Received: 14 January 2022 | Revised: 28 April 2022 | Accepted: 21 July 2022

DOI: 10.1002/qre.3188

RESEARCH ARTICLE

WILEY

On the identification of active factors in nonregular two-level designs with a small number of runs

Yngvild Hole Hamre | **John Tyssedal**

Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Correspondence
John Tyssedal, Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
Email: john.tyssedal@ntnu.no

Abstract

Nonregular two-level designs are attractive screening designs due to their good projection properties and flexible run sizes. In particular, the 12-run Plackett–Burman (PB) design has become quite popular. However, existing methods struggle with the identification of active factors when the number of active factors exceeds the projectivity of the designs. This is especially the case when interactions are present, the variance is high and the number of runs is small. In this paper, we propose a method for analysing nonregular two-level designs that particularly addresses the issues above. It exploits the projection properties of designs and is here applied on the 12-run PB design and the 16-run no-confounding (NC) designs. In the construction of the method, the use of test- and penalty-based procedures are avoided. Instead, the number of allowed terms in a model is restricted. The effectiveness of the method and comparison between designs are evaluated by simulations for different scenarios. Ways to evaluate the reliability of the screening procedure are pointed out. An example with real data is given to demonstrate how one might perform the analysis in practice.

KEYWORDS

capture frequency, factor screening, nonregular designs, projection properties, variable selection

1 | INTRODUCTION

In the first stage of an experiment, a large number of factors may have to be considered as potentially active. At that point, the main goal is to identify the ones that really influence the response. This is called factor screening. In most cases, the subspace of active factors is considerably smaller than the space of all factors. Box and Meyer[1] suggest 0.25 to be a reasonable prior probability for a factor to be active. Factors not identified to have an impact on the response are normally not considered afterwards. Good and reliable methods for determining which factors are influential are, therefore, crucial. Whilst screening often is considered a part of physical experimentation, it has also found its way into machine learning in order to reduce the dimension of the hyperparameter space (Lujan-Moreno et al.[2]).

The traditional choice of screening designs has been two-level fractional factorials, also called regular designs. They have orthogonal columns and exist for $\frac{1}{2^p}, p = 1, 2, \ldots k-1$ fractions of 2^k factorial designs, where k is the number of factors included in the design. The drawback of these designs is that effects may be fully aliased, making it difficult to separate the active effects from the rest. Nonregular two-level designs, in particular those introduced by Plackett and Burman,[3] have, therefore, become increasingly popular. Compared to regular designs, they have two particularly desirable properties. First, they project well onto lower dimensions.[4–6] Second, they seem to exist for all n that fulfil $n \bmod 4 = 0$, $n \geq 12$, thus they are far more flexible with regard to run sizes than regular designs. The alias structure may be complex, but the aliasing is often partial, making it possible to separate effects from each other. However, the partial aliasing between effects makes traditional analysis methods such as Lenth's method and normal and half-normal plots fall short, as they rely on the ability to totally separate contrasts from each other. Thus there is a need for other methods for factor screening when using nonregular designs.

There are two main strategies for analysing nonregular designs, effect-based and factor-based searches. Effect-based methods aim at identifying the significant effects. A linear model that can provide estimates of main effects and interactions is assumed to be an adequate approximation of the response. Strong or weak heredity is often a precept for choosing models, and also used to restrict the search. The strong heredity principle only allows a two-factor interaction in the model if both the main effects associated with the interaction are included. Weak heredity relaxes this requirement by only demanding that at least one of the main effects associated with the two-factor interaction is included. Examples of effect-based methods are the stepwise regression procedure proposed by Hamada and Wu,[7] the Bayesian stochastic search variable selection,[8] the modified least angle regression[9] and the simulated annealing model search.[10] The non-convex penalized least square described in Jin and Li[11] originally proposed by Fan and Li[12] and the Dantzig selector[13] represent effect-based methods that do not depend on the heredity principle.

A factor-based search aims at identifying the active factors, followed by an examination of the nature of the factor activity. A factor-based search is less-dependent on model assumptions, heredity included. The disadvantage of doing a factor-based search is a vulnerability for noise, as too much noise may lead to several candidate sets of active factors explaining the variation in the response equally well. Different factor-based search approaches have been suggested. Box and Meyer[1] proposed a Bayesian analysis with prior probabilities on factors being active, while Tyssedal and Samset[14] suggested a projection-based factor search, see also Kulachi and Box[15] and Tyssedal et al.[16] Tyssedal and Hussain[17] combined a projection-based factor search with forward selection, testing out the Akaike's Information criterion (AIC), the F-test and a particular criterion based on the change in the coefficient of determination, ΔR^2.

Both effect- and factor-based search methods have shown good performances when applied to specific examples. However, the proposed methods are often not tested out on more than a few models, and more frequently for three active factors than for four. Various success criteria have been used in simulations, among those the percentage of selected models being correct or partially correct. For a more complete list of such criteria, we refer to Tyssedal and Hussain.[17] The proposed procedure in this paper has similarities both to the one in Tyssedal and Hussain[17] and the one in Wolters and Bingham.[10] Like in Tyssedal and Hussain,[17] the objectives are to investigate how the amount of noise, the number of factors screened and the number of active factors affect the screening. But there are also important differences. Rather than using a panel, we will try out our procedure on a much wider range of models, and we will also avoid the use of stopping criteria. Instead, we will put restrictions on the number of allowed terms in the model, like Wolters and Bingham.[10] For comparison, their procedure is effect-based, our is factor-based. They use heredity to limit their search. We use projection models (to be explained later). Common in our and the two other procedures is that instead of focusing on identifying 'one correct model', for which experience has shown a rather low success probability, we will rather suggest reducing the number of possibly active candidate sets in several steps. An important feature of our procedure is that an evaluation of its reliability can be performed. This will be discussed in Section 5.

The designs used in the simulation studies are the 12-run Plackett–Burman (PB) design and the 16-run no-confounding (NC) designs for 6–8 factors introduced by Montgomery and Jones.[18] These are all orthogonal nonregular two-level designs having in common that only partial aliasing exists between main effects and interactions as well as between two-factor interactions. Also, they have similar projection properties onto three and four factors and hence are competitive alternatives to be considered for a screening when identifying up to four active factors is of interest.

We start this paper by introducing some concepts and the strategy for our factor-based search in Section 2. The proposed screening algorithm will be described in Section 3 and applied to a model from Tyssedal and Hussain[17] in Section 4. In Section 5, we present the results of a simulation study over a wide range of models followed by an application on real data in Section 6. Some concluding remarks are given in Section 7.

2 | IMPORTANT CONCEPTS AND STRATEGY

To ensure having a high chance of finding the correct active factors when only a few factors are assumed to be active, it is important that the screening design projects well onto lower dimensions. This property can by described by the *projectivity* of the design, as defined by Box and Tyssedal[4]:

> A $n \times k$ design with n runs and k factors each at two levels is said to be of projectivity P if the design contains a complete 2^P factorial in every possible subset of P out of the k factors, possibly with some points replicated.

If a design is of projectivity 3, all main effects and interactions corresponding to any choice of three factors can be estimated without bias if the remaining factors are inactive. If it can be assumed that main effects and low-order interactions can adequately model the response, estimating higher-order interactions may not be needed. A useful concept in such cases is generalized projectivity,[19] defined as:

> A $n \times k$ design with n runs and k factors each at two levels is said to be of generalized projectivity P_α, if for any selection of P columns of the design all factorial effects including up to α-factor interactions are estimable.

The 12-run PB design is a $P = 3$ design, but Wang and Wu[20] pointed out that it is possible to estimate the main effects and their two-factor interactions for any four factors, hence it is also a $P = 4_2$ design. By sacrificing the opportunity to fit the three-factor interaction, an additional factor is allowed to be included. The 16-run NC designs for six to eight factors share the same projectivity properties. A model including all main effects and interactions up to its projectivity either P or P_α will be called *a full projection model*, or FP-model for short. In the case of fitting a full projection model for the PB12 design assuming four active factors, the design will contain an intercept, four main effects and 6 two-factor interactions. Nearly all degrees of freedom are spent when fitting the full model, making it hard to assess the model fit and significance of each term. Having a procedure for selecting the subset of terms that should be included in the model without relying on significance tests would, therefore, be useful.

The term *candidate set* is used to denote a set of factors that potentially may be active. If, for instance, 11 experimental factors are included in the screening design, but only three are assumed to be active, there are $\binom{11}{3} = 165$ candidate sets of active factors before the screening. If the number of candidate sets can be reduced to 5 or 10 with the correct set of active factors included, standard regression techniques can be used to reduce the number further. In this process, the experimenter may look for the most parsimonious representation, use subject matter knowledge and the heredity principle. If there is still ambiguity, follow up runs can be added.

For a successful screening, it is important that when the number of candidate sets is reduced to a number r, the correct set of active factors is among those. The set consisting of r candidate sets of factors with the purpose of containing the correct set of active factors will be called the *capture set* of size r. To have a measure of how often this happens, we introduce the concept *capture frequency* $CF_r(i)$, defined as the number of simulations out of i in which the correct candidate set of factors is found in a capture set of size r selected by some criterion, see also Tyssedal and Hussain.[17] With the response values $y_i, i = 1, \dots, n$, we have used the mean square error $\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$, where $\hat{y}_i$ is the ith fitted response value and p is the number of terms in the model, intercept included. In this paper, a rate of 95% will be considered acceptable.

One of the most common strategies for doing variable selection is forward selection. The method starts with a minimal model, and a new term is added if it is the best among all candidates for which a test statistic exceeds a given threshold, or according to a chosen criterion. One challenge, in particular when using a F-test, is that in the beginning, the variation in the response caused by important terms that are not yet included will enlarge the error variance. The large error variance may hinder the inclusion of important effects, and in some cases, this might cause the algorithm to stop at an early stage. Wolters[21] reports on problems with criterion-based methods, among these is overfitting, see also Miller and Sitter[22] for a discussion about finding the appropriate penalty for such criteria. Another challenge is that spurious effects may be chosen to enter the model due to nonorthogonal effect columns.

Having too few terms or wrong terms will make the MSE a biased estimate of the response variance, and too many terms in a model may lead to some of the wrong candidate sets being able to explain the variation in the response equally well as the correct one. The method proposed in this paper tries to avoid these problems by using a selection strategy where for each estimated FP-model, one selects a predefined number of the effects with the largest coefficients in absolute value to be in a model that is then refitted to the data. Then all terms have an equal chance of entering the model, as they are

chosen simultaneously. This predefined number of effects, l, should be large enough for the reduced model to include all active effects in the candidate set with the correct active factors. The correct value for l is of course not known in advance. However, several values can be tried, and inherent in the procedure is a form of self-correction in that every candidate set in the capture set can be checked for their number of active terms. We think that the best way of doing this is to start with a low value of l and then gradually increase it by one in each step. This will be illustrated on a real example in Section 6. It is difficult to see how any test based or penalty-based procedure can offer the same opportunity. Another advantage is that the procedure is scale invariant. If all the response values are multiplied by a constant c all the estimated coefficients and the estimated σ will also be multiplied by c. Any ranking between coefficients and the MSEs of the candidate sets will be unchanged. No assumption about heredity is taken into account in this procedure. The heredity principle is not guaranteed to be valid, and we think it is better to see which candidate sets that are able to explain the variation in the response before we eventually discard some. The algorithm will be described in detail in Section 3.

3 | THE PROPOSED SCREENING ALGORITHM

The basic idea of the screening algorithm is to first do a rough selection of terms, utilizing the assumption that most often, only a small number of terms is needed to explain the response. The proposed screening algorithm is given by the following steps:

1. Given a set of n_t experimental factors, assume that n_a are active. Find all possible sets of n_a active factors, in total $k = \binom{n_t}{n_a}$.
2. For all k sets, fit the full projection model given the current design and n_a. The intercept is also included in the model.
3. Select the l terms corresponding to the largest coefficients in absolute value in the FP-model.
4. Refit the model with the selected terms and the intercept only. The refitted model will be referred to as the reduced model.
5. Store the MSE $= \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-l-1}$ for each reduced model.
6. Find the sets of active factors corresponding to the r smallest MSE.

As a result, the original k candidate sets of n_a active factors are reduced to r. To consider a set, a candidate set for n_a active factors is not affected by how many factors that are included in the reduced model. In practice, one will likely inspect the selected models to see which factors were actually chosen. As the algorithm assumes that the coefficients with the largest absolute value are the most important, it will from now on be referred to as the 'size-based method'. The emphasis will be to investigate for which values of r the active factors are among the final candidates in at least 95% of the cases.

It is our belief that starting with a coarse sorting in the beginning and proceeding with fine-tuning of the model is a rational approach, as reducing the number of candidate models makes it easier to compare and select a final model.

4 | A MOTIVATIONAL EXAMPLE

To have some impression of how well the algorithm suggested in Section 3 performs, it was first tested out on a model given by $Y = 2x_1 + 4x_3 + 2x_2x_3 + 2x_3x_4 + \epsilon, \epsilon \sim N(0, \sigma^2)$. This is a commonly used test model for screening procedures, see for example Hamada and Wu[7] or Wolters and Bingham.[10] It is also thoroughly investigated in Tyssedal and Hussain,[17] as their model 8 in the panel. The model has four active factors and four terms and obeys the weak heredity principle.

Tables 1 and 2 show the capture frequencies for the active factors when using the new size-based method, choosing the number of terms $l = 4$, 6 and 7, respectively. This is in line with Wolters and Bingham,[10] who suggest that $\frac{n}{3}$ is a reasonable estimate for the number of effects in the model, and that between $\frac{n}{3} + 2$ and $\frac{n}{3} + 4$ effects should be chosen in order to ensure finding the correct effects. Tables 1 and 2 show $\mathrm{CF}_r(1000)$ for r =1, 5 and 10. The choice of $i = 1000$ mimics Tyssedal and Hussain.[17] The results were found by using the design in Table 3, creating the responses based on the model, and then adding normally distributed noise with different variances. The proposed size-based method was used to test all possible candidate sets of four active factors having n_t experimental factors. The design used consists of the n_t first columns of the PB12 design in Table 3.

TABLE 1 $CF_r(1000)$ obtained from model 8 in Tyssedal and Hussain[17] varying σ^2, the size of the capture set, r, and the number of experimental factors, n_t, using $l = 4$ number of terms

	σ^2										
r	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
						$l = 4, n_t = 7$					
1	1000	1000	1000	1000	999	999	993	991	976	971	951
5	1000	1000	1000	1000	1000	1000	1000	1000	998	997	998
10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
						$l = 4, n_t = 9$					
1	1000	1000	1000	999	998	996	986	973	945	911	888
5	1000	1000	1000	1000	1000	1000	1000	997	998	991	988
10	1000	1000	1000	1000	1000	1000	1000	999	1000	998	999
						$l = 4, n_t = 11$					
1	1000	1000	1000	998	999	993	980	952	903	884	850
5	1000	1000	1000	1000	1000	1000	999	996	992	989	974
10	1000	1000	1000	1000	1000	1000	1000	1000	998	996	991

TABLE 2 $CF_r(1000)$ obtained from model 8 in Tyssedal and Hussain[17] varying σ^2, the size of the capture set, r, the number of experimental factors, n_t, and the number of terms, l

	σ^2										
r	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
						$l = 6, n_t = 7$					
1	1000	648	675	630	639	658	646	593	614	592	582
5	1000	1000	1000	1000	1000	998	999	992	989	986	970
10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	999
						$l = 6, n_t = 9$					
1	1000	658	656	641	670	604	632	568	541	551	468
5	1000	1000	1000	1000	998	992	989	973	951	943	905
10	1000	1000	1000	1000	1000	999	999	997	988	991	972
						$l = 6, n_t = 11$					
1	1000	528	529	526	527	487	470	486	425	398	403
5	1000	1000	1000	996	990	978	949	915	891	856	835
10	1000	1000	1000	1000	1000	997	982	972	953	947	932
						$l = 7, n_t = 7$					
1	0	531	542	534	522	500	500	516	512	463	455
5	1000	1000	1000	1000	1000	993	994	985	978	968	948
10	1000	1000	1000	1000	1000	1000	1000	1000	998	995	991
						$l = 7, n_t = 9$					
1	0	400	382	417	380	361	352	367	358	335	290
5	1000	897	877	886	851	854	830	829	806	793	748
10	1000	1000	1000	1000	998	988	978	964	944	928	917
						$l = 7, n_t = 11$					
1	0	238	271	242	250	248	228	220	222	193	214
5	0	657	693	651	716	671	660	645	619	579	578
10	0	884	908	891	913	872	871	846	831	788	791

TABLE 3 The 12-run PB design with 11 factors

A	B	C	D	E	F	G	H	I	J	K
1	1	−1	1	1	1	−1	−1	−1	1	−1
−1	1	1	−1	1	1	1	−1	−1	−1	1
1	−1	1	1	−1	1	1	1	−1	−1	−1
−1	1	−1	1	1	−1	1	1	1	−1	−1
−1	−1	1	−1	1	1	−1	1	1	1	−1
−1	−1	−1	1	−1	1	1	−1	1	1	1
1	−1	−1	−1	1	−1	1	1	−1	1	1
1	1	−1	−1	−1	1	−1	1	1	−1	1
1	1	1	−1	−1	−1	1	−1	1	1	−1
−1	1	1	1	−1	−1	−1	1	−1	1	1
1	−1	1	1	1	−1	−1	−1	1	−1	1
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

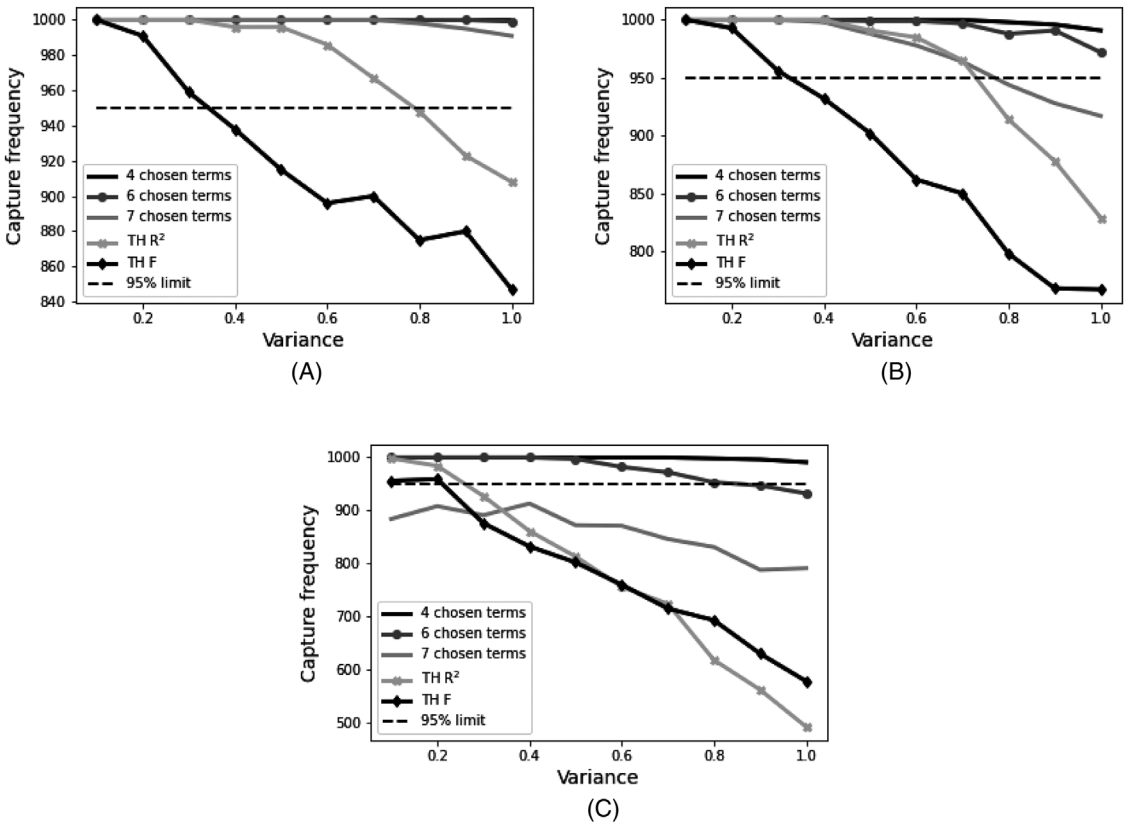


FIGURE 1 Plots of $CF_{10}(1000)$ against variance for comparison of the proposed size-based method and the methods from Tyssedal and Hussain,[17] for different numbers of experimental factors, n_t, and number of terms, l. The y-axis is different for each plot, but the black dashed lines show the 95% limit in all cases. (A) Seven factors in the design, (B) nine factors in the design, (C) 11 factors in the design

What is apparent from Tables 1 and 2 is that, at least for this model, our procedure may perform extremely well when $l = 4$ and also for $l = 6$ and $r = 10$. For $l = 4$, we obtained higher capture frequencies using $r = 1$ than Tyssedal and Hussain[17] obtained with $r = 10$. As expected, the number of experimental factors affects the performance. Even for $l = 6$ and $r = 5$, the results are good for quite high variances. For $l = 7$, the performance declines remarkably. In Figure 1, $CF_{10}(1000)$ is plotted against variance for comparing the size-based procedure with different l-values with the results obtained in Tyssedal and Hussain[17] with the ΔR^2-method and the F-test. It is easily seen that our proposed procedure outperforms the ΔR^2-method and the F-test method in all cases when $l = 4$ and 6. However, when $l = 7$, the ΔR^2-method

TABLE 4 The 16-run NC design with six factors

A	B	C	D	E	F
−1	−1	−1	−1	1	−1
1	−1	−1	−1	−1	−1
−1	1	−1	−1	−1	1
1	1	−1	−1	1	−1
−1	−1	1	−1	−1	1
1	−1	1	−1	1	1
−1	1	1	−1	1	1
1	1	1	−1	−1	−1
−1	−1	−1	1	−1	1
1	−1	−1	1	1	1
−1	1	−1	1	1	−1
1	1	−1	1	−1	1
−1	−1	1	1	1	−1
1	−1	1	1	−1	−1
−1	1	1	1	−1	−1
1	1	1	1	1	1

and the F-test have slightly higher capture frequencies for small values of σ^2. What is also apparent is that our procedure is more robust to increasing the variance than the ΔR^2-method and the F-test. As expected, the number of experimental factors affects the performance. The more experimental factors, the higher r should be used.

One problem appeared when choosing seven terms in the case of zero variance. The model with the correct factors was never the best, and when considering 11 experimental factors, it was not even among the 10 best. But the MSEs of the top 20 models were very similar, indicating that several sets of factors can explain this response equally well. Equivalent models sometimes occur when using the PB12 design due to the complex alias structure, making it more likely that some linear combinations are equivalent to the true model the more terms that are included. Therefore, only testing a small panel of models is not advisable, as the results may be strongly affected when such equivalent cases exist.

5 | A SIMULATION STUDY OF THE OVERALL PERFORMANCE

To assess the overall performance of our procedure, we have tried it out on a wide range of models. The designs used are six or more design columns from the 12-run PB design and the three 16-run NC designs given in Tables 4–6. For designs with 12 runs and n_t experimental factors, the n_t first columns from Table 3 will always be used. Note that the 12-run PB design has two different projections onto 5 and 6 dimensions. Table 3 is written in a form that contains the one preferred by Wang and Wu[20] in the first six columns. For all other dimensions, the projections are isomorphic.

The 16-run designs were chosen to examine how much gain in capture frequency that is obtained by using four more experimental runs. Also, their performance in a screening situation is, to our knowledge, not well tested out. The three NC designs presented in Tables 4–6 are for each number of factors just one out of several options. For six experimental factors, the design with the highest numbers of full 2^4 projections was chosen. It is made up of a 2^{5-1} design with generator E = ABCD and an additional factor column F generated as F = $\frac{1}{2}$(AD+ABD−CD+BCD), and can be found in Table 4. For seven and eight experimental factors, we use designs that are isomorphic to the ones proposed by Montgomery and Jones.[18] They can be found in Tables 5 and 6.

5.1 | A general procedure for testing the size-based method

The procedure was tested out through simulations for cases with both three and four active factors, using several model formats and various levels of noise. The models selected were submodels of the FP-models. Given the format and the

TABLE 5 The 16-run NC design with seven factors

A	B	C	D	E	F	G
−1	−1	−1	−1	1	−1	−1
1	−1	−1	−1	−1	−1	1
−1	1	−1	−1	−1	1	1
1	1	−1	−1	1	−1	−1
−1	−1	1	−1	−1	1	−1
1	−1	1	−1	1	1	1
−1	1	1	−1	1	1	−1
1	1	1	−1	−1	−1	1
−1	−1	−1	1	−1	1	1
1	−1	−1	1	1	1	−1
−1	1	−1	1	1	−1	1
1	1	−1	1	−1	1	−1
−1	−1	1	1	1	−1	1
1	−1	1	1	−1	−1	−1
−1	1	1	1	−1	−1	−1
1	1	1	1	1	1	1

TABLE 6 The 16-run NC design with eight factors

A	B	C	D	E	F	G	H
−1	−1	−1	−1	−1	1	1	1
1	−1	−1	−1	1	1	−1	1
−1	1	−1	−1	1	−1	1	−1
1	1	−1	−1	−1	−1	1	1
−1	−1	1	−1	1	−1	−1	1
1	−1	1	−1	−1	1	1	−1
−1	1	1	−1	−1	1	−1	−1
1	1	1	−1	1	−1	−1	−1
−1	−1	−1	1	1	1	−1	−1
1	−1	−1	1	1	−1	1	−1
−1	1	−1	1	−1	−1	−1	1
1	1	−1	1	−1	1	−1	−1
−1	−1	1	1	−1	−1	1	−1
1	−1	1	1	−1	−1	−1	1
−1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1

noise level, active factors and effects were drawn randomly, and the size of the effects drawn uniformly within specified intervals in each of 10,000 simulations. The full description of the procedure for testing the size-based method proposed in Section 3 is as follows:

1. Specify the format of the model: Number of active factors, n_a, number of candidate effects, n_e, number of main effects, n_m, minimum absolute value of the coefficients, b_{min}, maximum absolute value of the coefficients, b_{max}.
2. Specify the variance of the noise added to the response, σ^2.
3. Specify number of terms in the reduced model, l.
4. Draw the active factors, randomly distribute their effects between main effects and two-factor interactions. Draw the corresponding coefficients from a uniform distribution on the interval $[b_{min}, b_{max}]$, multiply with −1 or 1, drawn randomly with equal probability.

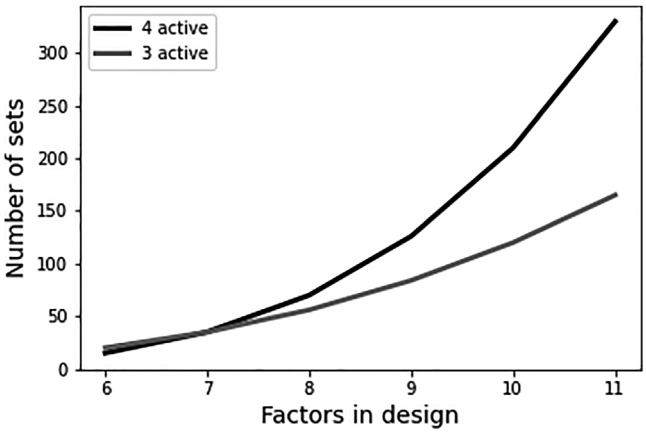


FIGURE 2 The number of possible candidate sets for three and four active factors as a function of the number of experimental factors

TABLE 7 $CF_1(10,000)$ for the 16-run NC-designs varying σ^2 in the case of three active factors. The simulated models have three main effects and three interaction effects and an absolute effect size between 1 and 3. All terms in the FP-model were chosen for the reduced models.

$CF_1(10,000)$ for the 16-run designs											
	σ^2										
n_t	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
6	10,000	9982	9968	9979	9964	9948	9940	9920	9912	9915	9899
7	10,000	9979	9968	9939	9955	9919	9898	9871	9857	9831	9840
8	10,000	9963	9931	9901	9855	9819	9803	9749	9698	9678	9651

5. Using the design considered, simulate responses adding error terms drawn from a normal distribution with mean zero and variance σ^2.
6. Apply the proposed algorithm and check if the correct set of active factors were used to construct any of the r reduced models with the smallest MSE.

In all cases, the $CF_r(10,000)$ for $r = 1, 5, 10$ and 15 was recorded. When $CF_r(10,000) = 10,000$ for all levels of σ^2, it is not presented in the result tables. Checking the performance for several levels of σ^2 is useful to give an indication of the most suitable size of the capture set. It is important to be aware of that the number of possible sets of active factors rapidly increases when the number of factors in the design increases. Figure 2 shows the number of sets as a function of the number of factors. For instance, when considering six factors in the design, there are only 20 possible sets of three factors, while if there are 11 factors in the design, there are 165. Thus being able to reduce the candidate set to 5, 10 or 15 is relatively more useful for designs with many experimental factors.

An important point to note about the simulations is that b_{min} was chosen as the value of the largest variance tested, while b_{max} was three times the value of the largest variance. If the response variance is much larger than the coefficients, it is believed to be very hard to find the correct model when using as few runs as 12 or 16.

5.2 | Identifying three active factors

First, the simplest case of three active factors was considered. As there are only seven terms in the full projection model, l was set to 7 and the mean square errors of the full projection models were compared. The simulated models were chosen to have three main effects and three interaction effects, all with coefficients with an absolute value between 1 and 3. The results were very good for both the 12- and 16-run designs. For the 16-run NC designs, the capture frequencies were almost always 10,000 when using $r = 5$, 10 and 15. The only exception was in the case of eight factors in the design and a variance of 1 when choosing the five best factors. Then the capture frequency was 9997. The results when choosing the very best model were also highly satisfactory, as shown in Table 7. When having a 95% chance of finding the correct model

TABLE 8 $CF_r(10000)$ for the PB12 design varying σ^2, the size of the capture set, r, and the number of experimental factors, n_t, in the case of three active factors. The simulated models have three main effects and three interaction effects and an absolute effect size between 1 and 3. All terms in the FP-model were chosen for the reduced models.

	σ^2										
r	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
						$n_t = 6$					
1	10,000	9886	9770	9684	9577	9485	9391	9250	9120	9031	8897
5	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9999	9997
						$n_t = 7$					
1	10,000	9822	9659	9430	9293	9144	8912	8763	8626	8453	8277
5	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9997	9996	9992	9992
						$n_t = 8$					
1	10,000	9655	9387	9048	8767	8588	8285	7981	7763	7599	7296
5	10,000	10,000	10,000	10,000	9999	9997	9994	9989	9983	9970	9939
10	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9999	9999	9998
						$n_t = 9$					
1	10,000	9522	9077	8652	8301	7953	7591	7306	6937	6670	6454
5	10,000	10,000	10,000	9997	9995	9988	9979	9947	9937	9915	9863
10	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9998	10,000	9990	9988
15	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9997	9999
						$n_t = 10$					
1	10,000	9370	8820	8271	7856	7407	7005	6629	6339	5967	5705
5	10,000	9997	9990	9993	9980	9961	9917	9886	9821	9730	9709
10	10,000	10,000	10,000	10,000	10,000	10,000	9998	9988	9987	9971	9969
15	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9998	9997	9991	9993
						$n_t = 11$					
1	10,000	9271	8606	8020	7448	6901	6599	6093	5749	5447	5101
5	10,000	9998	9980	9964	9930	9889	9805	9720	9639	9543	9408
10	10,000	10,000	10,000	10,000	10,000	9991	9994	9974	9956	9948	9924
15	10,000	10,000	10,000	10,000	10,000	9998	10,000	9995	9987	9988	9973

is considered good enough, selecting the best model in a search for three active factors is an acceptable strategy when using a 16-run design.

The 12-run PB design did, naturally, not perform as well as the 16-run designs, but when using $r = 5$, the model with the correct factors was almost always found in at least 95% of the cases. Thus being able to reduce the number of candidate sets down to five, using the proposed size-based method, is likely for the 12-run PB design even with 11 experimental factors. The results can be found in Table 8. To ease the comparison, the results corresponding to using $r = 1$ and $r = 5$ are plotted in Figure 3, for all the cases tested. It is easily seen that the 16-run designs perform better than the PB12 design for the same number of factors, and that the capture frequencies decrease with increasing number of experimental factors, as one would expect. Note that as 16-run designs with more than eight factors were not tested, only designs with six, seven and eight experimental factors can be fairly compared for 12 and 16 runs.

5.3 | Identifying four active factors

Besides some examples and the work of Tyssedal and Hussain,[17] there is to our knowledge limited information of how well the PB12 design performs when four factors are active. However, the above-mentioned work indicates that it is substantially more difficult to identify the right active factors when four are active compared to when three are. The first simulated models were specified to have four main effects and 2 two-factor interaction effects, all with coefficients with absolute values between 1 and 3. The reduced models included $l = 6$ terms. Results using the 16-run NC designs are pre-

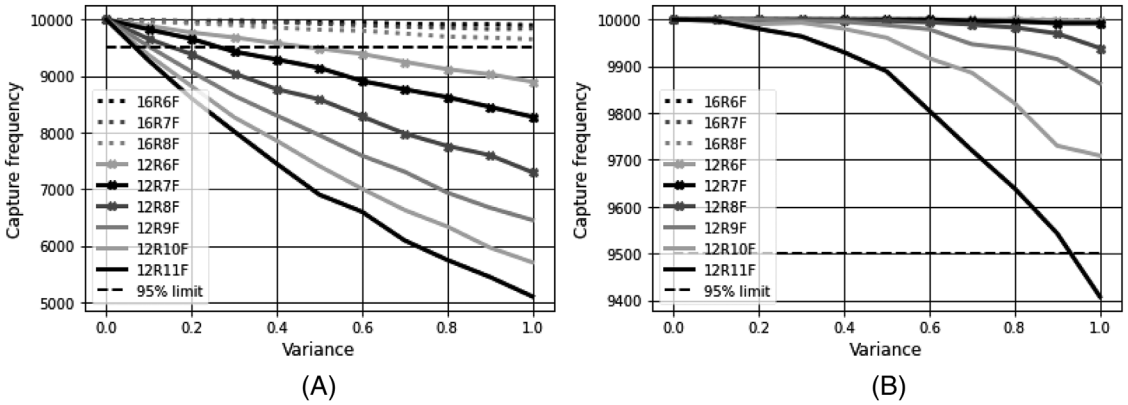


FIGURE 3 Plot of $CF_r(10,000)$ against variance varying the size of the capture set, r, and the number of experimental factors, n_t, for both 12- and 16-run designs having three active factors. The simulated models have three main effects and three interaction effects, and an absolute effect size between 1 and 3. All terms in the FP-model were chosen for the reduced models. R denotes the number of rows in the design, and F the number of factors. Note that the y-axis is different for each plot, but the black dashed lines show the 95% limit in all cases. (A) $CF_r(10,000)$ when $r = 1$, (B) $CF_r(10,000)$ when $r = 5$

TABLE 9 $CF_r(10,000)$ for the 16-run designs in the case of four active factors, varying σ^2 and the capture set size, r. The simulated models have four main effects and 2 two-factor interaction effects and an absolute effect size between 1 and 3. The number of terms in the reduced model is $l = 6$.

	σ^2										
r	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
					$n_t = 6$						
1	10,000	9862	9809	9776	9731	9673	9644	9567	9555	9496	9466
					$n_t = 7$						
1	10,000	9788	9645	9606	9535	9437	9319	9290	9219	9138	8993
5	10,000	9999	9999	10,000	9999	10,000	9997	9995	9996	9994	9992
					$n_t = 8$						
1	10,000	9616	9367	9085	9020	8828	8720	8496	8338	8180	8003
5	10,000	9995	9985	9962	9976	9952	9944	9918	9913	9883	9859
10	10,000	10,000	10,000	10,000	9999	9997	9997	9996	9994	9989	9991
15	10,000	10,000	10,000	10,000	10,000	10,000	10,000	9999	9999	10,000	9999

sented in Table 9. The capture frequencies when considering $r = 1$ decline quickly when the number of factors in the design is increased. This is reasonable, given that there exists 15 ways to choose four active factors among six candidate factors, and 70 ways to choose four active factors among eight candidate factors. Despite this, using $r = 5$ is sufficient for having a capture frequency well above 95% for all design sizes.

The results for the 12-run PB design with different numbers of factors in the design can be found in Table 10. In this case, using $r = 1$ for finding the active factors is not advisable, as the capture frequencies are then quite low. However, for reducing the number of candidate sets, the method yields satisfactory results in many cases. Including up to eight experimental factors in the design, using $r = 10$ yields a capture frequency above 95% in all but two cases. For more than eight factors in the design, $r = 10$ yields satisfactory results for low levels of noise. When suspecting a rather high variance, one may use $r = 15$ to improve the chances that the correct active factors are included in the capture set. For instance, when there are nine factors in the design, choosing $r = 15$ instead of $r = 10$ increases the maximal σ^2 for which the success probability is above 95% from 0.5 to 0.7.

To compare the general difference in performance for the 12- and 16-run designs, the results were plotted for the case of selecting the best and the five best models in Figure 4. When selecting the 10 best models, the results were very close to 10,000 for the 16-run designs, hence only results for the 12-run design were plotted in Figure 5. The plots leave little doubt that using 16-run designs are recommendable whenever possible, but using the 12-run designs with the same number of

TABLE 10 $CF_r(10,000)$ for the PB12 design in the case of four active factors with varying σ^2, capture set size r and number of experimental factors n_t. The simulated models have four main effects and 2 two-factor interaction effects and an absolute effect size between 1 and 3. The number of terms in the reduced model is $l = 6$.

	σ^2										
r	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
					$n_t = 6$						
1	10,000	9214	8771	8344	7875	7626	7267	6919	6639	6387	6117
5	10,000	9996	9984	9985	9968	9950	9943	9913	9878	9829	9773
10	10,000	10,000	9999	10,000	10,000	9999	10,000	9999	9996	9991	9993
					$n_t = 7$						
1	10,000	8624	8053	7417	6920	6477	5873	5606	5168	4891	4594
5	10,000	9968	9930	9868	9825	9714	9622	9462	9302	9179	8989
10	10,000	9999	9999	9995	9985	9964	9957	9943	9904	9855	9822
15	10,000	10,000	10,000	10,000	10,000	9996	9996	9988	9982	9973	9970
					$n_t = 8$						
1	10,000	8246	7338	6644	5953	5398	4834	4529	4093	3700	3554
5	10,000	9897	9786	9608	9412	9194	8921	8706	8433	8097	7862
10	10,000	9989	9968	9928	9884	9820	9749	9620	9535	9362	9224
15	10,000	9996	9992	9984	9959	9943	9912	9851	9814	9762	9656
					$n_t = 9$						
1	10,000	7849	6738	5800	5106	4559	3986	3553	3233	2774	2609
5	10,000	9834	9594	9313	8977	8632	8206	7840	7434	6883	6644
10	10,000	9964	9889	9819	9658	9542	9320	9089	8856	8540	8313
15	10,000	9994	9971	9944	9886	9826	9706	9596	9436	9248	9064
					$n_t = 10$						
1	10,000	7489	6220	5169	4458	3835	3371	2952	2524	2224	1986
5	10,000	9748	9414	8975	8507	7960	7562	7058	6532	6043	5634
10	10,000	9927	9798	9613	9410	9114	8842	8433	8098	7692	7372
15	10,000	9971	9921	9814	9688	9560	9375	9080	8851	8568	8372
					$n_t = 11$						
1	10,000	7119	5665	4555	3883	3284	2736	2370	2024	1746	1564
5	10,000	9608	9147	8526	7925	7267	6803	6136	5685	5161	4731
10	10,000	9866	9663	9344	8967	8561	8235	7748	7287	6833	6441
15	10,000	9942	9817	9664	9426	9149	8882	8529	8129	7773	7442

factors can also yield good results if one selects several candidate sets of active factors for further investigation and the variance is not too high.

5.4 | Testing different model specifications

Having demonstrated that the method works well for a given format for four active factors, it is interesting to see if the results are impacted by using different specifications for the simulated models. The models used in the previous section all had four main effects and 2 two-factor interactions. To check how the number and type of active effects affect the result, a panel of model types with four active factors and different specifications was tested:

1. Six active effects (four main effects, two two-factor interactions)
2. Six active effects (three main effects, three two-factor interactions)
3. Six active effects (two main effects, four two-factor interactions)
4. Four active effects (two main effects, two two-factor interactions)

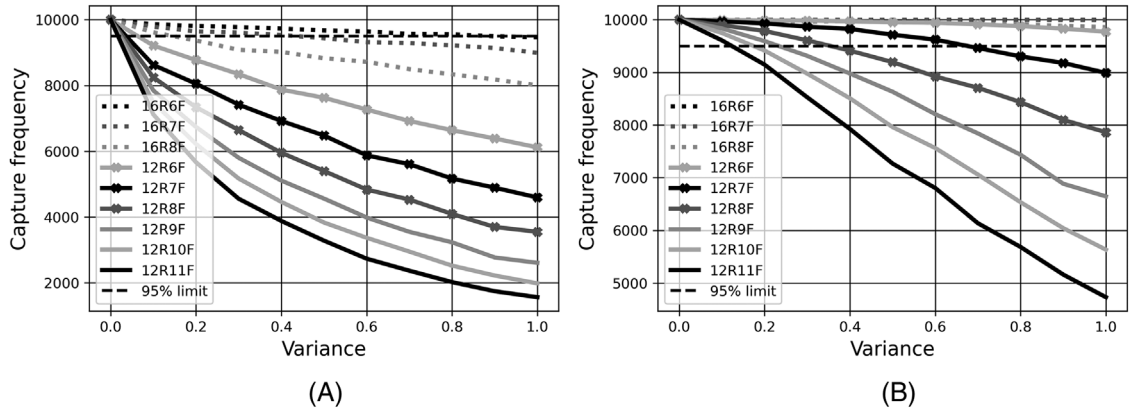


FIGURE 4 Plot of $CF_r(10,000)$ against variance for the 12- and 16-run designs with different numbers of factors in the design, using capture set size $r = 1$ and 5, respectively. The simulated models have four main effects and 2 two-factor interaction effects, and an absolute effect size between 1 and 3. The number of terms in the reduced model is $l = 6$. R denotes the number of rows in the design, and F the number of factors. The y-axis is different for each plot, but the black dashed lines show the 95% limit in all cases. (A) $CF_r(10,000)$ when $r = 1$, (B) $CF_r(10,000)$ when $r = 5$

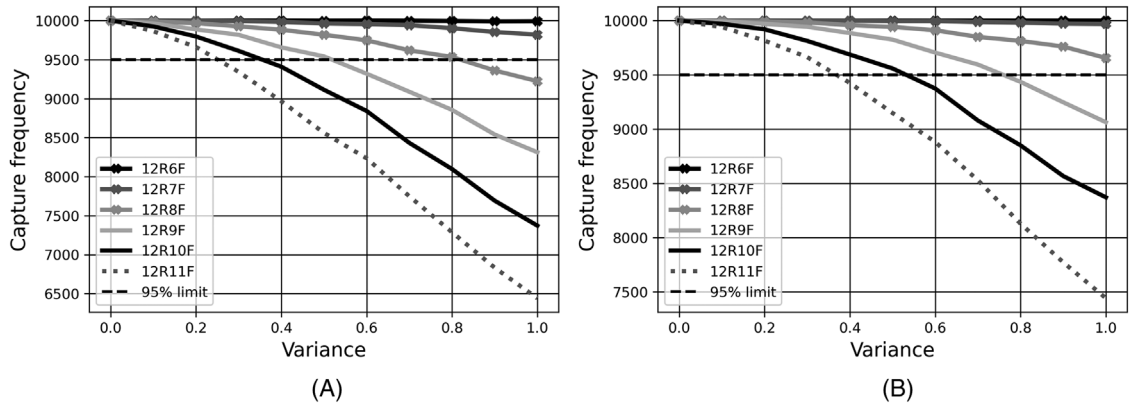


FIGURE 5 Plot of $CF_r(10,000)$ against variance for the 12-run PB design with different numbers of factors in the design, using capture set size $r = 10$ and $r = 15$, respectively. The simulated models have four main effects and 2 two-factor interaction effects, and an absolute effect size between 1 and 3. The number of terms in the reduced model is $l = 6$. R denotes the number of rows in the design, and F the number of factors. The y-axis is different for each plot, but the black dashed lines show the 95% limit in all cases. (aA) $CF_r(10,000)$ when $r = 10$, (B) $CF_r(10,000)$ when $r = 15$

The first specification is the one used in the previous section. The third is motivated from machine learning. When design of experiments is used for tuning of hyperparameters in algorithms like random forests, experience has shown that many two-factor interactions may appear in the screening phase, see Vatnedal.[23] All specifications were tested for different numbers of factors in the design, choosing $l = 6$ terms for the reduced models. The results for the 12-run designs can be found in Figure 6. In the plots, results are shown when choosing $r = 1$, and when choosing the r one would typically use for that model size (either 5, 10 or 15, depending on the capture frequency). The results seem to vary more when using $r = 1$ than when $r = 5, 10$ or 15. This is reassuring, as one would typically not choose only the best model. In general, the smallest model with only four active effects yields slightly better results than the models with six active effects, suggesting that sparse models make the active factors easier to find than large models. For the models with six active effects, the results are slightly worse for the models with three main effects and three two-factor interactions than the others.

The same panel of specifications was also tested for the 16-run nonregular NC design, using $l = 6$ in the reduced models, and the results can be seen in Figure 7. In this case, the sparsest models with only four active effects gave poor results when choosing $r = 1$. This might seem strange as this specification performed well in the 12-run case, and now it did not even yield a capture frequency above the 95% limit when the variance was zero. This is due to the aliasing pattern of the 16-run design. For the six and seven factor design both E = ABCD and F = $\frac{1}{2}$(AD+ABD−CD+BCD) are generators. When only

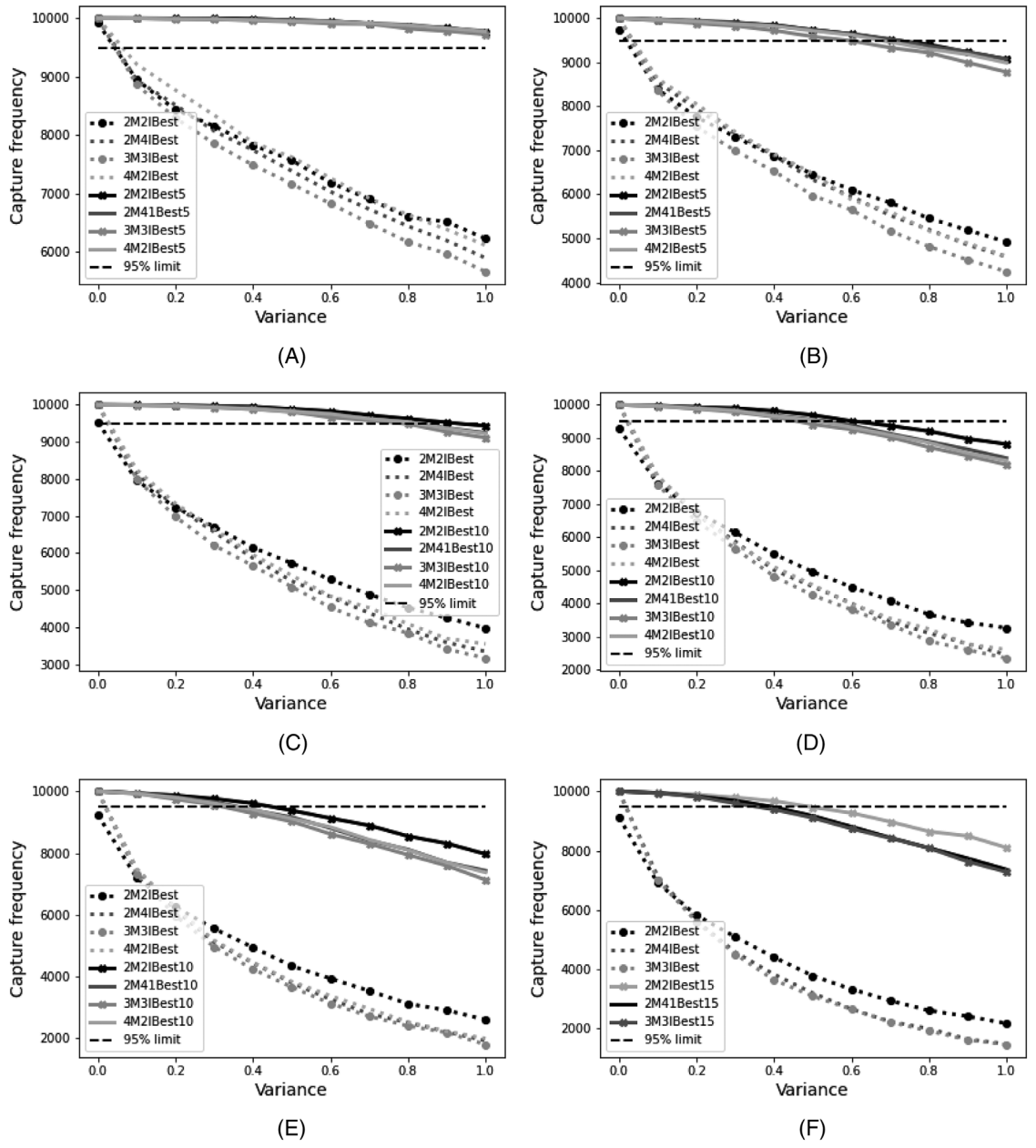


FIGURE 6 Plot of $CF_r(10,000)$ against variance for different model specifications using a PB12 design. 2M2IBest does for instance denote two main effects, 2 two-factor interactions and $r = 1$. In all cases, there were four active factors, and the absolute effect size was between 1 and 3. The number of terms in the reduced models is $l = 6$. (A) Six factors in the design, (B) seven factors in the design, (C) eight factors in the design, (D) nine factors in the design, (E) 10 factors in the design, (F) 11 factors in the design

four effects are active, but more effects are chosen for the reduced models, it is possible to construct alternative models, which are linearly equivalent to the true model.

For instance, if the true model has the active effects C, D, BC and DF. Then a linearly equivalent model can be constructed using the effects A, C, D, BC and AB. This is because $DF = \frac{1}{2}(A+AB-C+BC)$. But the plots also show that the correct model is found among the five best models almost equally often when there are four active effects as when there are six active effects. This effectively demonstrates that one should always consider choosing a candidate set of models for further investigation when using nonregular designs. Then one may proceed the analysis by testing reduced models with different values of l. If a small model has only a slightly higher MSE than a larger one with different active factors, it could indicate that the larger model is just another representation of the small one.

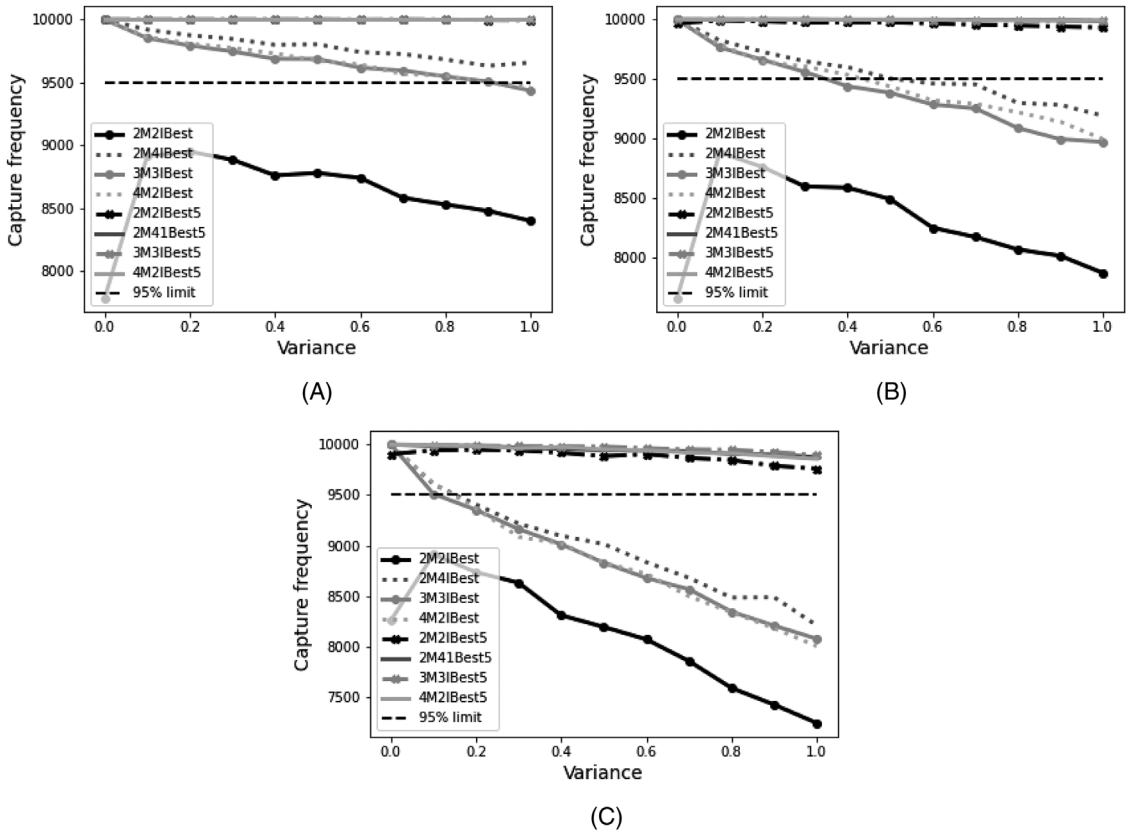


FIGURE 7 Plot of $CF_r(10,000)$ against variance for the 16-run design using different model specifications. 2M2IBest does for instance denote two main effects, 2 two-factor interactions and $r = 1$. In all cases, there were four active factors, and the absolute effect size was between 1 and 3. The number of terms in the reduced model is $l = 6$. (A) Six factors in the design, (B) seven factors in the design, (C) eight factors in the design

TABLE 11 Factors and levels for investigating the possible size of error variance in order to have a capture frequency of 95%

Symbol	Factor	Levels
A	Size of capture set, r	5, 10, 15
B	Number of excess terms in reduced model, $l - n_e$	0, 1, 2
C	Number of experimental factors, n_t	7, 9, 11
D	Number of terms in the model, n_e	4, 5, 6

5.5 | Evaluating the screening performance

From what is observed, factors like the size of the capture set, r, the number of terms in the reduced model, l, the number of experimental factors, n_t, and the number of terms in the true model, n_e, will affect the outcome of a screening. To investigate the effect of these factors for the 12 run PB design, a 3^4 experiment was conducted using the largest error variance for which a capture frequency above 95% can be obtained, from now on called the *capture variance*, as the response. Factors and levels are given in Table 11.

Capture frequencies based on 1000 simulations were used in the experiment. The models had four active factors and the number of terms, n_e, in the models was chosen to be 4, 5 and 6, always including 2 two-factor interactions. The data were analysed using the alternative analysis method given in Wu and Hamada,[24] page 287. Linear and quadratic effects were estimated by setting low, medium and high levels to (−1, 0, 1) and (1, −2, 1), respectively. No scaling to unit length was performed. A logarithmic transformation is often employed for variance modelling. However, in this case, the square root gave residuals better approximated to a normal distribution. Following the notation in Wu and Hamada,[24]

TABLE 12 Values of estimated capture standard deviations varying the size of the capture set, r, and the number of experimental factors, n_t. $l = n_e = 4$. The models have four active factors.

$r \backslash n_t$	7	9	11
5	1.02	0.73	0.60
10	1.43	0.92	0.72
15	1.85	1.11	0.85

the following model was estimated for the capture standard deviation with all terms being significant at a 5% level: $\hat{\sigma} = 0.766 + 0.263A_l - 0.114B_l - 0.366C_l - 0.134D_l + 0.055C_q + 0.012D_q + 0.018AB_{ll} - 0.147AC_{ll} - 0.023CD_{ll} + 0.022AC_{lq}$.

The subscripts l, q, ll and lq are used to denote linear, quadratic, linear-by-linear and linear-by-quadratic effects, respectively. The linear effects dominate together with the linear by linear interaction AC and the quadratic effect of C. As expected, the capture standard deviation is higher if r is high and if we have few experimental factors. It is an advantage that l is equal to n_e, and that the model has few terms. The linear-by-linear AC interaction has as a consequence that the effect of increasing r will decline when n_t increases. Table 12 gives capture standard deviation for different values of r and n_t with $l = n_e = 4$. Since the effect of increasing $l - n_e$ and n_e is mainly linear, it is rather easy to adjust for other values of these parameters.

We notice that the negative effect on the capture standard deviation of increasing n_t decreases when n_t increases. Overall, the results show that for the 12-run PB design, even with $n_t = 11$ and four active factors, it is in many cases possible to reduce the number of candidate sets down to 5. Only n_t is known in the beginning of the screening, but once performed, one will have values for l, and estimates for n_e and σ. The suitability of l can be checked against the models in the capture set. Since the procedure is scale invariant as pointed out in Section 2, the value of $\frac{\hat{\sigma}}{\hat{b}_{min}}$ can be used to check against the numbers in Table 12 and should, together with the model for $\hat{\sigma}$, provide useful information about what r to use in order to have a reasonable certainty that the 'correct' candidate set is captured. In this way, the reliability of the screening performance can be evaluated even if several important parameters are unknown from the beginning.

However, all the calculation are performed under the assumption that $\frac{b_{max}}{b_{min}} = 3$. It is meant to constitute a normal situation, but obviously this is not always true. In Figure 8, the capture frequencies are plotted for $\frac{b_{max}}{b_{min}} = \upsilon$ for $\upsilon = 1.5, 2.0, \ldots, 4$ in steps of 0.5, and for two models with four active factors. One of the models has four active effects and one has six. We notice that the capture frequency is decreasing with increasing σ^2 and n_t. For $\upsilon = 2$, the reduction in the estimated capture standard deviation compared to when $\upsilon = 3$ will be about 20%–30% for the model with four terms and 30%–40% for the model with six terms. For $\upsilon = 1.5$ these intervals are about 30-40% and 45%–70%. However, when running simulations with a given υ, the ratio between the largest and smallest coefficient will most likely be smaller than υ in absolute value. Therefore, numbers like the ones above and in Table 12 are a little pessimistic. We notice that for up to eight experimental factors, the method performs reasonably well even for quite high variances.

Figure 9 explains why the problem of identifying active factors is much harder when the effect range is small. Response values are simulated from a model with four active factors and six terms. No noise is added. Our procedure is then used to find the MSE of all the 126 candidate sets for nine experimental factors. The average of the nine candidate sets with a MSE closest to zero is plotted against υ. The smaller the υ the smaller the average, making it easier for other candidate sets than the correct one to be in the capture set. Increasing $l - n_e$ makes the problem of identifying the active factors harder.

6 | AN EXAMPLE WITH REAL DATA

Phoa et al.[25] reanalysed three real chemical experiments where the PB12 design was used, in order to demonstrate shortcomings of the traditional analysis approach. Here, their third example, an experiment regarding chemical characterization of grapes originally taken from Dopico-Garcia et al.,[26] will be considered. The response was the extraction of phenolic compounds, measured in area divided by amount of sample (see the original paper for details). Table 13 shows the factors and levels considered in the experiment, while Table 14 shows the experimental design and the corresponding responses. Note that the columns in the PB12 design in Table 14 are written in a different manner than the ones in Table 3.

Phoa et al.[25] found that the active factors were A, C and D. They proposed the following model: $\hat{Y} = 5.51 + 1.11C - 1.03D + 1.73AD$. Using the size-based method, assuming four active factors and including four, five and six terms in the reduced models yielded the same active factors, and factor F in addition. Using three terms in the reduced models was also tested,

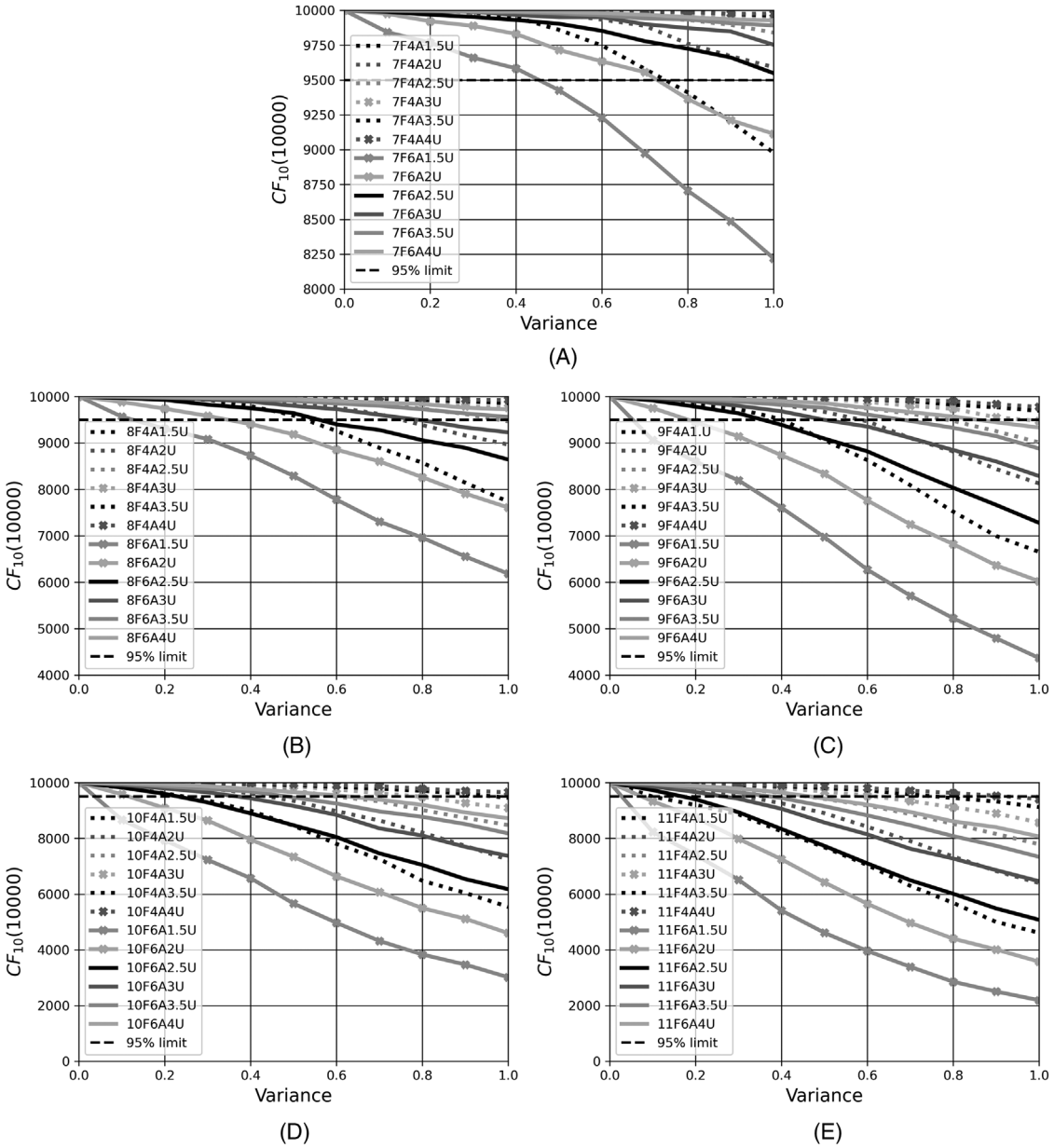


FIGURE 8 Plots of capture frequencies when testing different effect ranges. The minimum effect size was always 1, while the upper (U) was varied from 1.5 to 4. F is the number of factors in the design, while A is the number of active effects. There were always two active two-factor interactions. $l = n_e$ and $r = 10$. Note that the scale of the y-axis is different for each row. (A) Seven factors in the design, (B) eight factors in the design, (C) nine factors in the design, (D) 10 factors in the design, (E) 11 factors in the design

in which case the factors A, C and D were included in all candidate sets in accordance with the model chosen in Phoa et al.[25]

The active factors and the MSE for the five best reduced models when assuming three and four active factors can be found in Tables 15 and 16. Assuming three active factors, the factors A, C and D were always chosen. Note that for models with four, five and six terms, the MSE of the supposedly correct model is about half the size or less than the next best MSE. In the case of selecting three terms and assuming four active factors, the factors A, C and D are always included in the candidate sets, and all models yield the same MSE. It seems like a model with the same three factors is always chosen, despite the possibility of including an additional factor as long as there are only three terms. To consider whether there are three or four active factors, the models should be more thoroughly investigated.

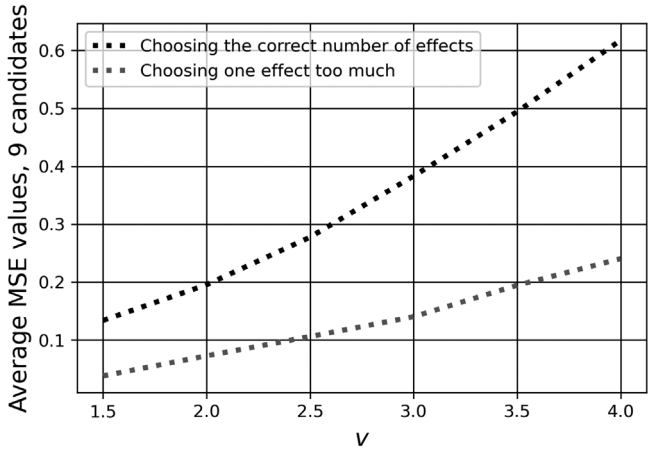


FIGURE 9 The average MSE of the nine best candidate models for different effect ranges. The values are based on 1000 iterations, using a design with nine factors, a minimum effect size of 1, and 0 variance. The simulated models had three active main effects and three active two-factor interactions.

TABLE 13 Factors and levels in the compound extraction experiment from Dopico-Garcia et al.[26]

Symbol	Factor	Unit	Low factor level (−)	High factor level (+)
A	Extraction solvent		Acid water	MeOH
B	Extraction volume	ml	50	250
C	Extraction time	min	5	20
D	Temperature	°C	40	50
E	Extraction type		Ultrasonic	Stirring
F	Sorbent type		EC	NEC
G	Elution solvent		EtOH	MeOH
H	Elution volume	ml	20	150

TABLE 14 Design matrix and responses for the real data from Dopico-Garcia et al.[26]

Run	A	B	C	D	E	F	G	H	Response Y
1	1	−1	1	−1	−1	−1	1	1	6.98
2	1	1	−1	1	−1	−1	−1	1	5.31
3	−1	1	1	−1	1	−1	−1	−1	9.67
4	1	−1	1	1	−1	1	−1	−1	6.45
5	1	1	−1	1	1	−1	1	−1	5.23
6	1	1	1	−1	1	1	−1	1	5.34
7	−1	1	1	1	−1	1	1	−1	4.03
8	−1	−1	1	1	1	−1	1	1	3.76
9	−1	−1	−1	1	1	1	−1	1	2.10
10	1	−1	−1	−1	1	1	1	−1	2.65
11	−1	1	−1	−1	−1	1	1	1	7.40
12	−1	−1	−1	−1	−1	−1	−1	−1	7.14

Different models assuming three and four active factors are given in Table 17, along with a list of terms having a p-value above 0.01, and the adjusted AIC, $\text{AIC}_a = n \ln \frac{\text{SSE}}{n} + \frac{2p(n+1)}{n-p}$. Here the sum of squared errors, SSE, is given by $\text{SSE}= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The difference between AIC and the AIC_a is that AIC_a punishes the addition of new terms more heavily than the AIC, for which the penalty is only $2p$. AIC_a is in particular considered suited for small sample sizes. In the case

TABLE 15 The active factors and their corresponding MSE for the five best models, when assuming three active factors and choosing 3, 4, 5 and 6 terms in the reduced models, respectively

(a) $l = 3$			(b) $l = 4$		
Rank	**Factors**	**MSE**	**Rank**	**Factors**	**MSE**
1	A C D	0.314364	1	A C D	0.243252
2	B C F	0.838144	2	B C F	0.553700
3	A D E	0.937919	3	A D E	0.599666
4	A D F	1.100919	4	E G H	0.817199
5	D E H	1.126051	5	A D F	0.894516
(c) $l = 5$			**(d) $l = 6$**		
Rank	**Factors**	**MSE**	**Rank**	**Factors**	**MSE**
1	A C D	0.162919	1	A C D	0.121591
2	A D E	0.523941	2	A D E	0.492138
3	B C F	0.531299	3	B C F	0.521731
4	E G H	0.560616	4	E G H	0.523125
5	A D F	0.723891	5	C E H	0.543581

TABLE 16 The active factors and their corresponding MSE for the five best models, when assuming four active factors and choosing 3, 4, 5 and 6 terms ion the reduced models, respectively

(a) $l = 3$			(b) $l = 4$		
Rank	**Factors**	**MSE**	**Rank**	**Factors**	**MSE**
1	ABCD	0.314	1	ACDF	0.123
2	ACDH	0.314	2	ACDG	0.243
3	ACDF	0.314	3	ACDE	0.243
4	ACDE	0.314	4	ABCD	0.283
5	ACDG	0.314	5	ACDH	0.283
(c) $l = 5$			**(d) $l = 6$**		
Rank	**Factors**	**MSE**	**Rank**	**Factors**	**MSE**
1	ACDF	0.055	1	ACDF	0.023
2	ABCD	0.163	2	ACDG	0.061
3	ACDH	0.163	3	ACDE	0.082
4	ACDE	0.177	4	DEGH	0.090
5	ABCF	0.190	5	ABCD	0.115

TABLE 17 Evaluation of different models with three and four active factors for the grapes data from Dopico-Garcia et al.[26] The intercept is not counted in the number of terms (T), as it is always included.

F	T	Model	AIC_a	p-value > 0.01
3	3	5.51+1.11C−1.03D+1.73AD	33.88	None
3	4	5.51+1.11C−1.03D+1.73AD−0.27CD	37.09	CD(0.20)
3	5	5.51−0.30A+1.11C−1.03D+1.73AD−0.37CD	41.08	A(0.14),CD(0.08)
3	6	5.43−0.30A+1.11C−1.03D+1.73AD−0.37CD−0.22ACD	50.77	ACD(0.25), A(0.13),CD(0.08)
4	3	5.51+1.11C−1.03D+1.73AD	33.88	None
4	4	5.51+1.30C−1.19D+1.79AD−0.50AF	28.95	AF (0.01)
4	5	5.51+1.26C−1.19D−0.28F+1.69AD−0.48AF	27.95	F (0.03)
4	6	5.51−0.18A+1.26C−1.19D−0.28F+1.69AD−0.48AF	30.66	A(0.05), F(0.01)
4	7	5.51−0.18A+1.24C−1.14D−0.33F+0.13AC+1.67AD−0.46AF	43.52	A(0.03), AC(0.10)

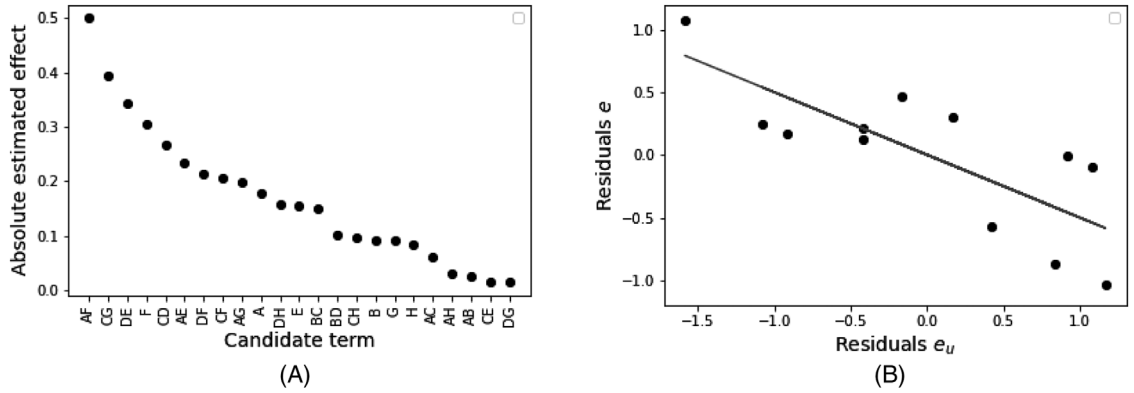


FIGURE 10 (A) An added variable Pareto plot for extending the model given in Phoa et al.[25] by one factor. (B) An added variable plot for the two-factor interaction AF. (A) All effects, (B) AF

of three active factors, AIC_a has a minimum for the original model, thus it seems like the best choice in that case. When allowing four active factors, AIC_a has a minimum for the model with five terms in the model. This is rather surprising, as it was not the model chosen by Phoa et al.[25] The difference from the originally chosen model is that the factor F is added through the main effect F and the interaction effect AF. Both effects are significant at a 5% level. In fact, the factor F is present in one or several terms in all models with four active factors and more than three terms. As shown in this example, the proposed screening method can be an effective start for performing model selection, as the candidate models are fitted as a part of the procedure.

From Table 15, it is clear that the analysis of these data very well might have ended concluding that the three factors A, C and D are the active ones. As a useful method for considering if additional factors should be added, we will now introduce the *added variable Pareto plot* (AVPP). Assume our current model is described by the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, with corresponding hat matrix $\boldsymbol{H}$. Adding one regressor variable, u, with corresponding design column $\boldsymbol{u}$, the new model becomes $\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{u}\beta_u + \boldsymbol{\epsilon}$. The least squares estimator for β_u is then given by $\hat{\beta_u}=\frac{\boldsymbol{u}'(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}}{\boldsymbol{u}'(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{u}}$. Estimating β_u for all terms u that extend the number of active factors by 1 may inform us if it is worth looking for more active factors. The corresponding estimated β_us may be ranked according to their absolute values and plotted in a Pareto plot to see what terms (or factors) that most likely should be added. Such an AVPP is shown in Figure 10A, where we have let u in turn be all main effects and two-factor interactions that extend the number of active factors by 1. The largest term in absolute value is the two-factor interaction AF, telling us that F may be the most important factor to add to A, C and D.

Figure 10B shows an added variable plot, as described in Abraham and Ledolter.[27] It works as follows: The residuals from the fitted model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ are given by $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$. Fitting $\boldsymbol{u}$ on $\boldsymbol{X}$ gives the residuals $\boldsymbol{e_u} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{u}$. The added variable plot is obtained by displaying $\boldsymbol{e}$ on the y-axis and $\boldsymbol{e_u}$ on the x-axis. A trend in the residuals would indicate that the variable should be added to the model. It can be shown that the slope in the scatterplot of the residuals is equal to the coefficient estimate of β_u when including u in the model, see Abraham and Ledolter[27] chapter 6.2.2 for more details. In this case, it is very clear that the two-factor interaction AF is a candidate to consider for entering the model. An AVPP could of course have been constructed using all possible candidate regressors, but the main point here is to illustrate a useful tool for knowing if all the important variables have been identified in a screening procedure.

7 | CONCLUDING REMARKS

In this paper, a new size-based approach for performing a factor-based search is proposed. The method is based on fitting the largest FP-model possible, then selecting the terms corresponding to the largest coefficients in absolute value and fitting a reduced model only including those. Then the subsets of factors in the reduced models yielding the r smallest MSE values are selected as candidate sets for being active. Using simulated models, where model coefficients were chosen

at random, the method was demonstrated to work well for the 12-run PB design and the 16-run NC designs assuming three and four active factors. The proposed method has the advantage of not relying on significance tests or a chosen criterion. However, an important parameter to decide upon and preferably vary is the number of terms chosen for the reduced model. An appropriate value for r can to some extent be chosen afterwards. Identifying four active factors turned out to be considerable harder than identifying three, but depending on the error variance, number of experimental factors and number of runs, a considerable reduction in all the possible candidate sets of factors being active was possible to obtain. Selecting the 10 models with the smallest MSE, the probability that the true set of active factors was included among these was found to be above 95% in most cases, except when using the 12-run PB design for a large number of factors and high levels of noise. Also, the problem of identifying active factors is considerable harder when the range of the coefficients values is small than when it is large.

Admittedly our method also relies on the assumption of factor sparsity and good projection properties of the design used. Being of both $P = 3$ and $P = 4_2$, the designs utilized in this paper guarantee the estimation of all main effects and interactions for any set of three factors and all main effects and two-factor interactions for any set of four. Srivastava[28] with his *search designs* also pointed out the necessity for a design to be able to discriminate among the estimated models and in the noiseless case the discrimination should be perfect. This is a strict requirement. A factor-based search already makes some restrictions on which models that can be fitted. However, any design found by choosing six columns from a 12-run PB design cannot guarantee the discrimination among models with three active factors having three main effects and three two-factor interactions, since at least 13 runs will be needed (see Cheng[29] and Morgan et al.[30]). Examples with six and seven factor NC designs where two different models with four active factors gave identical fit in the noiseless case is given in Section 5.4. This supports the arguments for reducing the number of possible active factors in several steps when designs like the ones in this paper are used. In practice, one would typically review the MSE of the candidate models after selecting the r best models. If there is a large gap in the MSE at some point, it might indicate that the correct factors can be found in a model, which is included in the subset of models corresponding to the smallest MSE values. These models can then be chosen for further analysis or review of alias patterns. As a help to check if additional factors should be considered active, we also suggested a method, the AVPP. An example with real data was included to demonstrate how to use the method and the AVPP in practice.

The proposed method can also be used for three level designs. As the number of different types of effects that may be included in the model then increases, one should consider carefully, which effects that are desirable to investigate and thereby also which designs to use. Designs like definitive screenings designs introduced by Jones and Nachtsheim[31] and orthogonal minimally aliased response surface designs proposed by Ares and Goos[32] allow the estimation of quadratic effects in addition to main effects and two-factor interactions. If a model with these terms included is estimable for all subset of factors of a given size, the method is straight forward applicable, see for instance Tyssedal and Chaudry[33] where several situations are simulated and the screening performance compared to two-level designs of similar size. For more on three level designs and projections properties, we refer to Xu et al.[34] and Alomair et al.[35]

Finally, we point out that when analysing nonregular two-level designs, it is always wise to use several methods in companion. For that purpose, we will in particular point to the graphical method proposed in Tyssedal and Niemi.[36] It can also be used to verify if our final proposed model is reasonable. It does not put any restriction on the number of active factors, though it works best on models with relatively few terms.

ACKNOWLEDGEMENTS

The authors are thankful to two anonymous referees for providing constructive comments and suggestions.

DATA AVAILABILITY STATEMENT

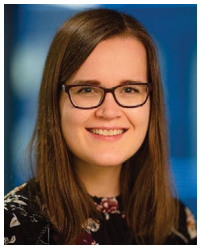
Data are simulated, except from the real data set given in the paper.

REFERENCES

1. Box GEP, Meyer RD. Finding the active factors in fractionated screening experiments. *J Qual Technol*. 1993;25(2):94-105. https://doi.org/10.1080/00224065.1993.11979432
2. Lujan-Moreno GA, Howard PR, Rojasa OG, Montgomery DC. Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study. *Expert Syst Appl*. 2018;109:195-205.

3. Plackett RL, Burman JP. The design of optimum multifactorial experiments. *Biometrika*. 1946;33(4):305-325. https://doi.org/10.1093/biomet/33.4.305
4. Box G, Tyssedal J. Projective properties of certain orthogonal arrays. *Biometrika*. 1996;83:950-955. https://doi.org/10.1093/biomet/83.4.950
5. Cheng C-S. Some projection properties of orthogonal arrays. *Ann Statist*. 1995;23(4):1223-1233. https://doi.org/10.1214/aos/1176324706
6. Lin DKJ, Draper NR. Projection properties of Plackett and Burman designs. *Technometrics*. 1992;34(4):423-428.
7. Hamada M, Wu CFJ. Analysis of designed experiments with complex aliasing. *J Qual Technol*. 1992;24(3):130-137. https://doi.org/10.1080/00224065.1992.11979383
8. Chipman H, Hamada M, Wu CFJ. A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*. 1997;39(4):372-381.
9. Ming Y, Joseph VR, Lin Y. Efficient variable selection approach for analyzing designed experiments. *Technometrics*. 2007;49(4):430-439. https://doi.org/10.1198/004017007000000173
10. Wolters MA, Bingham D. Simulated annealing model search for subset selection in screening experiments. *Technometrics*. 2011;53(3):225-237.
11. Jin DKJ, Li R. Analysis methods for supersaturated design: some comparisons. *J Data Sci*. 2003:249-260.
12. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Statist Assoc*. 2001;96(456):1348-1360. https://doi.org/10.1198/016214501753382273
13. Phoa FK, Pan Y, Xu H. Analysis of supersaturated designs via the Dantzig selector. *J Stat Plan Inference*. 2009;139(7):2362-2372. https://doi.org/10.1016/j.jspi.2008.10.023
14. Tyssedal J, Samset O. *Analysis of the 12 Run Plackett and Burman Design*. Technical Report No. 8. 1997.
15. Kulachi M, Box GEP. Catalysis of discovery and development in engineering and industry. *Qual Eng*. 2003;15(3):513-517.
16. Tyssedal J, Grinde H, Roestad C. The use of a 12.run Plackett–Burman design in the injection moulding of a technical plastic component. *Qual Reliab Eng Int*. 2006;22(6):651-657.
17. Tyssedal J, Hussain S. Factor screening in nonregular two-level designs based on projection-based variable selection. *J Appl Statist*. 2016;43(3):490-508. https://doi.org/10.1080/02664763.2015.1070805
18. Montgomery D, Jones B. Alternatives to resolution IV screening designs in 16 runs. *Int J Exp Design Process Optim*. 2010. https://doi.org/10.1504/IJEDPO.2010.034986
19. Evangelaras H, Koukouvinos C. On generalized projectivity of twolevel screening designs. *Stat Probab Lett*. 2004;68(4):429-434.
20. Wang JC, Wu CFJ. A hidden projection property of Plackett–Burman and related designs. *Stat Sin*. 1995;5:235-250.
21. Wolters MA. *Using Oversized Models to Find Active Variables in Screening Experiments*. Master's Thesis. Simon Fraser University; 2007.
22. Miller A, Sitter R. Using the folded-over 12-run plackett-burman design to consider interactions. *Technometrics*. 2001;43(1):44-55. https://doi.org/10.1198/004017001524043 18
23. Vatnedal R. *Optimizing Predictive Performance of Random Forests by Means of Design of Experiments and Resampling, with a Case-study in Credit Scoring*. Master's Thesis. NTNU; 2020.
24. Wu CFJ, Hamada MS. *Experiments: Planning, Analysis and Optimization*. 2nd ed. Wiley; 2009.
25. Phoa FKH, Wong WK, Xu H. The need of considering the interactions in the analysis of screening designs. Department of Statistics, UCLA; 2009. https://escholarship.org/uc/item/95x7k3c3. https://doi.org/10.1080/02664763.2015.1070805
26. Dopico-Garcia MS, Valentão P, Guerra L, Andrade PB, Seabra RM. Experimental design for extraction and quantification of phenolic compounds and organic acids in white "Vinho Verde" grapes. *Analytica Chimica Acta*. 2007;583(1):15-22. https://doi.org/10.1016/j.aca.2006.09.056
27. Abraham B, Ledolter J. *Introduction to Regression Modelling*. Duxbury Press; 2006.
28. Srivastava JN. Designs for searching non-neglible effects. International Symposium on Statistical Design and Linear Models, Amsterdam, New York. Elsevier Science Publishing Co., North-Holland Publishing Co.; 1975:507-520.
29. Cheng CS. Projection properties of factorial designs for factor screening. In: Dean AM, Lewis SM, eds. *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*. Springer Verlag; 2005:156-168.
30. Morgan JP, Gosh S, Dean AM. J.N. Srivastava and experimental design. *J Stat Plan Inference*. 2014;144:3-18. https://doi.org/10.1016/j.jspi.2012.09.007
31. Jones B, Nachtsheim C. A class of three-level designs for definitive screening in the presence of second-order effects. *J Qual Technol*. 2011;43:1-15.
32. Ares JN, Goos P. Enumeration and multicriteria selection of orthogonal minimally aliased response surface designs. *Technometrics*. 2020;62:21-36. https://doi.org/10.1080/00401706.2018.1549103
33. Tyssedal J, Chaudry MA. The choice of screening design. *Appl Stoch Models Bus Ind*. 2017;33:662-673.
34. Xu H, Cheng SW, Wu CFJ. Optimal projective three-level designs for factor screening and interaction detection. *Technometrics*. 2004;46(3):280-292. https://doi.org/10.1198/004017004000000310
35. Alomair MA, Georgiou SD, Aggarwal M. Projection properties of three-level screening designs. *Aust N Z J Stat*. 2020;62(4):407-425. https://doi.org/10.1111/anzs.12306
36. Tyssedal J, Niemi R. Graphical aids for the analysis of two-level nonregular designs. *J Comput Graph Statist*. 2014;23(3):678-699.

AUTHOR BIOGRAPHIES



Yngvild H. Hamre is a Data Scientist in DNB Bank ASA, currently pursuing an industrial Ph.D. in collaboration with the Norwegian University of Science and Technology (NTNU).

John Tyssedal is a Professor in statistics at the Norwegian University of Science and Technology (NTNU), Norway. His research includes design of experiments, statistical process control and time series. He is a member of the American Statistical Association and the European Network of Business and Industrial Statistics.

How to cite this article: Hamre YH, Tyssedal J. On the identification of active factors in nonregular two-level designs with a small number of runs. *Qual Reliab Eng Int*. 2022;38:4099–4121. https://doi.org/10.1002/qre.3188