ and AF are active since they appear consistently in the best models and have large estimates. Other effects that may also be active are the ME of factor B and the 2FI AG. In addition, the estimate corresponding to the contrast vector D_2 is large, which suggests that the four-level categorical factor D is active. Note that, whenever D_2 is included in the model, the contrast vectors D_1 and D_3 are also included. The parameter estimates corresponding to D_1 and D_3 are quite large in absolute value too, as witnessed by the rather dark band for these contrast vector in the raster plot.

The raster plot in Figure 3 suggests that the ME of factor D and the interaction AG are fully aliased because the interaction AG does not appear in the models with the three contrast vector corresponding to factor D, and vice versa. Two other effects that, when studying the raster plot, seem fully aliased are the ME of factor B and the interaction AC. The raster plot therefore confirms the aliasing patterns identified by Wu and Hamada (2009, ch. 7). More specifically, these authors point out that the ME of factor B and the interaction AC are fully aliased, and that the parameter corresponding to the contrast vector D_2 is fully aliased with the interaction AG. Therefore, it is more likely that the MEs of the factors B and D are active than the 2FIs involving the factor A.

Phadke (1986) declared the MEs of B, D, F, G and J, and the interaction GJ active, while Wu and Hamada (2009, ch. 7) concluded that the MEs of D, G and J, and the interactions GJ and AF were active. The results from our MIO approach confirm all the active effects discovered by Phadke, but they also suggest that the interaction AF is active, which is in line with the results of Wu and Hamada. Wu and Hamada attributed the influence of AF to a fully aliased interaction involving the categorical factor D and the two-level factor H, which is an interaction effect we did not consider in our analysis. Only a follow-up experiment would allow these two interactions to be de-aliased.

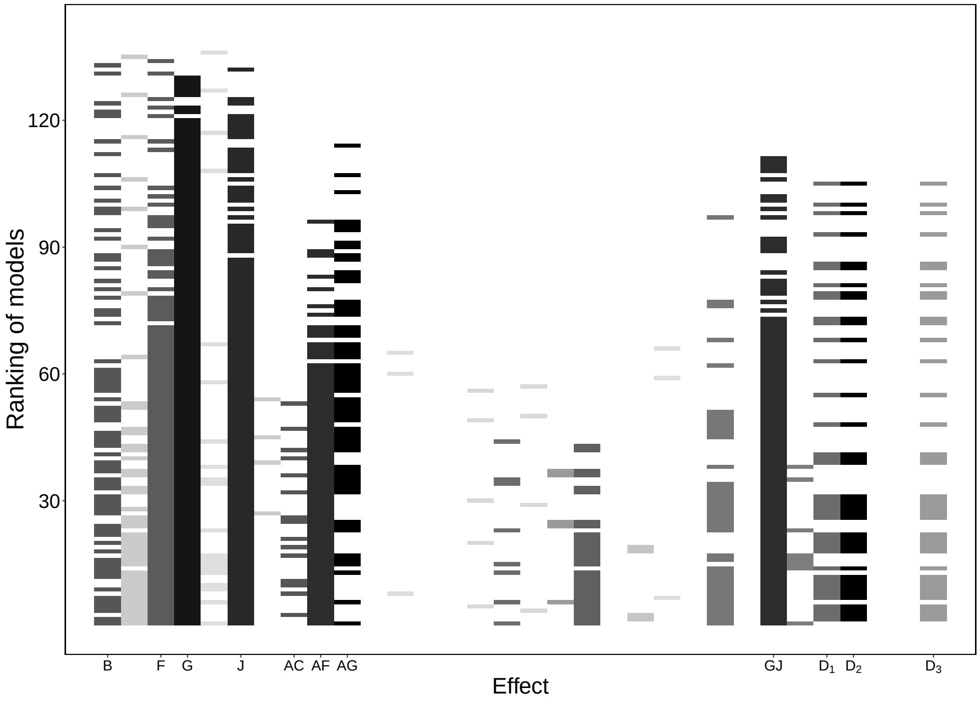


Figure 3: Raster plot obtained from MIO for the router-bit experiment in Example 3. The effects of the four-level factors are shown on the right. All models obey strong effect heredity.

4 Existing methods

In this section, we briefly discuss advantages and disadvantages of existing shrinkage and nonshrinkage methods to analyze data from screening designs. A detailed account of specific shrinkage and nonshrinkage methods as well as a comprehensive comparison between the benchmark methods and our MIO approach is included in Appendix A.

4.1 Shrinkage methods

Shrinkage methods (Hastie et al., 2009, ch. 3) perform model selection by biasing, or shrinking, some of the effect estimates toward zero. Although the primary focus of these methods is to build good predictive models, several authors use them for analyzing data from screening designs and thus for identifying active effects (Phoa et al., 2009; Marley and Woods, 2010; Draguljić et al., 2014; Weese et al., 2015; Errore et al., 2017). Two of the most popular shrinkage methods to analyze data from screening designs are the Dantzig selector (DS; Candes and Tao, 2007) and the LASSO (Tibshirani, 1999). In recent years, multiple extensions to these two methods have been presented. For instance, Yuan et al. (2009) and Liu et al. (2010) extended the LASSO and the DS, respectively, to deal with categorical factors. Extensions to the LASSO to impose effect heredity can be found in Yuan et al. (2007), Choi et al. (2010), and Bien et al. (2013).

Shrinkage methods include a tuning parameter t that controls the degree of shrinkage in the estimates and the complexity of the model. Models obtained for different values of t provide alternative interpretations of the data and imply that shrinkage methods posses Property 1. The fact that the DS and the LASSO can handle continuous and categorical factors implies that these methods also possess Property 4.

A drawback of shrinkage methods for analyzing data from screening designs is that they do not have the potential to reveal strong aliasing patterns due to the design. For instance, for a fixed value of the shrinkage parameter, these methods cannot generate alternative models supported by the data. Another drawback of these methods is that they do not allow any user-specified search constraint other than heredity constraints to be incorporated in the model search. The DS even does not allow heredity constraints to be taken into account. For all these reasons, shrinkage methods do not possess Properties 2 and 3.

4.2 Nonshrinkage methods

In contrast to shrinkage methods, nonshrinkage methods do not shrink the effect estimates towards zero. Two popular nonshrinkage methods for analyzing data from screening designs are forward selection (FS; Westfall et al., 1998) and SAMS (Wolters and Bingham, 2011). The FS approach possesses Property 4 because it can analyze data from screening designs involving continuous as well as categorical factors. However, this method lacks Properties 1 and 2 because it is not suitable to create a list of good models or to assess the aliasing of effects. Moreover, algorithms for FS, available in SAS 9.4, JMP 13 and the ‘leaps’ package in R, do not allow to incorporate search restrictions in the model search. So, these implementations of the FS approach do not posses Property 3.

SAMS utilizes a simulated annealing (SA) algorithm to find a large list of good models in terms of the RSS (usually 10,000), which are explored graphically. The SA algorithm constructs models that obey weak effect heredity with a fixed size of k which is 1, 2, 3 or 4 units larger than the maximum plausible size of the true model, $s_{\max}$. Like the MIO approach, SAMS uses raster plots to assess the aliasing caused by the experimental design and to detect the active effects. Due to its SA algorithm and its graphical aids, SAMS possesses Properties 1 and 2. However, the current implementation of SAMS cannot handle data from experiments involving categorical factors with three levels or more, and it does not allow model search restrictions other than heredity constraints for 2FIs or QEs to be imposed. Modifying the SA algorithm to overcome these shortcomings is not trivial, so that, at present, SAMS does not posses Properties 3 and 4.

5 A lean MIO approach for automatic model selection

The examples in Section 3 showed that the full MIO approach, involving the sequential algorithm to create the list of M best-fitting models, the user-specified search restrictions and the raster plot, possesses Properties 1–4 and permits the detection of the potentially active effects. Therefore, we strongly recommend the use of this full version of the MIO approach when analyzing data from screening designs. In some circumstances (for instance, when an experiment involves many responses), it might be desirable to perform the model selection and the identification of the active effects automatically. This is not our preferred choice as this approach lacks Properties 1 and 2, and therefore fails to recognize aliasing of effects, which is a major concern in the analysis of data from many screening designs.

In the event an automatic model selection is deemed necessary, a lean MIO approach would be to focus on the best-fitting model for each value of the model size parameter k and selecting a final model using an information criterion. In this section, we evaluate this approach using a simulation study involving two-level designs with 7 and 11 factors. We consider MEs and 2FIs, which amount to 28 and 66 effects for the 7- and 11-factor design, respectively. We compare our automatic selection approach to the DS, the LASSO with weak heredity constraints of Bien et al. (2013), and SAMS. We find the approach of Bien et al. (2013) to deal with heredity more appealing than the LASSO extensions of Yuan et al. (2007) and Choi et al. (2010) because it guarantees that the LASSO finds an optimal model for any shrinkage degree; see Appendix A.1.1 for details. We do not consider FS because it was outperformed by the DS in the simulation studies of Marley and Woods

(2010) and Draguljić et al. (2014). More specifically, both studies showed that, compared to the DS, FS tends to miss many inactive effects when analyzing data from two-level screening experiments.

Our lean MIO approach is based on best-subset selection. To the best of our knowledge, we are the first to compare the performance of best-subset selection to the DS, the LASSO and SAMS for screening experiments using simulations. Hitherto, this was computationally infeasible because solving the best-subset selection problem was computationally demanding. Now, our MIO approach reduces the computational burden of best-subset selection to the extent that its performance can be studied using simulations.

We first detail our simulation protocol and the automatic model selection procedures we compared. Then, we discuss the performance of the automatic selection procedures to correctly identify the active effects.

5.1 Simulation protocol

Our simulation protocol uses two-level designs for 7 and 11 factors with 20 and 40 runs, respectively. More specifically, we use the 7-factor design 20.7.1 and the 11-factor design 40.11.1a from Mee et al. (2017). These designs have ME contrast vectors that are orthogonal to each other and permit the estimation of almost all models including all MEs and up to seven 2FIs. In the 7-factor design, some 2FI contrast vectors are correlated with other 2FI contrast vectors as well as with ME contrast vectors. More specifically, there are 105 pairs of one ME contrast vector and one 2FI contrast vector with an absolute correlation of 0.2. In addition, there are 99 pairs of 2FI contrast vectors with an absolute correlation of 0.2 and six pairs of 2FI contrast vectors with an absolute correlation of 0.6. In the 11-factor design, only pairs of 2FI contrast vectors are correlated. More specifically, there are 936 pairs of 2FI contrast vectors with an absolute correlation of 0.2 and 54 pairs of 2FI contrast vectors with an absolute correlation of 0.6 in this design. Therefore, the 7- and 11-factor designs represent the typical screening situation in which certain pairs of effects are highly aliased, others are only aliased to a small extent or yet others are not aliased at all.

For each design, each of our 1,000 simulations consisted of the following steps:

1. We randomly selected the number of active MEs, m, from the set $\{2, 3, 4, 5\}$ and the number of active 2FIs, g, from the set $\{1, 2, 3, 4, 5, 6, 7\}$.

2. We randomly selected m columns from the design matrix and associated these with the m active MEs chosen. Next, we randomly selected g 2FI columns of the model

matrix $\mathbf{X}$ subject to the constraint that weak effect heredity is satisfied.

3. We generated the true values, β_u, for the active effects using two scenarios:
   - 'Equal' scenario: The absolute values for the $m + g$ active effects are randomly sampled (with replacement) from the set $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$. A '+' or '$-$' sign is randomly assigned to each sampled value.
   - 'Unequal' scenario: The absolute values for the m active MEs are randomly sampled (with replacement) from the set $\{2, 2.5, 3, 3.5\}$. The absolute values for the g active 2FIs are randomly sampled (with replacement) from the set $\{0.5, 1, 1.5, 2\}$. A '+' or '$-$' sign is randomly assigned to each sampled value.
4. We generated a response vector $\mathbf{y}$ using the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the matrix involving only the contrast vectors corresponding to the active effects, $\boldsymbol{\beta}$ contains the simulated coefficients β_u and $\epsilon_i \sim N(0, 1)$.
5. We performed an automatic model selection using the DS, the LASSO, SAMS and the lean version of our MIO approach.

The ranges of 2–5 active MEs and 1–7 active 2FIs cover a wide range of situations that are likely to occur in practice. In one scenario, which we call the 'Equal' scenario, the active MEs and 2FIs are assumed to have sizes that are comparable. In the other scenario, which we call the 'Unequal' scenario, the active MEs are generally larger than the active 2FIs. Mee et al. (2017) used a similar simulation protocol to compare two-level screening designs.

In our simulation, the settings of the automatic model selection procedures were as follows:

- DS and LASSO: We used the automatic selection procedure of Draguljić et al. (2014) discussed in Appendix A.1.3 with values of the shrinkage parameter t ranging from 0 to $t_0 = \|\mathbf{X}^T\mathbf{y}\|_\infty$. We selected the best model according to the corrected Akaike information criterion:

  $$\text{cAIC} = n \log\left(\frac{\text{RSS}}{n}\right) + \frac{2n\tilde{p}}{n - \tilde{p} - 1},$$

  where RSS denotes the residual sum of squares and $\tilde{p}$ denotes the number of nonzero parameters in the model. Next, we computed the OLS estimates of the best model and retained the effects whose estimates have an absolute value larger than $\gamma = 0.5$, the smallest size of an active effect (in absolute value) in the simulations.

- SAMS: We used the settings recommended by Wolters and Bingham (2011). In particular, we used $s_{\max} = 12$ as the maximum number of active effects in the simulations. For each simulated data set, we generated 10,000 models with $k = s_{\max} + 3$ terms and selected the model with the highest entropy.

- MIO: We used Algorithm 1 with $M = 1$ and $k_{\max} = s_{\max}$, and with the weak heredity constraints (12) added to the MIO formulation. As with the DS and the LASSO, we selected the best model according to the cAIC. From this model, we retained the effects whose estimates have an absolute value larger than $\gamma = 0.5$. To limit the computational burden, we imposed a maximum of 60 seconds for Gurobi to search for each model. This means that the MIO models we used in our simulation study may not be optimal. Our preliminary tests showed that Gurobi usually finds the best models with at most 10 terms in less than 60 seconds. For larger models, Gurobi requires a few more extra minutes to certify optimality.

5.2 Results

We use three measures to compare the automatic model selection methods: sensitivity, false discovery rate (FDR) and type-I error rate. The sensitivity is the proportion of active effects that are successfully detected. The FDR is the proportion of effects declared active that are actually inactive. The type-I error rate is the proportion of inactive effects that are incorrectly declared active. Obviously, the sensitivity should be maximized, while the FDR and the type-I error rate should be minimized. We obtained empirical distributions of these measures for each model selection method using the 1,000 simulations for each combination of design and scenario ('Equal' or 'Unequal'). One of our results is that all model selection methods had type-I error rates below 0.1 for both scenarios with the 7-factor 20-run design, except for the lean MIO approach. For this method, the type-I error rates were below 0.15 with a median type-1 error rate of 0.08 for both scenarios. So, the benchmark methods provide slightly smaller type-I error rates than our lean MIO approach for the 7-factor design. Another result was that almost all type-I error rates, including those for the lean MIO approach, were well below 0.05 in both scenarios with the 11-factor 40-run design. Therefore, all model selection methods are comparable when considering the type-I error rate in the 11-factor design.

Figure 4 compares the sensitivity and the FDR of the four model selection methods under investigation for the 7-factor 20-run design. The left panel of the figure shows the

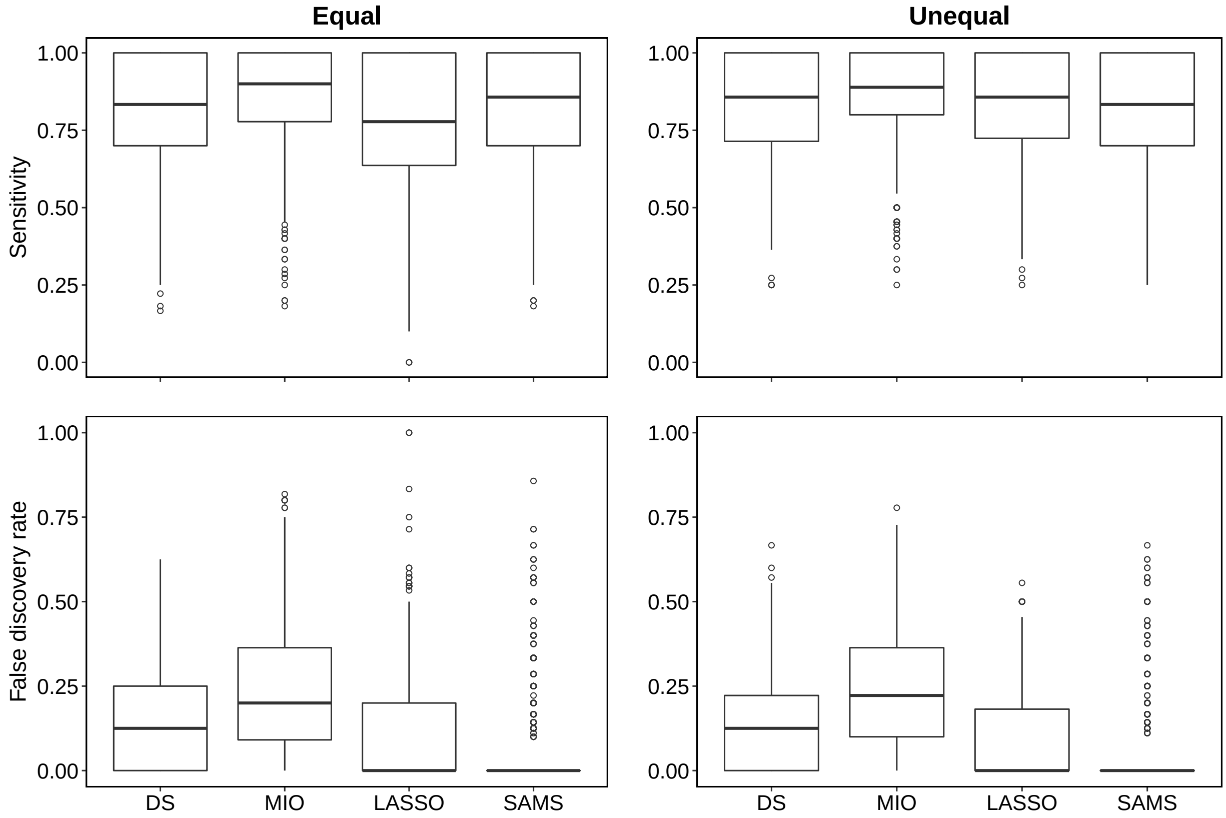


Figure 4: Performance measures for screening strategies for the 7-factor 20-run orthogonal design. ‘Equal’ refers to a scenario in which all active effects are obtained from the same distribution, while ‘Unequal’ refers to a scenario in which active MEs are generally larger than active 2FIs.

simulation results for the ‘Equal’ scenario, in which the active effect are obtained from the same distribution. The right panel shows the simulation results for the ‘Unequal’ scenario, in which the active MEs are generally larger than the active 2FIs.

The top left panel of Figure 4 shows that the lean MIO approach outperforms the benchmark methods in terms of the sensitivity for the 7-factor design under the ‘Equal’ scenario. As a matter of fact, the lean MIO approach is the only one for which the median sensitivity values is as high as 0.9. This means that the lean MIO approaches reaches sensitivity values larger than 0.9 more frequently than the benchmark model selection methods. Also, for 75% of the simulated data sets, the lean MIO approach had a sensitivity larger than 0.78. The bottom left panel of Figure 4 shows that the lean MIO approach has a slightly larger FDR than the DS and the LASSO. For more than 75% of the simulated data sets, the FDRs for these three methods were below 0.4. The SAMS method has a zero

FDR for nearly all simulated data sets, but it has a smaller sensitivity than the lean MIO approach.

The top right panel of Figure 4 shows that, for 75% of the simulated data sets, the sensitivity values of the lean MIO approach were above 0.8 for the 7-factor design under the 'Unequal' scenario. In terms of sensitivity, the lean MIO approach outperforms the three benchmark methods, but it comes at the expense of a larger FDR than the benchmark methods.

Figure 5 compares the sensitivity and the FDR of the four model selection methods under investigation for the 11-factor 40-run design. The figure's top panels show that the lean MIO approach outperforms the benchmark methods in terms of the sensitivity, under both the 'Equal' and the 'Unequal' scenario. In both scenarios, the MIO approach has sensitivity values larger than 0.8 for the vast majority of the simulated data sets. Moreover, for more than 50% of the simulated data sets, its sensitivity values equal 1. In both the 'Equal' and the 'Unequal' scenario, the MIO approach, the DS and the LASSO performed similarly in terms of the FDR. Most of the FDR values for these three methods were smaller than 0.2. As with the 7-factor design, SAMS performs extremely well in terms of the FDR, but this good performance is accompanied by a smaller sensitivity than the lean MIO approach in both the 'Equal' and the 'Unequal' scenario.

Overall, the simulation results show that lean version of the MIO approach, in which only the best-fitting model for each k value is retained, is a good strategy to correctly identify active effects. The lean MIO approach outperformed the benchmark methods in terms of the sensitivity. SAMS had lower FDR values than the lean MIO approach and the two other benchmark approaches. For screening experiments, however, we prefer a higher sensitivity over a low FDR because it is generally considered less desirable to miss active effects than to have false positives. This is due to the fact that follow-up experiments will typically reveal that some of the effects initially declared to be active are in fact inactive. Factors that are declared inactive are generally dropped from a study, so that declaring factors to be inactive is irrevocable. Therefore, if an automatic selection of active effects is needed, we recommend the lean MIO approach.

Note that the DS, the LASSO and the lean MIO approach in our simulations involved a threshold γ to select the active effects and to control the FDRs and type-I error rates. As Phoa et al. (2009), Marley and Woods (2010), Draguljić et al. (2014) and Mee et al. (2017), we used a γ value based on the distributions of the coefficients of the active effects. Therefore, our simulation results, as well as those in the papers cited, may show an overly

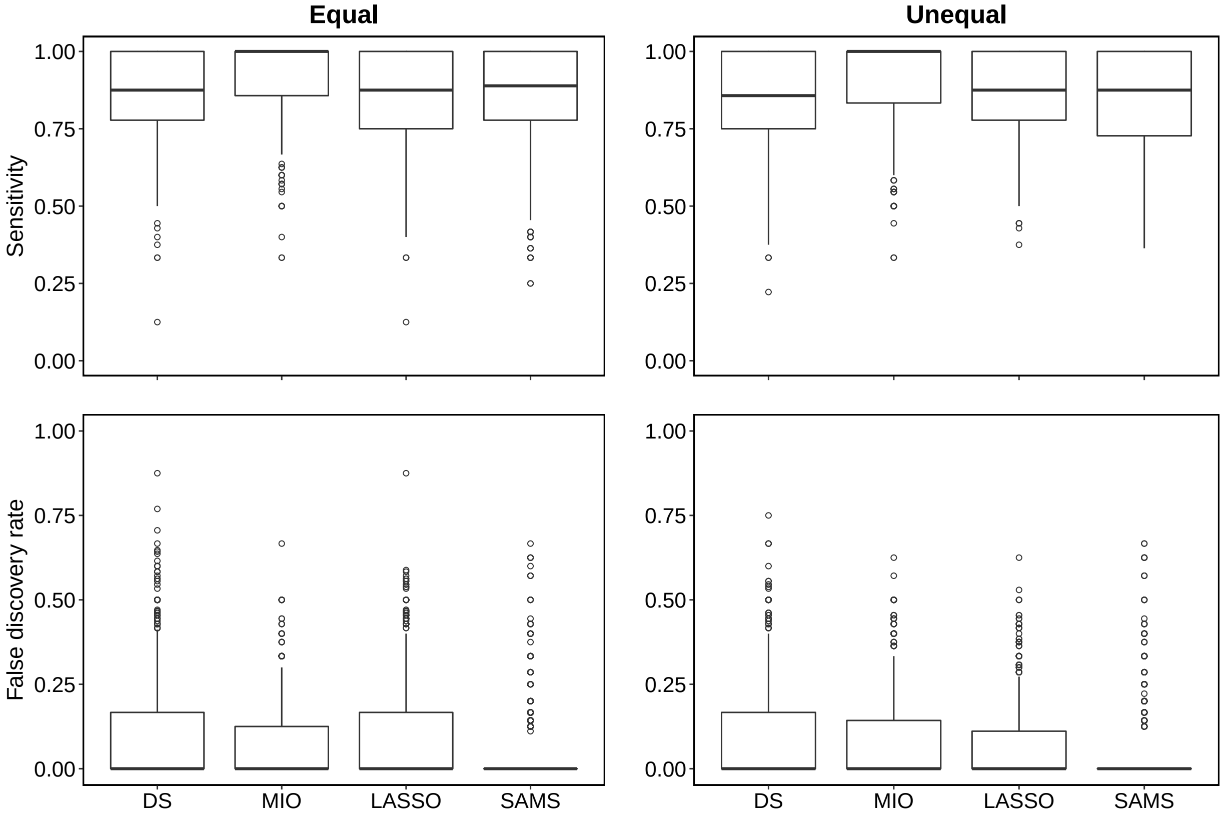


Figure 5: Performance measures for screening strategies for the 11-factor 40-run orthogonal design. 'Equal' refers to a scenario in which all active effects are obtained from the same distribution, while 'Unequal' refers to a scenario in which active MEs are generally larger than active 2FIs.

optimistic performance of the methods. Our simulation study did not involve any pruning for the models obtained by SAMS, because that would lead to an even smaller sensitivity for this method.

The automatic model selection methods in our simulation study should not be considered a substitute for an expert analysis. In practice, an expert analysis based on the full MIO approach rather than its lean version or any of the other automatic model selection methods will allow a more informed decision on the active effects than any automatic model selection. An expert analysis will lead to a much better sensitivity and FDR than the ones displayed in Figures 4 and 5.

6 Concluding remarks

In this paper, we proposed a MIO approach to analyze data from screening designs. The full MIO approach, involving a sequential algorithm to create lists of best-fitting models of different sizes, the user-specified search restrictions and the raster plot, possesses Properties 1–4, which are generally considered desirable for a method of analysis. Unlike the benchmark methods we discussed, the MIO approach permits the identification of the potentially active effects while revealing the aliasing of effects due to the screening design used. Our examples in Section 3 showed that model aliasing is a major concern in the analysis of data from screening designs. The MIO approach also provides a flexible framework to take into account subject-matter expertise through linear constraints that can be added to the best-subset selection problem, and it guarantees that the best-fitting models satisfying various user-specified search restrictions will be found. Moreover, the MIO approach renders best-subset regression feasible and relies on least-squares estimation. For all these reasons, the MIO approach should be appealing to practitioners running screening experiments in any branch of science and in industry. Our simulation study demonstrated that a lean, automated version of the MIO approach also has the potential to provide valuable information, since it was able to detect most of the active effects.

Our MIO model selection approach greatly benefits from the dramatic improvements in computational hardware and from the theoretical and algorithmic advances in mixed integer optimization in recent years. For instance, our search for the best model in Examples 1 and 2 took less than 30 and 45 minutes, respectively. For Example 3, the list of best-fitting models was found in less than a minute. The other currently available algorithms for best-subset selection are either not computationally feasible or require more than three hours to find the best models. Our implementation of the MIO approach used the solver Gurobi v6.5.2. Just like the CPLEX solvers, Gurobi solvers use state-of-the-art insights from the operations research literature. Newer and future versions of Gurobi and CPLEX will certainly speed up the MIO approach even further. For instance, according to the developers of the solver, Gurobi v7.5 is 220% faster than previous versions for solving mixed integer optimization problems involving a quadratic objective function such as the MIO problem (Gurobi Optimization, 2017).

Finally, our MIO approach provides a new benchmark to assess the performance of best-subset selection in simulations. It would then be interesting to compare best-subset selection to the Dantzig selector and the LASSO in more complex simulation scenarios such as those described in Draguljić et al. (2014).

Appendices

A Qualitative comparison to benchmark methods

In this section, we discuss existing benchmark methods to analyze data from screening designs. We first introduce the shrinkage methods and then the nonshrinkage methods. We discuss the advantages and the disadvantages of each type of methods and compare them to the MIO approach.

A.1 Shrinkage methods

Shrinkage methods (Hastie et al., 2009, ch. 3) perform model selection by biasing, or shrinking, some of the effect estimates toward zero. Here, we discuss two popular shrinkage methods for analyzing data from screening designs, namely the Dantzig selector and the LASSO. We also discuss extensions of these methods to impose effect heredity and to deal with categorical factors.

A.1.1 LASSO

The least absolute shrinkage and selection operator (LASSO; Tibshirani, 1999) is a penalized regression method that performs model selection by setting some of the effect estimates to zero. The LASSO estimator $\hat{\boldsymbol{\beta}}$ is the solution to the following optimization problem:

$$\min_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \sum_{u=1}^{p} x_{iu} \hat{\beta}_u \right)^2 + t \sum_{u=1}^{p} |\hat{\beta}_u|, \tag{26}$$

where t is a strictly positive tuning parameter (usually referred to as a shrinkage parameter). The objective function in the LASSO is usually referred to as a penalty function and combines the residual sum of squares (RSS) with the sum of the absolute estimates (i.e., the l_1-norm of $\hat{\boldsymbol{\beta}}$). Because it includes the l_1-norm of $\hat{\boldsymbol{\beta}}$, the objective function involves a stimulus to shrink estimates to zero. The larger the tuning parameter t, the larger is the stimulus to set estimates to zero and the fewer are the effects considered active. Thus, the parameter t controls the model complexity. The optimization problem in (26) is convex, meaning that an optimal solution is guaranteed. The LARS (Efron et al., 2004) algorithm is often used to solve the problem in (26).

In recent years, multiple extensions to the original LASSO have been presented. For instance, Yuan et al. (2009) extended LASSO to deal with categorical factors. Yuan et al. (2007), Choi et al. (2010) and Bien et al. (2013) presented extensions to the LASSO to impose effect heredity. A weakness in the approach of Yuan et al. (2007) is that a second-order effect and the corresponding main effects (MEs) are either all included in the model simultaneously, or none of the effects is included. The method therefore does not allow the user to find out whether either one of the MEs or the second-order effect is active, or both. Choi et al. (2010) impose effect heredity in the LASSO by parametrizing the second-order effects as products of the corresponding MEs. This results in a nonconvex problem, so that the optimal solution to their modification of the LASSO cannot be guaranteed. To impose weak or strong heredity, we find the approach of Bien et al. (2013) the most appealing because it preserves the LASSO problem's convexity. The approach is available in the 'hierNet' package in R. The actual implementation of the LASSO with heredity constraints uses an elastic net (Zou and Hastie, 2005) involving two penalty functions, the l_1-norm of $\hat{\boldsymbol{\beta}}$ as well as the sum of the squared $\hat{\beta}_u$ values. This modification ensures that the optimal solution to the LASSO problem with heredity constraints is unique.

A.1.2 Dantzig selector

The Dantzig selector (DS; Candes and Tao, 2007) is a shrinkage method in which the estimator $\hat{\boldsymbol{\beta}}$ is the solution to the following convex optimization problem:

$$\min_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} \sum_{u=1}^{p} |\hat{\beta}_u| \quad \text{subject to } \|\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\|_\infty \leq t, \tag{27}$$

where $\|\mathbf{x}\|_\infty$ denotes the largest absolute element of $\mathbf{x}$ (i.e., the l_∞-norm of $\mathbf{x}$). The DS minimizes the l_1-norm of $\hat{\boldsymbol{\beta}}$, subject to a constraint on the maximum absolute element of the vector $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Ideally, the vector $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is zero, which means that the residuals do not contain any information on $\mathbf{y}$ that is contained in $\mathbf{X}$. In the event it exists, the ordinary least squares estimator achieves this ideal. The DS produces an alternative estimator involving fewer nonzero parameter estimates and a nonzero $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ vector. The extent to which the vector $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ can differ from zero is controlled by the tuning parameter t. More specifically, the maximum absolute element of $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is bounded by t. So, statistically speaking, the DS essentially seeks the most sparse estimator or, alternatively, the most parsimonious model that produces residuals containing little information about the factor effects on the response. In other words, the DS seeks a parsimonious model that fits the data well. The problem in (27) can be reformulated as a linear program and solved using Gurobi, CPLEX or the 'lpSolve' package in R. While it is possible to study categorical factors using the DS (Liu et al., 2010), there is no way to impose weak or strong heredity.

A.1.3 Shrinkage methods versus MIO

Shrinkage methods can be viewed as surrogates for best-subset selection. They approximate the problem in (1) using a convex formulation that can be efficiently solved using readily available linear and quadratic optimization algorithms. Shrinkage methods possess Property 4 since they can handle continuous and categorical factors. These methods include a tuning parameter t that controls the degree of shrinkage in the estimates and the complexity of the model. Because shrinkage methods are commonly used to build good predictive models, the value of t is often chosen by minimizing a prediction error sum of squares for a hold-out sample. Unlike in the MIO approach, where k represents the model size, the tuning parameter t in the shrinkage methods does not have a simple intuitive interpretation.

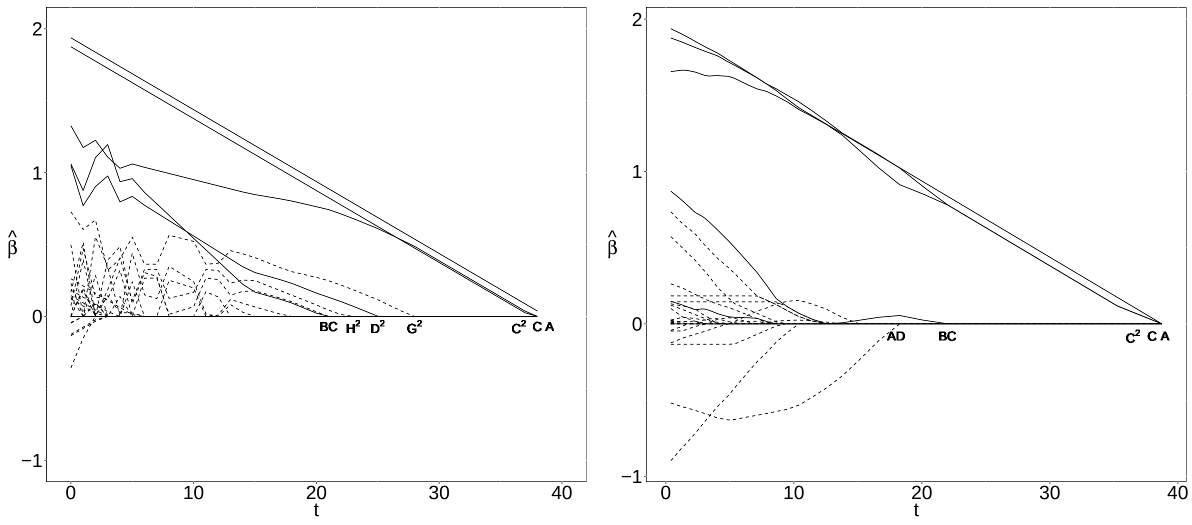
For screening experiments, the shrinkage in the estimates forces the model selection and estimation to be performed in three steps (Phoa et al., 2009). The first step consists of finding models for different values of t. These models provide alternative interpretations of the data and imply that shrinkage methods possess Property 1. In the second step, the models are evaluated using a profile plot, which displays the values of t on the horizontal axis and the effect estimates for each model on the vertical axis. The plot is read from right to left, starting from the models with large degrees of shrinkage and ending with models involving little shrinkage. The effects whose estimates remain 'large' for a wide range of t values are declared active. The last step in the model selection and estimation procedure is to correct for any bias by computing the ordinary least squares (OLS) estimates of the active effects. In contrast with this three-step approach, the MIO approach produces the OLS estimates directly and guarantees that the models found are the best ones in terms of the RSS value. As a result, the MIO approach performs model selection and estimation simultaneously.

A drawback of model selection using shrinkage methods is that the decision of whether an effect is active or not is based on biased estimates. This bias and the aliasing in the screening design may render the detection of the active effects in the profile plots very difficult. This is illustrated in the following example.

Example 4. (Continuation of Example 1 in the main text.) We analyzed the data from the synthetic experiment using the DS and using the LASSO with weak heredity for the second-order effects, as in Bien et al. (2013). Figure 6 shows the profile plots for the effect estimates from both methods.

Figure 6a shows that the estimates of the MEs of the factors A and C and the quadratic effect (QE) of factor C remain large for all values of t when using the DS. Therefore, they are correctly identified as active. As t goes to zero, the estimates of the second-order effects G^2, D^2, H^2, GJ and BC become larger than zero. Due to the shrinkage, the estimates of D^2 and BC only 'stand out' when t is smaller than 10. The DS does not provide any evidence that the two-factor interaction (2FI) CD is active because the profile plot is noisy for small values of the shrinkage parameter t. This is a common picture when there is a complex aliasing in the effects. Based on the DS, we would detect five of the six truly active effects, and declare three inactive effects to be active.

Similarly, the LASSO correctly indicates that the MEs of the factors A and C and the QE of factor C are active because their estimates are large for most values of t; see Figure 6b. The profile plot for the LASSO also suggests that the 2FIs BC and AD, which obey



(a) Dantzig Selector

(b) LASSO

Figure 6: Profile plots for the Dantzig Selector and the LASSO with heredity constraints for the second-order effects, for Example 1. The truly active effects (the main effects of the factors A and C, the two-factor interactions BC and CD, and the quadratic effects C^2 and D^2) are shown using solid lines.

weak heredity, are active. Note that the profile plot suggests that BC should be the fourth effect to add to the model, but, due to the shrinkage, its estimate remains virtually zero until t is smaller than 10. It is therefore tempting to declare the interaction BC to be inactive. The profile plot in Figure 6b does not suggest any other effect as being active. Therefore, the LASSO would only identify four of the six truly active effects, and declare one inactive effect to be active (namely AD).

Example 4 shows that the DS and the LASSO tend to bring inactive effects into the final model. Zou and Hastie (2005), Mazumder et al. (2011) and Mazumder and Radchenko (2017) demonstrated that this is typical for these methods in the event some of the effects are strongly aliased, such as the second-order effects in the DSD in Table 1 in the main text. The MIO approach is able to overcome the complex aliasing in the DSD by revealing strongly aliased sets of effects in the raster plots (Property 2) and by allowing user-specified restrictions in the model search (Property 3).

One attractive feature of the MIO approach is that its objective function value has a simple, statistically meaningful interpretation, and that the objective function values of all models in the list generated by the sequential MIO algorithm can be compared directly (regardless of whether the models possess the same size k). The objective function values of the shrinkage methods do not possess a useful statistical interpretation and cannot be used to compare models obtained for different values of the tuning parameter t. Moreover, for a fixed value of t, shrinkage methods cannot generate alternative models supported by the data. Therefore, these methods do not have the potential to reveal strong aliasing patterns due to the experimental design. Also, the DS does not allow any user-specified search constraint to be taken into account, and the LASSO does not allow constraints other than heredity constraints. For all these reasons, shrinkage methods do not possess Properties 2 and 3. The weaknesses of the DS and the LASSO also apply to screening strategies based on other shrinkage methods not discussed here, such as ridge regression (Hoerl and Kennard, 1970), the nonnegative garrote (Breiman, 1995) and SCAD (Fan and Li, 2001).

Phoa et al. (2009) proposed an alternative procedure to perform model selection automatically when using shrinkage methods. This procedure involves a fixed strictly positive threshold γ to declare an effect active and uses the corrected Akaike and Bayesian information criteria (Claeskens and Hjort, 2008). For each value of the tuning parameter t, the model is pruned by only retaining the effects whose estimates have an absolute value larger than γ. In this way, most of the inactive effects that are brought into the model by the shrinkage method are removed. The list of reduced models obtained for different values of t is then evaluated using an information criterion and the best model is reported. Marley and Woods (2010) and Weese et al. (2015) used this automatic selection procedure to assess the performance of shrinkage methods using simulations. Draguljić et al. (2014) proposed a variation of this automatic selection procedure. They evaluate the models obtained for different values of t using an information criterion and calculate the OLS estimates of the best model. Finally, they declare an effect as active if its OLS estimate has an absolute value larger than γ. Therefore, this selection procedure relies on the shrinkage methods to force some effects to zero and applies the threshold γ to unbiased estimates of the selected effects. Using simulations, Phoa et al. (2009), Marley and Woods (2010), Draguljić et al. (2014) and Weese et al. (2015) showed that their respective automatic selection procedures work well as they tend to correctly select the active effects.

A.2 Nonshrinkage methods

In contrast to shrinkage methods, nonshrinkage methods do not shrink the effect estimates towards zero. Here, we discuss two popular nonshrinkage methods for analyzing data from screening designs, namely forward selection and simulated annealing model search (SAMS).

A.2.1 Forward selection

Forward selection (FS; Westfall et al., 1998) is a well-known surrogate for best-subset selection. FS is a heuristic method that starts with the null model and then adds the most significant effect at each step according to an F-test. The process finishes when no more effects are significant. The main advantages of this approach is that it is computationally fast and produces OLS estimates for the selected effects. Moreover, it is easy to impose heredity when using FS, and to deal with categorical factors.

A.2.2 SAMS

SAMS (Wolters and Bingham, 2011) utilizes a simulated annealing (SA) algorithm to find a large list of good models in terms of the RSS (usually 10,000), which are explored graphically. The SA algorithm constructs models that obey weak effect heredity with a fixed size of k which is 1, 2, 3 or 4 units larger than the maximum plausible size of the true model, $s_{\max}$. The algorithm also delivers the OLS estimates for the selected effects. As the MIO approach, SAMS uses raster plots to assess the aliasing caused by the experimental design and to detect the active effects.

To perform model selection automatically, SAMS uses an entropy criterion that measures the degree of 'surprise' at seeing a set of effects occur with the observed relative frequency in the list of models. The effects in the set with the highest entropy are declared active. Draguljić et al. (2014) showed that SAMS in combination with the entropy criterion tends to correctly identify the active MEs and 2FIs from two-level screening experiments.

A.2.3 Nonshrinkage methods versus MIO

Nonshrinkage methods perform model selection using heuristic algorithms that optimize the RSS and deliver the OLS estimates of the selected effects. The heuristic algorithms are computationally fast and permit effect heredity to be imposed. Unlike the MIO approach, these algorithms do not guarantee that the best-fitting models will be found.

The FS approach possesses Property 4 because it can analyze data from screening designs involving continuous as well as categorical factors. However, this method lacks Properties 1 and 2 because it is not suitable to create a list of good models or to assess the aliasing of effects. Algorithms for FS are available in SAS 9.4, JMP 13 and the 'leaps' package in R. From these algorithms, only the one in JMP 13 allows heredity restrictions to be imposed in the search. None of the algorithms allows other restrictions to be incorporated. So, these implementations of the FS do not possess Property 3. In addition, Marley and Woods (2010) and Draguljić et al. (2014) showed that FS tends to miss many active effects when analyzing data from screening experiments.

Due to its SA algorithm and its graphical aids, SAMS possesses Properties 1 and 2. However, the current implementation of SAMS cannot handle data from experiments involving categorical factors with three levels or more, and it does not allow model search restrictions other than heredity constraints for 2FIs or QEs to be imposed. Modifying the SA algorithm to overcome these shortcomings is not trivial, so that, at present, SAMS does not posses Properties 3 and 4.

Another weakness of SAMS is that, in some situations, the large list of models generated by the SA algorithm may not clearly support a single set of active effects in the raster plot. To overcome this weakness, Wolters and Bingham (2011) proposed clustered raster plots that use K-means clustering methods (Hastie et al., 2009, ch. 13) to visualize alternative sub-models supported by the data. However, independent runs of the clustering algorithm may suggest notably different sub-models and thus different sets of active effects. For instance, for the data in Example 2, Mee (2013, sec. 3.2) showed that one clustered raster plot supported three sub-models including the MEs of A, B, E, G, I and J, and the 2FIs AD, BE and GI, whereas another clustered plot supported five sub-models including the MEs of A, B, E, G, and I, and the 2FIs AD, AM, BE, BJ, EH and GI. The MIO approach does not suffer from these weaknesses because its raster plot is based on a relatively small list of best-fitting models.

Acknowledgments

The research that led to this paper was financially supported by the Flemish Fund for Scientific Research FWO. Part of the simulation study was carried out at the HPC core facility CalcUA at the University of Antwerp. The authors are grateful to José Núñez Ares and Nha Vo-Thanh for drawing their attention to the work of Bertsimas.

References

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

Beale, E. M. L. and Forrest, J. J. H. (1976). Global optimization using special ordered sets. *Mathematical Programming*, 10:52–69.

Bertsimas, D. and King, A. (2016). An algorithmic approach to linear regression. *Operations Research*, 64:2–16.

Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44:813–852.

Bertsimas, D. and Weismantel, R. (2005). *Optimization Over Integers*. Dynamic Ideas Press.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A LASSO for hierarchical interactions. *The Annals of Statistics*, 41:1111–1141.

Bixby, R. (2012). A brief history of linear and mixed-integer programming computation. *Documenta Mathematica. Extra Volume: Optimization Stories*, pages 107–121.

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

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384.

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

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

Choi, N., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105:354–364.

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Draguljić, D., Woods, D. C., Dean, A. M., Lewis, S. M., and Vine, A.-J. E. (2014). Screening strategies in the presence of interactions. *Technometrics*, 56:1–16.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32:407–499.

Errore, A., Jones, B., Li, W., and Nachtsheim, C. J. (2017). Using definitive screening designs to identify active first- and second-order factor effects. *Journal of Quality Technology*, 49:244–264.

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.

Furnival, G. and Wilson, R. (1974). Regression by leaps and bounds. *Technometrics*, 16:499–511.

Gurobi Optimization, I. (2017). Gurobi 7.5 performance benchmarks. Available at http://www.gurobi.com/pdfs/benchmarks.pdf. Accessed 1 February 2018.

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 edition.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Jones, B. and Nachtsheim, C. J. (2011). A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology*, 43:1–15.

Lenth, R. (1989). Quick and easy analysis of unreplicated experiments. *Technometrics*, 31:467–473.

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

Liu, H., Zhang, J., Jiang, X., and Liu, J. (2010). The group Dantzig selector. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 461–468.

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.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106:1125–1138.

Mazumder, R. and Radchenko, P. (2017). The discrete Dantzig selector: Estimating sparse linear models via mixed integer linear optimization. *IEEE Transactions on Information Theory*, 63:3053–3075.

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

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

Miller, A. (2002). *Subset Selection in Regression.* Chapman & Hall/CRC, 2nd edition.

Miller, A. and Sitter, R. R. (2004). Choosing columns from the 12-run Plackett-Burman design. *Statistics and Probability Letters*, 67:193–201.

Natarajan, B. . (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–324.

Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society Series A*, 140:48–77.

Nelder, J. (1994). The statistics of linear models: back to basics. *Statistics and Computing*, 4:221–234.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course.* Springer.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161.

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

Phadke, M. S. (1986). Design optimization case studies. *AT&T Technical Journal*, 65:51–68.

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

Schoen, E. D. and Mee, R. W. (2012). Two-level designs of strength 3 and up to 48 runs. *Journal of the Royal Statistical Society Series C*, 61:163–174.

Schoen, E. D., Vo-Thanh, N., and Goos, P. (2017). Two-level orthogonal screening designs with 24, 28, 32, and 36 runs. *Journal of the American Statistical Association*, 112:1354–1369.

Tibshirani, R. (1999). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 52:267–288.

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

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

Wolters, M. A. and Bingham, D. (2011). Simulated annealing model search for subset selection in screening experiments. *Technometrics*, 53:225–237.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*. Wiley, 2nd edition.

Xu, H., Cheng, S.-W., and Wu, C. (2004). Optimal projective three-level designs for factor screening and interaction detection. *Technometrics*, 46:280–292.

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

Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3:1738–1757.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67:301–320.