Doctoral theses at NTNU, 2024:220

Yngvild Hole Hamre

Blocking and analyzing screening designs

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of Philosophiae Doctor
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of Science and Technology

Yngvild Hole Hamre

Blocking and analyzing screening designs

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2024

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-326-8032-0 (printed ver.)
ISBN 978-82-326-8031-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2024:220

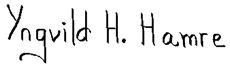
Printed by NTNU grafisk senter

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philiosophiae Doctor (PhD) at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. The work was carried out partly at DNB and partly at the Department of Mathematical Sciences during the years 2020-2024. Being an industrial PhD, it was funded by the Research Council of Norway and DNB Bank ASA, the latter in which I am employed.

I am very grateful to DNB for giving me the opportunity to pursue a PhD while remaining at the bank, and for continued support through major changes to the project. My supervisor, John Tyssedal, has also been of outmost importance during these years. I would not have chosen to pursue, or fulfill, a PhD without his patience and excellent supervision. I would also like to thank my co-supervisor at DNB, Jacob Laading, for helpful guidance in times when motivation was needed, as well as my colleagues in the bank for a much-appreciated working environment in years mostly spent on this solitary project.

Last, but not least, I would like to thank my family for support in the years leading up to, and during, this thesis. My beloved husband, Kristian, for proofreading and discussing, and for taking good care of our son Jonas, whose presence is a large source of joy and need for improved efficiency. My parents, for being loving, supporting and nurturing a love of learning. My brother, for keeping me company in Trondheim while Covid restrictions limited our social and professional lives. And my sister, whose enthusiasm for getting pupils through school luckily was extended to include me.



Yngvild Hole Hamre

Oslo, 28.02.2024

Contents

I Introduction **1**

1 Design of Experiments 3
2 Potential applications in DNB 13
3 Paper summaries 26
4 Bibliography 30

II Research papers **41**

1 Preserving projection properties when regular two-level designs are blocked **43**
John Sølve Tyssedal and Yngvild Hole Hamre.
Published in Journal of Statistical Planning and Inference, volume 221, 2022, pages 266-280.

2 Preserving projection properties when two-level screening designs are blocked **61**
Yngvild Hole Hamre and John Sølve Tyssedal.
Submitted to Metrika for review.

3 On the identification of active factors in nonregular two-level designs with a small number of runs **111**
Yngvild Hole Hamre and John Sølve Tyssedal
Published in Quality and Reliability Engineering International, Volume 38(8), 2022, pages 4099-4121.

4 A decoupling method for analyzing foldover designs **137**
Yngvild Hole Hamre and John Sølve Tyssedal
Submitted to QREI for second review.

Part I

Introduction

Introduction

The main topics of this thesis are how to analyze nonregular screening designs and how to block both regular and nonregular two-level screening designs. These are topics at the heart of the field Design of Experiments, thus an introduction to the historical background and relevant terminology will be given in Section 1. Selected examples of applications with relevance for DNB will be given in Section 2.

1 Design of Experiments

The human mind is inclined to search for explanations. We observe and analyze our surroundings continuously, and strive to rationalize, and sometimes influence, our observations. The field of statistics provides tools and frameworks for analyzing structured information. In many cases we have little or no control of the observed process and are at the mercy of shifting circumstances when the data is collected. But in some lucky cases, the circumstances can be controlled, thus one may ensure observing the desired conditions and results thereof. The field of statistics concerned with planning and conducting controlled gathering of data and its subsequent analysis is called Design of Experiments (DoE). The godfather of DoE is Ronald Fisher, whose work at the agricultural research institution Rothamstead Experimental Station in the 1920's inspired the development of methods for improving experimentation (Bodmer, 2003). Sowing and harvesting are slow processes that require large areas of land. Trying to improve the process requires structured gathering of data to utilize the resources efficiently. The approaches are however general, and his book "The Design of Experiments" (Fisher, 1935) is considered a cornerstone of DoE. The field has found applications in a wide variety of settings in which data is costly to collect, but the experimental settings can be controlled.

A key problem in DoE for industrial applications is to plan experiments for investigating an unknown response surface (Myers et al., 2016). Given a response y and a set of potentially influential factors $x_1, x_2, ..., x_k$, the experimenter may wonder which factors affect the response, how the nature of the process can be described and how the response may be optimized. Consider for instance the process of baking a cake without a predefined recipe. Then the response y may be the amount of cake eaten by the guests, or in more professional settings, the total score given by a tasting panel. Potentially influential factors may be the amount of flour, sugar, eggs and other ingredients, as well as temperature and baking time. One way to assess the impact of each factor is by varying one factor at a time (OFAT). This is often a chosen strategy as it ensures control of which factor affected the response and how much, but it has some obvious drawbacks (Czitrom, 1999). First of all, testing one factor at a time requires conducting a huge number of experiments if there are more than just a few factors. Moreover, there may be interaction effects between the factors that will not be discovered when only varying one factor at a time.

1.1 Two-level screening designs

The starting point of an experimentation with many potentially influential factors is often a screening, in which the goal is to identify the factors affecting the response, called the active factors, for further experimentation. One can then begin by conducting a two-level screening design and analyze this under the assumption that the unknown response surface can be coarsely approximated by a model consisting of main effects and interactions, which is used to identify the active factors. These can then be tested more thoroughly using more levels and thereby enabling more advanced models, for instance including quadratic terms.

1.1.1 Regular designs

A typical starting point for screening is a two-level factorial design, in which a high and a low level is chosen for each factor. The corresponding values are coded as 1 and -1 in the matrix representing the experimental design. Consider for instance the simple case of baking a four-ingredient cake consisting of sugar, eggs, flour and milk (Lee, 2022). The six factors under consideration can then be the amount of sugar, the amount of eggs, the amount of flour, the amount of milk, oven temperature and baking time. The corresponding levels may be chosen as in Table 1.1.

Table 1.1: High and low factor levels for the cake experiment.

Factor	Factor name	Unit of measurement	Low level (-1)	High level (1)
A	Sugar	g	100	200
B	Eggs		2	6
C	Flour	g	50	150
D	Milk	dl	1	2
E	Temperature	degrees Celsius	150	200
F	Baking time	minutes	40	60

As the experiment includes $k = 6$ experimental factors with two levels each, testing all possible combinations will require 2^6=64 experimental runs. Then the experiment is called a full factorial experiment and enables the estimation of all effects corresponding to the six factors. The first eight runs of the resulting experiment can be seen in Table 1.2, where the combination in the first row represents baking a cake consisting of 100 grams of sugar, 2 eggs, 50 grams of flour and 1 deciliter of milk for 40 minutes at 150 degrees Celsius.

Table 1.2: The eight first rows of the 2^6 design.

Run	A	B	C	D	E	F
1	-1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	-1	-1
3	- 1	1	-1	- 1	-1	-1
4	1	1	-1	- 1	-1	-1
5	-1	-1	1	-1	-1	-1
6	1	-1	1	-1	-1	-1
7	- 1	1	1	- 1	-1	-1
8	1	1	1	- 1	-1	-1

The resulting model is given as $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\mathbf{X}$ is a design matrix with $n = 64$ rows and $p = 64$ columns if an intercept and all possible effects are to be included in the model. The number of effects for such a model is given by $1 + \binom{k}{1} + \binom{k}{2} + ... + \binom{k}{k-1} + \binom{k}{k}$. Interaction effects are found by including the Hadamard product of the corresponding factor columns. The regression coefficients $\hat{\boldsymbol{\beta}}$ can then be calculated as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$. This is the usual way to find coefficients for linear models. In DoE, it is common to use the term "effect" to denote the

change in response when a factor is changed from the low level to the high level. The effect of A is therefore twice the regression coefficient of A. For balanced designs, the effect of A can also be calculated as the difference between the mean response for the runs where A is at the high level and the mean response for the runs where A is at the low level.

It is common to assume that higher-order interactions are unlikely to be active. In this case, one would for instance expect the two-factor-interaction EF (temperature and baking time) to affect the results, whereas the six-factor interaction between all the included factors is much harder to interpret. If one is willing to ignore the higher-order interactions, one can reduce the number of runs needed for the experiment by allocating factors to higher-order interactions columns. For instance assigning the factor E to the column given by ABC and F to the column BCD. Then the interaction effects ABC and BCD are impossible to separate from the main effects E and F respectively, but if higher-order effects are negligible, that is not a problem. Using two higher-order interactions to define factor columns, the number of runs is reduced to 2^{6-2}, and the resulting design is a fraction of the original full factorial design. It is therefore called a fractional factorial design. The eight first rows of the design can be found in Table 1.3. Note how the signs in columns E and F have changed compared to the 2^6 design in Table 1.2.

Table 1.3: The eight first rows of the 2^{6-2} design.

Run	A	B	C	D	E=ABC	F=BCD
1	-1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	1	-1
3	- 1	1	-1	- 1	1	1
4	1	1	-1	- 1	-1	1
5	-1	-1	1	-1	1	1
6	1	-1	1	-1	-1	1
7	- 1	1	1	- 1	-1	-1
8	1	1	1	- 1	1	-1

In general, a 2^{k-p} design consists of a $\frac{1}{2^p}$ fraction of the 2^k design. The interactions assigned to factor columns, in this case E=ABC and F=BCD, are called generators for the design. The complete defining relation consists of all terms that equal the identity column. In this case, I=ABCE=BCDF=I^2=ADEF, so I=ABCE=BCDF=ADEF is the complete defining relation. The terms in the defining relation are referred

to as words. The aliasing structure for an effect can then be found by multiplying it with all terms in the defining relation. The resolution R of a design equals the shortest word in the defining relation (Box et al., 2005). If the shortest word is ABCE, the resolution is IV. Then two-factor interactions are aliased with each other, and main effects with three-factor interactions. The higher the resolution, the more lower-order effects can be estimated, which is a desirable property.

A related property which is very useful when assessing the capabilities of a screening design is projectivity, defined by Box and Tyssedal (1996) as "*A 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 for instance a design is of projectivity $P = 3$, all main effects and interactions can be estimated for any combination of three active factors. In some cases, not requiring the higher-order interactions to be estimable can increase the number of active factors for which the lower-order effects can be estimated. Evangelaras and Koukouvinos (2004) therefore defined generalized projectivity as "*A 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 from the design all factorial effects including up to* α*-factor interactions are estimable*". Then using a $P = 3_2$ design, all main effects and two-factor-interactions can be estimated for any combination of three active factors. For regular designs, the projectivity is given by $R - 1$.

In both full factorial and fractional factorial designs, all design columns are orthogonal to each other, so the effects can either be estimated independently of each other or are completely aliased. Designs with this property are called regular designs. Being able to independently estimate the effects eases the analysis. The drawback of regular designs is that the run sizes are not very flexible, and the aliasing between effects can yield poor projectivity properties.

1.1.2 Nonregular designs

To overcome the challenges faced by regular designs, one may consider using nonregular designs, in which there might be partial aliasing between effects. Allowing aliasing between effects enables more flexible run sizes and better projectivity properties, but the designs become harder to analyze, as all methods used for regular designs are not valid for the nonregular ones. There are several classes of nonregular designs, and

the ones used in the papers will be introduced here.

Plackett-Burman designs

Plackett-Burman (PB) designs were introduced right after World War II by Plackett and Burman (1946). They are efficient screening designs which can be used to study up to $n-1$ factors in n runs, n being a multiple of 4 and less than or equal to 100. When n is a power of 2, they coincide with fractional factorial designs, but for run sizes such as 12, 20 and 24, they are interesting alternatives. The main effects in the designs are orthogonal to each other, but there may be complex aliasing between main effects and two-factor interactions, and between other lower-order effects. This can be a challenge when analyzing the designs. One option is then to conduct the foldover runs as well, as will be explained in Section 1.1.3.

Many Plackett-Burman designs can easily be constructed by rotating a row vector one step for each run. Let + and - denote the levels 1 and -1. The row vector for the 12-run design can then be given by [+,+,-,+,+,+,-,-,-,+,-]. This can be rotated by moving all elements one step to the right and placing the rightmost element at the beginning of the new vector. To construct the design, this must be done 10 times, creating 11 different row vectors. The final row in the design is a vector consisting of only minus entries. The corresponding vectors for the 20-run and 24-run designs are [+,+,-,-,+,+,+,+,-,+,-,+,-,-,-,-,+,+,-] and [+,+,+,+,+,+,-,+,-,+,+,-,-,+,+,-,-,+,-,+,-,-,-,-], respectively.

No-confounding designs

The no-confounding (NC) designs is another class of efficient two-level screening designs, which have been found for 16, 20 and 24 runs. They are orthogonal designs with no complete aliasing between main effects and two-factor interactions. For all but the 9-14 factor 16-run designs, they have the desirable property that no two-factor interactions are completely aliased with each other. Along with flexible sizes, this makes them attractive alternatives to the regular designs with the same number of runs. The drawback of the NC designs is some partial aliasing between main effects and two-factor interactions. The class of designs started with the NC16 designs with 6-8 runs introduced in Jones and Montgomery (2010), which were found by choosing the subsets of columns from the 16-run orthogonal arrays presented in Hall (1961) that minimize the $E(s^2)$ and trace($\mathbf{AA^T}$) criteria. This approach minimizes

the sum of squared off-diagonal elements of $\mathbf{X^T X}$ as well as the total bias in the design, see Myers et al. (2016) for details. Subsequently, the class of designs was expanded with NC16 designs with 9-14 factors in Jones et al. (2015), 20-run NC designs with 6 to 12 factors in Stone et al. (2017b), and 24-run NC designs with 7 to 12 factors in Stone et al. (2017a).

Definitive screening designs

Another class of designs for which the main effects may be orthogonal is the definitive screening designs (DSDs), which were introduced by Jones and Nachtsheim (2011) and are thoroughly described in Myers et al. (2016). As they consist of three-level factor columns with the values -1, 0 and 1, they have the desirable property of being able to screen for quadratic effects as well as main effects and two-factor interactions. DSDs exist for all numbers of factors $k \geq 4$, with a size of $2k + n_c$ runs, where n_c is the number of center runs. The $2k$ runs that are not center runs form a foldover. Thus the main effects are not aliased with the two-factor interactions or the quadratic effects, but for odd k there is a small amount of partial aliasing between the main effects themselves. This is not the case for even k, so using a design with the latter property is recommended. A possibility for getting orthogonality for odd k at the expense of two extra runs is to select the design which has one more factor and drop the excess factor. Regardless of the orthogonality properties of the main effects, the two-factor interactions are partially aliased with each other. The quadratic effects are also partially aliased with each other, and with the two-factor interactions.

OMARS designs

A recently introduced broad class of screening designs that also include the orthogonal DSDs are the orthogonal minimally aliased response surface (OMARS) designs introduced by Ares and Goos (2020). They are three-level orthogonal designs with levels -1, 0 and 1, and can therefore be used to screen for quadratic effects in addition to the usual main effects and interactions. The main effects in OMARS designs are required to be orthogonal to other main effects and to all second-order effects. The designs exist for a wide range of sizes, making them very flexible. Most, but not all, OMARS designs have the foldover property described in Section 1.1.3, with center runs added.

1.1.3 Foldover property and mirror image pairs

Many efficient screening designs, for instance the PB designs, have partial aliasing between effects that can complicate their analysis. If there is ambiguity in the identification of active lower-order effects, one alternative is to conduct foldover runs and do a follow-up analysis. Let $\mathbf{X}$ denote the original design matrix with n runs and $k+1$ columns. The foldover design is created by reversing the signs of all entries and adding the resulting runs to the original design. Thus for each run in the original design, a run with the opposite signs is included, and together, these make up what is called a mirror image pair. In addition to this, a new intercept column must be included to enable estimation of the constant term. The resulting design matrix $\mathbf{X_f}$ is then given by $\mathbf{X_f} = \begin{bmatrix} 1 & \mathbf{X} \\ 1 & -\mathbf{X} \end{bmatrix}$. This matrix has $2n$ rows and $k+2$ columns, thus folding over enables estimation of one more factor compared to the original design. For regular designs of odd resolution R, their foldovers will have resolution $R+1$ (Box and Wilson, 1951). A desirable property of all foldover designs is that the main effects and two-factor interactions are orthogonal to each other, making the effects easier to separate. More specifically, the odd and even effects can be divided in two orthogonal subspaces, a property that is utilized for proposing a new analysis method for foldover designs in Paper 4. Another useful property of foldover designs is that the runs form mirror image pairs. In Paper 1 and 2, this property is used to find candidate blocks for blocking both regular and nonregular designs.

1.1.4 Blocking

When all runs of an experimental design cannot be performed under homogeneous conditions, one should consider including block effects in the design to account for the varying conditions. Recall the baking example in Section 1.1.1. If there are two ovens in the kitchen, the chef might want to use both to complete the experimentation faster. But the ovens are not guaranteed to behave identically. To ensure that this does not disturb the results, a block effect representing the ovens may be included in the design. Note that it is common to assume that there are no interactions between the block effect(s) and treatment effects of the design. If that is the case, it may severely complicate the analysis (Myers et al., 2016). Thus that assumption should be thoroughly discussed before conducting the experiment.

The recommended way of blocking regular designs has traditionally been to assign the block defining contrast(s) to higher-order interaction(s). This ensures orthogonality between the block effect(s) and the treatment effects, but may severely worsen the projectivity properties of the design. A combined word length pattern may also be used to rank blockings of regular two-level designs. This has been investigated for instance by Sitter et al. (1997), Chen and Cheng (1999), Zhang and Park (2000) and Zhao et al. (2013). It is common to require that the blocks effect(s) are orthogonal to the main effects. Limiting the confounding between the blocks effect(s) and the two-factor interactions is therefore a natural focus. Two examples are Cheng and Mukerjee (2001) and Sun et al. (1997). The first proposed a criterion based on minimizing the number of two-factor interactions that are confounded with block effect(s) or aliased with main effects, while uniformly distributing these interactions over the alias sets. The latter focused on the number of clear main effects and two-factor interactions (that is, not confounded with other main effects, two-factor interactions or block effects) as well as word length patterns when suggesting blocks.

For nonregular designs, the focus has often been to limit the amount of partial aliasing between block effects and lower-order effects by minimizing generalized minimum aberration criteria. The much-cited work by Cheng et al. (2004) proposed a definition of word-length patterns for nonregular designs, as well as two different minimum aberration criteria for blocking nonregular designs. This resulted in four different optimality criteria which were used to find blocking schemes for 12-, 16- and 20-run orthogonal designs. Park et al. (2007) built upon that work to suggest arrangements in three blocks. Ou et al. (2011) investigated simultaneous employment of blocking and foldover, while Schoen et al. (2013) considered five different criteria based on generalized word length patterns to find blocks for orthogonal resolution III designs. Blocking of three-level nonregular designs has also been researched, for instance by Ares and Goos (2023), who used linear programming to find block effect(s) for the OMARS designs that are orthogonal to the main effects and limit the confounding with second-order effects.

Limiting the confounding between block effect(s) and second-order effects is a well-justified approach for finding candidate blocks for both regular and nonregular designs. It might however result in candidate blocks that negatively impact the projectivity properties of the design. In particular for blocked screening designs, we believe that being able

to estimate as many effects as possible for a subset of factors up to a certain size is a desirable property. In Paper 1 and 2, we therefore introduce new candidate blocks that preserve the projectivity of several popular regular and nonregular screening designs as well as possible, and compare the results with existing proposed ways of blocking. To achieve the best projectivity possible, we allowed partial confounding between block effects and interaction effects, and in some cases even between block effects and main effects. This is often not accepted in previous work on blocking, and the resulting impact on the standard deviations of the effects is therefore included. To ensure that the effects are well-estimated within the current projectivity, the block yielding the highest minimum D_s-efficiency when considering all combinations of active factors was chosen. D_s-efficiency is defined as $D_{s,eff} = \frac{[\frac{\mathbf{X}^t\mathbf{X}}{\mathbf{X}_b^t\mathbf{X}_b}]^{\frac{1}{s}}}{n}$, where $\mathbf{X}$ is the entire blocked design (including interactions), $\mathbf{X}_b$ are the block columns and s is the number of effects of interest. See the papers for further details on the criterion.

1.1.5 Variable selection

When performing a screening experiment with a large number of experimental factors, identifying the ones actually affecting the response is important to limit the scope for further experimentation. For regular designs, there are several methods based on utilizing the orthogonality of the effects, such as the half-normal plot (Daniel, 1976) and Lenth's method (Lenth, 1989). These are not applicable for nonregular designs, for which other methods must be chosen. To ease the analysis by reducing the number of potential effects, one may start by considering if heredity assumptions should be imposed. Assuming weak heredity, an interaction will only be included in the model if at least one of the corresponding main effects is present. A stricter assumption is strong heredity, which requires all main effects involved with the interaction to be present. Assuming heredity may be reasonable in some cases, but if falsely assumed it can severely impact the result, as is demonstrated in Paper 4.

There are two different main approaches for analyzing nonregular screening designs: Effect-based and factor-based searches. Effect-based searches aim at identifying the correct active effects, that is, to determine the linear model that best approximates the response. To limit the search space for the effects, heredity assumptions are often used for these methods, for instance the stepwise regression procedure (Hamada and Wu, 1992),

modified least angle regression (Yuan et al., 2007; Ming et al., 2007) and simulating annealing model search (Wolters and Bingham, 2011). There are also some effect-based methods for which heredity is not assumed, examples being the non-convex penalized least squares first proposed by Fan and Li (2001), and the Dantzig selector (Candes and Tao, 2007). If the design has a foldover structure, this may be utilized when attempting to identify the active effects, as the odd and even effects can then be arranged in two orthogonal subspaces. This was first proposed for PB designs in Miller and Sitter (2001) and for non-orthogonal designs in Miller and Sitter (2005). These methods inspired Jones and Nachtsheim (2017) to suggest a similar approach for analyzing DSDs, while Hameed et al. (2023) proposed a version for OMARS designs that only requires orthogonality between main effects and second-order effects. Utilizing the foldover structure for an effect-based search based on creating two new responses is the topic of Paper 4.

Factor-based searches instead aim at identifying the active factors, which should then be further investigated to find the best approximate model. Some methods within that category are the Bayesian analysis procedure proposed by Box and Meyer (1993) and the projection-based factor searches proposed by Tyssedal and Samset (1997), Kulachi and Box (2003) and Tyssedal and Hussain (2016). The latter combined a projection-based factor search with forward selection, an idea which is further developed in Paper 3, using restrictions on the size of the model instead of stopping criteria.

2 Potential applications in DNB

This thesis was partly funded by the largest bank in Norway, DNB, and the observant reader may ask herself what the applications of screening designs are within DNB. The author, Yngvild Hole Hamre, is employed in DNB as a Data Scientist in Personal Banking (PB), working in a department mainly focused on developing prediction models for personalizing the communication towards customers. A cornerstone in this regard is a framework called $(PM)^2$ for rapidly developing prediction models by reusing relevant code. Originally, the scope of this project was using screening for identification of important hyperparameters in machine learning models and applying response surface methodology for finding good settings for those. In the research phase of that work, we discovered a new class of marketing models called uplift models, aimed at identifying the customers that are most likely to be impacted by a

marketing activity (for instance a phone call or an email). To enhance the benefit of the project, we therefore decided to test hyperparameter tuning for an uplift model, such that a new modelling framework could be included in $(PM)^2$. To be able to fit an uplift model, data was collected from several campaigns. Unfortunately, the resulting models were poor and neither suited for practical use nor inclusion in research papers. Lots of lessons were learned in terms of planning collection of structured data, and a summary of these were presented at the Math Meets Industry conference in Trondheim in 2022.

While working on the uplift part of the project, departmental changes along with the implementation of automated Bayesian optimization in the $(PM)^2$ framework led to a decreased need for alternative methods for optimizing hyperparameters. It was therefore decided to focus on theoretical parts of DoE for the remainder of the project period, which is reflected in the included research papers. The wise words of R. Fisher stating that *"the best time to plan an experiment is after you've done it"* (Box et al., 2005) certainly applies to planning a PhD-project as well. We do however believe that DoE may find applications in many fields relevant for DNB and will therefore give a brief overview of potential use cases in this chapter. First, an introduction to XGBoost and different hyperparameter optimization approaches is given in Section 2.1. Then online controlled experiments, online active learning and conjoint analysis are discussed in Section 2.2, 2.3 and 2.4, respectively.

2.1 XGBoost and tuning hyperparameters

XGBoost is a highly optimized machine learning algorithm introduced by Chen and Guestrin (2016). It has become very popular due to the ability to yield good results for large tabular data sets in a reasonable time span, as it is optimized for performance. The ability to automatically handle missing values is another valuable property of the algorithm. For these reasons, it is the algorithm used in $(PM)^2$.

XGBoost is a parallelized boosting framework most commonly using regression trees as base learners. Regression trees being base learners means that the model is an additive combination of regression trees, i.e. $\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i)$. In iteration t, $f_t(\mathbf{x})$ is decided by minimizing a loss function $L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$, where n is the number of observations, l is a differentiable convex loss function, f_t is a regression tree, and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda||\mathbf{w}||^2$ is a regularization term penalizing complexity. T is the number of leaves in the trees, $\mathbf{w}$ is a vector of

leaf weights, and γ and λ are regularization parameters. New trees are added greedily, using gradient descent to minimize the loss function.

An important aspect of the algorithm is to choose hyperparameter settings, both with regards to the loss function and to the tree fitting procedure. The choice of hyperparameters has a large impact on the structure of the model, and thereby on the results. This is the case also for other machine learning methods, for instance neural nets, in which the hyperparameters can be the number of hidden layers in the network, the number of neurons in each layer, and so on. Unfortunately, as the shape of the search space is unknown and very likely not convex, no optimization procedure is guaranteed to find an optimum value. The question then is how to find an acceptable value within a reasonable period of time.

Historically, the most common tuning strategies have been grid search and random search, where the latter has been improved in a procedure called Hyperband. The conceptually more complex Bayesian optimization procedure and the related Tree Parzen Estimator are also becoming increasingly popular, as they use previously gathered data to select the next settings to evaluate. One strategy which has not received much attention is Response Surface Methodology, a well-known strategy within the statistics community for exploring the unknown relationship between possible explanatory variables and a response they might affect. Investigating the potential for this strategy is therefore an interesting possibility for utilizing DoE within machine learning. This will be elaborated after a brief introduction to the different tuning strategies.

2.1.1 Grid search, random search and Hyperband

Grid search, the most intuitive tuning strategy, is a simple approach where the user specifies a range of interesting values for each hyperparameter and proceeds by testing all combinations of these values. This approach ensures that the area of interest is evenly covered, but does not offer the opportunity to investigate the most promising parts of the search space more closely than others. In addition, using a grid limits the number of values which is tested for each hyperparameter, despite requiring many runs to test all settings. Random search is an alternative method which can be used to limit the search and enable evaluation of several distinct values for each hyperparameter. One may draw the hyperparameters under consideration from different probability distributions, or simply use a uniform distribution with the same range as

would have been applied for the grid search. In Bergstra and Bengio (2012), random search was shown to outperform grid search in most of the cases tested.

The Hyperband-algorithm, introduced in Li et al. (2016), focuses on speeding up random search by using early stopping when the settings do not seem promising. It uses successive halving as a subroutine, in which a budget (time, iterations, etc.) is allocated to the settings one wishes to evaluate, and after spending some of the budget, the worst half of the settings are eliminated. This process is repeated until there is only one setting left. The Hyperband-strategy can also be applied to other algorithms than random search. Falkner et al. (2018) used the Tree Parzen Estimator (TPE), a special Bayesian optimization, in the algorithm instead of random search. This was shown to outperform both standard Bayesian optimization and Hyperband using random search in a wide range of examples. The procedure is available in the HpBandSter package in Python (Falkner, 2018).

2.1.2 Bayesian optimization

Bayesian optimization is a method designed to select the hyperparameter setting to evaluate based on the previously evaluated settings, rather than drawing settings randomly or exploring a grid. There have been several studies showing that this strategy outperforms random search, for instance Bergstra et al. (2011) and Turner et al. (2021). Thus Bayesian optimization has grown increasingly popular, and is available through several implementations in Python, for instance the BayesOpt package introduced by Martinez-Cantin (2014). A comprehensive overview of recent advantages and open problems in the field can be found in the newly published work by Wang et al. (2023).

As described in the tutorial in Frazier (2018), Bayesian optimization is a good choice when one wants to optimize a continuous, derivative-free black-box function with respect to input with 20 or less dimensions, and evaluation of the function is costly. This typically applies to tuning of hyperparameters when using algorithms as XGBoost to model large data sets. The idea of Bayesian optimization is to build a surrogate for the objective function $f(\mathbf{x})$. It is often performed assuming that the response one wants to optimize was sampled from a Gaussian process. A posterior distribution for this function is made and updated using all previously tested settings, and the next setting to evaluate is found for instance as the one that yields the highest value for an acquisition

function (often expected improvement) according to the posterior. The advantage of this approach is therefore that it uses all available information, not just local approximations, to suggest new promising points to evaluate.

Formally, as explained in Brochu et al. (2010), the problem can be formulated as finding a global maximum of $f = f(\mathbf{x})$. Here, noiseless observations are assumed for simplicity. Let the previously observed data be denoted $D_{1:t} = (\mathbf{x}_1, y_1)...(\mathbf{x}_t, y_t) = \{\mathbf{x}_{1:t}, \mathbf{f}_{1:t}\}$, where y_i are values from the objective function. Bayesian optimization assumes that $y_1, ..., y_t$ are drawn from a multivariate distribution, for instance assuming that f is normally distributed with mean $\mu(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. According to Bayes theorem, $p(y|x) \propto p(x|y)p(y)$, where $p(y|x)$ is the posterior distribution of y given x, and $p(y)$ is the prior distribution of y. The posterior at $\mathbf{x}$ is therefore given by $p(f|D_{1:t}) \propto p(D_{1:t}|f)p(f)$. Using a Gaussian process prior with the squared exponential function as the covariance function, the predictive distribution of $f_{t+1}(\mathbf{x})$ is given by $p(f_{t+1}(\mathbf{x})|D_{1:t}) = N(\mu_t(\mathbf{x}), \sigma_t^2(\mathbf{x}))$. Other priors may be used, but the Gaussian process prior is popular as it yields a known Gaussian process distribution to the posterior function.

Let the best observation so far be given by $f(\mathbf{x}^+)$. The next point to evaluate is then found by maximizing an acquisition function, for instance the expected improvement compared to $f(\mathbf{x}^+)$ using the predictive distribution $p(f_{t+1}(\mathbf{x})|D_{1:t})$. With $\mathbf{x}_{t+1}$ being the next point, y_{t+1} can be found from the objective function, the predictive distribution is updated, and the procedure can continue. It is also possible to use other acquisition functions for finding the next setting to evaluate, for instance taking into account that one wants to balance maximizing the expected improvement with exploring areas with high variance.

The Tree Parzen Estimator (TPE) introduced in Bergstra et al. (2011) is another Bayesian optimization algorithm that utilizes a surrogate function to suggest new settings based on expected improvement. Instead of utilizing a Gaussian process prior to model the posterior $p(y|\mathbf{x})$, TPE aims to model $p(y|\mathbf{x})$ via $p(\mathbf{x}|y)$ and $p(y)$, as $p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$. The algorithm got the "Tree"-part of the name from the fact that the algorithm can handle a tree-structured search space. A popular package in which this procedure is available is hyperopt, based on Bergstra et al. (2013). A detailed description of the procedure and recommended settings can be found in Watanabe (2023).

2.1.3 Response surface methodology

As described in Myers et al. (2016),

> Response surface methodology (RSM) is a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes.

Response surface methodology was originally developed to find optimal settings in industrial applications such as chemical investigations and in agriculture, but has since found wide-spread use in a variety of areas in which collection of data is expensive in terms of resources such as time, money, materials or processing power. The goal is to optimize a response surface of unknown shape, which is potentially influenced by several input variables.

The optimizing procedure is often performed in a sequential manner. First, the variables that actually influence the response are identified through conducting a screening experiment and fitting a simple model. The goal is to limit the search space before conducting more experiments. In the second phase, the aim is to decide whether the investigated area is close to the optimum. Then one will typically conduct one or more additional experiments, developed to fit first-order models, and use steepest ascent/descent to move in the direction of possibly better results. When the process is close to the optimum, the final phase starts. Then one may want to fit a second-order model to determine the settings yielding the optimum value, thus an experiment suited for fitting more complex models must be conducted. An advantage of this approach compared to other optimization strategies is that it yields possibly valuable information about the relationships between the important hyperparameters and the response, and how the hyperparameters interact with each other.

This approach has similarities with Bayesian optimization, as both methods are sequential and utilize previously gathered data to fit surrogate functions and suggest new settings to evaluate. While Bayesian optimization uses all the previously gathered points and aims to find a global optimum, RSM focuses on finding a local optimum, possibly only using data gathered close to that area. There has been some previous use of RSM for tuning, one of the early works being Staelin (2003), that used a DOE-inspired method to tweak the boundaries and the resolution of a grid search when tuning a support vector machine. This methodology was successfully applied for a least squares SVM classifier used for credit evaluation by Yu et al. (2011). More recently, Lujan-Moreno

et al. (2018) successfully tested using RSM for tuning the hyperparameters of random forest, and Pannakkong et al. (2022) compared results when using RSM and grid search to tune an artificial neural network, a support vector machine and a deep belief network. The use of RSM was shown to find equally good hyperparameter settings as grid search, reducing the number of runs needed by 97.79%, 97.81%, and 80.69%, respectively. Comparison with more efficient, state-of the art algorithms is therefore an interesting path of further research.

RSM was developed in a time when computational resources were scarce, thus the strategy relies heavily on a human experimenter choosing the experimental plan and deciding the next steps in each phase. This requires more expert knowledge than using an automated procedure. A possible starting point could be to first limit the number of hyperparameters using a screening experiment, thereafter perform Bayesian optimization on the chosen hyperparameters. Choosing only a few hyperparameters for Bayesian optimization is likely to reduce the computation time substantially. Combining DoE, RSM and Bayesian optimization was tested in Schau-Hansen (2022), where RSM and Bayesian optimization was first applied separately, but also combined in two different ways. First using a fractional factorial design as the starting point for selecting the hyperparameters to tune, then using RSM to find the area in which a central composite design was placed and used as an initial grid for Bayesian optimization, with the aim of possibly improving the exploitation. The latter seemed to yield the most stable optimization. An interesting paper in this regard is also Sunder et al. (2022), in which another hybrid strategy between RSM and Bayesian optimization is proposed. They use DSDs to sample the search space and perform a test to consider the complexity of the black box-function. RSM is then used for optimization if a second-order approximation seems appropriate, Bayesian optimization otherwise.

2.2 Online controlled experiments

Choosing the best layout for the message you want to convey is important in any kind of visual communication. Online controlled experiments, in its simplest form called A/B-testing, is a popular approach to achieve this in areas such as webpage design and marketing. An A/B-test can be considered a simple controlled experiment, assessing whether alternative A or alternative B should be preferred, answered by allocating users randomly to one of the options and considering some summary statistic after the experiment has ended. In DNB, A/B-tests are fre-

quently conducted to improve the webpages. Then first a hypothesis is made about what might affect the response, which is usually defined as clicking an element. The hypothesis may for instance be that addressing the customer directly ("book your appointment") leads to more clicks than a passive formulation ("book appointment"). The alternative text is implemented as version B, which is then compared with the original, version A. The duration of each test depends on the number of visitors on the webpage, as the tests are usually run until 95% confidence about the result is achieved or the estimated sample size is reached. To account for different user behavior on different days, the tests are most often run for at least a week.

Experiments with more than one factor are often referred to as multivariate testing. A famous example of such is the webpage experiment conducted by the Obama campaign team, in which the use of a factorial experiment to improve the webpage led to an estimated increase in donations of about 60 million dollars (Siroker, 2010). Additional examples where experimental design has boosted sales can be found in Almquist and Wyner (2001). They point out that conducting a series of A/B-tests is not very effective and encourage marketers to utilize design of experiments to assess how the components of a marketing campaign influence consumer behavior.

There are some concerns that differentiate online controlled experiments from traditional experiments, as pointed out by Haizler and Steinberg (2020): They are often real-time experiments with the goal of selling a product or creating interest. Thus there is a cost related to testing the sub-optimal alternative, often referred to as "regret", which will be defined in Section 2.2.2. Sample sizes tend to be large, and the response is often a summary statistic such as click-through-rate (CTR) or conversion rate (CR). These rates are in many cases low, but a slight increase might represent a substantial increase in earnings. Moreover, there might be time trends, originating for instance from active users logging in more often, or users living in different time zones. Examples of successful controlled experiments and advice on how to conduct them are presented in Kohavi et al. (2008), whose authors have extensive practical experience from experimentation at Amazon, Microsoft, Dupont and NASA. Of special relevance for DNB is also the fractional factorial marketing experiment described in Krutsick (2012), where the aim of the online retailer was to construct the optimal email for reactivating customers who had not purchased from the store the past 12 months.

A challenge related to online controlled experiments is that they require the establishment of a technical infrastructure as well as an experimental culture within the organization. An overview of potential benefits, challenges and best practices when doing so is given in Bojinov and Gupta (2022). Furthermore, an introduction to the most prominent issues in the field can be found in Gupta et al. (2019), which summarizes the discussion from the first Practical Online Controlled Experiments Summit (held in 2018). There, practitioners from 13 companies with extensive use of large-scale online experiments, such as Netflix, LinkedIn, Google and Microsoft, were gathered to share experiences. Some of the challenges they point out are estimating long-term effects of interventions without running the experiment for an extensive period, and choosing the overall evaluation criterion of a test and assessing its drawbacks. Other discussed issues were evaluating interactions between experiments as well as network effects, establishing an engineering culture that aids trustworthy experimentation and ensuring good data quality.

2.2.1 Multi-platform testing

In marketing, one often communicates with the customers across different platforms. DNB does for instance use SMS, email, the mobile app and the home page for digital marketing. When creating a new message to communicate, it may therefore be useful to adjust it to different platforms. Either because the individual customers should get the message in their preferred communication channel, or because it should be distributed in several channels simultaneously to increase visibility. Having a similar layout across the platforms may make the message easier to recognize and remember, but practical considerations such as screen size and format of the web banner can put constraints on the layout. The impact of different design elements may also differ based on the channel. The customer may for instance be more willing to read a longer text after opening an email with a given topic stated in the header than when visiting the mobile app primarily to pay a bill. Sadeghi et al. (2019) has taken such considerations into account and propose using a new class of experiments called "sliced factorial designs", which is suited for multivariate, multi-platform experiments. One possible approach in this setting could be to assign the different platforms to different blocks, as described in Section 1.1.4. Traditional blocking methods do however rely on the assumption that there are no interaction effects between the block effects and the treatment effects. This assumption is likely to be violated in a multi-platform setting, as in the example with email vs.

mobile app. The sliced designs therefore take a different approach, aiming to facilitate the estimation of treatment-platform interactions. The concepts of resolution and minimum aberration are extended to sliced designs and used to create algorithms to generate sliced designs with good properties. The approach is demonstrated through an experimental study testing the layout of an email for two platforms (laptop and phone/tablet), showing that analyzing the data while ignoring the slice factor leads to not identifying any active effects, while when including the slice factor, it is discovered that the optimal layouts differ for the two platforms.

2.2.2 Multi-armed bandits

Traditional online controlled experiments consist of an initial exploration stage gathering data from the different alternatives, followed by an exploitation stage where the best alternative is chosen. An alternative way of exploring the difference between several opportunities when the observations are conducted in a sequential manner is to consider the experimentation as a multi-armed bandit (MAB) problem, where a trade-off between exploration and exploitation can be continuously made. Then each option is considered one arm of the bandit, and for each observation, which arm to pull is a decision based on the information gathered from the previous observations. See for instance Burtini et al. (2015), a survey paper in which the multi-armed bandit is motivated and described, and a wide variety of algorithms compared. Strategies allocating observations to different arms often focus on minimizing the regret, that is, the reduction in expected total reward induced by experimenting instead of always choosing the best option (as it is not known). There are several ways in which regret can be defined. In Haizler and Steinberg (2020), it is given as $\text{Regret}_T = \sum_{t=1}^{T}(\mu - E(y_t))$, where T is the total number of observations, μ is the optimal reward and $E(y_t)$ is the expected outcome for observation t (which typically changes over time, as more information is gathered).

The drawback of minimizing regret is that it reduces the statistical power to estimate the treatment effects corresponding to each arm, as discussed in Simchi-Levi and Wang (2022), which give an overview of literature on the topic and suggest a framework to create Pareto-optimal multi-armed bandit experiments for a given level of the trade-off between minimizing regret and maintaining statistical power. The trade-off was also investigated by Haizler and Steinberg (2020), who suggest combining blocked fractional factorial designs and Thompson sampling (a much-used sam-

pling technique for MAB) to secure sufficient exploitation of all factors while achieving a lower regret than when solely using a fractional factorial design or a one-factor-at-a-time strategy. Their proposed method also yields a lower negative bias in the estimated posterior probabilities than when only using the Thompson sampling. In their simulations, they focus on identifying main effects, and use the block generators for regular designs recommended by Wu and Hamada (2009). An interesting approach for further work on the topic could be to include interactions in the set of potentially active effects, and test the blocks recommended in Paper 1 and 2.

2.3 Online active learning

One field of research which has become more important with the increasing amounts of data available is online active learning, in which the goal is to choose the most informative data point to label from a stream of unlabeled data. This can be useful if the data is costly to label, for instance if the labeling must be performed by manual inspection. The labeled data points can then for instance be used to model the underlying process. A thorough survey of online active learning can be found in Cacciarelli and Kulahci (2023). In Cacciarelli et al. (2022), the D-optimality criterion commonly used to assess design matrices in DoE is utilized in a new approach for performing online active learning with linear regression models. A version of the proposed algorithm which is more robust to outliers is introduced in Cacciarelli et al. (2023).

In DNB, there are many streams of unlabeled data, for instance data from the customers' interactions with the chatbot and transaction data. A setting where the stream of transactions is assessed is fraud detection. Online active learning in a fraud detection setting has been discussed by Carcillo et al. (2017) and Carcillo et al. (2018). An important property of fraud detection is that labeling of the data cannot be done randomly, but coincides with assessment of potentially fraudulent transactions. Assessing many transactions that are not likely to be fraudulent would be very costly in terms of lost true positive cases. The accuracy of a fraud detection system can therefore be measured as the precision over the top k alerted transactions or credit cards, where k is the budget that can be assessed during the time period under consideration. In Carcillo et al. (2017), they present several online active learning techniques for fraud detection and assess them using a real data set with 12 million transactions from Wordline, a company specialized in transaction services. They showed that the baseline method "Highest Risk Querying"

(choosing the transactions with the highest estimated probability of being fraudulent) could be improved by up to five percent by combining it with Stochastic Semi-supervised learning, in which some of the transactions are labeled as non-fraudulent (without manual inspection, either using the model or drawn randomly), and the labeled data points (both positives and negatives) are used to retrain the model. In Carcillo et al. (2018), similar topics are investigated, along with visualizations of the impact of active learning on the distribution of the training set.

2.4 Conjoint analysis

When developing new products or services, an important question is how to assess the customers' preferences in order to prioritize the right properties. The preferences may be investigated using conjoint analysis (note that some researchers strictly distinguish discrete choice experiments from conjoint analysis (Louviere et al., 2010). For simplicity, a distinction will not be made here). The general idea is to present several carefully chosen concept profiles with different levels of the attributes under consideration, and make the respondents either rank, rate or choose between the profiles. A review of the field can be found in Agarwal et al. (2015), in which they focus on three popular approaches. The first is choice-based conjoint analysis, in which the participants repeatedly choose between several concept profiles. This used to be analyzed using a multinomial logit model, but in recent years, the Hierarchical Bayes methodology has become increasingly popular, as it allows estimation of a utility function for each respondent, enabling simulation of different outcomes. The second approach is menu-based choice, in which the subject chooses several individually priced features from a list of options. The third approach is best/worst conjoint analysis, in which the subjects are asked to choose the best and worst attribute level for each concept profile.

The experimental design, in terms of number of concept profiles and the attribute levels used for each, is an important part of conducting a conjoint analysis. As for traditional design of experiments, efficiencies can be assessed to ensure achieving as much information as possible from a given number of profiles. An interesting question regarding design efficiency within conjoint analysis is whether customers are able to make choices that reflect their utility function properly when the alternatives are very different from each other. It is easy to state whether you prefer property A or property B given that all other properties are kept equal, but when all properties are varied at the same time, one may choose

to use a simpler mental decision model, such as always choosing the cheapest option. This dilemma was investigated in Flynn et al. (2015), in which an end-of-life survey was conducted using two designs with different efficiencies.

Conjoint analysis is highly relevant for DNB, which delivers many different value propositions to the customers in which properties must be prioritized. There are for instance several customer programs with different benefits offered to the relevant adult customers. Those between 18 and 28 years old qualify for membership in the program UNG (Young Adults), which offers advantages such as discounts on popular festivals, car rental services, legal advice and insurance (DNB, 2024). Similarly, Pluss is for everyone, SAGA for those with high income and/or high net worth, and Private Banking for those with the highest net worth. To increase customer satisfaction and loyalty, it is important to ensure that the advantages are attractive in the target group. When evaluating the customer programs in 2021, a choice-based conjoint analysis was conducted to assess the preferences of relevant customers. A total of 1011 customers were included in the study, and repeatedly asked to choose between 4 concept profiles with combinations of properties from 8 different main categories. An example of two of the profiles can be found in Table 2.1. Based on the survey, DNB could assess how attractive the offers were for different customer groups, which supported prioritization in the continued work on improving the programs.

Table 2.1: Example of concept profiles for the customer program analysis.

Mortgage interest	0.03% discount	0.01% discount
Monthly offer	15% discount on Power	25% discount on Home&Cottage
Special advantage	ID security	Customer service priority
Counselling	Saving and investments	Loan and daily economy
Product advantage	Mortgage: 50% discount on Google Home	Boat insurance: Free boat driving course
Customer service	Self-service	Self-service
Yearly price	295 NOK	495 NOK
Tickets/discounts	Free tickets to sports events	None

Another use case for conjoint analysis is understanding which aspects of the bank that are most important to the customers. To assess this, a yearly best/worst (also called MaxDiff) conjoint analysis is conducted, in which the customers are repeatedly presented with 4 different features (out of 21 in total), among which they must choose which they appreciate the most and the least. This is then complimented with a quarterly survey about bank satisfaction, where the customers rate their satisfaction with the bank, as well as different statements about the bank (for instance, how much they agree with the phrase "My bank is easy to get in touch with"). Customer satisfaction is then modelled as a function of the different statements, yielding an importance to each of them. Based on the best/worst conjoint analysis and the satisfaction regression model, a combined analysis of the customers' prioritizations is conducted. This is a valuable tool in deciding what the bank should focus on improving, for instance whether one should hire more customer service agents for personal counseling or more engineers for chatbot development. Regularly performing these analyses also enables assessing trends over time, for instance highlighting how increased interest rates makes the customers change their preferences.

3 Paper summaries

The main part of the thesis consists of four papers. Paper 1 and 2 are focused on finding alternative ways to block regular and nonregular two-level screening designs, respectively. Paper 3 and 4 are concerned with analysis of nonregular two-level screening designs.

3.1 Paper 1

The first paper focuses on alternative ways of blocking regular two-level fractional factorial designs. Traditionally, the proposed way to arrange the designs in two blocks has been to use a higher-order interaction column (Wu and Hamada, 2009). This ensures orthogonality between the block effect and the main effects, but often leads to a large decrease in projectivity. We believe that projectivity properties are very important in screening settings, as one is then guaranteed the possibility to estimate (possibly up to some limit) the effects corresponding to the few active factors. Alternative blocks were therefore tested for 16-, 32- and 64-run designs with good projection properties. The blocks were required to be orthogonal to the main effects and were assessed by considering the resulting maximal minimum D_s-efficiencies considering all possible

combinations of active factors within the highest possible projectivity. In many cases, blocks resulting in better projectivity properties than the traditionally preferred blocks were found, with high corresponding D_s-efficiencies. To assess the impact of partial aliasing between blocks and interactions effects, the change in standard deviations induced by the blocking was also investigated.

The number of possible blockings rapidly increases with the number of runs. We only considered arrangement in two, four and eight blocks of equal size. Considering arrangement of a design with $2t$ runs into two blocks, there are $\frac{\binom{2t}{t}}{2!}$ block candidates. For 16 runs, there are 6435 candidates, while for 32 runs, there are 300540195 candidates. Clearly, the combinatorial explosion makes it infeasible to test all blocking candidates for large designs. Thus several alternative strategies for generating reasonable candidate blocks were tested. The first one was based on allocating the runs belonging to a mirror image pair to the same block (the MIP-strategy), a strategy enabled by the foldover structure of the fractional factorial designs. This strategy was first proposed by Jacroux (2009). The second strategy was based on constructing blocking schemes by doubling a blocked design, and the third strategy on using columns from Hadamard matrices containing the design under consideration. The fourth and final strategy was to arrange an existing blocking into more blocks, for instance utilizing an arrangement in two blocks to find an arrangement in four blocks.

3.2 Paper 2

The second paper continues the work in Paper 1, focusing on finding blocks that preserve the projection properties while obtaining high D_s-efficiencies. First, some fractional factorial designs that were not included in Paper 1 were considered in order to complete the results for those designs. Then we shifted the focus to testing blocking arrangements for nonregular two-level screening designs, in particular the 16-run no-confounding (NC) designs and 12-, 20- and 24-run Plackett-Burman (PB) designs and their foldovers. For the 16- to 24-run designs, all possible blocks were tested to ensure finding the best block. Results for the 16- and 20-run designs were compared to the blocking arrangements suggested by Cheng et al. (2004), a much-cited work on blocking nonregular designs that focused on achieving generalized minimum aberration blocked designs. We showed that when prioritizing high projectivity, better projectivity properties could often be obtained, especially for ar-

rangements in more than two blocks if allowing a small degree of confounding between the block effect(s) and the main effects. This has often not been done in the literature, as having clear main effects has been considered a priority. Thus the experimenter should be cautious if considering using these blocks. To support an informed choice, we provide information about the change in standard deviations for the main effects resulting from the confounding.

For the foldover designs, the combinatorial explosion prohibited testing all possible blocks except for the PB_{12+12} design. Strategies utilizing mirror image pairs were therefore applied. As in Paper 1, the MIP-strategy was tested. In this paper we also tested placing runs belonging to a mirror image pair in different blocks, as proposed by Ou et al. (2011). Moreover, testing placement of the original design in one block and the foldover in another was tested, as well as using an unassigned column if all columns from the design were not included. This often yielded optimal or close to optimal blocks. For arranging foldover designs in more than two blocks, the MIP-strategy was explored, as well as using a blocking for the original design to arrange the foldover runs in new blocks, and thereby expanding the number of blocks. Testing different strategies enabled suggesting blocks for different blocking scenarios, such as folding over a blocked design and expanding a blocked design with foldover runs which must be run in new blocks.

In Paper 1 and 2, we demonstrated that projection properties of blocked designs could in many cases be improved by choosing other blocks than the ones usually suggested, which tend to focus on minimum aberration. There is a large potential for further work on blockings that focus on projection properties and D_s-efficiencies. For instance, one could have tested all different orthogonal arrays of different sizes, instead of focusing on designs with good properties. Moreover, it would be very useful to assess whether blocks can be found that are good all-round candidates across different projectivities. Another interesting approach could be to test restricting the amount of confounding between blocks and interactions.

3.3 Paper 3

In the third paper, we propose a new method for analyzing nonregular screening designs. Identifying the active factors for nonregular designs can be challenging, especially if the design has few runs, there is a high variance and interactions are active. Strategies for analyzing nonregular

designs can be divided into two main classes, effect-based and factor-based searches. Effect-based approaches aim at identifying the correct active effects, resulting in a final model that is a satisfying approximation of the underlying response surface. Factor-based approaches aim at identifying the active factors, which can then be investigated in further experimentation. Our proposed method belongs to the latter class.

One problem that may arise when testing many candidate models is that several candidate sets of active factors may explain the variation in the response equally well. The chance of this increases the more terms that are included in the model. Our proposed method is therefore based on utilizing projection models, but limiting the number of effects that are included in the final candidate models. For each candidate set of a given size, we fit the full projection (FP) model, that is, a model with all main effects and interactions included, up to the limit given by the projectivity of the design. Then the l terms with the largest corresponding coefficient values are chosen from the FP model, and the reduced model with only these terms included is fitted. The corresponding MSE is stored for each reduced model, resulting in a candidate set of models with the lowest MSE-values. The size of this candidate set can be chosen beforehand, or for instance by inspecting the best models and choosing the cutoff where a large increase in MSE occurs.

The method was also compared to the R^2- and F-test-based methods proposed in Tyssedal and Hussain (2016), which it outperformed in most cases. An important feature was increased robustness in the presence of large variance. The performance of the method was further assessed in a simulation study, testing situations with 3 and 4 active factors for the 16-run NC designs with 6-8 factors and the 12-run PB design. To test the performance for a wide range of cases and reduce the risk of linear dependence, we drew a new model of the specified format in each iteration instead of using the common approach of assessing only a small panel of models. The results made it clear that using 16 runs is recommended whenever possible, as it increased the chance of including the correct active factors among the final models substantially.

3.4 Paper 4

The fourth paper also considers the analysis of nonregular designs, but this time with emphasis on designs with a foldover structure. A special property of foldover designs is that their effects can be separated in two orthogonal subspaces, for the odd and even effects, respectively. We

introduce a new analysis method called *the decoupling method*, utilizing the two subspaces to create two new responses which can be analyzed separately of each other.

The idea of analyzing foldover designs in two steps was first proposed for PB designs by Miller and Sitter (2001), which followed up with a similar method suited for nonorthogonal designs in Miller and Sitter (2005). The main idea is to estimate and select main effects in the first step, possibly using unassigned columns to estimate the variance. Then two-factor interactions are included in the second step, by applying an all-possible-subsets procedure. This work inspired Jones and Nachtsheim (2017) to introduce a related analysis method tailored for DSDs, in which fake factors are added to get a variance estimate in the first step, and an F-test used to select second-order effects in the second step. A generalized version of this was proposed in Hameed et al. (2023), utilizing properties of the recently introduced OMARS designs (Ares and Goos, 2020). They avoid the use of fake factors and are able to utilize additional degrees of freedom in some cases. These methods perform well, but have some drawbacks, for instance assuming that only main effects and two-factor interactions are active. There is also a risk that the variance estimate in the second step is affected by selecting the wrong main effects in the first step. Furthermore, residuals cannot be assessed in the first step, as they are possibly affected by active two-factor interactions.

These drawbacks are overcome by the decoupling method, which yields a valid variance estimate in each step and enables assessment of residuals. Moreover, it is possible to test for presence of higher-order interactions, and we suggest a procedure for including those. After the new responses have been created for each step, standard statistical methods can be applied, so the method is easily available for practitioners. The proposed method is assessed and compared to existing ones in a simulation study, where we also demonstrate the huge impact of heredity assumptions.

4 Bibliography

Agarwal, J., W.S. DeSarbo, N.K. Malhotra, and V.R. Rao (2015), "An interdisciplinary review of research in conjoint analysis: Recent developments and directions for future research." *Customer Needs and Solutions*, 2, 19–40.

Almquist, E. and G./Harvard Business Review Wyner (2001), "Boost your marketing ROI with ex-

perimental design." https://hbr.org/2001/10/boost-your-marketing-roi-with-experimental-design. Accessed: 02.01.2024.

Ares, J. N. and P. Goos (2023), "Blocking OMARS designs and definitive screening designs." *Journal of Quality Technology*, 55, 489–509, URL https://doi.org/10.1080%2F00224065.2023.2196035.

Ares, J.N. and P. Goos (2020), "Enumeration and multicriteria selection of orthogonal minimally aliased response surface designs." *Technometrics*, 62, 21–36.

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kegl (2011), "Algorithms for hyper-parameter optimization." In *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, eds.), volume 24, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Bergstra, J. and Y. Bengio (2012), "Random search for hyper-parameter optimization." *Journal of Machine Learning Research*, 13, 281–305, URL http://jmlr.org/papers/v13/bergstra12a.html.

Bergstra, J., D. Yamins, and D. Cox (2013), "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures." In *Proceedings of the 30th International Conference on Machine Learning* (Sanjoy Dasgupta and David McAllester, eds.), volume 28 of *Proceedings of Machine Learning Research*, 115–123, PMLR, Atlanta, Georgia, USA, URL https://proceedings.mlr.press/v28/bergstra13.html.

Bodmer, W. (2003), "RA Fisher, statistician and geneticist extraordinary: a personal view." *International Journal of Epidemiology*, 32, 938–942, URL http://dx.doi.org/10.1093/ije/dyg289.

Bojinov, I. and S. Gupta (2022), "Online experimentation: Benefits, operational and methodological challenges, and scaling guide." *Harvard Data Science Review*, URL http://dx.doi.org/10.1162/99608f92.a579756e.

Box, G. and J. Tyssedal (1996), "Projective properties of certain orthogonal arrays." *Biometrika*, 83, 950–955, URL https://doi.org/10.1093%2Fbiomet%2F83.4.950.

Box, G.E.P, W.G. Hunter, and J.S. Hunter (2005), *Statistics for Experimenters*, 2 edition. Wiley, New York.

Box, G.E.P. and R.D. Meyer (1993), "Finding the active factors in fractionated screening experiments." *Journal of Quality Technology*, 25, 94–105.

Box, G.E.P. and K.P. Wilson (1951), "On the experimental attainment of optimum conditions." *Journal of the Royal Statistical Society*, 13, 1–45.

Brochu, E., V. M. Cora, and N. de Freitas (2010), "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning." URL `https://arxiv.org/abs/1012.2599`.

Burtini, G., J. Loeppky, and R. Lawrence (2015), "A survey of online experiment design with the stochastic multi-armed bandit."

Cacciarelli, D. and M. Kulahci (2023), "Active learning for data streams: a survey." *Machine Learning*, 113, 185–239, URL `http://dx.doi.org/10.1007/s10994-023-06454-2`.

Cacciarelli, D., M. Kulahci, and J. S. Tyssedal (2022), "Stream-based active learning with linear models." *Knowledge-Based Systems*, 254, 109664, URL `http://dx.doi.org/10.1016/j.knosys.2022.109664`.

Cacciarelli, D., M. Kulahci, and J. S. Tyssedal (2023), "Robust online active learning." *Quality and Reliability Engineering International*, URL `http://dx.doi.org/10.1002/qre.3392`.

Candes, E. and T. Tao (2007), "The Dantzig selector: Statistical estimation when p is much larger than n." *The Annals of Statistics*, 35, URL `http://dx.doi.org/10.1214/009053606000001523`.

Carcillo, F., Y.-A. Le Borgne, O. Caelen, and G. Bontempi (2017), "An assessment of streaming active learning strategies for real-life credit card fraud detection." In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, URL `http://dx.doi.org/10.1109/dsaa.2017.10`.

Carcillo, F., Y.-A. Le Borgne, O. Caelen, and G. Bontempi (2018), "Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization." *International Journal of*

Data Science and Analytics, 5, 285–300, URL http://dx.doi.org/10.1007/s41060-018-0116-z.

Chen, H. and C.-S. Cheng (1999), "Theory of optimal blocking of 2^{n-m} designs." *The Annals of Statistics*, 27, URL http://dx.doi.org/10.1214/aos/1017939246.

Chen, T. and C. Guestrin (2016), "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, URL https://doi.org/10.1145%2F2939672.2939785.

Cheng, C.-S. and R. Mukerjee (2001), "Blocked regular fractional factorial designs with maximum estimation capacity." *The Annals of Statistics*, 29, URL http://dx.doi.org/10.1214/aos/1009210551.

Cheng, S.-W., W. Li, and K. Q. Ye (2004), "Blocked nonregular two-level factorial designs." *Technometrics*, 46, 269–279, URL https://doi.org/10.1198%2F004017004000000301.

Czitrom, V. (1999), "One-factor-at-a-time versus designed experiments." *The American Statistician*, 53, 126, URL http://dx.doi.org/10.2307/2685731.

Daniel, C. (1976), *Applications of Statistics to Industrial Experimentation*. Wiley.

DNB (2024), "Ung (Young Adults)." URL https://www.dnb.no/en/customer-loyalty-programme/ung.

Evangelaras, H. and C. Koukouvinos (2004), "On generalized projectivity of two-level screening designs." *Statistiscs & Probability Letters*, 68, 429–434, URL https://doi.org/10.1016%2Fj.spl.2004.04.011.

Falkner, S. (2018), "Welcome to HpBandSter's documentation!" URL https://automl.github.io/HpBandSter/build/html/index.html#.

Falkner, S., A. Klein, and F. Hutter (2018), "BOHB: Robust and efficient hyperparameter optimization at scale." URL https://arxiv.org/abs/1807.01774.

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

Fisher, R. (1935), *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Flynn, T. N., M. Bilger, C. Malhotra, and E. A. Finkelstein (2015), "Are efficient designs used in discrete choice experiments too difficult for some respondents? A case study eliciting preferences for end-of-life care." *PharmacoEconomics*, 34, 273–284, URL http://dx.doi.org/10.1007/s40273-015-0338-z.

Frazier, P. I. (2018), "A tutorial on Bayesian optimization." URL https://arxiv.org/abs/1807.02811.

Gupta, S., R. Kohavi, D. Tang, Y. Xu, R. Andersen, E. Bakshy, N. Cardin, S. Chandran, N. Chen, D. Coey, M. Curtis, A. Deng, W. Duan, P. Forbes, B. Frasca, T. Guy, G. W. Imbens, G. Saint Jacques, P. Kantawala, I. Katsev, M. Katzwer, M. Konutgan, E. Kunakova, M. Lee, M.J. Lee, J. Liu, J. McQueen, A. Najmi, B. Smith, V. Trehan, L. Vermeer, T. Walker, J. Wong, and I. Yashkov (2019), "Top challenges from the first practical online controlled experiments summit." *SIGKDD Explor. Newsl.*, 21, 20–35, URL https://doi.org/10.1145/3331651.3331655.

Haizler, T. and D. M. Steinberg (2020), "Factorial designs for online experiments." *Technometrics*, 63, 1–12, URL http://dx.doi.org/10.1080/00401706.2019.1701556.

Hall, M. Jr. (1961), "Hadamard matrix of order 16." *Jet Propulsion Laboratory Research Summary*, 21–26.

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

Hameed, M. S. I., J. Nunez, and P. Goos (2023), "Analysis of data from ortogonal minimally aliased response surface designs." *Journal of Quality Technology*, 55, 366–384.

Jacroux, M. (2009), "Blocking in two-level non-regular fractional factorial designs." *Journal of Statistical Planning and Inference*, 139, 1215–1220, URL https://doi.org/10.1016%2Fj.jspi.2008.08.001.

Jones, B. and D. C. Montgomery (2010), "Alternatives to resolution IV screening designs in 16 runs." *International Journal of Experimental Design and Process Optimisation*, 1, 285, URL https://doi.org/10.1504%2Fijedpo.2010.034986.

Jones, B. and C. J. Nachtsheim (2017), "Effective design-based model selection for definite screening designs." *Technometrics*, 59, 319–323.

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

Jones, B., S. M. Shinde, and D. C. Montgomery (2015), "Alternatives to resolution III regular fractional factorial designs for 9–14 factors in 16 runs." *Applied Stochastic Models in Business and Industry*, 31, 50–58.

Kohavi, R., R. Longbotham, D. Sommerfield, and R. M. Henne (2008), "Controlled experiments on the web: survey and practical guide." *Data Mining and Knowledge Discovery*, 18, 140–181, URL `http://dx.doi.org/10.1007/s10618-008-0114-1`.

Krutsick, R. (2012), "Finding the winning combination: An application of multivariate testing from digital marketing." In *Proceedings of the 2012 SAS Global Forum Conference*, : SAS Institute Inc, URL `https://support.sas.com/resources/papers/proceedings12/206-2012.pdf`.

Kulachi, M. and G. E. P. Box (2003), "Catalysis of discovery and development in engineering and industry." *Quality Engineering*, 15, 513–517.

Lee, J. (2022), "4 ingredient magic cake (no butter or oil)." URL `https://kirbiecravings.com/4-ingredient-magic-cake/`.

Lenth, R. V. (1989), "Quick and easy analysis of unreplicated factorials." *Technometrics*, 31, 469–473, URL `http://dx.doi.org/10.1080/00401706.1989.10488595`.

Li, L., . Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar (2016), "Hyperband: A novel bandit-based approach to hyperparameter optimization." URL `https://arxiv.org/abs/1603.06560`.

Louviere, J. J., T. N. Flynn, and R. T. Carson (2010), "Discrete choice experiments are not conjoint analysis." *Journal of Choice Modelling*, 3, 57–72, URL `http://dx.doi.org/10.1016/s1755-5345(13)70014-9`.

Lujan-Moreno, G. A., P. R. Howard, O.G. Rojas, and D.C. Montgomery (2018), "Design of experiments and response surface methodology to

tune machine learning hyperparameters, with a random forest case-study." *Expert Systems with Applications*, 109, 195–205.

Martinez-Cantin, R. (2014), "Bayesopt: A Bayesian optimization library for nonlinear optimization, experimental design and bandits." *Journal of Machine Learning Research*, 15, 3915–3919, URL `http://jmlr.org/papers/v15/martinezcantin14a.html`.

Miller, A. and R.R. Sitter (2001), "Using the Folded-Over 12-Run Plackett-Burman Design to Consider Interactions." *Technometrics*, 43, 44–55.

Miller, A. and R.R. Sitter (2005), "Using folded-over nonorthogonal designs." *Technometrics*, 47, 502–513.

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

Myers, R. H, D. C Montgomery, and C. M. Anderson-Cook (2016), *Response Surface Methodology*, 4 edition. Wiley, New Jersey.

Ou, Z., H. Qin, and H. Li (2011), "Optimal blocking and foldover plans for nonregular two-level designs." *Journal of Statistical Planning and Inference*, 141, 1635–1645.

Pannakkong, W., K. Thiwa-Anont, K. Singthong, P. Parthanadee, and J. Buddhakulsomsiri (2022), "Hyperparameter tuning of machine learning algorithms using response surface methodology: A case study of ANN, SVM, and DBN." *Mathematical Problems in Engineering*.

Park, D.-K., H.-S. Kim, and H.-K. Kang (2007), "Classification rule for optimal blocking for nonregular factorial designs." *Communications for Statistical Applications and Methods*, 14, 483–495.

Plackett, R.L. and J.P. Burman (1946), "The design of optimum multifactorial experiments." *Biometrika*, 33, 305–325.

Sadeghi, S., P. Chien, and N. Arora (2019), "Sliced designs for multi-platform online experiments." *Technometrics*, 62, 387–402, URL `http://dx.doi.org/10.1080/00401706.2019.1647288`.

Schau-Hansen, H. (2022), "Improving direct response modelling through hyperparameter optimization of extreme gradient boosting and random forests."

Schoen, E. D., B. Sartono, and P. Goos (2013), "Optimal blocking for general resolution-3 designs." *Journal of Quality Technology*, 45, 166–187, URL https://doi.org/10.1080%2F00224065.2013.11917924.

Simchi-Levi, D. and C. Wang (2022), "Multi-armed bandit experimental design: Online decision-making and adaptive inference." *SSRN Electronic Journal*, URL http://dx.doi.org/10.2139/ssrn.4224969.

Siroker, D./Optimizely (2010), "Obama's 60 million doller experiment." https://www.optimizely.com/no/insights/blog/how-obama-raised-60-million-by-running-a-simple-experiment/. Accessed: 02.01.2024.

Sitter, R. R., J. Chen, and M. Feder (1997), "Fractional resolution and minimum aberration in blocked 2^{n-k} designs." *Technometrics*, 39, 382–390, URL http://dx.doi.org/10.1080/00401706.1997.10485157.

Staelin, C. (2003), "Parameter selection for support vector machines. technical report hpl-2002-354." https://www.hpl.hp.com/techreports/2002/HPL-2002-354R1.pdf.

Stone, B. B., D. C. Montgomery, R. T. Silvestrini, and B. Jones (2017a), "No-confounding designs with 24 runs for 7-12 factors." *International Journal of Experimental Design and Process Optimisation*, 5, 151, URL http://dx.doi.org/10.1504/ijedpo.2017.087583.

Stone, B. B., D. C. Montgomery, R. T. Silvestrini, and B. Jones (2017b), "No-confounding designs with 20 runs— alternatives to resolution IV screening designs." *Quality and Reliability Engineering International*, 33, 1861–1872, URL http://dx.doi.org/10.1002/qre.2150.

Sun, D. X., C. F. J. Wu, and Y. Chen (1997), "Optimal blocking schemes for 2^n and 2^{n-p} designs." *Technometrics*, 39, 298, URL http://dx.doi.org/10.2307/1271134.

Sunder, G., T. A. Albrecht, and C. J. Nachtsheim (2022), "Robust sequential experimental strategy for black-box optimization with application to hyperparameter tuning." *Quality and Reliability Engineering International*, 38, 3992–4014, URL http://dx.doi.org/10.1002/qre.3181.

Turner, T., D. Eriksson, M. J. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. M. Guyon (2021), "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis

of the black-box optimization challenge 2020." In *Neural Information Processing Systems*, URL https://api.semanticscholar.org/CorpusID:233324399.

Tyssedal, J. and S. Hussain (2016), "Factor screening in nonregular two-level designs based on projection-based variable selection." *Journal of Applied Statistics*, 43, 490–508, URL http://dx.doi.org/10.1080/02664763.2015.1070805.

Tyssedal, J. and O. Samset (1997), "Analysis of the 12 run Plackett-Burman design." *Technical Report no. 8*.

Wang, X., Y. Jin, S. Schmitt, and M. Olhofer (2023), "Recent advances in Bayesian optimization." *ACM Computing Surveys*, 55, 1–36, URL http://dx.doi.org/10.1145/3582078.

Watanabe, S. (2023), "Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance." *ArXiv*, abs/2304.11127, URL https://api.semanticscholar.org/CorpusID:258291728.

Wolters, M.A. and D. Bingham (2011), "Simulated annealing model search for subset selection in screening experiments." *Technometrics*, 53, 225–237, URL http://www.jstor.org/stable/23210399.

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

Yu, L., X. Yao, S. Wang, and K.K. Lai (2011), "Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection." *Expert Systems with Applications*, 38, 15392–15399, URL http://dx.doi.org/10.1016/j.eswa.2011.06.023.

Yuan, M., V. R. Joseph, and Y. Lin (2007), "An efficient variable selection approach for analyzing designed experiments." *Technometrics*, 49, 430–439, URL http://dx.doi.org/10.1198/004017007000000173.

Zhang, R. and D.K. Park (2000), "Optimal blocking of two-level fractional factorial designs." *Journal of Statistical Planning and Inference*, 91, 107–121, URL http://dx.doi.org/10.1016/s0378-3758(00)00133-6.

Zhao, S., P. Li, and R. Karunamuni (2013), "Blocked two-level regular factorial designs with weak minimum aberration." *Biometrika*, 100, 249–253, URL http://dx.doi.org/10.1093/biomet/ass061.

Part II

Research papers

Paper 1

Preserving projection properties when regular two-level designs are blocked

John Sølve Tyssedal and Yngvild Hole Hamre.

Published in *Journal of Statistical Planning and Inference*, volume 221, 2022, pages 266-280.

Journal of Statistical Planning and Inference 221 (2022) 266–280

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Preserving projection properties when regular two-level designs are blocked

John Tyssedal [a,*], Yngvild Hole Hamre [a,b]

[a] Department of Mathematical Sciences, NTNU, Norway
[b] DNB, Norway

ARTICLE INFO

Article history:
Received 20 August 2020
Received in revised form 4 May 2022
Accepted 8 May 2022
Available online 16 May 2022

Keywords:
Aliasing
Doubling
D_s-efficiency
Hadamard matrices
Mirror image pair runs

ABSTRACT

Regular two-level designs are useful and popular screening designs, but if they need to be run in blocks, their projection properties can dramatically deteriorate. Interactions may be fully confounded with block defining contrast(s), causing uncertainty in the identification of active factors. In this paper, we demonstrate alternative ways of blocking two-level regular designs such that their projective properties can be preserved or just weakly affected at the expense of just a small decrease in efficiency. Thereby, we can estimate effects we are normally interested in even if the design is blocked. Common regular two-level designs with 16, 32 and 64 runs are considered.

1. Introduction

Regular fractional factorial two-level designs, usually denoted as 2^{k-p} designs, are among the most popular experimental plans used for screening. They are characterized by the property that any two effects can be estimated independently of each other or are fully aliased which makes their analysis rather straight forward. Their properties and usefulness are nicely explained in several textbooks like Box et al. (2005); Wu and Hamada (2009) and Montgomery (2019).

Regular two-level designs can be constructed by allocating k–p factors to the principal factor columns in a full factorial with 2^{k-p} runs and the remaining p factor(s) to some higher order interaction column(s). This causes the full aliasing of effects. Resolution is an important concept to describe the aliasing. *By definition, a two-level design is of resolution R if no p-factor effect is aliased with any other effect containing less than R–p factors* (Box and Hunter, 1961). The aliasing can be derived from the defining relation which is the set of all columns that are equal to the identity column. For instance, if factor D is assigned to a three-factor interaction column, say ABC, it creates a word in the defining relation given by $ABCD$. Any effect associated with one, two or three of these four letters will then be aliased with the effect associated with the remaining one(s). In general, the defining relation has $2^p - 1$ words and to each design a wordlength pattern can be constructed of the form $W = (A_3, A_4, \ldots, A_k)$, where A_i is the number of words of length i. It is assumed that no main effects are aliased with each other. The higher the resolution, the better, but designs with the same resolution can have unequal numbers of fully aliased effects. For given k and p, let d_1 and d_2 be two designs. If r is the smallest integer for which $A_r(d_1) \neq A_r(d_2)$ and $A_r(d_1) < A_r(d_2)$, d_1 is said to have less aberration than d_2. If no design has less aberration than d, it is said to have minimum aberration (Fries and Hunter, 1980). This is a useful way to rank regular two-level designs.

* Corresponding author.
E-mail address: john.tyssedal@ntnu.no (J. Tyssedal).

https://doi.org/10.1016/j.jspi.2022.05.001
0378-3758/

A much cited rule for experimental work is: "Block what you can and randomize what you can't". Blocking is an effective way to improve the efficiency of a design when not all the experimental runs can be performed under homogeneous conditions. For regular two-level designs the general rule of blocking is to assign one or a set of higher order interaction column(s), named block defining contrast(s) as block factor(s), and then associate the distinct level combinations in the column(s) with different blocks. A good blocking scheme should have as few as possible lower order interactions confounded with the block effects, and an additional wordlength pattern $W_b = (A_{2b}, A_{3b}, \ldots, A_{kb})$ can be constructed, where A_{ib} is the number of ith order interactions confounded with the block effect(s).

In order to rank blocked regular two-level designs, a combined wordlength pattern may be constructed. Several ways of doing this have been proposed and discussed in the literature (Sitter and M, 1997; Chen and Cheng, 1999; Zhang and Park, 2000; Mukerjee and Wu, 2006; Cheng and Tsai, 2009; Xu and Mee, 2010; Zhao et al., 2013). Cheng and Mukerjee (2001) proposed and studied a criterion for blocking based on the alias pattern of the interactions with the purposes of maximizing the number of two-factor interactions that are neither aliased with main effects nor confounded with blocks and at the same time distributing the interactions over the alias sets as uniformly as possible.

Sun et al. (1997) used the two wordlength patterns in addition to two other criteria, the number of clear main effects and the number of clear two-factor interactions, to come up with good blocking schemes. A main effect was called clear if it was not aliased with any two-factor interaction and any block effect. Similarly, a two-factor interaction was called clear if it was not aliased with any main effect, any other two-factor interaction and any block effect. Based on these four criteria, the concept of admissibility of blocking schemes was introduced, as a way to rule out bad designs. This criterion was further explored in Mukerjee and Wu (1999).

In this paper we will mainly be concerned with using blocked regular two-level designs for screening purposes. Screening is about separating out the normally few factors, from potentially many, that can explain most of the variation in the response. Projection properties concern how good a design is when restricted to a subset of factors. It is therefore important for a screening design to have good projection properties. Box and Tyssedal (1996) defined projectivity of two-level designs as follows: *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.* For regular two-level designs the projectivity is always given by $P = R - 1$. For more on projectivity, factor sparsity and screening, we refer to Box and Tyssedal (2001) and Tyssedal (2008).

However, recommended schemes for blocking two level regular designs may cause the projection properties of the blocked designs to deteriorate. For instance, the two-level resolution V (projectivity $P = 4$) design for five factors in 16 runs, denoted as the 2_V^{5-1} design, becomes a projectivity $P = 1$ design when blocked in two blocks using the recommended scheme.

We will in this paper present alternative ways of blocking regular designs, such that projection properties will be preserved either fully or to some extent. Reflected in our focus on screening, all the designs chosen to be blocked have good projections properties. Only blocks having an equal number of runs in each block will be considered. The designs to be blocked will have from 16 to 64 runs and will, given k and p, be minimum aberration designs taken from Wu and Hamada (2009), pages 253–257. Also, whenever we write *recommended* way of blocking, we will from now on refer to the way of blocking given there on pages 260–263. With reference to the tables presented in Sun et al. (1997), the blocked designs found in Wu and Hamada (2009) are the ones that rank best according to the number of clear main effects for 16 run designs. For designs with more than 16 runs, the number of clear two-factor interactions is used as ranking criterion.

This paper is organized as follows. In Section 2, we explain what is meant by projectivity of blocked regular two-level designs and introduce the criterion used to discriminate between different ways of blocking. A motivational example is given in Section 3. Section 4 is devoted to strategies for finding good blocking schemes, and in Section 5 we present specific ways of blocking the designs and the projectivity that is possible to obtain in each case. Several ways of blocking regular two-level designs are discussed and compared in Section 6. Concluding remarks are given in Section 7.

2. Projectivity concepts and evaluation criterion

If many design factors are active, one may have to give up on estimating all higher order interactions. Evangelaras and Koukouvinos (2004) introduced the concept of generalized projectivity for two level designs 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 from the design all factorial effects including up to α-factor interactions are estimable.* In line with that, Hussain and Tyssedal (2016) defined the projectivity of blocked two-level designs as: *A blocked two-level design is of projectivity P or P_α if for any selection of P-columns the intercept and all factorial effects up to including P-factor interactions or α-factor interactions are estimable, respectively.*

Let the model for the expected response $\mathbf{Y}$ in a blocked experiment be written as

$$E[\mathbf{Y}] = \mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_b\boldsymbol{\beta}_b. \tag{1}$$

Here $\mathbf{Y}$ is a vector of n observations, $\mathbf{X}_e$ is a $n \times (p+1)$ matrix containing a column for the intercept and the effect columns (main effects and interactions under consideration), $\boldsymbol{\beta}_e = [\beta_0, \beta_1, \ldots, \beta_p]^t$ is the $(p+1)$ dimensional vector containing the intercept and the regression coefficients that are half the corresponding effects, $\mathbf{X}_b$ is a $n \times b$ matrix containing the columns for the blocks and $\boldsymbol{\beta}_b = [\beta_1^*, \ldots, \beta_b^*]^t$ is the corresponding vector of the coefficients for the b block effects. While

Table 1
The 2_{IV}^{8-4} design with two alternative ways of blocking.

Run	A	B	C	D	E = ABC	F = ABD	G = ACD	H = BCD	B_b	B_b^*
1	−1	−1	−1	−1	−1	−1	−1	−1	1	1
2	1	−1	−1	−1	1	1	1	−1	−1	1
3	−1	1	−1	−1	1	1	−1	1	−1	1
4	1	1	−1	−1	−1	−1	1	1	1	−1
5	−1	−1	1	−1	1	−1	1	1	1	1
6	1	−1	1	−1	−1	1	−1	1	−1	−1
7	−1	1	1	−1	−1	1	1	−1	−1	−1
8	1	1	1	−1	1	−1	−1	−1	1	−1
9	−1	−1	−1	1	−1	1	1	1	1	−1
10	1	−1	−1	1	1	−1	−1	1	−1	−1
11	−1	1	−1	1	1	−1	1	−1	−1	−1
12	1	1	−1	1	−1	1	−1	−1	1	1
13	−1	−1	1	1	1	1	−1	−1	1	−1
14	1	−1	1	1	−1	−1	1	−1	−1	1
15	−1	1	1	1	−1	−1	−1	1	−1	1
16	1	1	1	1	1	1	1	1	1	1

the definition of projectivity informs what is possible to estimate, it does not tell how well the effects are estimated. A useful criterion in that context is the *D-optimality* criterion defining a D-optimal design as the one for which $\frac{|\mathbf{X}^t\mathbf{X}|}{n^p}$ is maximized. Here $\mathbf{X}=[\mathbf{X}_e, \mathbf{X}_b]$ is the design matrix. $|\mathbf{X}^t\mathbf{X}|$ denotes the determinant of $\mathbf{X}^t\mathbf{X}$ and is inversely proportional to the square of the volume of the confidence region of the regression coefficients and thereby directly related to estimating the effects.

When a design is blocked, it may be argued that a precise estimation of the corresponding block effect(s) is not as important as for the effects of the factors under investigation (Atkinson and Donev, 1992, page 106). For $\mathbf{X}=[\mathbf{X}_e, \mathbf{X}_b]$ we have $\mathbf{X}^t\mathbf{X}=\begin{bmatrix}\mathbf{X}_e^t\mathbf{X}_e & \mathbf{X}_e^t\mathbf{X}_b\\ \mathbf{X}_b^t\mathbf{X}_e & \mathbf{X}_b^t\mathbf{X}_b\end{bmatrix}$, and the covariance matrix for the least square estimators for $\boldsymbol{\beta}_e$ is proportional to the upper left $(p+1)\times(p+1)$ submatrix of $\left[\mathbf{X}^t\mathbf{X}\right]^{-1}$ which is equal to $\left[\mathbf{X}_e^t\mathbf{X}_e-\left(\mathbf{X}_e^t\mathbf{X}_b\right)\left(\mathbf{X}_b^t\mathbf{X}_b\right)^{-1}\left(\mathbf{X}_b^t\mathbf{X}_e\right)\right]^{-1}$.

It can be shown that $\left|\mathbf{X}_e^t\mathbf{X}_e-\left(\mathbf{X}_e^t\mathbf{X}_b\right)\left(\mathbf{X}_b^t\mathbf{X}_b\right)^{-1}\left(\mathbf{X}_b^t\mathbf{X}_e\right)\right|=\frac{|\mathbf{X}^t\mathbf{X}|}{|\mathbf{X}_b^t\mathbf{X}_b|}$. A design that maximizes $\frac{|\mathbf{X}^t\mathbf{X}|}{|\mathbf{X}_b^t\mathbf{X}_b|}$ is said to be D_s-optimal (Atkinson and Donev, 1992, page 106). The subscript s refers to the subset of s columns for which we are really interested in estimating the effects and equals $p+1$ for the case given above. It is then natural to define D_s-efficiency as

$$D_{s,eff}=\frac{\left[\frac{|\mathbf{X}^t\mathbf{X}|}{|\mathbf{X}_b^t\mathbf{X}_b|}\right]^{\frac{1}{s}}}{n}. \tag{2}$$

where $\frac{1}{s}$ takes care of the increase in the determinant that occurs by increasing s. If for a given way of blocking $D_{s,eff}=0$ for one set of P-factors where s-1 is the number of all factorial effects up to P-factor interactions or α-factor interactions, it is clear that also $|\boldsymbol{X}^t\boldsymbol{X}|$ has to be zero. Then the inverse of $\boldsymbol{X}^t\boldsymbol{X}$ does not exist and projectivity P or P_α is not obtained, respectively. Otherwise, it is natural to choose a candidate from the ones with good overall values for $D_{s,eff}$ considering both the minimum, maximum and average values over all projections onto P factors.

3. A motivational example, the 2_{IV}^{8-4} design arranged in two blocks

The 2_{IV}^{8-4} design is a resolution IV design and hence of projectivity $P=3$, and all its projections onto three dimensions contain a fully replicated 2^3 design. The principal fraction of this design is given in Table 1, denoting the four principal factor columns as A, B, C and D. The generators for the four additional columns, E, F, G and H, are also given, together with two possible block defining contrasts. One is the recommended one, B_b, according to Wu and Hamada (2009). The other contrast, B_b^*, represents a possible alternative.

The recommended blocking scheme for arranging the design in two blocks is to let B_b= AB be the block defining contrast (any other two-factor interaction could have been chosen). The defining relation for the 2_{IV}^{8-4} design consists of 14 four letter words and one eight letter word. As a result, the four two-factor interactions AB, CE, DF and GH are fully confounded with the block defining contrast and 24 out of the 56 projections onto three dimensions are affected in the sense that one two-factor interaction cannot be estimated without being fully confounded with the block effect. Thus, when blocked the recommended way of blocking, the design becomes a $P=1$ design.

Now, if we let B_b^* be the block defining contrast, we will find that for 48 projections onto three dimensions $D_{s,eff}=0.917$, and $D_{s,eff}=1$ for the eight others. Here $D_{s,eff}$ is obtained letting $\mathbf{X}_e$ be a 16×8 matrix containing a column for

Table 2
The projectivity of some regular designs when they are blocked.

Design	Screen	Screen when blocked the recommended way
2^{8-4}_{IV}	(16, 8, 3)	(16, 8, 1, 2)
2^{5-1}_{V}	(16, 5, 4)	(16, 5, 1, 2)
2^{6-1}_{VI}	(32, 6, 5)	(32, 6, 2, 2), (32, 6, 1, 4)
2^{7-2}_{IV}	(32, 7, 3)	(32, 7, 2, 2), (32, 7, 1, 4)
2^{8-3}_{IV}	(32, 8, 3)	(32, 8, 2, 2), (32, 8, 1, 4)
2^{9-4}_{IV}	(32, 9, 3)	(32, 9, 1, 2)
2^{8-2}_{V}	(64, 8, 4)	(64, 8, 2, 2), (64, 8, 2, 4)

the intercept, the effect columns for the three main effects and their interactions and $\mathbf{X}_b = B^*_b$. Hence all the main effects and their interactions are now estimable for any three factors, and the design is of projectivity $P = 3$ when run in two blocks.

The notation (n, k, P) screen is used to describe a two-level projectivity P design with n runs and k factors used for screening. When run in b blocks of equal size, we will denote it a (n, k, P, b) screen. Table 2 summarizes what happens with the projectivity of some regular two-level designs when they are blocked the recommended way. As can be observed a rather severe loss in projectivity is common for many of them.

4. Blocking strategies

If a design with $n = 2t$ runs is to be run in two blocks with an equal number of runs in each, this can be done in $\frac{\binom{2t}{t}}{2!}$ possible ways, and for $t \geq 4$ and a multiple of two there are $\frac{\binom{2t}{t/2}\binom{2t-t/2}{t/2}\binom{t}{t/2}}{4!}$ possibilities for blocking into four blocks. For example, a 16 run design can be arranged in two and four blocks in 6435 and 2,627,625 possible ways respectively, and a 32 run design can be arranged in two blocks in 300,540,195 possible ways. Many of these block alternatives will have undesirable properties, and therefore some restrictions should be placed on the alternatives to investigate. We will restrict the blocking contrast(s) to be orthogonal to the main effect contrasts. The combinatorial explosion illustrated above also asks for strategies that can provide good blocking alternatives without investigating all possibilities. Several such strategies now follow.

Strategy S1. *Allocate mirror image pair runs to the same block*

A full factorial two-level design in n runs consists of $\frac{n}{2}$ mirror image pair runs. This is also true for some of its fractions. For example, if the factor columns in a 2^3 design are denoted A, B and C, we may add a fourth factor column D = ABC and obtain a 2^{4-1}_{IV} design. The generator of this fraction is then D = ABC and the defining relation is I = ABCD, a column of only 1's. The defining relation here consists of one word of length four, and the design consists of all level combinations for which their entry-wise product is 1. For a given level combination in the design, it is then evident that its mirror image run is also included in the design, and that this must be the case when the defining relation is a word of even length. Depending on the degree of fractioning and which fraction is chosen, the defining relation may consist of several words with a plus or a minus in front of them. The corresponding design will then consist of all level combinations that satisfy the defining relation. For a given level combination in the design, its mirror image run will also satisfy the defining relation and be included if all these words are of even length. Furthermore, if all level combinations in a design are mirror image pair runs, there can be no word of odd length in the defining relation, since a level combination that satisfies this constraint will exclude its mirror image run. For designs for which it can be used, this way of blocking assures the blocking contrast(s) to be orthogonal to main effect and odd factor interaction contrasts and reduces the number of blocking alternatives. When blocked in two blocks the reduction is from 6435 to 35 for a 16 run design and from 300,540,195 to 6435 for a 32 run design. Since the loss in efficiency obtained by blocking is caused by effects being confounded with the block defining contrast(s), the use of $D_{s,eff}$ to evaluate the alternatives obtained by this strategy act as a way to reduce the confounding between even factor interaction contrasts and the block defining contrast(s).

Strategy S2. *Construct blocking schemes by doubling a blocked design*

Hadamard matrices are $n \times n$ orthogonal matrices consisting of 1's and -1's. They exist for $n = 1$ and 2 and otherwise apparently for n a multiple of 4. Without loss of generality we may let the first column consist of 1's. The remaining columns will then have equally many 1's and -1's.

Several resolution IV designs can be constructed from Hadamard matrices. Let $\boldsymbol{S}_1 = 1$ and let

$$\boldsymbol{S}_n = \begin{bmatrix} \boldsymbol{S}_{n/2} & \boldsymbol{S}_{n/2} \\ \boldsymbol{S}_{n/2} & -\boldsymbol{S}_{n/2} \end{bmatrix}, n = 2, 4, 8, \ldots . \tag{3}$$

These matrices are called Sylvester type Hadamard matrices and will provide us with all saturated regular designs, with an intercept column included. For $n = 2^{2+t}$, $t = 1, 2, 3, \ldots$ the designs $\boldsymbol{D}_n = \begin{bmatrix} \boldsymbol{S}_{n/2} \\ -\boldsymbol{S}_{n/2} \end{bmatrix}$ are resolution IV designs in n runs and $n/2$ factors. Clearly, for every level combination its mirror image run will also be included. For a given regular design $\boldsymbol{D}_n$ with n runs and k factors, $k = 1, 2, \ldots, n-1$, a regular two-level design with $2n$ runs and $2k$ ($2k+1$ if an intercept column is included in $\boldsymbol{D}_n$) factors can be constructed by doubling as

$$\boldsymbol{D}_{2n} = \begin{bmatrix} \boldsymbol{D}_n & \boldsymbol{D}_n \\ \boldsymbol{D}_n & -\boldsymbol{D}_n \end{bmatrix}. \tag{4}$$

If $\boldsymbol{D}_n$ is a resolution IV design, $\boldsymbol{D}_{2n}$ is also a resolution IV design.

Now suppose a regular design $\boldsymbol{D}_n$ with k factors has been blocked in two blocks of equal size $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ such that $\boldsymbol{D}_n = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \end{bmatrix}$. The design constructed by doubling is then

$$\boldsymbol{D}_{2n} = \left[\begin{array}{cc} \boldsymbol{B}_1 & \boldsymbol{B}_1 \\ \boldsymbol{B}_2 & \boldsymbol{B}_2 \\ \hline \boldsymbol{B}_1 & -\boldsymbol{B}_1 \\ \boldsymbol{B}_2 & -\boldsymbol{B}_2 \end{array}\right] \tag{5}$$

and has $2k$ factors. As it is written, it may look like it gives a way to block $\boldsymbol{D}_{2n}$ in two blocks, but two-factor interactions will be fully confounded with the block defining contrast no matter what the resolution is, and hence it will become a $P = 1$ design when blocked. However, both the configurations

$$\boldsymbol{D}^{*}_{2n} = \left[\begin{array}{cc} \boldsymbol{B}_1 & \boldsymbol{B}_1 \\ \boldsymbol{B}_1 & -\boldsymbol{B}_1 \\ \hline \boldsymbol{B}_2 & \boldsymbol{B}_2 \\ \boldsymbol{B}_2 & -\boldsymbol{B}_2 \end{array}\right] \tag{6}$$

and

$$\boldsymbol{D}^{**}_{2n} = \left[\begin{array}{cc} \boldsymbol{B}_1 & \boldsymbol{B}_1 \\ \boldsymbol{B}_2 & -\boldsymbol{B}_2 \\ \hline \boldsymbol{B}_1 & -\boldsymbol{B}_1 \\ \boldsymbol{B}_2 & \boldsymbol{B}_2 \end{array}\right] \tag{7}$$

represent valid ways of blocking a design with $2k$ factors in $2n$ runs into two blocks without having two-factor interactions fully confounded with the block effect.

The idea can be directly extended to four blocks and beyond. Suppose the regular design $\boldsymbol{D}_n$ is arranged into four blocks such that $\boldsymbol{D}_n = \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \boldsymbol{B}_3 \\ \boldsymbol{B}_4 \end{bmatrix}$. Then a possible blocking arrangement in four blocks for a design with $2n$ runs and $2k$ factors is:

$$\boldsymbol{D}_{2n} = \left[\begin{array}{cc} \boldsymbol{B}_1 & \boldsymbol{B}_1 \\ \boldsymbol{B}_4 & -\boldsymbol{B}_4 \\ \hline \boldsymbol{B}_3 & -\boldsymbol{B}_3 \\ \boldsymbol{B}_2 & \boldsymbol{B}_2 \\ \hline \boldsymbol{B}_3 & \boldsymbol{B}_3 \\ \boldsymbol{B}_4 & \boldsymbol{B}_4 \\ \hline \boldsymbol{B}_2 & -\boldsymbol{B}_2 \\ \boldsymbol{B}_1 & -\boldsymbol{B}_1 \end{array}\right]. \tag{8}$$

There are $\frac{\binom{8}{2}\binom{6}{2}\binom{4}{2}}{4!} = 105$ ways to arrange the eight submatrices in four blocks. Nine of these, the ones obtained from creating two blocks from $\begin{bmatrix} B_i & B_i \end{bmatrix}$, $i = 1, 2, 3, 4$ and two blocks from $\begin{bmatrix} B_i & -B_i \end{bmatrix}$, $i = 1, 2, 3, 4$, will have two-factor interactions fully confounded with block effects. The 96 remaining all represent ways of arranging a design in four blocks without having two-factor interactions fully confounded with block effects. This may be a very effective way to obtain good blocking alternatives. If $\boldsymbol{D}_n$ has been blocked by allocating mirror image pair runs to the same block, the blocks in $\boldsymbol{D}_{2n}$ will also consist of mirror image pair runs and the number of blocks will be a subset of all blocks obtained by using strategy S1. For in that case, arranging $\boldsymbol{D}_{2n}$ in two blocks using the blocking of $\boldsymbol{D}_n$, it follows from (2) that the minimum value of $D_{s,eff}$ for $\boldsymbol{D}_{2n}$ will be the same as for $\boldsymbol{D}_n$ considering projections onto the same dimension. Therefore, the blocks used for blocking $\boldsymbol{D}_{2n}$ should be the best ones obtained from blocking $\boldsymbol{D}_n$. The folding technique used in the blocking schemes is similar to the one suggested in Tyssedal and Samset (2010) for constructing supersaturated designs with good projection properties. For some subsets of factors this technique will reduce the partial confounding between effects and the block defining factor(s), and the average $D_{s,eff}$ for $\boldsymbol{D}_{2n}$ is expected to be higher than the one for $\boldsymbol{D}_n$. This is also observed for the cases where this method has been used.

Strategy S3*: Use of Hadamard matrices*

Given that our design columns are included in a Hadamard matrix, the additional columns are all potential columns for arranging the design in blocks. There are for instance five different Hadamard matrices for $n = 16$ giving rise to five different 16 run designs. Four of those contain the 2^{8-4}_{IV} design and three contain the 2^{5-1}_{V} design (Box and Tyssedal, 2001). The use of Hadamard matrices for arranging designs in blocks is particularly useful for blocking non-regular designs, and in our case when there are words of odd length in the defining relation. The method may be very effective as will be demonstrated. However, a combinatorial explosion of the number of Hadamard matrices up to equivalence occurs for $n = 32$ (Kharagani and Tayfeh-Rezaie, 2012). The number is then 13710027. For $n > 32$, the number of Hadamard matrices is to our knowledge not known. Hence checking all blocking possibilities becomes infeasible when n grows.

Strategy S4*: Arranging blocks in new blocks*

For some designs already blocked in two blocks, one may arrange each block in two new blocks to obtain a blocking scheme for four blocks. This may particularly be useful if the design has been obtained by doubling. For instance, the 128 run 2^{64-57}_{IV} design can be constructed from the 2^{32-26}_{IV} design by doubling and arranged in two blocks using strategy S2. If each of these blocks can be arranged in two blocks, we have a way to block the 2^{64-57}_{IV} design in four blocks.

The importance of the different strategies given here will change as the number of runs increases. If all words in the defining relation are of even length, strategy S1 may be the preferred choice for the number of runs equal to 16 and 32. For designs with more runs, S2 and S4 become more important, if they can be applied, providing us with a more feasible number of candidate-blocks to investigate. S2 may also be used when not all words in the defining relation are of even length, depending on how the design is constructed. Otherwise, in that case, S3 is our suggested strategy. Even though there is a combinatorial explosion in the number of possible Hadamard designs when $n = 32$, it is our experience that not that many Hadamard matrices need to be examined to obtain blocking alternatives that function well.

5. Blocking regular two-level designs with 16, 32 and 64 runs

5.1. 16 run designs

The 2^{8-4}_{IV} design in Table 1 is a $P = 3$ design and out of the 6435 possible ways to block it in two blocks, 6028 will preserve the projectivity while 407 will not. However, its defining relation has only words of length four and eight and thus it consists of eight mirror image pair runs, making strategy S1 suitable for blocking, leaving us with only 35 possible alternatives to check. As commented in Section 3, the recommended strategy is to use a two-factor interaction column, which will also place mirror image pair runs in the same block. Now, all two-factor interactions are aliased in strings of four, such that all the 28 two-factor interactions can be arranged in 7 sets where all two-factor interactions within one set are fully aliased. Thereby, seven of the 35 possibilities have as a result that three other two-factor interactions are fully aliased with the block effect. All the other 28 alternatives give the same minimum, maximum and average values for $D_{s,eff}$, as given in Table 5, when estimating all main effects and interactions for any three factors and thus provide us with (16,8,3,2) screens. It turns out that these are also the 28 with the highest minimum and average $D_{s,eff}$ among all the 6435 possibilities. For all of them, 48 out of the 56 projections onto three factors will have a $D_{s,eff} = 0.917$ and eight will have a $D_{s,eff} = 1$ when divided into two blocks. The alternative block defining contrast in Table 1 is generated as $B^*_b = \frac{1}{2}$ [AD + BD + CD-DE]. Hence there is only partial confounding with two-factor interactions. Using B^*_b as block generator, the 8 projections with a $D_{s,eff} = 1$ are {ABC}, {ABE}, {ACE}, {BCE}, {DFG}, {DFH}, {DGH} and {FGH}.

The 2^{5-1}_{V} design with defining relation I = ABCDE is a $P = 4$ screen, but it is also a $P = 5_2$ screen. Obviously, none of these properties are possible to preserve when it is blocked. The recommended way of arranging the design in two blocks is to use a two-factor interaction column as block generator. Thereby the blocked design becomes a projectivity $P = 1$ design, a rather dramatic loss in projection properties. The resolution of this design is five and thus it is not built up from mirror image pairs. All the 6435 possible ways to run the 2^{5-1}_{V} design in two blocks were checked out. It turned

Table 3
The 2_V^{5-1} design with two alternative ways of blocking.

Run	A	B	C	D	E	B_b	B_b^*
1	−1	−1	−1	−1	1	1	−1
2	1	−1	−1	−1	−1	−1	−1
3	−1	1	−1	−1	−1	−1	1
4	1	1	−1	−1	1	1	−1
5	−1	−1	1	−1	−1	1	1
6	1	−1	1	−1	1	−1	1
7	−1	1	1	−1	1	−1	1
8	1	1	1	−1	−1	1	−1
9	−1	−1	−1	1	−1	1	1
10	1	−1	−1	1	1	−1	1
11	−1	1	−1	1	1	−1	−1
12	1	1	−1	1	−1	1	1
13	−1	−1	1	1	1	1	−1
14	1	−1	1	1	−1	−1	−1
15	−1	1	1	1	−1	−1	−1
16	1	1	1	1	1	1	1

out that for 60 of those the block generator is only partially confounded with two-factor interactions. For projections onto three factors, all of these block generators perform equally well, providing us with (16,5,3,2) screens. Their $D_{s,eff}$ values are given in Table 5. The design is given in Table 3 with a recommended and a suggested block generator denoted B_b and B_b^* respectively.

The recommended block generator is $B_b = \text{AB}$ and the suggested one is $B_b^* = \frac{1}{2}[\text{AD} + \text{AE} + \text{CE-CD}]$. When projected onto four factors, it is possible to estimate all main effects and two-factor interactions with a $D_{s,eff} = 0.939$ for four out of five projections, using B_b^* as the block defining contrast. With B_b as block generator this is possible for only two of the five projections. For the fifth projection, the one that consists of the factors A, C, D, and E, one is only guaranteed to estimate three two-factor interactions. However, excluding one of the four two-factor interactions AD, AE, CD and CE from the model, one may estimate five of the six two-factor interactions for this projection with a $D_{s,eff} = 0.871$. In Table 5 this design is denoted a $(16, 5, 4_{1+3}, 2)$ screen for projections onto four factors.

5.2. 32 run designs

In Table 4 the 2_{IV}^{16-11} design is given, with generators for the 11 factor columns F-Q.

It has only words of even length in its defining relation and hence allow the use of strategy S1. But it can also be constructed by doubling the 2_{IV}^{8-4} design, and thereby strategy S2 can also be applied for arranging in two blocks. Using strategy S1, there are 6435 possible blocks to consider, but leaving out those that correspond to using a two-factor interaction or a four-factor interaction as block generator, there are only 6420 left. All of these make the blocked design a (32, 16, 3, 2) design, but 1380 of these have a minimum $D_{s,eff} = 0.88$. The remaining 5040 blocks all gave the same distribution of $D_{s,eff}$, with minimum, maximum and average values given in Table 5. One of these blocking alternatives is given by B_2^* in Table A.1 in Appendix. It can be generated as follows:

$$B_2^* = 0.25\text{AC} - 0.25\text{AD} + 0.25\text{AE} - 0.25\text{BC} + 0.5\text{CE} + 0.5\text{DE} - 0.25\text{ACDE} - 0.25\text{BCDE}.$$

Based on the 28 equally good ways of arranging the 2_{IV}^{8-4} design in two blocks, it is possible to exploit strategy S2 to block the 2_{IV}^{16-11} design. Using both $\boldsymbol{D}_{2n}^*$ and the $\boldsymbol{D}_{2n}^{**}$, there are only 56 blocks to check, and they are all equally good according to $D_{s,eff}$. Their minimum, maximum and average values for $D_{s,eff}$ are also given in Table 5. We notice that the average $D_{s,eff}$ is only slightly smaller, despite that the number of investigated alternatives being less than 1% of those with strategy S1.

There are 2,627,625 possible ways to block a design with 32 runs in four blocks using strategy S1 and 2,098,336 of these will give us a (32, 16, 3, 4) screen. Among these, 715,680 have the highest minimum $D_{s,eff} = 0.834$ and about equal average. The one given in Table A.1 in Appendix is one out of 50400 of these that also has a maximum $D_{s,eff} = 1$. The two block defining contrasts are given by:

$$B_{41}^* = -0.5\text{BD} - 0.5\text{CD} - 0.5\text{DE} + 0.5\text{BCDE} \quad \text{and}$$
$$B_{42}^* = -0.25\text{AD} - 0.5\text{AE} + 0.25\text{BD} - 0.5\text{BE} - 0.25\text{CD} - 0.25\text{DE} + 0.25\text{ABCD} + 0.25\text{ABDE} + 0.25\text{ACDE} - 0.25\text{BCDE}$$

Due to their reasonably good projection properties and since there exist recommended blocking schemes for these designs, the 2_{VI}^{6-1}, 2_{IV}^{7-2}, 2_{IV}^{8-3} and the 2_{IV}^{9-4} designs have also been explored for blocking alternatives that may preserve projections properties. Using recommended blocking schemes, not even the 2_{VI}^{6-1} design gives a $P = 3$ screen when arranged in two blocks.

The 2_{VI}^{6-1} design with generator F = ABCDE is a $P = 5$ design, which is clearly not obtainable when it is blocked. It can be constructed from the 2_{IV}^{16-11} design by removing all factors assigned to three-factor interaction columns. Blocked

Table 4
The 2_{IV}^{16-11} design with generators.

					ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE	ABCDE
A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
1	−1	−1	−1	−1	1	1	1	1	1	1	−1	−1	−1	−1	1
−1	1	−1	−1	−1	1	1	1	−1	−1	−1	1	1	1	−1	1
1	1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	−1	−1
−1	−1	1	−1	−1	1	−1	−1	1	1	−1	1	1	−1	1	1
1	−1	1	−1	−1	−1	1	1	−1	−1	1	1	1	−1	1	−1
−1	1	1	−1	−1	−1	1	1	1	1	−1	−1	−1	1	1	−1
1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	−1	1	1	1
−1	−1	−1	1	−1	−1	−1	−1	1	−1	1	1	−1	1	1	1
1	−1	−1	1	−1	1	1	1	−1	1	−1	1	−1	1	1	−1
−1	1	−1	1	−1	1	1	1	1	−1	1	−1	1	−1	1	−1
1	1	−1	1	−1	−1	−1	−1	−1	1	−1	−1	1	−1	1	1
−1	−1	1	1	−1	−1	1	1	−1	1	−1	−1	1	1	−1	1
1	−1	1	1	−1	−1	1	1	1	−1	1	−1	1	1	−1	−1
−1	1	1	1	−1	−1	1	1	−1	1	−1	1	−1	−1	−1	1
1	1	1	1	−1	1	−1	−1	1	−1	1	1	−1	−1	−1	−1
−1	−1	−1	−1	1	−1	1	1	−1	1	1	−1	1	1	1	1
1	−1	−1	−1	1	1	−1	−1	1	−1	−1	−1	1	1	1	−1
−1	1	−1	−1	1	1	−1	−1	−1	1	1	1	−1	−1	1	−1
1	1	−1	−1	1	−1	1	1	1	−1	−1	1	−1	−1	1	1
−1	−1	1	−1	1	1	1	1	1	−1	1	1	−1	1	−1	−1
1	−1	1	−1	1	−1	−1	−1	−1	1	−1	1	−1	1	−1	1
−1	1	1	−1	1	−1	−1	−1	1	−1	1	−1	1	−1	−1	1
1	1	1	−1	1	1	1	1	−1	1	−1	−1	1	−1	−1	−1
−1	−1	−1	1	1	−1	1	1	−1	−1	−1	1	1	−1	−1	−1
1	−1	−1	1	1	1	−1	−1	−1	−1	1	1	1	−1	−1	1
−1	1	−1	1	1	1	−1	−1	1	1	−1	−1	−1	1	−1	1
1	1	−1	1	1	−1	1	1	−1	−1	1	−1	−1	1	−1	−1
−1	−1	1	1	1	1	1	1	−1	−1	−1	−1	−1	−1	1	1
1	−1	1	1	1	−1	−1	−1	1	1	1	−1	−1	−1	1	−1
−1	1	1	1	1	−1	−1	−1	−1	−1	−1	1	1	1	1	−1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

the recommended way in two blocks, it becomes a $P = 2$ design and even a $P = 1$ design when arranged in four. The word in its defining relation is of even length, and strategy S1 can be used. The 5040 blocking alternatives found to have the lowest confounding with two-factor interactions for the 2_{IV}^{16-11} design provided us with both $(32, 6, 3, 2)$, $(32, 6, 4, 2)$ and $(32, 6, 5_3, 2)$ screens, and as $(32, 6, 4, 2)$ screens they are all equally good. Some preferences of blocking can be made for the $(32, 6, 5_3, 2)$ screen if also four factor interactions are of concern. Details can be found in Hamre (2018). Strategy S1 was also used for arranging the 2_{VI}^{6-1} design in four blocks. Considering projections onto three, four and five factors separately, one may choose different block defining contrasts in each case. Out of the many alternatives (Hamre, 2018), we have chosen to present two block defining contrasts with reasonably good performance for five active factors or less. The block defining contrasts for arranging the design in two and four blocks are given in Table A.1 in Appendix, and they may be obtained as follows:

$$B_2^* = 0.25\text{AB} + 0.25\text{AC} + 0.25\text{BD} + 0.25\text{BE} + 0.25\text{CD} + 0.25\text{CE} + 0.5\text{DE} - 0.25\text{ABDE} - 0.25\text{ACDE} - 0.5\text{BCDE}$$

$$B_{41}^* = -0.5\text{AC} - 0.25\text{AE} - 0.5\text{BD} - 0.25\text{BE} - 0.25\text{CE} - 0.25\text{DE} + 0.25\text{ABCE} + 0.25\text{ABDE} + 0.25\text{ACDE} + 0.25\text{BCDE}$$

$$B_{42}^* = 0.25\text{AB} + 0.25\text{AC} + 0.25\text{AE} - 0.25\text{BD} - 0.25\text{CD} - 0.5\text{CE} - 0.25\text{DE} - 0.25\text{ABCE} - 0.5\text{ACDE} + 0.25\text{BCDE}$$

The corresponding $D_{s,eff}$ values are given in Table 5.

The projection properties of the minimum aberration 2_{IV}^{7-2}, 2_{IV}^{8-3} and 2_{IV}^{9-4} designs when arranged in two and four blocks the recommended ways, are given in Table 2. When blocked in two blocks, the projectivity is two for the 2_{IV}^{7-2} and the 2_{IV}^{8-3} designs, and it is further reduced to one when arranged in four. The 2_{IV}^{9-4} design is even a $(32, 8, 1, 2)$ screen. None of the designs have a defining relation consisting of only even words. Strategy S3 is therefore a possible alternative trying to obtain good blocking schemes. We have used two 32 run Hadamard matrices, had.32.t1 and had.32.t3, from the web page "A library of Hadamard matrices" (Sloane, access 2018), hereafter referred to as M_1 and M_2, respectively. For both matrices all rows with a -1 in the first column of M_1 and M_2 were multiplied with -1, and thereafter the rows were rearranged such that the principle factor columns A, B, C, D and E could be identified among the 32 columns. The additional factor columns were generated by their generators:

F = ABC and G = ABDE for the 2_{IV}^{7-2} design

F = ABC, G = ABD and H = ACDE for the 2_{IV}^{8-3} design

Table 5
Projection properties and values of $D_{s,eff}$ for the design blocked.

Design	Strategy	Screen	Min $D_{s,eff}$	Max $D_{s,eff}$	Mean $D_{s,eff}$
2_{IV}^{8-4}	S1 + all possibilities	(16, 8, 3, 2)	0.917	1	0.929
2_{V}^{5-1}	S3 + all possibilities	(16, 5, 3, 2)	0.917	1	0.934
2_{V}^{5-1}	S3 + all possibilities	$(16, 5, 4_{1+3}, 2)$	0.814	1	0.952
2_{IV}^{16-11}	S1	(32, 16, 3, 2)	0.917	1	0.970
2_{IV}^{16-11}	S2	(32, 16, 3, 2)	0.917	1	0.967
2_{IV}^{16-11}	S1	(32, 16, 3, 4)	0.834	1	0.908
2_{VI}^{6-1}	S1	(32, 6, 3, 2)	0.943	0.983	0.971
2_{VI}^{6-1}	S1	(32, 6, 4, 2)	0.917	0.982	0.959
2_{VI}^{6-1}	S1	$(32, 6, 5_3, 2)$	0.948	0.963	0.958
2_{VI}^{6-1}	S1	$(32, 6, 5_{3+2}, 2)$	0.906	0.966	0.939
2_{VI}^{6-1}	S1 + S4	(32, 6, 3, 4)	0.865	0.949	0.911
2_{VI}^{6-1}	S1	$(32, 6, 4_3, 4)$	0.863	0.919	0.893
2_{VI}^{6-1}	S1	$(32, 6, 5_3, 4)$	0.866	0.866	0.866
2_{IV}^{7-2}	S3	(32, 7, 3, 2)	0.917	1	0.982
2_{IV}^{7-2}	S3	(32, 7, 3, 4)	0.917	1	0.939
2_{IV}^{8-3}	S3	(32, 8, 3, 2)	0.917	1	0.981
2_{IV}^{8-3}	S3	(32, 8, 3, 4)	0.853	1	0.929
2_{IV}^{9-4}	S3	(32, 9, 3, 2)	0.917	1	0.982
2_{IV}^{9-4}	S3	(32, 9, 3, 4)	0.853	1	0.925
2_{IV}^{32-26}	S2	(64, 32, 3, 2)	0.917	1	0.987
2_{IV}^{32-26}	S2	(64, 32, 3, 4)	0.913	1	0.987
2_{V}^{8-2}	S3	(64, 8, 3, 2)	0.965	1	0.992
2_{V}^{8-2}	S3	(64, 8, 4, 2)	0.958	1	0.986
2_{V}^{8-2}	S3	(64, 8, 3, 4)	0.931	1	0.982
2_{V}^{8-2}	S3	(64, 8, 4, 4)	0.917	1	0.966
2_{V}^{8-2}	S3	(64, 8, 3, 8)	0.834	1	0.918
2_{V}^{8-2}	S3	(64, 8, 4, 8)	0.808	0.917	0.887

F = ABC, G = ABD, H = ACD and J = BCDE for the 2_{IV}^{9-4} design.

The remaining 26 columns in the rearranged Hadamard matrices were then tested as potential block generators. For all three designs, M_1 yielded good results for arranging in two blocks, while M_2 provided good results for arranging in four. Block generators orthogonal to all main effects contrasts were found, and projection properties were preserved. Several equally good alternatives exist, Hamre (2018). One for each design and each blocking scheme is given in Table A.2 in Appendix. $D_{s,eff}$ values are given in Table 5.

5.3. 64 run designs

The 2_{IV}^{32-26} design can be constructed by doubling the 2_{IV}^{16-11} design given in Table 3. Hence arranging the 2_{IV}^{32-26} design in two and four blocks can be done using strategy S2, exploiting the preferred ways of blocking the 2_{IV}^{16-11} design. All the 5040 blocks that turned out equally good using strategy S1 for the 2_{IV}^{16-11} design, were tested for both configurations $\boldsymbol{D}_{2n}^{*}$ and $\boldsymbol{D}_{2n}^{**}$. With the same configuration, all the tested blocks gave the same distributions of $D_{s,eff}$ with $\boldsymbol{D}_{2n}^{**} = \left[\begin{array}{cc} \boldsymbol{B}_1 & \boldsymbol{B}_1 \\ \boldsymbol{B}_2 & -\boldsymbol{B}_2 \\ \hline \boldsymbol{B}_1 & -\boldsymbol{B}_1 \\ \boldsymbol{B}_2 & \boldsymbol{B}_2 \end{array}\right]$ being slightly better. Two possible alternatives for $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are given in Table A.1 in Appendix. $D_{s,eff}$ values are given in Table 5.

For arranging the 2_{IV}^{16-11} design in four blocks there are 40320 equally good alternatives as measured by D_s-efficiency. All of these could then be tested for all the 96 possible configurations and for all 4960 possible projections onto three dimensions. We chose to only use 10 of these blocks.

For all these only the configuration mattered and four were better than the others. One of these is given below with blocks $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{B}_3$ and $\boldsymbol{B}_4$ taken from Table A.1. $D_{s,eff}$-values are given in Table 5.

$$\left[\begin{array}{cc} \boldsymbol{B}_1 & \boldsymbol{B}_1 \\ \hline \boldsymbol{B}_4 & \boldsymbol{B}_4 \\ \boldsymbol{B}_4 & -\boldsymbol{B}_4 \\ \hline \boldsymbol{B}_3 & \boldsymbol{B}_3 \\ \boldsymbol{B}_3 & -\boldsymbol{B}_3 \\ \hline \boldsymbol{B}_2 & -\boldsymbol{B}_2 \\ \boldsymbol{B}_2 & \boldsymbol{B}_2 \\ \hline \boldsymbol{B}_1 & -\boldsymbol{B}_1 \end{array}\right].$$

The 2_V^{8-2} design with generators G = ABCD and H = ABEF is of odd resolution. Strategy S3 was used to block the design in 2, 4 and 8 blocks. The 64 run Hadamard matrices used to obtain orthogonal main effects blocking were the ones obtained by doubling the 32 run Hadamard matrices M1, M2 and had32.t3 (Sloane, access 2018) after rearranging. All of them contained the 6 principal factor columns A, B, C, D, E and F. The 57 remaining columns were all tested as block alternatives, assuring that main effects were orthogonal to the blocks. Several possibilities turned out to perform equally well, see Hamre (2018). Table A.3 in Appendix provide ways to block the design in 2, 4 and 8 blocks. The $D_{s,eff}$ values are given in Table 5.

Table 5 summarizes the projection properties and the minimum, maximum and average value of the $D_{s,eff}$ criterion for the designs that are investigated. We observe that in many cases the projection properties are preserved when these designs are blocked. In others the projection properties obtained are much better than when the recommended way of blocking is used. The price one has to pay is a slight decrease in $D_{s,eff}$.

The decrease in $D_{s,eff}$ is related to an increase in the standard deviations of the effect estimators. For a given projection onto P factors, the covariance matrix for all coefficient estimators (block effect(s) included) is given by $\sigma^2\left[\mathbf{X}^T\mathbf{X}\right]^{-1}$. If $\hat{\beta}_i$ is the effect coefficient estimator with the largest variance given by $\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)_{ii}^{-1}$ and $\hat{\beta}_j$ the one with the smallest variance given by $\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)_{jj}^{-1}$, the ratio of their standard deviations is $SD_e\text{-}ratio = \sqrt{\frac{(\mathbf{X}^T\mathbf{X})_{ii}^{-1}}{(\mathbf{X}^T\mathbf{X})_{jj}^{-1}}}$. Similarly, if $\hat{\beta}_b$ is the block defining contrast with the largest variance, one may define $SD_b\text{-}ratio = \sqrt{\frac{(\mathbf{X}^T\mathbf{X})_{bb}^{-1}}{(\mathbf{X}^T\mathbf{X})_{jj}^{-1}}}$. In Table 6 these ratios are computed for each design and for two projections, one giving the smallest $D_{s,eff}$ and one giving the largest. Since main effects columns are orthogonal to the block defining contrasts, their corresponding estimators always attain the smallest variance. These ratios therefore give the increase in standard deviations due to some interactions being partially confounded with the block defining contrast(s). As observed, this increase is normally larger for projections having the smallest $D_{s,eff}$ and then often in the range 20%–40%. As intended, the increase in standard deviations is smaller for the factor effects than for the block effects in most cases.

6. Discussion and comparison of various ways of evaluating blocked regular designs

Most traditional blocking schemes for regular design are evaluated and ranked from the two wordlength patterns W and W_b. Common for all the blocking schemes is that the designs are at least of resolution III, and that main effects are unconfounded with block effects. Assuming that three-factor and higher order interactions are negligible, a rather common assumption when blocking schemes are evaluated, the relevant terms to consider are A_3, A_4 and A_{2b}.

Aliased effects can be partitioned into alias sets. For instance, from the word ABC three alias sets can be constructed, A and BC, B and AC and C and AB. Let n_{mb} be the number of two-factor interactions that are either aliased with main effects or confounded with blocks and let $M_1, \ldots, M_f$ be the alias sets that do not contain such effects. With k main effects the number of two-factor interactions in these alias sets sums to $\binom{k}{2} - n_{mb}$. Distributing these two-factor interactions as uniformly as possible over $M_1, \ldots, M_f$ is related to minimizing A_4 (Cheng et al., 1999). A small A_4 is also necessary in order to have many clear two-factor interactions. Minimization of $3A_3 + A_{2b}$ amounts to minimizing n_{mb}.

Sequentially minimizing the sequence $3A_3, A_4, A_{2b}$ (Cheng and Wu, 2002) might then support what we have denoted the recommended way of blocking, emphasizing many clear two-factor interactions, while the same sequential operation on $3A_3 + A_{2b}, A_4$ fits well into the way of blocking two-level designs proposed in Cheng and Mukerjee (2001), and mentioned in the introduction. It is interesting to note that these sequential operations lead to two different blocking schemes for a 2^{5-1} design arranged in two blocks. The recommended one, which is a 2_V^{5-1} design with defining relation I=ABCDE and the two-factor interaction AB used as block factor, and one which is a 2_{IV}^{5-1} design with defining relation

Table 6
Standard deviation ratios for projections with smallest and largest $D_{s,eff}$.

Design	Screen	Min $D_{s,eff}$	$SD_e^{\min}$-ratio	$SD_b^{\min}$-ratio	Max $D_{s,eff}$	$SD_e^{\max}$-ratio	$SD_b^{\max}$-ratio
2_{IV}^{8-4}	(16, 8, 3, 2)	0.917	1.2247	1.4142	1	1	1
2_{V}^{5-1}	(16, 5, 3, 2)	0.917	1.2247	1.4142	1	1	1
2_{V}^{5-1}	$(16, 5, 4_{1+3}, 2)$	0.814	1.4142	2	1	1	1
2_{IV}^{16-11}	(32, 16, 3, 2)	0.917	1.2247	1.4142	1	1	1
2_{IV}^{16-11}	(32, 16, 3, 2)	0.917	1.2247	1.4142	1	1	1
2_{IV}^{16-11}	(32, 16, 3, 4)	0.834	1.3904	1.5275	1	1	1
2_{VI}^{6-1}	(32, 6, 3, 2)	0.943	1.128	1.206	0.983	1.035	1.069
2_{VI}^{6-1}	(32, 6, 4, 2)	0.917	1.4142	2	0.982	1.0408	1.1547
2_{VI}^{6-1}	$(32, 6, 5_3, 2)$	0.948	1.4142	2	0.963	1.2910	1.6330
2_{VI}^{6-1}	$(32, 6, 5_{3+2}, 2)$	0.906	2.2361	4	0.966	1.2910	1.6330
2_{VI}^{6-1}	(32, 6, 3, 4)	0.865	1.318	1.318	0.949	1.0742	1.0742
2_{VI}^{6-1}	$(32, 6, 4_3, 4)$	0.863	1.512	1.604	0.919	1.291	1.4142
2_{VI}^{6-1}	$(32, 6, 5_3, 4)$	0.866	1.8257	2.4495	0.866	1.8257	2.4495
2_{IV}^{7-2}	(32, 7, 3, 2)	0.917	1.2247	1,4142	1	1	1
2_{IV}^{7-2}	(32, 7, 3, 4)	0.917	1.2247	1,4142	1	1	1
2_{IV}^{8-3}	(32, 8, 3, 2)	0.917	1.2247	1,4142	1	1	1
2_{IV}^{8-3}	(32, 8, 3, 4)	0.853	1.2247	1,4142	1	1	1
2_{IV}^{9-4}	(32, 9, 3, 2)	0.917	1,2247	1,4142	1	1	1
2_{IV}^{9-4}	(32, 9, 3, 4)	0.853	1.2247	1,4142	1	1	1
2_{IV}^{32-26}	(64, 32, 3, 2)	0.917	1.2247	1,4142	1	1	1
2_{IV}^{32-26}	(64, 32, 3, 4)	0.913	1.1704	1.2163	1	1	1
2_{V}^{8-2}	(64, 8, 3, 2)	0.965	1.1547	1.1547	1	1	1
2_{V}^{8-2}	(64, 8, 4, 2)	0.958	1.2247	1.4142	1	1	1
2_{V}^{8-2}	(64, 8, 3, 4)	0.931	1.1547	1.1547	1	1	1
2_{V}^{8-2}	(64, 8, 4, 4)	0.917	1.2247	1.4142	1	1	1
2_{V}^{8-2}	(64, 8, 3, 8)	0.834	1.5492	1.5492	1	1	1
2_{V}^{8-2}	(64, 8, 4, 8)	0.808	1.5492	1.5492	0.917	1,2247	1.4142

Table 7
Estimation capacity sequence for three blocking schemes for 2^{5-1} designs.

Design	$E_1(d)$	$E_2(d)$	$E_3(d)$	$E_4(d)$	$E_5(d)$	$E_6(d)$	$E_7(d)$	$E_8(d)$	$E_9(d)$
$2_{IV}^{5-1,ABD}$	10	42	96	129	102	44	8	0	0
$2_{V}^{5-1,AB}$	9	36	84	126	126	84	36	9	1
$2_{V}^{5-1,B_b^*}$	10	45	120	209	246	195	100	30	4

I=ABCE with ABD used as block factor. The first scheme has 9 clear two-factor interactions and projectivity $P = 1$. The second scheme has only 4 clear two-factor interactions but projectivity $P = 3_2$.

It is not obvious how blocking schemes with only partial confounding can be fairly evaluated with criteria constructed from W and W_b. Cheng and Mukerjee (2001) pointed out that their criterion could be tied to maximum estimation capacity, see also Chen and Cheng (1999) and Cheng et al. (1999). For a blocked two-level design d with k factors, let $E_u(d)$ be the number of models with k main effects and u two-factor interactions that can be estimated by d. Then an estimation capacity sequence $(E_1(d), E_2(d), \ldots)$ can be constructed and used for comparison of blocked designs. For two designs d_1 and d_2, if $E_u(d_1) \geq E_u(d_2)$ for all u with strict inequality for some, then d_1 dominates d_2. Let the two blocking schemes for the 2^{5-1} designs be denoted $2_V^{5-1,AB}$ and $2_{IV}^{5-1,ABD}$ respectively. A comparison of the estimation capacity sequence for these two schemes and our blocking scheme, denoted $2_V^{5-1,B_b^*}$ is given in Table 7.

The $2_{IV}^{5-1,ABD}$ dominates the $2_V^{5-1,AB}$ with respect to estimation capacity for $u \leq 4$, but for $u \geq 5$, it is opposite. However, both these schemes are clearly dominated by the $2_V^{5-1,B_b^*}$. The reason why is easily explainable. Let us just compare the $2_V^{5-1,AB}$ and the $2_V^{5-1,B_b^*}$. There are enough degrees of freedom for estimating nine two-factor interactions together with five

main effects, the intercept and the block effect. For the $2_V^{5-1,AB}$ we then get $E_u(d) = \binom{9}{u}$, $1 \le u \le 9$. B_b^* is constructed from four two-factor interaction columns and partially confounded with the corresponding effects, which means that only three of them can be estimated at the same time. Leaving out one, say AD, we have a set of nine two-factor interactions to freely choose from. For given $1 \le u \le 9$ this gives $\binom{9}{u}$ different models that can be estimated. In addition, we also need to count the ones where AD is forced to be in the model.

Words of length four in the defining relation give alias sets consisting of more than one two-factor interaction. Let us assume there are a such alias sets and let A_c be the collection of these. Further let c be the number of clear two-factor interactions. When possible and needed, the recommended way of blocking uses one or more two-factor interaction column(s) corresponding to effects in A_c as block defining contrast(s) to have as many clear two-factor interactions as possible. To get a better understanding of how this affects the estimation capacity sequence, let us look at the 2_{IV}^{7-2} design run in four blocks. The design has $a = 3$ with two members in each and $c = 15$. Blocked the recommended way, ACD and BCD are used as block factors. These are chosen such that their interaction effect is a two-factor interaction in one of the alias sets in A_c, leaving only $a-1 = 2$ alias sets to consider for estimation capacity. For a given u, let $n_{(u|y,z)}$ be the number of models that can be estimated using y alias sets in A_c and z of the clear two-factor interactions. Further define $n_{(0|\cdot,\cdot)} = 1$. Then to obtain $n_{(u|2,15)}$we may count the ones when one of the clear two-factor interactions is taken out, and then add the ones we get when this interaction is forced to be in the model. Hence, we have $n_{(u|2,15)} = n_{(u|2,14)} + n_{(u-1|2,14)}$. If instead the two block factors had been chosen such that their interaction effect was one of the 15 clear two-factor interactions as in Xu and Lau (2006), the number of models that can be estimated had become $n_{(u|3,14)} = n_{(u|2,14)} + 2n_{(u-1|2,14)} > n_{(u|2,15)}$. Hence the recommended way of blocking, giving priority to having as many clear two-factor interactions as possible, does not necessarily give a blocking scheme with the best estimation capacity.

Blocking the 2_{IV}^{7-2} design in four blocks the way we propose, one of the block factors is constructed from four clear two-factor interaction columns. The two others can be expressed as a linear combination of two-factor interaction and higher order interaction columns, a situation not uncommon when strategy S3 is used. Leaving out one clear two-factor interaction, say AD, we get that the number of estimated models for a given u equals $n_{(u|3,14)}$ + the ones we get when AD is forced to be in the models. Hence it dominates both the schemes given above.

We have checked all our proposed blocking schemes. The one for the 2_{VI}^{6-1} design stands out as an exception. Blocking this design in two and four blocks the recommended way make use of three-factor interaction columns. Arranged in four blocks one of the two-factor interactions will be fully confounded with a block effect. We have used strategy S1 which for this design, as explained in Section 4, will cause block factor effects to be partially confounded with two-factor interaction, and for our scheme 10 such for each block factor. As a result, the recommended way of blocking dominates our schemes with respect to estimation capacity for $u \ge 10$ when the design is run in two blocks. Arranged in four blocks, our way of blocking will as a maximum allow the estimation of 12 two-factor interactions together with 6 main effects, while the recommended way will allow up to 14. Another strategy, for instance S3, might have given blocking schemes with better estimation capacities. In all other cases our schemes were equally good or better than the recommended ones, even though estimation capacity was not a criterion for choosing scheme and strategy.

From what we have seen, different criteria may lead to different blocking schemes with different projection properties. Maximizing the number of clear two-factor interactions may come in conflict with maximizing the estimation capacity. Our blocking schemes were constructed with the purpose of being able to estimate as many effects as possible for a subset of factors up to a certain size, which differs from the motivation behind the recommended way of blocking and the method suggested by Cheng and Mukerjee (2001). It is therefore encouraging that our schemes in most cases seem to do that well with respect to estimation capacity and are comparable with the recommended way of blocking when it comes to the number of two-factor interactions that can be estimated in a model.

The partial confounding exploited in our blocking schemes impacts the efficiency with which the effects can be estimated for some models. When blocked in two blocks the recommended way, the 2_{VI}^{6-1} design is a $P = 6_2$ design, and the 2_{IV}^{7-2} and 2_{IV}^{8-3} designs both have projectivity $P = 3_2$. The 2_V^{8-2} design is a $P = 8_2$ design when blocked in two and four blocks. Hence, if the respective number of main effects and two-factor interactions are considered adequate for modelling the response, the recommended way of blocking represents a good alternative in these cases, having a $D_{s,eff} = 1$ for the effects of interest. However, if factor sparsity is a reasonable assumption, research shows that the better projection properties a design has, the more robust the screening will be with respect to assumptions about the model, amount of noise and distortion of effects too small to be detected (Tyssedal and Chaudhry, 2017; Chaudhry and Tyssedal, 2019; Chaudhry, 2019).

7. Concluding remarks

We have demonstrated that many regular two-level designs with 16, 32 and 64 runs can be blocked such that their projection properties are preserved or only weakly affected. Different strategies have been used depending on how the designs are constructed. These include letting mirror image pair runs be in the same block, exploiting that blocking of a n run design can be utilized to block a design in $2n$ runs and taking advantage of the rich selection of different Hadamard

Table A.1
Possible ways to arrange the 2_{IV}^{16-11} and the 2_{VI}^{6-1} designs in two and four blocks in order to have good projection properties.

2_{IV}^{16-11}					2_{VI}^{6-1}		
B_2^*	Blocks	B_{41}^*	B_{42}^*	Blocks	B_2^*	B_{41}^*	B_{42}^*
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	−1
1	$\boldsymbol{B}_1$	−1	1	$\boldsymbol{B}_3$	1	1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	1
1	$\boldsymbol{B}_1$	−1	1	$\boldsymbol{B}_3$	1	−1	−1
−1	$\boldsymbol{B}_2$	1	−1	$\boldsymbol{B}_2$	−1	1	1
−1	$\boldsymbol{B}_2$	1	1	$\boldsymbol{B}_1$	−1	1	1
1	$\boldsymbol{B}_1$	1	−1	$\boldsymbol{B}_2$	1	−1	1
−1	$\boldsymbol{B}_2$	1	1	$\boldsymbol{B}_1$	−1	1	−1
1	$\boldsymbol{B}_1$	1	1	$\boldsymbol{B}_1$	−1	−1	1
−1	$\boldsymbol{B}_2$	1	1	$\boldsymbol{B}_1$	−1	1	−1
−1	$\boldsymbol{B}_2$	1	−1	$\boldsymbol{B}_2$	−1	1	1
−1	$\boldsymbol{B}_2$	1	−1	$\boldsymbol{B}_2$	−1	1	1
−1	$\boldsymbol{B}_2$	−1	1	$\boldsymbol{B}_3$	−1	1	−1
−1	$\boldsymbol{B}_2$	−1	1	$\boldsymbol{B}_3$	1	−1	1
−1	$\boldsymbol{B}_2$	−1	1	$\boldsymbol{B}_3$	1	−1	1
−1	$\boldsymbol{B}_2$	−1	1	$\boldsymbol{B}_3$	−1	1	−1
−1	$\boldsymbol{B}_2$	1	−1	$\boldsymbol{B}_2$	−1	1	1
−1	$\boldsymbol{B}_2$	1	−1	$\boldsymbol{B}_2$	−1	1	1
−1	$\boldsymbol{B}_2$	1	1	$\boldsymbol{B}_1$	−1	1	−1
1	$\boldsymbol{B}_1$	1	1	$\boldsymbol{B}_1$	−1	−1	1
−1	$\boldsymbol{B}_2$	1	1	$\boldsymbol{B}_1$	−1	1	−1
1	$\boldsymbol{B}_1$	1	−1	$\boldsymbol{B}_2$	1	−1	1
−1	$\boldsymbol{B}_2$	1	1	$\boldsymbol{B}_1$	−1	1	1
−1	$\boldsymbol{B}_2$	1	−1	$\boldsymbol{B}_2$	−1	1	1
1	$\boldsymbol{B}_1$	−1	1	$\boldsymbol{B}_3$	1	−1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	1
1	$\boldsymbol{B}_1$	−1	1	$\boldsymbol{B}_3$	1	1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	−1
1	$\boldsymbol{B}_1$	−1	−1	$\boldsymbol{B}_4$	1	−1	−1

Table A.2
Possible ways to run the 2_{IV}^{7-2}, 2_{IV}^{8-3} and the 2_{IV}^{9-4} designs in two and four blocks such that projection properties are preserved.

2_{IV}^{7-2}			2_{IV}^{8-3}			2_{IV}^{9-4}		
B_2^*	B_{41}^*	B_{42}^*	B_2^*	B_{41}^*	B_{42}^*	B_2^*	B_{41}^*	B_{42}^*
−1	−1	1	1	1	1	−1	−1	1
−1	1	−1	1	−1	−1	1	1	−1
−1	1	1	1	1	−1	−1	−1	1
1	1	−1	−1	1	1	−1	1	1
1	1	1	−1	1	−1	1	−1	−1
1	−1	−1	−1	−1	1	1	1	−1
−1	−1	1	1	1	1	−1	−1	−1
−1	−1	−1	−1	1	−1	1	1	1
1	1	1	1	−1	1	1	1	1
1	−1	1	1	−1	1	1	−1	−1
−1	−1	−1	1	1	1	1	1	1
−1	−1	−1	−1	−1	1	−1	−1	1
−1	−1	1	−1	−1	−1	1	1	−1
−1	1	1	−1	−1	−1	−1	−1	−1
1	1	−1	1	1	−1	−1	1	−1
1	1	−1	−1	−1	−1	−1	−1	1
1	−1	−1	−1	−1	−1	1	1	1
1	1	1	−1	1	1	−1	−1	−1
1	1	−1	−1	−1	1	1	−1	−1
1	1	1	1	−1	−1	1	1	−1
−1	1	−1	1	−1	1	−1	−1	1
−1	−1	1	1	1	−1	−1	1	1
−1	−1	−1	−1	−1	−1	1	1	−1
−1	−1	1	1	−1	1	−1	−1	1

(*continued on next page*)

Table A.2 (*continued*).

2_{IV}^{7-2}			2_{IV}^{8-3}			2_{IV}^{9-4}		
B_2^*	B_{41}^*	B_{42}^*	B_2^*	B_{41}^*	B_{42}^*	B_2^*	B_{41}^*	B_{42}^*
−1	1	−1	−1	1	−1	−1	−1	1
−1	−1	−1	−1	1	−1	−1	1	−1
−1	−1	1	−1	−1	−1	−1	1	−1
−1	−1	1	1	1	−1	1	−1	−1
1	−1	−1	1	1	1	−1	1	1
1	1	−1	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 A.3
Possible blocking generators for blocking the 2_V^{8-2} design in two, four and eight blocks.

Row	B_2^*		B_{41}^*		B_{42}^*		B_{81}^*		B_{82}^*		B_{83}^*	
1/33	−1	−1	−1	1	−1	−1	−1	−1	−1	1	−1	1
2/34	−1	1	1	1	−1	1	1	−1	1	1	1	1
3/35	−1	−1	−1	−1	−1	1	−1	−1	−1	1	1	−1
4/36	1	−1	1	−1	1	1	1	−1	−1	−1	1	1
5/37	1	1	−1	1	1	1	1	1	1	−1	−1	1
6/38	1	−1	1	1	1	−1	1	−1	−1	−1	−1	−1
7/39	1	1	−1	−1	1	1	1	1	1	−1	1	−1
8/40	−1	1	1	−1	−1	1	1	−1	1	−1	−1	−1
9/41	1	1	1	−1	−1	1	1	−1	−1	1	1	1
10/42	1	−1	−1	−1	1	1	1	1	−1	1	−1	1
11/43	1	1	1	1	1	−1	1	−1	1	1	1	1
12/44	−1	1	−1	1	1	1	1	1	−1	1	1	−1
13/45	−1	−1	1	−1	1	−1	1	−1	1	−1	−1	−1
14/46	−1	1	−1	−1	−1	−1	−1	−1	1	−1	−1	1
15/47	−1	−1	1	1	−1	1	1	−1	−1	−1	−1	−1
16/48	1	−1	−1	1	−1	−1	−1	−1	1	−1	1	−1
17/49	−1	1	1	1	1	1	−1	1	−1	−1	−1	−1
18/50	1	1	1	−1	1	−1	1	1	1	−1	1	−1
19/51	1	−1	−1	1	1	1	−1	1	1	1	−1	−1
20/52	1	1	−1	−1	−1	1	1	1	1	−1	−1	1
21/53	1	−1	1	1	−1	−1	−1	1	1	1	1	1
22/54	−1	−1	1	−1	−1	1	−1	−1	−1	1	1	−1
23/55	−1	1	−1	1	−1	−1	−1	1	−1	−1	1	1
24/56	−1	−1	−1	−1	1	−1	−1	−1	−1	1	−1	1
25/57	1	−1	−1	−1	1	−1	−1	−1	1	−1	1	−1
26/58	−1	−1	−1	1	−1	−1	−1	1	1	1	−1	−1
27/59	−1	1	1	−1	−1	1	−1	−1	1	−1	−1	1
28/60	−1	−1	1	1	−1	−1	−1	1	−1	−1	−1	−1
29/61	−1	1	−1	−1	−1	1	1	1	−1	1	1	−1
30/62	1	1	−1	1	1	1	−1	1	−1	−1	1	1
31/63	1	−1	1	−1	1	−1	1	1	−1	1	−1	1
32/64	1	1	1	1	1	1	−1	1	1	1	1	1

matrices that exist. We have restricted our blocking to cases with an equal number of runs in the same block and kept the block defining contrast(s) orthogonal to main effect contrasts. Blocking alternatives have been assessed using the $D_{s,eff}$ criterion. Following these strategies has made it possible to obtain blocking schemes with better projection properties than can be achieved using higher order interactions columns from regular designs. Unintended, our blocking schemes also seem to have good estimation capacity, sometimes better than other schemes, and are comparable with the blocking schemes given in Wu and Hamada (2009) when it comes to how many two-factor interactions that can be estimated in a model.

When the number of runs increases there are enormously many possible blocking alternatives and investigating all of them becomes almost infeasible in time. For the two 16 run designs we were able to examine all possible blocking arrangements. It was encouraging that strategy S1 led us directly to the best ones for the 2_{IV}^{8-4} design, according to our criteria, by only considering less than 1% of the alternatives. The best way of blocking may also depend on how well one wants to estimate the effects when the design is projected onto different dimensions. Each choice of dimension may prefer different block defining contrast(s). It is our experience, however, that searching strategies, such as S2, that only consider a rather small subset of all blocking candidates, come up with almost equally good blocking alternatives as when investigating more numerous ways of blocking. We believe that the reason why S1 and S2 perform that well is because alias reduction is inherent in both.

The use of Hadamard matrices, strategy S3, represent an alternative way of blocking that has a potential to be successful in preserving projection properties when blocking all regular designs, and particularly for those that does not fit into the

design structures for using S1 and S2. An exhaustive search is also here close to infeasible when n increases, due to the combinatorial explosion of Hadamard matrices that occurs for $n = 32$. Fortunately, it seems like high values for minimum and average $D_{s.eff}$ can be obtained by only searching through a few Hadamard matrices.

Preserving projection properties when blocking has a price. Some effects will be estimated with less efficiency due to partial confounding of interactions with block generators. We think that for a factor-based screening this drawback is well compensated for by the possibility to separate estimates of interactions and block effects. We also believe that the methods presented here can be useful for experimental work with other restrictions on randomization than blocking, such as split-plot experimentation. Especially full confounding between subplot interactions and whole plot interactions can be avoided assigning whole plot factors to columns here used as block defining contrasts.

Acknowledgements

The author thanks the associate editor and two anonymous referees for their insightful comments which have led to improvement of this article.

Appendix

See Tables A.1–A.3.

References

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Design. Clarendon Press, Oxford.
Box, G.E.P., Hunter, J.S., 1961. The 2^{k-p} fractional factorial designs, part I. Technometrics 3 (3), 311–352.
Box, G.E.P., Hunter, W.G., Hunter, J.S., 2005. Statistics for Experimenters, second ed. Wiley New York.
Box, G., Tyssedal, J., 1996. Projective properties of certain orthogonal arrays. Biometrika 83 (4), 950–955.
Box, G.E.P., Tyssedal, J., 2001. Sixteen run designs of high projectivity for factor screening. Commun. Stat. - Simul. Comput. 30 (2), 217–228.
Chaudhry, M.A., 2019. Robustness of Screening Designs with a Small Number of Runs (Doctoral theses at NTNU). p. 162.
Chaudhry, M.A., Tyssedal, J.S., 2019. Assessing some aspects of factor screening with non-normal responses. In: Applied Stochastic Models in Business and Industry. pp. 1–16. http://dx.doi.org/10.1002/asmb.2444.
Chen, H., Cheng, C.S., 1999. Theory of optimal blocking of 2^{n-m} designs. Ann. Statist. 27, 1948–1973.
Cheng, C.-S., Mukerjee, R., 2001. Fractional factorial designs with maximum estimation capacity. Ann. Statist. 29 (2), 530–548.
Cheng, C.-S., Steinberg, D.M., Sun, D.X., 1999. Minimum aberration and maximum estimation capacity. J. R. Stat. Soc. 61 (1), 85–93.
Cheng, C.-S., Tsai, P.W., 2009. Optimal two-level regular fractional factorial block and split-plot designs. Biometrika Trust 96 (1), 83–93.
Cheng, S.W., Wu, C.F.J., 2002. Choice of optimal blocking schemes in two-level and three-level designs. Technometrics 44 (3), 269–277.
Evangelaras, H., Koukouvinos, C., 2004. On generalized projectivity of two-level screening designs. Statist. Probab. Lett. 68 (4), 429–434.
Fries, A., Hunter, W.G., 1980. Minimum aberration 2^{k-p} designs. Technometrics 22 (4), 601–608.
Hamre, Y.H., 2018. Preserving Projection Properties when Regular Two-Level Designs are Blocked (Master's thesis). Department of Mathematical Sciences, NTNU.
Hussain, S., Tyssedal, J., 2016. Projection properties of blocked non-regular two-level designs. Qual. Reliab. Eng. Int. 32 (8), 3011–3021.
Kharagani, H., Tayfeh-Rezaie, B., 2012. Hadamard matrices of order 32. J. Comb. Des. 21 (5), 212–221.
Montgomery, D.C., 2019. Design and Analysis of Experiments, ninth ed. Wiley.
Mukerjee, R., Wu, C.F.J., 1999. Blocking in regular fractional factorials: A projective geometric approach. Ann. Statist. 27 (4), 1256–1271.
Mukerjee, R., Wu, C.F.J., 2006. A Modern Theory of Factorial Designs. Springer, New York.
Sitter, J., M, Feder., 1997. Fractional resolution and minimum aberration in blocked 2^{n-k} designs. Technometrics 39 (3), 382–390.
Sloane, N.A., A library of Hadamard matrices. http://neilsloane.com/hadamard/. (Accessed 2018).
Sun, D.X., Wu, C.F.J., Chen, Y., 1997. Optimal blocking schemes for 2^n and 2^{n-p} designs. Technometrics 39 (3), 298–307.
Tyssedal, J.S., 2008. Projectivity in experimental designs. In: Encyclopedia of Statistics in Quality and Reliability.
Tyssedal, J.S., Chaudhry, M.A., 2017. The choice of screening design. Appl. Stoch. Models Bus. Ind. 33 (6).
Tyssedal, J., Samset, O., 2010. Supersaturated designs of projectivity p = 3 or near p = 3. J. Statist. Plann. Inference 140 (4), 1021–1029.
Wu, C.F.J., Hamada, M.S., 2009. Experiments: Planning, Analysis and Optimization, second ed. Wiley, New York.
Xu, H., Lau, S., 2006. Minimum aberration blocking schemes for two- and three-level fractional factorial designs. J. Statist. Plann. Inference 136 (11), 4088–4118.
Xu, H.Q., Mee, R.W., 2010. Minimum aberration blocking schemes for 128-run designs. J. Statist. Plann. Inference 140, 3213–3229.
Zhang, R., Park, D., 2000. Optimal blocking of two-level fractional factorial designs. J. Statist. Plann. Inference 91 (1), 107–121.
Zhao, S., Li, P., Karunamuni, R., 2013. Blocked two-level regular factorial designs with weak minimum aberration. Biometrika Trust 100 (1), 249–253.

Paper 2

Preserving projection properties when two-level screening designs are blocked

Yngvild Hole Hamre and John Sølve Tyssedal

Submitted to Metrika for review.

This paper is submitted for publication and is therefore not included.

Paper 3

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

Yngvild Hole Hamre and John Sølve Tyssedal

Published in *Quality and Reliability Engineering International*, Volume 38(8), 2022, pages 4099-4121.

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 nonconvex 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

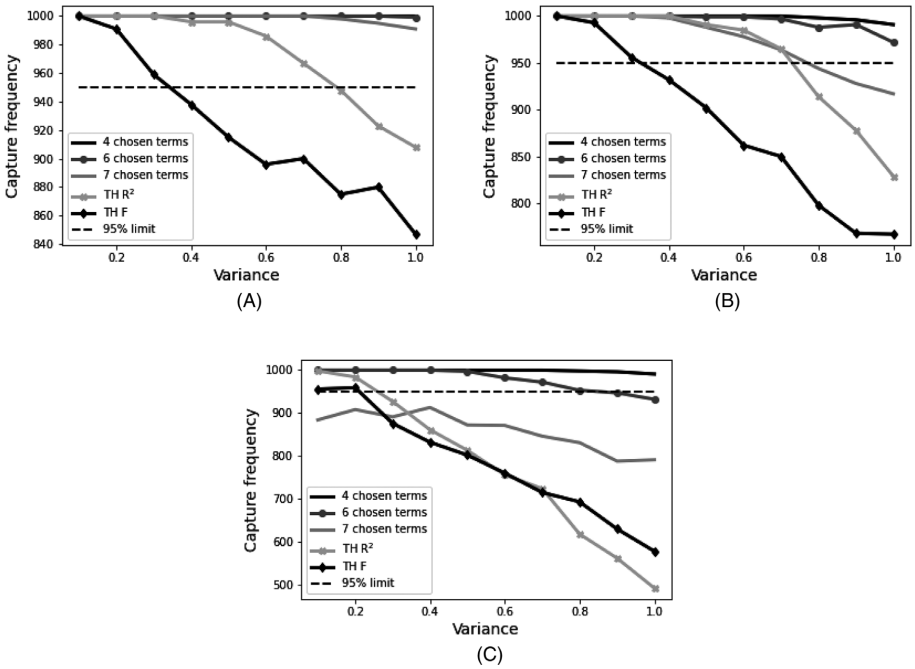


FIGURE 1 Plots of $CF_{10}(1000)$ against variance for comparison of the proposed size-based method and the methods from Tyssedal and Hussain,[17] for different numbers of experimental factors, n_t, and number of terms, l. The y-axis is different for each plot, but the black dashed lines show the 95%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

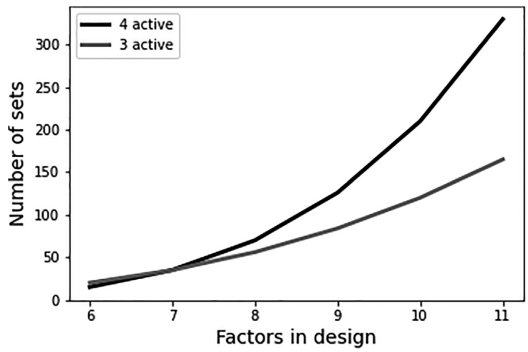


FIGURE 2 The number of possible candidate sets for three and four active factors as a function of the number of experimental factors

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-

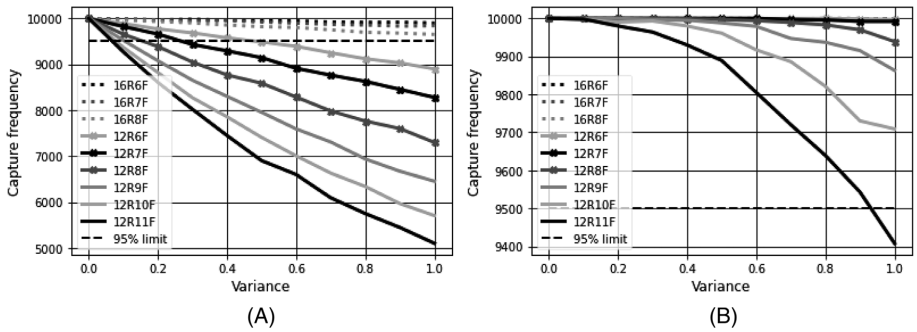


FIGURE 3 Plot of $CF_r(10,000)$ against variance varying the size of the capture set, r, and the number of experimental factors, n_t, for both 12- and 16-run designs having three active factors. The simulated models have three main effects and three interaction effects, and an absolute effect size between 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)

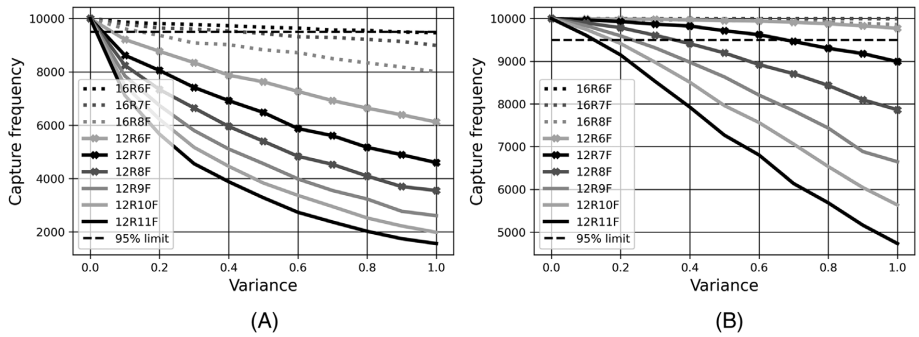


FIGURE 4 Plot of $CF_r(10,000)$ against variance for the 12- and 16-run designs with different numbers of factors in the design, using capture set size $r = 1$ and 5, respectively. The simulated models have four main effects and 2 two-factor interaction effects, and an absolute effect size between 1 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 $r = 5$

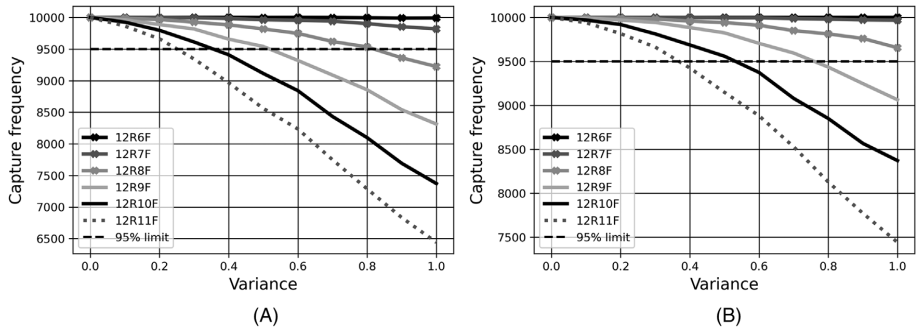


FIGURE 5 Plot of $CF_r(10,000)$ against variance for the 12-run PB design with different numbers of factors in the design, using capture set size $r = 10$ and $r = 15$, respectively. The simulated models have four main effects and 2 two-factor interaction effects, and an absolute effect size between 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

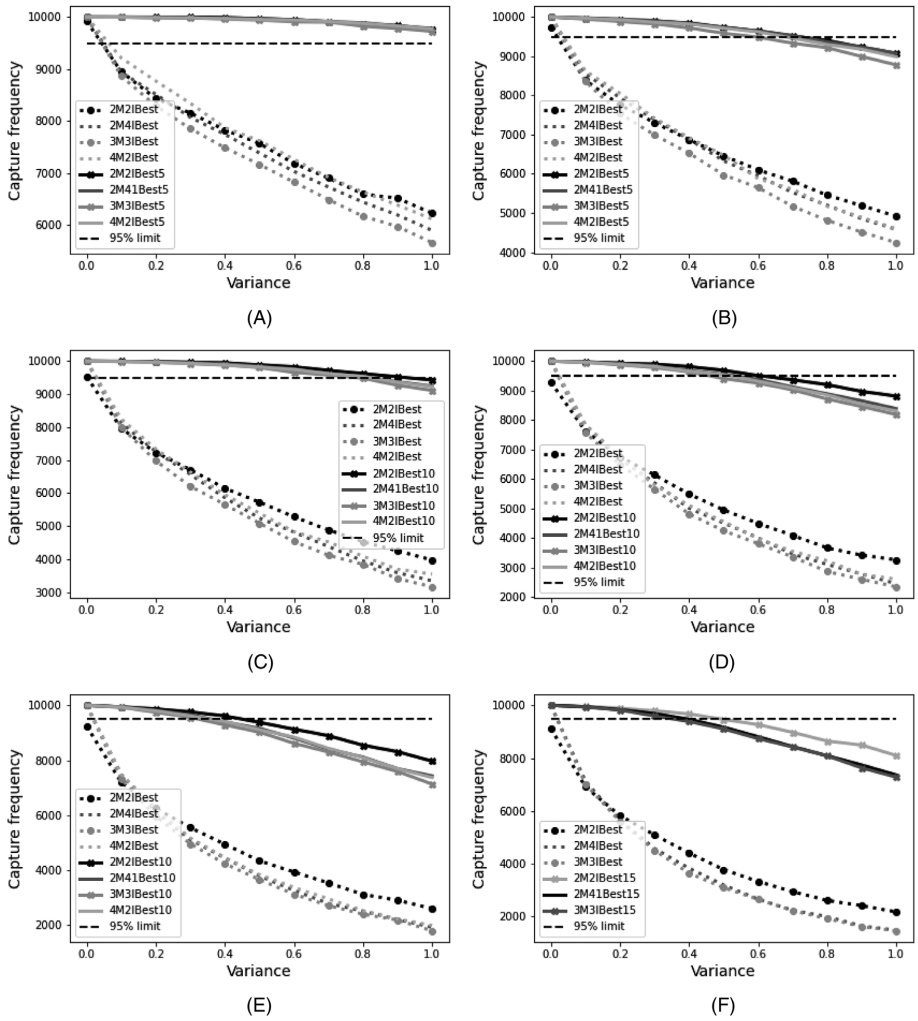


FIGURE 6 Plot of $CF_r(10,000)$ against variance for different model specifications using a PB12 design. 2M2IBest does for instance denote two main effects, 2 two-factor interactions and $r = 1$. In all cases, there were four active factors, and the absolute effect size was between 1 and 3. The number 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.

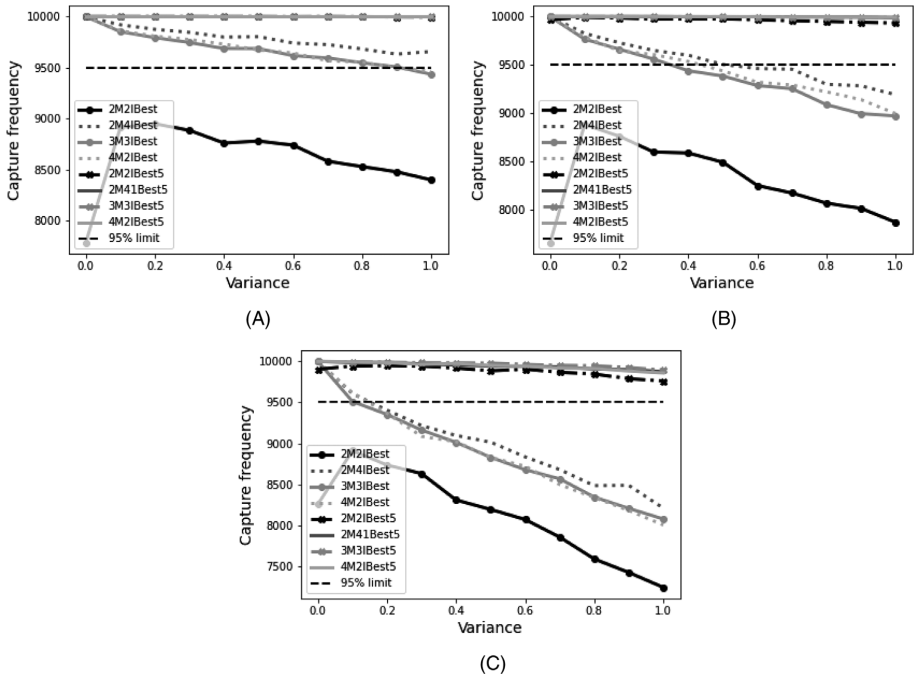


FIGURE 7 Plot of $CF_r(10,000)$ against variance for the 16-run design using different model specifications. 2M2IBest does for instance denote two main effects, 2 two-factor interactions and $r = 1$. In all cases, there were four active factors, and the absolute effect size was between 1 and 3. The number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

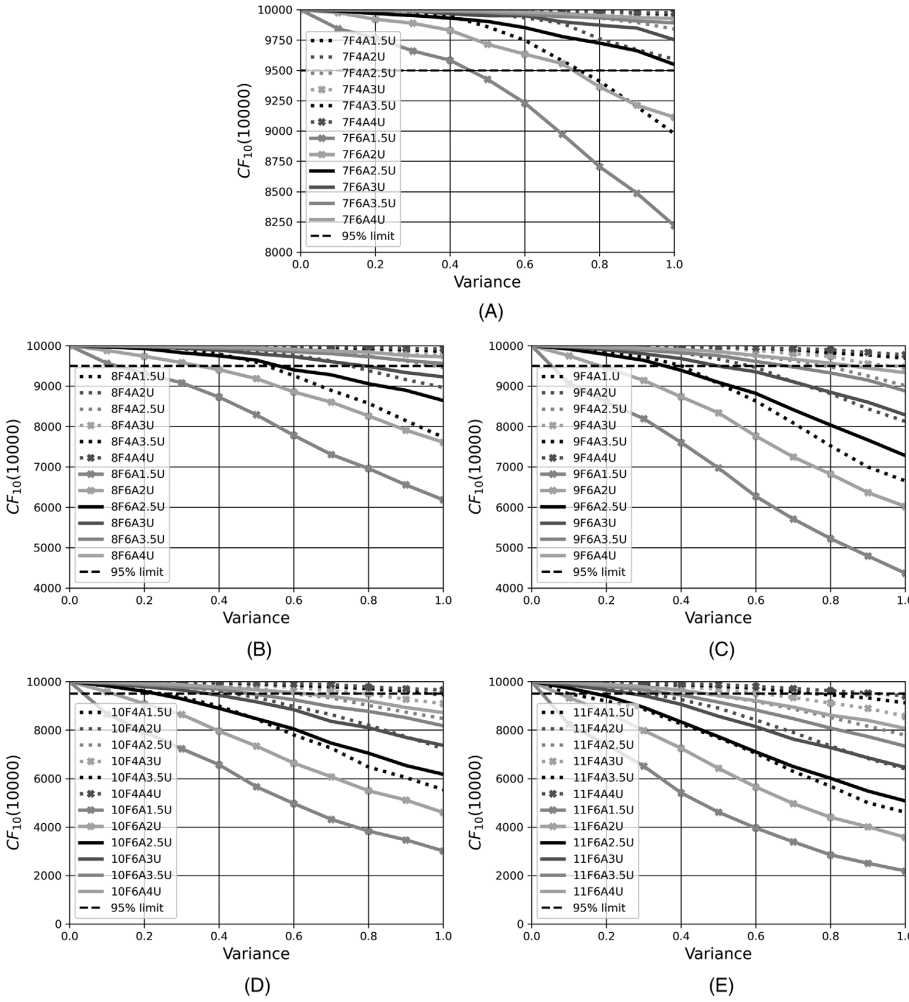


FIGURE 8 Plots of capture frequencies when testing different effect ranges. The minimum effect size was always 1, while the upper (U) was varied from 1.5 to 4. F is the number of factors in the design, while A is the number of active effects. There were always two active two-factor interactions. $l = n_e$ 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.

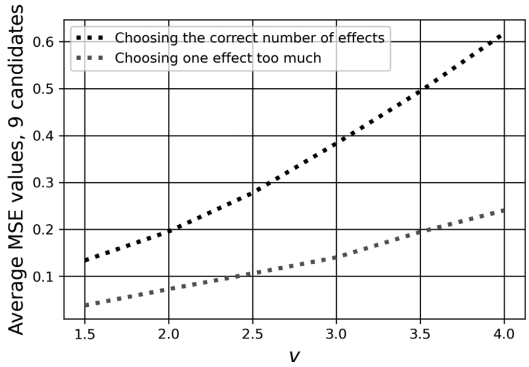


FIGURE 9 The average MSE of the nine best candidate models for different effect ranges. The values are based on 1000 iterations, using a design with nine factors, a minimum effect size of 1, and 0 variance. The simulated models had three active main effects and three active two-factor interactions.

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_α	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)

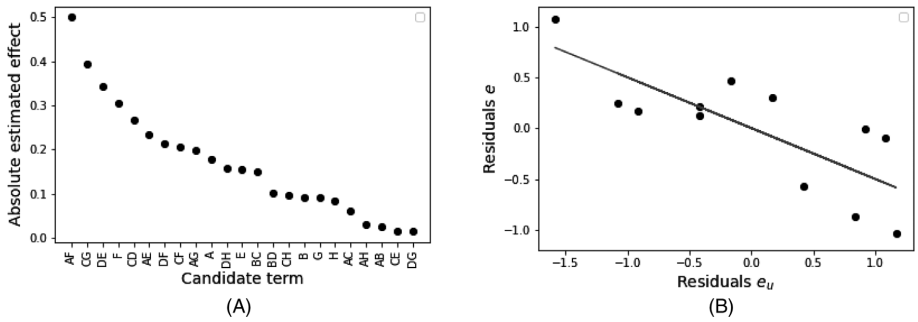


FIGURE 10 (A) An added variable Pareto plot for extending the model given in Phoa et al.[25] by one factor. (B) An added variable plot for the two-factor interaction AF. (A) All effects, (B) AF

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/10.1198/00401700152404318
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

Paper 4

A decoupling method for analyzing foldover designs

Yngvild Hole Hamre and John Sølve Tyssedal

Submitted to QREI for second review. Results presented at the 2023 ENBIS conference in Valencia.

A decoupling method for analyzing foldover designs

Yngvild Hole Hamre[1,2] and John Sølve Tyssedal[2]
yngvild.hole.hamre@dnb.no john.tyssedal@ntnu.no

[1]Norwegian University of Science and Technology

[2]DNB Bank ASA

February 28, 2024

Abstract

Foldover designs often have attractive properties. Among these is that the effects can be divided into two orthogonal subspaces, one for odd effects and one for even effects. In this paper, we introduce a new method for analyzing foldover designs called *the decoupling method* that exploits this trait. Utilizing mirror image pair runs, two new responses are created, where each of them is only affected by effects in one of the orthogonal subspaces. Thereby the analysis of odd and even effects can be performed in two independent steps, enabling use of standard statistical procedures and formal testing of the presence of higher-order interactions. The method is demonstrated on real data from a foldover of a 12-run Plackett-Burman (PB) design, and further evaluated through a simulation study, in which the decoupling method is compared to existing analysis methods. To get a thorough understanding of the properties, both a PB design and an OMARS design are used, and different design sizes and heredity scenarios considered. The method is especially suited for screening, as it yields high power for detecting the active effects.

1 Introduction

Folding over is a strategy that can be used to create designs with desirable properties. In particular, it ensures de-aliasing of certain effects. In this article, a foldover design denotes a design where the signs of all entries in the original design have been reversed, forming runs that are mirror images of the original ones, and then added to the original design's run.

Now let $\mathbf{X}$ be a $n \times p$ design matrix for k two-level factors with levels -1 and 1 with an intercept column included and $p = k + 1$. The foldover design matrix $\mathbf{X_f}$ is then constructed as $\mathbf{X_f} = \begin{bmatrix} \mathbf{1} & \mathbf{X} \\ \mathbf{1} & -\mathbf{X} \end{bmatrix}$, where a new intercept column must be included, as the original intercept column from $\mathbf{X}$ now is a $2n \times 1$ column where the first n and last n entries are 1 and -1, respectively. The original intercept column may therefore be used to include an additional factor. The foldover design $\mathbf{X_f}$ has $k + 2$ columns and $2n$ rows. For a regular two-level design of odd resolution R, the foldover design will be of resolution $R + 1$ (Box and Wilson, 1951). In general, a design is of resolution R if no p-factor effect is aliased with an effect with less than $R - p$ factors (Box et al., 1978). The properties of non-regular designs may better be described by their projection properties than by their resolution. A design of projectivity P can estimate all main effects and interactions for all sets of P factors (Box and Tyssedal, 1996), while a design of generalized projectivity P_α can estimate all main effects and interactions up to order α for all sets of P factors (Evangelaras and Koukouvinos, 2004). Non-regular two-level designs in which all factor columns are orthogonal to each other are called orthogonal. Examples of two-level orthogonal screening designs are the non-regular Plackett-Burman (PB) designs (Plackett and Burman, 1946), whose complex alias patterns make them candidates for folding over. When designs are folded over, the main effects and two-factor interactions can be estimated independently of each other, and the projectivity properties may improve. In his articles, Cheng (1995, 1998) showed that if the number of runs, n, is not a multiple of 8, the foldover of an orthogonal projectivity $P = 3$ two-level design with n runs is of projectivity $P = 4$ and generalized projectivity $P = 5_2$.

Defining the mirror image of the factor setting 0 to be 0 in runs having at least one other level, there are also three-level screening designs that have the foldover property. One example is the class of definitive screening designs (DSDs) (Jones and Nachtsheim, 2011), which are foldover designs with center runs added. They are popular screening designs, as they require a small number of runs while yielding the possibility to estimate quadratic effects. The DSDs with an even number of factors are orthogonal by design, and they can be made orthogonal for an odd number of factors as well by choosing a DSD with one more factor and two more rows and then omitting one of the factor columns. The recently introduced class of orthogonal minimally aliased response surface (OMARS) designs (Ares and Goos, 2020) include the orthogonal DSDs, as well as Box-Behnken designs (BBDs) and central composite designs (CCDs). As implicated by their name, the OMARS designs have orthogonal factor columns. The majority of them have the foldover property, and they exist for a wider range of sizes than the DSDs. Orthogonal factor columns along with a foldover structure eases the identification of active main effects. Foldover designs can however be constructed from non-orthogonal designs as well, for instance many of the minimum run resolution IV designs (Webb, 1968).

In this paper, we will focus on the analysis of foldover designs, possibly with center runs added. Based on the possibility of dividing odd and even effects into two orthogonal

subspaces, Miller and Sitter (2001) proposed a two-step procedure for finding active main effects and two-factor interactions in foldovers of PB designs. The procedure may also be used for other orthogonal designs with orthogonality between main effects and two-factor interactions. It will hereafter be referred to as the MS method and can be summarized as follows: In step 1, all the main effects are estimated, and standard methods are used to select significant effects. If columns for which no factor is assigned are available, hereafter denoted unassigned columns, a model independent variance for the effects is obtainable by calculating an assumed main effect for each of these, squaring them and averaging. This works due to orthogonal main effect columns and under the assumption of no odd effects of order three or higher. In step 2, the significant main effects from step 1 are included in the model together with the two-factor interactions under consideration (weak heredity may be imposed). An all-possible-subsets procedure with the main effects from step 1 forced into the model is suggested to find the active two-factor interactions. With some slight altering, Miller and Sitter (2005) demonstrated that the method could also be applied to designs with non-orthogonal main effects.

Inspired by the MS method, Jones and Nachtsheim (2017) developed a two-step procedure tailored for DSDs with fake factors added. This procedure will be referred to as the JN method. In the first step, they fit a main effects model to the data and identify the active main effects using a variance estimate based on the fake factors and eventual center runs. Then the variance estimate is updated using the inactive main effects and utilized in an F-test for selecting second-order effects (two-factor interactions and quadratic effects) in the second step, where the response is the residuals from the main effects model.

To accommodate analysis of OMARS designs, Hameed et al. (2023) introduced a generalized version of the JN method in which fake factors are not needed, and in some cases, more degrees of freedom are available for the variance estimate. This is because they use residuals from the full second-order model to estimate the variance in the first step, and they find the degrees of freedom using the rank of the corresponding design matrix. This method will be referred to as the HAG method. They point out that their method does not require a foldover structure, as it is applicable for any design with main effects that are orthogonal to the second-order effects. This also holds for the MS and JN methods. As Hameed et al. (2023) demonstrated that the HAG method is overall more powerful than the JN method, and they share a lot of the same features, the JN method will not be further considered in this paper.

The before-mentioned procedures have been demonstrated to work well, especially when the coefficient sizes are large compared to their standard deviations, but there are some drawbacks:

1. They all assume that only main effects and second-order effects are active. There is no suggested way of testing this assumption.
2. The MS method requires unassigned main effect columns in order to evaluate the significance of main effects in the presence of interactions.

3. For the MS and HAG method, undiscovered active main effects in step 1 will affect the estimated error variance in step 2.
4. For the MS and HAG method, the residuals in step 1 will be affected by active second-order interactions, making it impossible to assess them properly. The analysis in step 2 will be affected by any undiscovered active main effects.
5. For the MS method, if the foldover of a design is non-orthogonal, only R^2 or a modification thereof is recommended to use for model selection in step 1.
6. For the HAG method, an F-test is conducted in step 2. This requires selecting a significance level to decide when to stop. Adjusting this value may affect the results substantially, as demonstrated in Tyssedal and Hussain (2016).

To overcome these drawbacks, we introduce a different approach for analyzing foldover designs, called the *decoupling method*, or the DC method for short. The method is applicable to foldovers of both orthogonal and non-orthogonal designs (also for estimating quadratic effects, depending on the design). It is based on creating two new responses, Y_O and Y_E, where Y_O is only affected by odd effects and Y_E only by even effects. The analyses of these responses can be performed in any order, and choosing the wrong effects in one step will not affect the results found in the other step. However, following the hierarchy principle, it will be natural to start with Y_O. This will enable assumption of heredity, if desirable. When assuming weak heredity, any interaction must contain at least one active main effect, while when assuming strong heredity, a higher-order interaction is only considered if all the corresponding main effects are active. Either of the assumptions will introduce a dependency between the steps.

The theory behind the DC method is presented in Section 2, where we also explain how the drawbacks given above are avoided. Examples of designs with a foldover structure are given in Section 3. An example of applying the method to real data from a foldover PB design is given in Section 4, followed by results from an extensive simulation study assessing the performance on both a foldover PB design and an OMARS design in different heredity scenarios in Section 5. Finally, concluding remarks are given in Section 6.

2 Decoupling of models and methodology

2.1 Decoupling of models

From the n first runs in the matrix $\mathbf{X_f}$, let a model matrix be constructed as $\mathbf{X_m} = \begin{bmatrix}\mathbf{X_O} & \mathbf{X_E}\end{bmatrix}$, where $\mathbf{X_O}$ contains the columns that correspond to the odd effects of interest, and $\mathbf{X_E}$ the columns corresponding to the even effects of interest, including the intercept. As the n last runs in a foldover design matrix are obtained by reversing the signs for main effect columns, these will have opposite signs for odd effects. Even effects, on the other hand, will have the same signs, as reversing the signs of factors included in a multiplication with an even number of terms leaves the corresponding column unchanged. Thus, the

model matrix for the foldover design may be written as

$$\mathbf{X_{fm}} = \begin{bmatrix} \mathbf{X_O} & \mathbf{X_E} \\ -\mathbf{X_O} & \mathbf{X_E} \end{bmatrix}.$$

Furthermore, let $\boldsymbol{\beta^T} = \begin{bmatrix} \boldsymbol{\beta_O} & \boldsymbol{\beta_E} \end{bmatrix}$ be the corresponding regression coefficients, which, except for the intercept, are half the corresponding effect sizes. Then the $2n \times 1$ response vector $\mathbf{Y}$ for the full model can be written as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_O} \\ -\mathbf{X_O} \end{bmatrix} \boldsymbol{\beta_O} + \begin{bmatrix} \mathbf{X_E} \\ \mathbf{X_E} \end{bmatrix} \boldsymbol{\beta_E} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}$ is assumed to be $N(0, \sigma^2 \mathbf{I}_{2n\times 2n})$. $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are $n \times 1$ vectors containing the n first and the n last response variables, respectively. Similarly, let $\boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \end{bmatrix}$, where $\boldsymbol{\epsilon_1}$ and $\boldsymbol{\epsilon_2}$ are defined the same way. Then $\mathbf{Y_1} = \mathbf{X_O}\boldsymbol{\beta_O} + \mathbf{X_E}\boldsymbol{\beta_E} + \boldsymbol{\epsilon_1}$ and $\mathbf{Y_2} = -\mathbf{X_O}\boldsymbol{\beta_O} + \mathbf{X_E}\boldsymbol{\beta_E} + \boldsymbol{\epsilon_2}$. Define two new response vectors $\mathbf{Y_O}$ and $\mathbf{Y_E}$ as follows:

$$\mathbf{Y_O} = \frac{\mathbf{Y_1} - \mathbf{Y_2}}{2}$$
$$\mathbf{Y_E} = \frac{\mathbf{Y_1} + \mathbf{Y_2}}{2}$$

and let $\boldsymbol{\epsilon_O} = \frac{\boldsymbol{\epsilon_1} - \boldsymbol{\epsilon_2}}{2}$ and $\boldsymbol{\epsilon_E} = \frac{\boldsymbol{\epsilon_1} + \boldsymbol{\epsilon_2}}{2}$. Then obviously

$$\mathbf{Y_O} = \mathbf{X_O}\boldsymbol{\beta_O} + \boldsymbol{\epsilon_O} \tag{2}$$

and

$$\mathbf{Y_E} = \mathbf{X_E}\boldsymbol{\beta_E} + \boldsymbol{\epsilon_E}. \tag{3}$$

Further $\boldsymbol{\epsilon_O} \sim N(0, \frac{\sigma^2}{2}\mathbf{I}_{n\times n})$, $\boldsymbol{\epsilon_E} \sim N(0, \frac{\sigma^2}{2}\mathbf{I}_{n\times n})$ and $\mathrm{Cov}(\mathbf{Y_O}, \mathbf{Y_E}) = \mathrm{E}[\boldsymbol{\epsilon_O}\boldsymbol{\epsilon_E}^T] = \mathbf{0}_{n\times n}$. This follows since $\mathrm{E}[\epsilon_{Oi}\epsilon_{Ej}] = 0$, $i \neq j$ and $\mathrm{E}[\epsilon_{Oi}\epsilon_{Ei}] = \mathrm{E}[\frac{\epsilon_i^2 - \epsilon_{i+n}^2}{4}] = 0$, $i = 1, 2, ..., n$. Hence, assuming independent and identically normally distributed errors, $\mathbf{Y_O}$ and $\mathbf{Y_E}$ are independent random vectors, and model (2) and model (3) can be used to identify the odd and even effects independently of each other. Given which odd and even effects that are under consideration, estimators for $\boldsymbol{\beta_O}$ and $\boldsymbol{\beta_E}$ are $\hat{\boldsymbol{\beta}}_{\boldsymbol{O}} = (\mathbf{X_O}^T\mathbf{X_O})^{-1}\mathbf{X_O}^T\mathbf{Y_O}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{E}} = (\mathbf{X_E}^T\mathbf{X_E})^{-1}\mathbf{X_E}^T\mathbf{Y_E}$, assuming that the inverses exist. The covariance matrices are $\frac{\sigma^2}{2}(\mathbf{X_O}^T\mathbf{X_O})^{-1}$ and $\frac{\sigma^2}{2}(\mathbf{X_E}^T\mathbf{X_E})^{-1}$, respectively.

Alternatively, these estimators could have been obtained from the foldover design with model matrix $\mathbf{X_{fm}}$, as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{O}} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{E}} \end{bmatrix} = (\mathbf{X_{fm}}^T\mathbf{X_{fm}})^{-1}\mathbf{X_{fm}}^T\mathbf{Y} = \begin{bmatrix} (\mathbf{X_O}^T\mathbf{X_O})^{-1}\mathbf{X_O}^T\frac{(\mathbf{Y_1} - \mathbf{Y_2})}{2} \\ (\mathbf{X_E}^T\mathbf{X_E})^{-1}\mathbf{X_E}^T\frac{(\mathbf{Y_1} + \mathbf{Y_2})}{2} \end{bmatrix}$$

and hence

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{O}} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{E}} \end{bmatrix} = \begin{bmatrix} (\mathbf{X_O}^T\mathbf{X_O})^{-1}\mathbf{X_O}^T\mathbf{Y_O} \\ (\mathbf{X_E}^T\mathbf{X_E})^{-1}\mathbf{X_E}^T\mathbf{Y_E} \end{bmatrix},$$

with the corresponding covariance matrix for the coefficient estimators given by

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X_{fm}}^T\mathbf{X_{fm}})^{-1} = \frac{\sigma^2}{2}\begin{bmatrix} (\mathbf{X_O}^T\mathbf{X_O})^{-1} & 0 \\ 0 & (\mathbf{X_E}^T\mathbf{X_E})^{-1} \end{bmatrix}.$$

For the same model the coefficient estimates will be identical whether we use the full foldover design or base our analysis on model (2) and model (3) separately, and with known residual variance the inference would be identical. It is important to note that no assumption about orthogonal design columns is necessary for the above results to be true.

There are, however, differences in the identification of active factors depending on whether the full foldover design is used or the decoupled models (2) and (3). Let a linear model be written as

$$\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon},$$

where $\mathbf{X_1}$ is a $n \times r_1$ matrix, $\mathbf{X_2}$ is a $n \times r_2$ matrix, $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the respective coefficient vectors and $\boldsymbol{\epsilon}$ a vector of uncorrelated and identically distributed errors. Then we know that if only the model

$$\mathbf{Y} = \mathbf{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon}^*$$

is fitted, the least squares estimator for $\boldsymbol{\beta_1}$ will be biased by an amount $\mathbf{A}\boldsymbol{\beta_2}$, where $\mathbf{A} = (\mathbf{X_1}^T\mathbf{X_1})^{-1}\mathbf{X_1}^T\mathbf{X_2}$ is the alias matrix. For the least squares estimator for the variance, $\hat{\sigma}^{*2}$, we have

$$E(\hat{\sigma}^{*2}) = \sigma^2 + \frac{\boldsymbol{\beta_2}^T(\mathbf{X_2} - \mathbf{X_1}\mathbf{A})^T(\mathbf{X_2} - \mathbf{X_1}\mathbf{A})\boldsymbol{\beta_2}}{n - r_1},$$

see for instance Draper and Smith (1998), page 239.

If there are p_o odd effects of interest and we fit the model $\mathbf{Y} = \begin{bmatrix} \mathbf{X_O} \\ -\mathbf{X_O} \end{bmatrix} \boldsymbol{\beta_O} + \boldsymbol{\epsilon}^*$, such that $\mathbf{X_1} = \begin{bmatrix} \mathbf{X_O} \\ -\mathbf{X_O} \end{bmatrix}$, $\boldsymbol{\beta_1} = \boldsymbol{\beta_O}$, $\mathbf{X_2} = \begin{bmatrix} \mathbf{X_E} \\ \mathbf{X_E} \end{bmatrix}$ and $\boldsymbol{\beta_2} = \boldsymbol{\beta_E}$, we have for the least squares estimator of the variance, $\hat{\sigma}_o^2$, that

$$E(\hat{\sigma}_o^2) = \sigma^2 + \frac{\boldsymbol{\beta_E}^T(2\mathbf{X_E}^T\mathbf{X_E})\boldsymbol{\beta_E}}{n - p_o},$$

since the alias matrix in this case will be a matrix of zeros. Hence, if there are active even effects, the estimated error variance will be inflated and affect statistical procedures used for variable selection and model checking, including penalty-based methods, p-values and residual plots. This also affects step 1 of the procedures in Miller and Sitter (2001, 2005). If only main effects and two-factor interactions are active and the design is orthogonal, one way to get an unbiased estimate of the error variance is to use unassigned columns if such exist. This was the approach suggested by Miller and Sitter (2001). If there are not enough unassigned columns for variance estimation, they suggest using normal plots and Lenth's method as the general procedure for selecting effects. If the factor columns are non-orthogonal or if odd effects of order greater than 1 exist, none of these methods will work properly. Moreover, unidentified main effects in step 1 will inflate the error variance and affect step 2 of the procedures in Miller and Sitter (2001, 2005) in the same way as described for step 1. Unidentified main effects will also affect step 2 in Hameed et al. (2023), as they use a variance estimate in which the inactive main effects are included.

2.1.1 Choice of selection criterion

In a best subset selection, several criteria, including penalty-based ones, are often evaluated to decide upon the best or best few models. Typical such criteria are R^2_{adj}, AIC, BIC and AIC_c. R^2_{adj} is defined as $1-(1-R^2)\frac{n-1}{n-p}$, where $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$. Here $\hat{y}_i$ is the fitted value for y_i and, as before, n is the number of runs and p the number of parameters (variance excluded) in the model. AIC is defined as $2p^* - 2\ln(\hat{L})$, where $\hat{L}$ is the maximum of the likelihood function and $p^* = p + 1$. BIC is given by $p^*\ln(n) - 2\ln(\hat{L})$. For small sample sizes, Burnham and Anderson (2004) recommend using $\text{AIC}_\text{c} = \text{AIC} + \frac{2p^{*2}+2p^*}{n-p^*-1}$. Note that for linear regression, the error variance is also counted as a parameter when using both AIC, AIC_c and BIC, therefore $p^* = p + 1$.

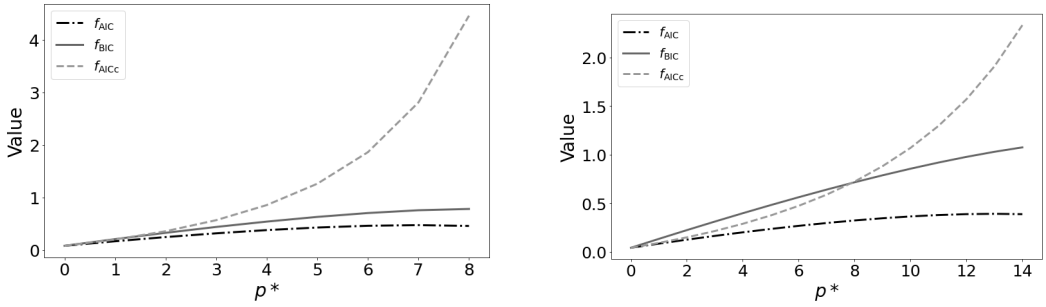
For linear regression with normally distributed errors, AIC, BIC and AIC_c can be rewritten as follows when constant terms are removed from the expressions (as in Banks and Joyner (2017)):

$\text{AIC} = n[\ln(\hat{\sigma}_e^2) + \frac{2p^*}{n} - \ln(\frac{n}{n-p^*+1})] = n[\ln(\hat{\sigma}_e^2) + f_{\text{AIC}}(n, p^*)]$

$\text{AIC}_\text{c} = n[\ln(\hat{\sigma}_e^2) + \frac{2p^*}{n-p^*-1} - \ln(\frac{n}{n-p^*+1})] = n[\ln(\hat{\sigma}_e^2) + f_{\text{AIC}_\text{c}}(n, p^*)]$

$\text{BIC} = n[\ln(\hat{\sigma}_e^2) + \frac{p^*\ln(n)}{n} - \ln(\frac{n}{n-p^*+1})] = n[\ln(\hat{\sigma}_e^2) + f_{\text{BIC}}(n, p^*)]$.

Since maximizing R^2_{adj} is equivalent to minimizing $\ln(\hat{\sigma}_e^2)$, where $\hat{\sigma}_e^2$ is given by $\hat{\sigma}_e^2 = \frac{\sum_i (y_i - \hat{y})^2}{n-p}$, f_{AIC}, f_{AIC_c} and f_{BIC} show how the respective criteria penalize the addition of parameters compared to R^2_{adj}. As illustrated in Figure 1, AIC_c penalizes the number of parameters much more than AIC and BIC for $n = 12$. For $n = 24$, the behavior of the criteria is different when considering the same number of parameters. Since all criteria depend on the number of observations, the decision about adding parameters may therefore differ depending on the method used.



(a) $n = 12$ (b) $n = 24$

Figure 1: The scaled penalty terms for AIC, BIC and AIC_c for $n = 12$ and $n = 24$.

Furthermore, applying these criteria may be suboptimal for the MS method, as it can face problems with inflated variances and the penalty being dependent on the number of main effects chosen to be included from step 1. For folded over non-orthogonal designs, Miller and Sitter (2005) suggested using the coefficient of determination, R^2, or a modification

thereof, the proportion of the total sum of squares that is explainable by the odd effects, in step 1. R^2 is a criterion recommended to be used with caution (Montgomery and Peck (1982), page 33, Walpole et al. (2012), page 408) since it will always increase when factors are added. We think most practitioners will prefer to base best subset selection on more than just one criterion. If the DC method is used, all standard statistical procedures for variable selection and model checking in linear regression models are available.

2.1.2 F-test for detecting interactions of order greater than two

The two decoupled models (2) and (3) provide us with independent estimates of the same error variance which are unbiased if the correct active effects are included in both. These estimates can be pooled to get more degrees of freedom and thereby obtain better inference. The pooled estimate is given by $\frac{n_{f1}\hat{\sigma}_1^2+n_{f2}\hat{\sigma}_2^2}{n_{f1}+n_{f2}}$, where $\hat{\sigma}_r^2$ is the variance estimate in step $r, r = 1, 2$, and n_{fr} is the corresponding number of degrees of freedom. However, the variance estimates can also be used in another way. Suppose model (2) is used to identify p_o main effects, but some three-factor interactions are active too. The estimated error variance will then be inflated. Therefore, if the estimated error variance from model (2) is much larger than the one from model (3), it is an indication that odd effects of order greater than one are active. Let $\hat{\sigma}_o^2$ be the least squares variance estimator from model (2) and let $\hat{\sigma}_e^2$ be the least squares variance estimator from model (3). A formal test can be performed using the test statistic

$$F = \frac{\hat{\sigma}_o^2}{\hat{\sigma}_e^2}.$$

Under the assumption of independent and identically normally distributed errors, the two estimators are independent, hence F will be Fisher distributed with $n - p_o$ and $n - p_e$ degrees of freedom. Here p_o and p_e are the number of odd and even effects. If, on the other hand, the variance estimate from model (3) is much larger than the one we get from model (2), higher-order even effects may be present.

2.2 Adding center runs to foldover designs

Center runs may be added to foldover designs for various reasons, like testing for lack of fit, obtaining a model free estimate of the error variance, or, as for DSDs, to improve the estimation efficiency if quadratic effects are present. Using the notation introduced in Section 2.1, adding n_c center runs $[\mathbf{X_C}\ \mathbf{X_C^*}]$ with corresponding response $\mathbf{Y_C}$, the $(2n + n_c) \times 1$ vector $\mathbf{Y}$ for the full model can be written as:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \\ \mathbf{Y_C} \end{bmatrix} = \begin{bmatrix} \mathbf{X_O} & \mathbf{X_E} \\ -\mathbf{X_O} & \mathbf{X_E} \\ \mathbf{X_C} & \mathbf{X_C^*} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_O} \\ \boldsymbol{\beta_E} \end{bmatrix} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is assumed to be $N(0, \sigma^2\mathbf{I}_{(2n+n_c)\times(2n+n_c)})$. Then $\mathbf{X_C}$ is a $n_c \times p_o$ matrix of zeroes, while $\mathbf{X_C^*}$ is a $n_c \times p_e$ matrix with a column of 1's for the estimation of the intercept and zeroes otherwise. Hence

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_{\boldsymbol{O}}\\ \hat{\boldsymbol{\beta}}_{\boldsymbol{E}}\end{bmatrix} = \begin{bmatrix}(\mathbf{X_O}^T\mathbf{X_O})^{-1}\mathbf{X_O}^T\mathbf{Y_O}\\ \left[\mathbf{X_E}^T\mathbf{X_E} + \frac{(\mathbf{X_C^*})^T\mathbf{X_C^*}}{2}\right]^{-1}\left[\mathbf{X_E}^T\mathbf{Y_E} + \frac{(\mathbf{X_C^*})^T\mathbf{Y_C}}{2}\right]\end{bmatrix},$$

Thus odd and even effects can still be estimated separately, and center runs belong to the even effects. The $(n + n_c) \times 1$ response vector for step 2 in the DC method then becomes

$$\begin{bmatrix}\mathbf{Y_E}\\ \mathbf{Y_C}\end{bmatrix} = \begin{bmatrix}\mathbf{X_E}\\ \mathbf{X_C^*}\end{bmatrix}\boldsymbol{\beta}_{\boldsymbol{E}} + \boldsymbol{\epsilon}. \tag{4}$$

where $\boldsymbol{\epsilon}$ is a $(n + n_c) \times 1$ vector. Since the variance for the variables in Y_E and Y_C differ, weighted least squares regression, as described in Fahrmeir et al. (2013), should be used for estimating the coefficients. The covariance matrix is given by

$$\text{Cov}\begin{bmatrix}\mathbf{Y_E}\\ \mathbf{Y_C}\end{bmatrix} = \sigma^2\mathbf{W}^{-1} = \sigma^2\begin{bmatrix}\mathbf{W_E} & \mathbf{0}\\ \mathbf{0} & \mathbf{W_C}\end{bmatrix}^{-1},$$

where $\mathbf{W_E}$ is a $(n \times n)$ diagonal matrix with $w_E = 2$ along the diagonal, and $\mathbf{W_C}$ is a $(n_c \times n_c)$ diagonal matrix with $w_C = 1$ along the diagonal. The corresponding weighted least squares estimator for $\boldsymbol{\beta}_{\boldsymbol{E}}$ is given by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{E}} = \left[\left[\mathbf{X_E}^T\ (\mathbf{X_C^*})^T\right]\mathbf{W}\begin{bmatrix}\mathbf{X_E}\\ \mathbf{X_C}\end{bmatrix}\right]^{-1}\left[\mathbf{X_E}^T\ (\mathbf{X_C^*})^T\right]\mathbf{W}\begin{bmatrix}\mathbf{Y_E}\\ \mathbf{Y_C}\end{bmatrix} =$$

$$\left[\mathbf{X_E}^T\mathbf{X_E} + \frac{(\mathbf{X_C^*})^T\mathbf{X_C^*}}{2}\right]^{-1}\left[\mathbf{X_E}^T\mathbf{Y_E} + \frac{(\mathbf{X_C^*})^T\mathbf{Y_C}}{2}\right]$$

Using weighted least squares to find the even effects, the selection criteria must be adjusted accordingly. Then, according to Banks and Joyner (2017), $\text{AIC}_{\text{cWLS}} = (n + n_c)\ln[\frac{\sum_{i=1}^{n+n_c} w_i(y_i-\hat{y}_i)^2}{n+n_c}] + \frac{2(p_e+1)(n+n_c)}{(n+n_c-p_e-2)}$, where $p_e + 1$ is the number of estimated parameters, including the error variance (note that constants were removed from the expression).

An alternative for an even number c of center runs is to consider them as $\frac{c}{2}$ mirror image pair runs and treat them as described in Section 2.1. Their c degrees of freedom will then be equally shared between odd and even effects.

2.3 The decoupling method

The basic idea of the screening algorithm is to first use the original response to create two new response vectors for which the expected values rely on the odd and even effects, respectively. Then these effects can be found separately of each other, while obtaining variance estimates that are not inflated by effects in the other subspace.

When using a foldover design of size $2n$, the algorithm is given by the following actions:

1. Order the foldover design matrix and corresponding response values such that the mirror image pairs are indexed by i and $i + n$ respectively, $i = 1, 2, ..., n$.
2. From the response values $\mathbf{y}$, create the response vectors $\mathbf{y_O}$ and $\mathbf{y_E}$, with elements given by $y_{O,i} = \frac{y_i - y_{i+n}}{2}$ and $y_{E,i} = \frac{y_i + y_{i+n}}{2}, i = 1, 2, ..., n$.

3. Step 1: If computationally feasible, use best subset selection to select main effects, using model (2). Select the number of parameters for the final model based on R^2_{adj}, AIC_c or another suitable criteria. If one desires to investigate three-factor interactions as well, one can proceed by forcing the selected main effects into the model and use best subset selection to consider three-factor interactions.
4. Step 2: Use best subset selection to select even effects (intercept, two-factor interactions, quadratic effects and if needed higher order even effects) accommodated by the design , using model (3), or model (4) in the presence of center runs.
5. Combine the effects selected in step 3 and 4 into a final model and use the whole data set when fitting it. This will yield the same coefficient estimates as found in Step 1 and Step 2, but a new variance estimate with more degrees of freedom, thus one may consider performing further assessment of the model.

Notes on the selection procedure

Due to best subset selection being computationally demanding, alternative strategies may be useful when analyzing very large designs. When the factor columns are orthogonal and there are no available degrees of freedom, normal or half-normal plots (Daniel, 1959) can be used for a coarse initial sorting. Having selected a candidate subset of effects, forward selection may be applied to check if additional main effects should be added. Backward elimination can be applied to see if any of the initially added effects should be removed. When there are available degrees of freedom when fitting the initial model, one may apply backward elimination directly instead of using half-normal plots. This is the approach used in Section 5. An interesting line of research using integer programming optimization techniques for best subset selection (Bertsimas et al., 2016; Vazquez et al., 2020) has shown impressing abilities compared to the common brute force approach. Thus considerations regarding large designs are likely to be reduced as such methods become more widespread.

2.4 The decoupling method with F-test

An alternative option inspired by Jones and Nachtsheim (2017) and Hameed et al. (2023) is to use an F-test to decide which even effects should be included. In step 1, selecting the main effects, two different variance estimates can be found. If there are degrees of freedom available, one may start by fitting the full main effects model and use the corresponding MSE for the variance estimate $\hat{\sigma}^2_{ind}$, which is independent of the choice of main effects. If one will allow for dependence, one may replace this estimate with the variance estimate from the model with only the main effects found active, which will be called $\hat{\sigma}^2_{dep}$. The resulting algorithm is as given in 2.3, except for step 2, which is modified as follows:

Use an F-test to select second-order effects. Start by using the F-statistic $F_0 = \frac{RSS_0/(n-1)}{\hat{\sigma}^2_*}$ to assess whether any second-order interactions should be added. Here $\hat{\sigma}^2_*$ is given by either $\hat{\sigma}^2_{ind}$ or $\hat{\sigma}^2_{dep}$. RSS_0 is the RSS from the intercept-only model. The critical value is then $F_{\alpha,n-1,df_*}$, where df_* is $n - n_f$ if $\hat{\sigma}^2_{ind}$ is used, $n - n_f + n_{in}$ if $\hat{\sigma}^2_{dep}$ is used. Here n_f denotes the number of experimental factors and n_{in} the number of inactive main effects.

If $F_0 > F_{\alpha,n,df_*}$, proceed by fitting all models with one second-order effect, and choose the model with the lowest RSS, denoted RSS_1. The F-statistic is then $F_1 = \frac{RSS_1/df_1}{\hat{\sigma}_*^2}$, where df_1=$n-2$ (as an intercept is included), with corresponding critical value F_{α,df_1,df_*}. If $F_1 > F_{\alpha,df_1,df_*}$, proceed to test the best model with two second-order effects, and so on. Let F_i and RSS_i denote the F-statistic and RSS corresponding to the best model with i second-order effects. If $F_i < F_{\alpha,df_i,df_*}$, the procedure should be ended and the currently considered second-order effects chosen.

Checking for three-factor interactions and higher order odd and even effects as described in Section 2.1.2 will not be possible using the DC method with an F-test since the variance estimates are then already used for detecting second-order effects. An alternative option can be to consider for instance residuals plots and QQ plots after each step.

3 Foldover of non-regular designs

Orthogonal non-regular designs

Having an orthogonal design matrix $\mathbf{X}$ eases the identification of main effects, as their estimates are then independent of each other. Orthogonality between the main effects is therefore often prioritized when considering which design to use. Important orthogonal designs are the two-level PB designs and the three-level OMARS designs. The main focus in this paper is foldovers of PB designs and OMARS designs with a foldover structure, but the proposed method is also directly applicable for non-orthogonal foldover designs.

The Plackett-Burman designs are popular screening designs, as they can accommodate up to $n-1$ factors in n runs, n being a multiple of 4 and ≤ 100. If $n = 2^k, k = 2, 3, ..., 6$, they coincide with the regular fractional factorial two-level designs, otherwise they are non-regular. Orthogonal non-regular two-level designs normally have very good projection properties (Tyssedal, 2008), but their alias patterns may be rather complex. A computer search conducted by Tyssedal and Samset (1999) showed that all non-regular PB designs can be folded over to become $P = 4$ designs, except when n=40, 56, 88 and 96.

One of the most used non-regular two-level designs is the 12-run PB (PB_{12}) design. It is a $P = 3$ design, but each main effect is aliased with all two-factor interactions for which it is not involved, which complicates the analysis of the design. Using its foldover, from now on referred to as PB_{12+12}, following the notation in Miller and Sitter (2001), this problem is resolved. The design is presented in Table 1. Since it is of projectivity $P = 4$ and $P_\alpha = 5_2$, it is clearly more flexible than the PB_{12} design regarding estimating models including 4 or 5 active factors. In total, the PB_{12+12} design allows for 12 orthogonal factor columns with 12 degrees of freedom, from which 66 two-factor interaction columns can be constructed. There are 11 degrees of freedom available for two-factor interactions. All main effects can be estimated independently of each other. Two-factor interaction columns are orthogonal if they share a common factor. If not, they are partially aliased by an amount of

$\frac{1}{3}$ in absolute value. The main effect columns are orthogonal to all even effect columns, but they may be aliased with higher-order odd effect columns. Active three-factor interactions may therefore affect the estimates of the main effects. More specifically, any three-factor interaction $2\beta_{xyz}$ introduces a bias of $\pm\frac{2}{3}\ \beta_{xyz}$ on all main effects which are not x, y or z.

Table 1: The PB_{12+12} design.

Run	A	B	C	D	E	F	G	H	I	J	K	L
1	1	1	-1	1	1	1	-1	-1	-1	1	-1	1
2	1	-1	1	1	1	-1	-1	-1	1	-1	1	1
3	-1	1	1	1	-1	-1	-1	1	-1	1	1	1
4	1	1	1	-1	-1	-1	1	-1	1	1	-1	1
5	1	1	-1	-1	-1	1	-1	1	1	-1	1	1
6	1	-1	-1	-1	1	-1	1	1	-1	1	1	1
7	-1	-1	-1	1	-1	1	1	-1	1	1	1	1
8	-1	-1	1	-1	1	1	-1	1	1	1	-1	1
9	-1	1	-1	1	1	-1	1	1	1	-1	-1	1
10	1	-1	1	1	-1	1	1	1	-1	-1	-1	1
11	-1	1	1	-1	1	1	1	-1	-1	-1	1	1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
13	-1	-1	1	-1	-1	-1	1	1	1	-1	1	-1
14	-1	1	-1	-1	-1	1	1	1	-1	1	-1	-1
15	1	-1	-1	-1	1	1	1	-1	1	-1	-1	-1
16	-1	-1	-1	1	1	1	-1	1	-1	-1	1	-1
17	-1	-1	1	1	1	-1	1	-1	-1	1	-1	-1
18	-1	1	1	1	-1	1	-1	-1	1	-1	-1	-1
19	1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1
20	1	1	-1	1	-1	-1	1	-1	-1	-1	1	-1
21	1	-1	1	-1	-1	1	-1	-1	-1	1	1	-1
22	-1	1	-1	-1	1	-1	-1	-1	1	1	1	-1
23	1	-1	-1	1	-1	-1	-1	1	1	1	-1	-1
24	1	1	1	1	1	1	1	1	1	1	1	-1

The orthogonal minimally aliased response surface (OMARS) designs (Ares and Goos, 2020) constitute another class of orthogonal designs. These are three-level designs and can therefore be used to screen for quadratic effects as well as main effects and two-factor interactions. Main effect columns are required to be orthogonal to each other and to the second-order effect columns. The orthogonal DSDs are also OMARS design, but unlike the DSDs, the OMARS designs may have main effect columns which contain more than 2 zeros in addition to the ones in the center runs. The OMARS designs are also more flexible with regards to run size than the DSDs. The vast majority of the OMARS designs consists of a foldover design and possibly center runs, and for these the DC method can be applied. A

27-run OMARS design which will later be used in the simulation study can be found in Table 2.

Run	A	B	C	D	E	F	G	H
1	-1	-1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	1	1	1	0	0
3	-1	-1	1	-1	0	0	1	1
4	-1	1	-1	0	0	1	1	0
5	-1	1	0	-1	0	1	0	-1
6	-1	1	0	0	-1	0	-1	1
7	-1	0	1	1	1	0	-1	0
8	-1	0	1	0	1	-1	0	-1
9	-1	0	0	1	-1	-1	1	1
10	0	-1	-1	-1	1	0	0	1
11	0	-1	1	0	-1	1	1	-1
12	0	-1	0	1	-1	1	-1	0
13	0	0	-1	1	0	-1	1	-1
14	1	1	1	1	1	1	1	1
15	1	1	1	-1	-1	-1	0	0
16	1	1	-1	1	0	0	-1	-1
17	1	-1	1	0	0	-1	-1	0
18	1	-1	0	1	0	-1	0	1
19	1	-1	0	0	1	0	1	-1
20	1	0	-1	-1	-1	0	1	0
21	1	0	-1	0	-1	1	0	1
22	1	0	0	-1	1	1	-1	-1
23	0	1	1	1	-1	0	0	-1
24	0	1	-1	0	1	-1	-1	1
25	0	1	0	-1	1	-1	1	0
26	0	0	1	-1	0	1	-1	1
27	0	0	0	0	0	0	0	0

Table 2: The 8-factor 27-run foldover OMARS design.

Non-orthogonal designs

When the primary goal of a screening experiment is to identify the active main effects with a secondary goal of assessing a small number of two-factor interactions, Miller and Sitter (2005) pointed out that a foldover of a non-orthogonal design matrix can be advantageous compared to an orthogonal resolution III design. Such foldover designs do not have orthogonal main effect columns, but they may have very flexible run sizes without a large efficiency loss in estimating effects. The decoupling method is applicable for non-orthogonal designs as well. This is an advantage compared to the methods proposed by Miller and Sitter, who suggest different approaches for orthogonal and non-orthogonal designs. When using the decoupling method for non-orthogonal designs, the only concern is that the standard analysis methods based on orthogonality assumptions such as Lenth's method and normal plots must be avoided. But for instance residual plots can still be assessed, and tests for higher-order interactions applied.

4 An example with a PB_{12+12} design

As pointed out in Section 3, active three-factor interactions introduce a bias in estimated main effects when a PB design is used. An advantage of the DC method is the possibility to assess residuals after selecting main effects. One may also perform an F-test for the presence of higher-order interactions. To illustrate how the analysis can be conducted and to compare it with other analysis methods, a real example where active three-factor interactions were found in the original analysis will now be presented.

In Mønnes (2012), data from a metal cutting experiment originally performed by Garzon (2000) was analyzed. The goal of the experiment was to identify the effects that affect the surface finish of the metal, using a full-factorial experiment with 6 factors and 64 runs, where each combination of factors was repeated 8 times. Choosing only a subset of the runs, it is possible to get 6 factors from a PB_{12+12} design with corresponding responses. There exist two different projections of the PB_{12} design onto 6 factors and thereby also of its foldover. The rows were chosen to match the projection without a mirror image pair run as preferred by Wang and Wu (1995), and the data can be found in Table 4 in Appendix 7. The response for each combination is the mean of the inverted response for the eight repetitions, as used in Mønnes (2012), where also several other transformations of the response were tested and compared. In that paper, they identified the effects D, E, F, CD, CE, CF, DE, DF, EF, ADF, BCD, CDE and DEF as active (significant on a 5% level) when fitting a model with all main effects, two-factor interactions and three-factor interactions to the transformed data, using all higher order interactions to estimate the noise. They point out that the significance might be affected by multiple testing. These were the same effects that were used as target values in Mønnes et al. (2007) when analyzing the original data, and correspond to in total 6 active factors, A, B, C, D, E and F.

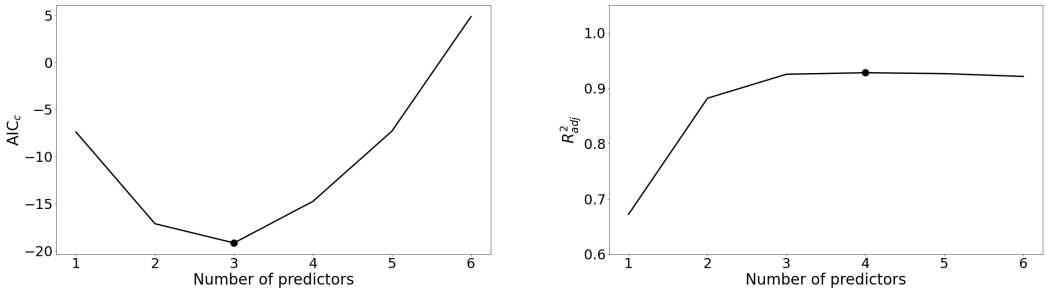
In the next sections, the data will be analyzed using the DC method (with and without an F-test in step 2), the MS method and the HAG method, to illustrate the use of each method and compare results. The MS method is performed as described in Miller and Sitter (2001), assuming that interactions of an order greater than two are not present, while the DC method is performed as suggested in Section 2.3. The HAG method is performed as described in Hameed et al. (2023). A summary of the resulting models is given in Table 3 in Section 4.4.

4.1 Analysis using the DC method

As the data set is small, enabling computationally demanding methods, best subset selection was used to find the best model of each size. The selection criteria chosen were R^2_{adj} and AIC_c, as AIC_c penalizes having many parameters the most when $n = 12$, while R^2_{adj} penalizes the least, as shown in Section 2.1.1. When these two criteria are in close agreement, it will increase our confidence in the number of parameters suggested. Note that the p-values found in this analysis are conditional on the chosen models and must therefore be interpreted with caution. We mainly comment on them if they are of a magnitude

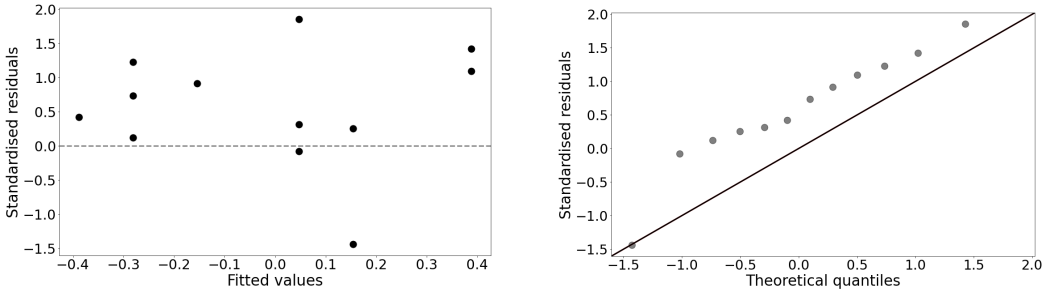
indicating that the corresponding effect may be inert.

Plots showing AIC_c and R^2_{adj} for best subset selection in step 1 are given in Figure 2. AIC_c has its minimum for a model with 3 predictors, namely D, E and F. R^2_{adj} is in close agreement with AIC_c and quite high. The chosen predictors are the same as in the original analysis for the full 2^6 design. In Figure 3, the corresponding residual plot and QQ plot are presented. Neither the residuals, being mainly positive, nor the QQ plot indicate that this model for the odd effects is satisfactory. One possibility is that three-factor interactions are present in the true model.



(a) AIC_c (b) R^2_{adj}

Figure 2: Performance criteria for the best main effects models of different sizes when analyzing the metal cutting data using the DC method.

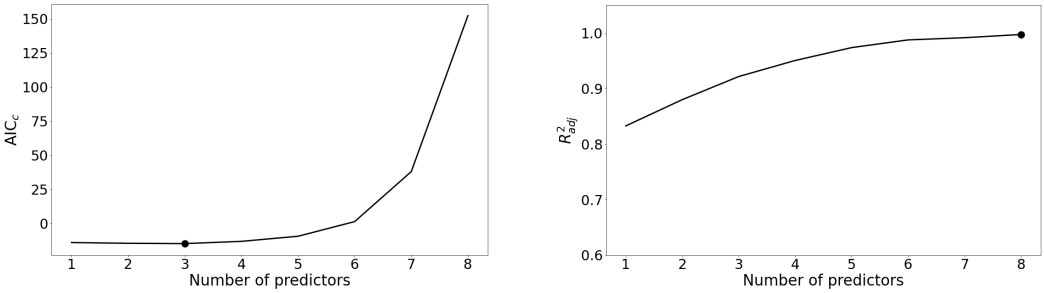


(a) Residual plot (b) QQ plot

Figure 3: Residual plot and QQ plot for evaluating the residuals for the chosen main effects model when analyzing the metal cutting data using the DC method.

In step 2, best subset selection was used to identify two-factor interactions. No assumptions about heredity were made. The corresponding graphs can be seen in Figure 4. The pattern is less clear than for the main effects, but AIC_c has its minimum for 3 two-factor interactions. These are AD, DE and DF. Comparing with the original 2^6 experiment, DE and DF are included, while CD, CE, CF and EF are left out. The residual plot and QQ plot (not shown) indicate no severe lack of fit and are in reasonable agreement with the assumption of normally distributed data.

R^2_{adj}, however, achieves its maximum when the number of predictors is 8 and is rather small when only 3 are considered, indicating a more complex model. The chosen interactions then become AB, AD, BC, BE, CD, CF, DE and DF, all with p-values less than 0.05. We observe that 4 of them are considered active in the original paper and 4 are not. The corresponding residual plot and QQ plot (not shown) look acceptable. As can be seen in Figure 4, the AIC_c is extremely high in this case, so there is a risk that the model with 8 two-factor interactions is overfitted. One way to proceed could then be to consider coefficient sizes and p-values together with system knowledge to choose which terms to keep. In case of ambiguity, follow-up runs could be considered.

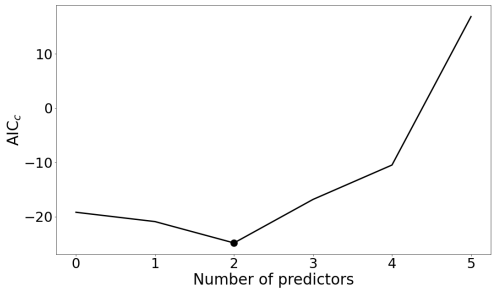


(a) AIC_c (b) R^2_{adj}

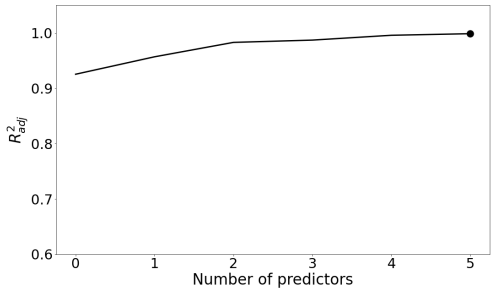
Figure 4: Performance criteria for the best two-factor interaction effects models of different sizes when analyzing the metal cutting data using the DC method. Note that the intercept is not included in the number of predictors.

The variance estimates from the odd effects model and the sparse even effects model were in this case of the same size, 0.00504 and 0.00483 respectively, and an F-test would not indicate the presence of three-factor interactions. Using the model with 8 two-factor interactions suggested by R^2_{adj}, the F-statistic becomes 32.1 with a one-sided p-value of 0.0079 for a null hypothesis of equal variances. Thus the sparse two-factor interaction model may be missing some parameters, enlarging the variance. This underlines the importance of having several tools available for model checking.

Based on the residual plots for the main effects model, best subset selection was used to look for three-factor interactions. The chosen main effects were forced to be in the model. The corresponding plots can be seen in Figure 5. AIC_c reaches its minimum when 2 three-factor interactions are included, and the R^2_{adj} is high and has a breaking point for 2 three-factor interactions. The chosen interactions are ADF and DEF, both included in the original model from the full factorial (together with BCD and CDE). The residual plot and QQ plot are shown in Figure 6. They are clearly improved compared to the ones in Figure 3 for the main effects only model.

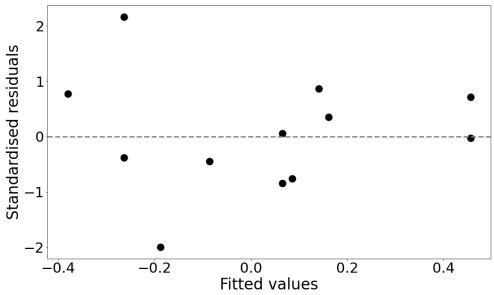


(a) AIC$_c$

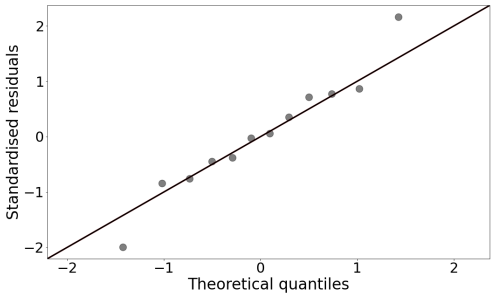


(b) R$^2_{adj}$

Figure 5: Performance criteria for the best three-factor interaction effects models of different sizes when analyzing the metal cutting data using the DC method. Note that the main effects forced into the model are not included in the number of predictors.



(a) Residual plot



(b) QQ plot

Figure 6: Residual plot and QQ plot for the model with odd effects and the chosen three-factor interactions included when analyzing the metal cutting data using the DC method.

The results deviate slightly from the analysis in Mønnes (2012), the largest difference being that considering AIC$_c$, the DC method suggested a sparser model with 3 less two-factor interactions and 2 less three-factor interactions as an acceptable alternative. As a subset of the original data set was used for the analysis, the original model chosen by Mønnes (2012) is not necessarily the best fit. It is therefore interesting to assess the performance of the original model when using the reduced data set. When fitting the model with the effects found to be significant in the original analysis to the reduced data set, the AIC$_c$ was -12.01 and R^2_{adj} 0.985, and the three-factor interactions BCD and CDE had large p-values, 0.849 and 0.234, respectively. Neither the residual plot nor the QQ plot for this model (not shown) indicated a severe lack of fit.

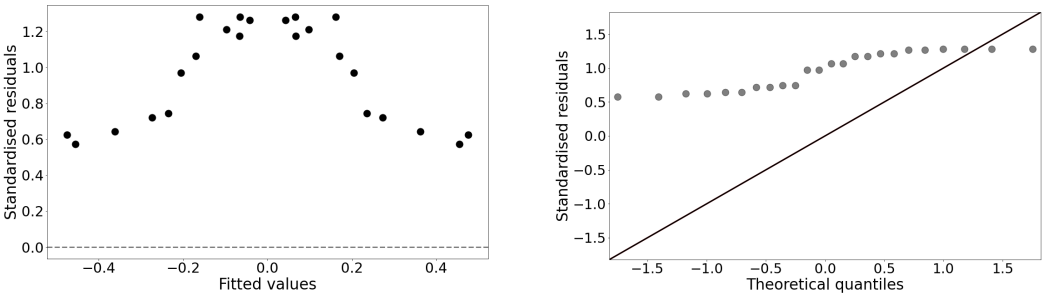
Using an F-test in step 2

As suggested in Section 2.4, F-tests may be used to select second-order interactions. Testing this approach (with $\alpha = 0.2$, as used in Hameed et al. (2023)) resulted in the selection of DE when using $\hat{\sigma}^2_{ind}$ in the denominator, and DE and DF when using $\hat{\sigma}^2_{dep}$. Thus using an

F-test results in a sparser final model. Note that an F-test cannot be used to assess the need for three-factor interactions in this case, so the residual plot and QQ plot for step 1 (Figure 3) would have to be evaluated when considering if three-factor interactions are missing. Plots for assessing residuals of the final models are not included here.

4.2 Analysis using the MS method

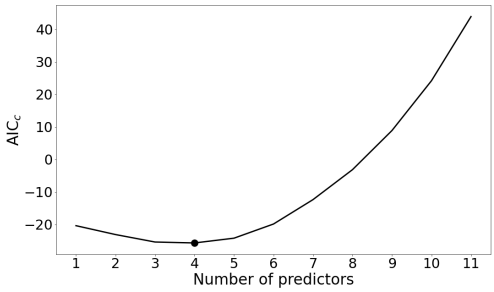
The first step of the MS method is to estimate all the main effects. As there were only 6 experimental factors, but the design has 12 columns, 6 unassigned columns were available for variance estimation. These were also included in the design matrix, so that 12 main effects were fitted. The variance is given by $\hat{\sigma}^2_{coeff} = \frac{\sum_{i \in U} \hat{\beta}_i^2}{n_u}$, where U is the set of unassigned columns, and n_u is the number of elements in U. Using standard t-tests, the main effects D, E and F were found to be significant. These were the same effects as were chosen in the original article and by the DC method. Plots for assessing the residuals are shown in Figure 7. Both plots show a strange behavior and underline the difficulties with model checking when errors are strongly affected by active effects that are not included. In addition, it is difficult to know which effects that might cause the problem.



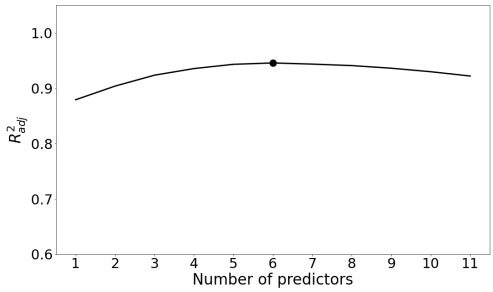
(a) Residual plot (b) QQ plot

Figure 7: Residual plot and QQ plot for evaluating the residuals for the chosen main effects model when analyzing the metal cutting data using the MS method.

In the second step, best subset selection with an intercept and the chosen main effects D, E and F forced into the model was used to find the final model. No assumptions about heredity were made, so all two-factor interactions were assessed as long as the resulting model matrix had full rank. Plots showing the AIC_c and R^2_{adj} for the best models of different sizes can be seen in Figure 8. The model with the lowest AIC_c includes 4 two-factor interactions, CE, CF, DE and DF. The resulting AIC_c and R^2_{adj} are -25.66 and 0.936, respectively. The residual and QQ plot can be found in Figure 9. In this case, the residual plot still shows sign of a curved trend, while the QQ plot does not indicate deviation from normality.

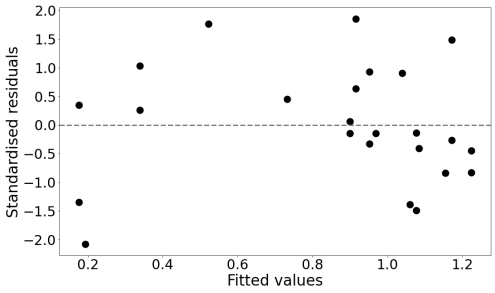


(a) AIC_c

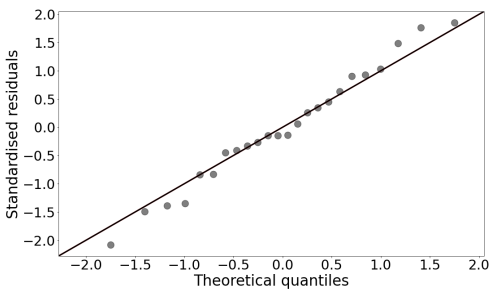


(b) R^2_{adj}

Figure 8: Performance criteria for the best two-factor interaction effects models of different sizes when analyzing the metal cutting data using the MS method. Note that the intercept is not included in the number of predictors.



(a) Residual plot



(b) QQ plot

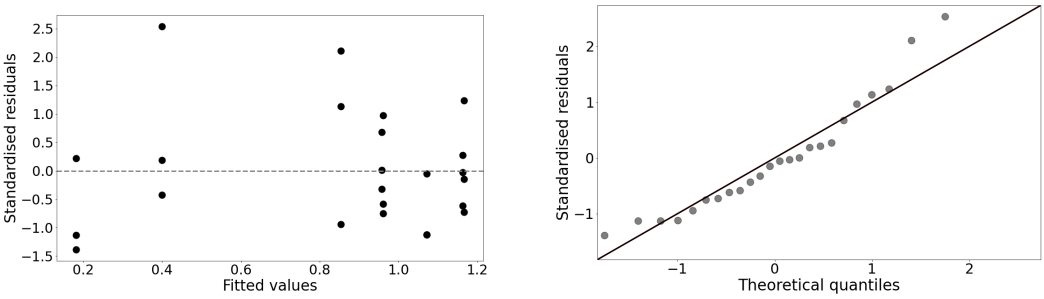
Figure 9: Residual plot and QQ plot for the model with main effects and the 4 chosen two-factor interactions when analyzing the metal cutting data using the MS method.

Had 6 interactions been chosen, as indicated by R^2_{adj}, they would have been the same ones as in Mønnes (2012). Among them, only CD gets a large p-value (0.216). The AIC_c is -19.83 and R^2_{adj} is 0.946. As the procedure is based on assuming that all three-factor interactions and higher-order interactions are negligible, these are the final candidate models. The R^2_{adj} and AIC_c values are slightly inferior compared to the results from the models chosen using the DC method.

4.3 Analysis using the HAG method

For the PB_{12+12} design, the value of the estimated variance in step 1 of MS and HAG is equal. The same main effects are therefore chosen using both methods, and as the same data is used, the residual plots are equal. In step 2, an F-test was conducted to find the active two-factor interactions. The resulting interactions were DE and DF. Those were the interactions with the highest estimated absolute values when using the MS method. The corresponding AIC_c and R^2_{adj} values are -23.08 and 0.904. The residual plot and QQ plot for the final model, shown in Figure 10, have some large positive residuals indicating possible lack of fit.

158



(a) Residual plot (b) QQ plot

Figure 10: Residual plot and QQ plot for the model with main effects and the 2 chosen two-factor interactions when analyzing the metal cutting data using the HAG method.

4.4 Summary for the final models

A summary of the results for all tested methods is given in Table 3. The lowest AIC_c-value is obtained by the sparse DC model, followed by the MS model. The highest R^2_{adj}-values were obtained by the largest models, which is to be expected, as R^2_{adj} favors large models more than AIC_c, as discussed in Section 2.1.1. The sparse DC model balances a low AIC_c with a rather high R^2_{adj} and may therefore be a viable choice in this situation. This example was made to highlight the differences between the suggested methods, in a real life setting all models could of course have been investigated in even more detail.

Table 3: Results from the PB_{12+12}-example.

Method	**ME**	**2FI**	**3FI**	**Factors**	AIC_c	R^2_{adj}
Original	D,E,F	CD,CE,CF DE,DF,EF	ADF,BCD CDE,DEF	6	-12.01	0.985
DC large	D,E,F	AB, AD, BC, BE CD, CF, DE, DF	ADF,DEF	6	-15.86	0.987
DC sparse	D,E,F	AD, DE, DF	ADF,DEF	4	-28.14	0.952
DC indep F	D,E,F	DE	ADF, DEF	4	-18.62	0.899
DC dep F	D,E,F	DE, DF	ADF, DEF	4	-22.94	0.928
MS	D,E,F	CE, CF, DE, DF	Not assessible	4	-25.66	0.936
HAG	D,E,F	DE,DF	Not assessible	3	-23.08	0.904

5 A simulation study

To verify the proposed method and compare it to the alternative analysis strategies, we conducted a simulation study using the 24-run PB_{12+12} design in Table 1 and the 27-run OMARS design in Table 2. As the performance of the methods may be dependent on the size of the design, the number of experimental factors used was 6 or 8, assigned to the first 6 or 8 columns of the designs. The DC method, the DC method with an independent and dependent F-test, the HAG method and the MS method were all used to find the active effects, except when using the 8-factor OMARS design. Then the MS method could not be applied due to lack of unassigned columns. There are multiple other analysis strategies that could have been applied, but Hameed et al. (2023) already demonstrated that the HAG method was as good as or better than popular methods such as stepwise model selection, hierNet (Bien et al., 2013) and the Dantzig selector (Candes and Tao, 2007). Projection-based methods, like the ones proposed in Tyssedal and Samset (1997), Kulachi and Box (2003), Tyssedal (2008), Tyssedal and Hussain (2016) and Hamre and Tyssedal (2022), are not feasible when a large number of factors may be active, as in the simulated models. The MS method was included in the comparison as it was not assessed by neither Jones and Nachtsheim (2017) nor Hameed et al. (2023), and being the originally proposed method, it may serve as a benchmark. For all methods, power and type 1 error were used to assess the results. Power is the average proportion of active effects that were identified, while type 1 error is the average proportion of inactive effects that were chosen.

To assess the impact of weak heredity, four scenarios were used in the simulation study:
1) No heredity used in the drawn models, no heredity assumed in the analysis of models.
2) Weak heredity used in the drawn models, no heredity assumed in the analysis of models.
3) No heredity used in the drawn models, weak heredity assumed in the analysis of models.
4) Weak heredity used in the drawn models, weak heredity assumed in the analysis of models.

For all scenarios, the models had 6 active factors, distributed between 4 main effects and 2 two-factor interactions, all of which were randomly chosen (possibly imposing heredity on the interactions, depending on the scenario). The constant had an absolute value of 2, while the absolute value for each of the other coefficients was randomly drawn from a standard exponential distribution and added to a constant value, referred to as "baseline", of either 0.5 or 1. The sign of each coefficient and the intercept was randomly drawn as well. This is a strategy inspired by Hameed et al. (2023). Noise from a standard normal distribution was added to the corresponding responses. The baselines of 0.5 and 1 were chosen to ensure that failures in identifying correct main effects and two-factor interactions should occur. This enables highlighting the differences in performance and the impact of heredity assumptions. Each combination of design, size, baseline and heredity scenario was tested using 1000 iterations. The MS and HAG methods were performed as described in Miller and Sitter (2001) and Hameed et al. (2023), respectively.

Note that two-factor interactions were the only second-order effects used when draw-

ing and assessing models in the first part of the simulation study. This choice enables comparison of the performance across the designs, but when using the OMARS design, quadratic effects could have been included as well. To demonstrate the ability to select quadratic effects, an extension of the simulation study in which quadratic effects were included will be presented in Section 5.4.

The DC method was performed as follows:

1. The two decoupled response vectors, $\mathbf{y_O}$ and $\mathbf{y_E}$, were constructed. Model (2) was used to find the significant main effects using backward elimination with significance level 0.05. Backward elimination was used instead of best subset selection to reduce computation time.
2. Model (3) was used to search for two-factor interactions, using best subset selection with AIC_c (the same criterion as for the MS method), always including an intercept.

The procedure for the DC method with F-test was:

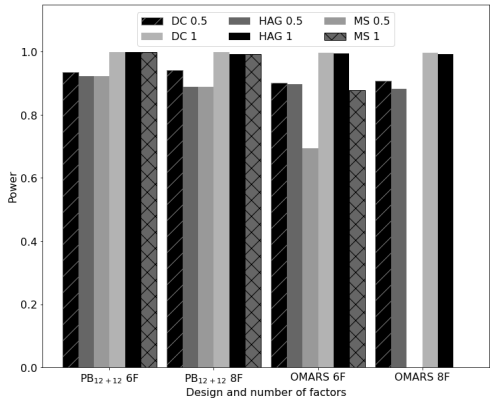
1. The two decoupled response vectors, $\mathbf{y_O}$ and $\mathbf{y_E}$, were constructed. Model (2) was used to find the significant main effects using backward elimination, with significance level 0.05. The MSE from the full main effects model was used to estimate the variance $\hat{\sigma}^2_{ind}$. Thereafter $\hat{\sigma}^2_{dep}$ was estimated using the MSE from the model with only active main effects.
2. Model (3) was used to search for two-factor interactions, using best subset selection with RSS as the selection criterion, always including an intercept. In each step, an F-test was used to decide whether a larger model should be evaluated or not, as described in Section 2.4. As in Hameed et al. (2023) the significance level was $\alpha = 0.2$. Both $\hat{\sigma}^2_{ind}$ and $\hat{\sigma}^2_{dep}$ was used, to see the effect of dependence between the steps.

When the 27-run OMARS design was used, the DC procedures were altered as prescribed in Section 2.2 to take the center run into account. Since best subset was used to select the interactions both in the MS and DC method, the number of second-order effects that could be chosen was limited to 4 for all methods. Had more interactions been allowed, the powers and type 1 errors could have been higher.

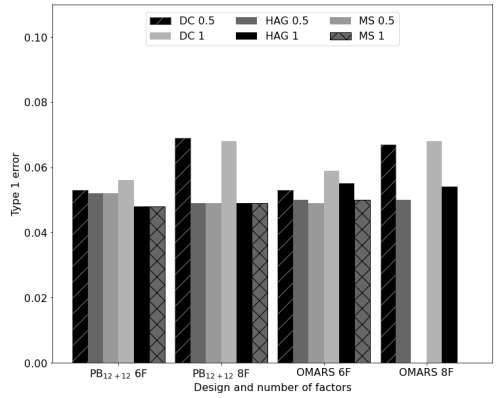
5.1 Results for main effects

The results for the selection of main effects can be found in Figure 11. As the identification of main effects is not affected by heredity assumptions, the values in Figure 11 were found by averaging across the heredity scenarios. The variance estimates in the first step of the HAG and MS methods were equal when using the PB_{12+12} design with 6 and 8 factors. In that case, there were 6 and 4 unassigned columns, respectively, used by the MS method for variance estimation, while when using 6 factors from the 27-run OMARS design, there were only 2 unassigned columns. This seems to have severely affected the power of the MS method. In general, the first step of the DC method consistently performs better than the first step of the MS and HAG methods, so using backward elimination where the variance estimate is updated in each step improves the power. We observe that the type

1 errors of the DC method are in some cases higher than for the others. However, in a screening situation, unidentified factors are often not considered for further investigation. It is therefore important that the set of active identified effects include the correct ones, and power will be the most informative measure.



(a) Power

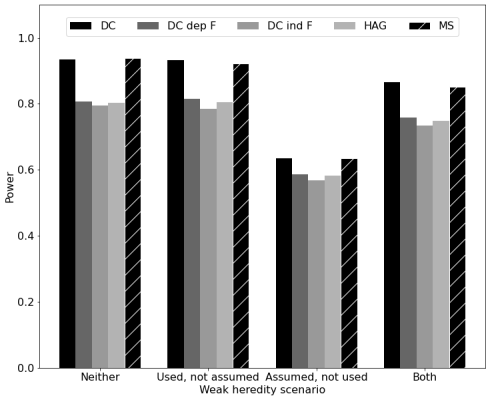


(b) Type 1 error

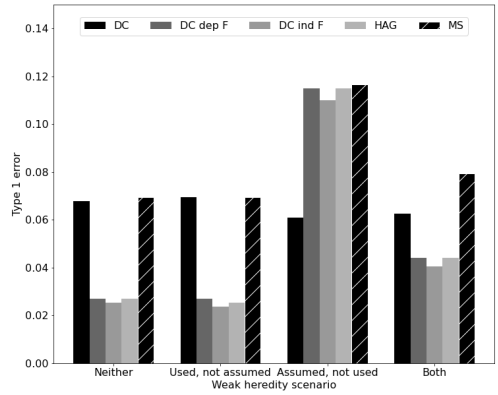
Figure 11: Power and type 1 error for selecting main effects using the DC, HAG and MS methods. Note that for the PB_{12+12} design, HAG and MS yielded the same results. DC 0.5 means that a baseline of 0.5 was used when drawing the coefficients. The bars are ordered following the order of the legend items, from left to right and from top to bottom.

5.2 Results for two-factor interactions, using a PB_{12+12} design

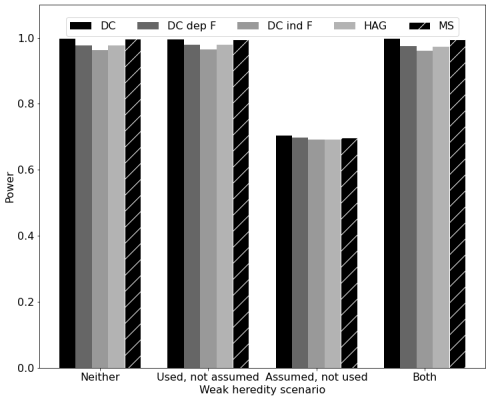
Plots with results for selecting two-factor interactions when using the foldover PB design can be found in Figure 12 and 13. When a baseline of 1 was used, the results were quite similar for all methods. Thus the 0.5 baseline case is more interesting, as the active effects are then harder to identify, making the differences between the methods more evident. With 6 experimental factors in the design, the MS and DC methods yielded very similar powers, substantially higher than the F-test based methods. These methods clearly select fewer interactions, as they result in a lower type 1 error, unless weak heredity is falsely assumed. In that case, the F-test based methods tend to select the maximum number of allowed interactions more often, resulting in high type 1 errors. The MS method also yielded a high type 1 error in that scenario. This might be because the added penalty term of the $\mathrm{AIC_c}$ is lower for higher n, as shown in Figure 1. When the correct interactions are not available for selection, there may be no model with a substantially lower $\hat{\sigma}_e^2$ than the others, and then model size penalization becomes relatively more important, making the MS method likely to choose more effects than the DC method.



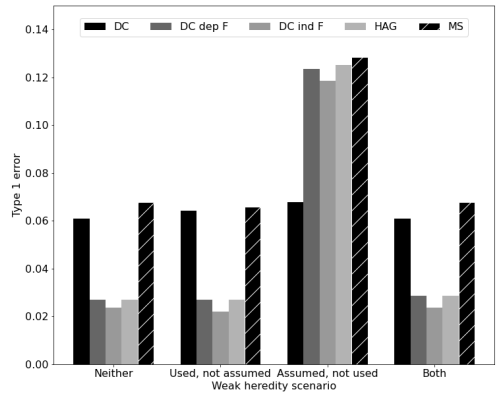
(a) Power, using a baseline of 0.5.



(b) Type 1 error, using a baseline of 0.5.



(c) Power, using a baseline of 1.

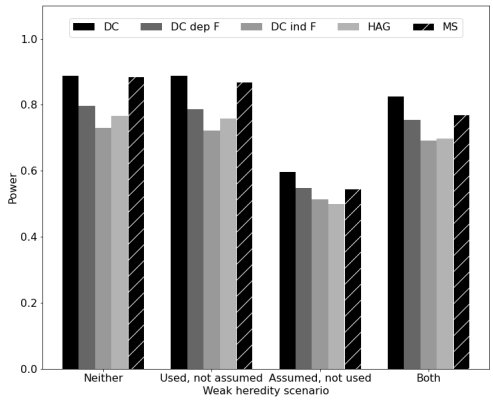


(d) Type 1 error, using a baseline of 1.

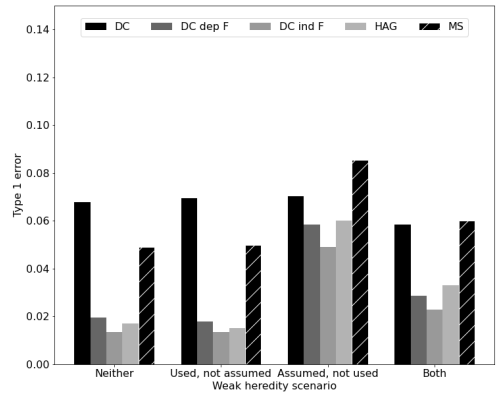
Figure 12: Power and type 1 error for selecting two-factor interactions when a PB_{12+12} design with 6 factors was used and 2 two-factor interactions were active.

An interesting observation regarding the heredity assumptions is the similar performance of the methods for the two scenarios "weak heredity neither used nor assumed" ("Neither" in the plots) and "Used, not assumed". When weak heredity is both used and assumed, on the other hand, the power is substantially reduced for all methods when a baseline of 0.5 is used. This is logical, as in that case, the correct main effects are not always chosen, and thus the true 2FI are not always available for selection. When falsely assuming weak heredity, the correct interactions may not be possible to choose even when the correct main effects are chosen, making the results even worse. It is clearly risky to assume weak heredity even when it seems very reasonable. If one has a strong belief that heredity is present, an alternative is to avoid heredity assumptions when fitting interactions, and rather include the main effects corresponding to the chosen interactions in the full model and do further evaluation to consider reducing it. Another option when applying the DC method is to consider the residuals for the second-order effects to evaluate whether heredity may have been falsely assumed. If the residuals have a strange pattern, it might be due to lack of correct second-order effects, and one can test including effects not fulfilling the

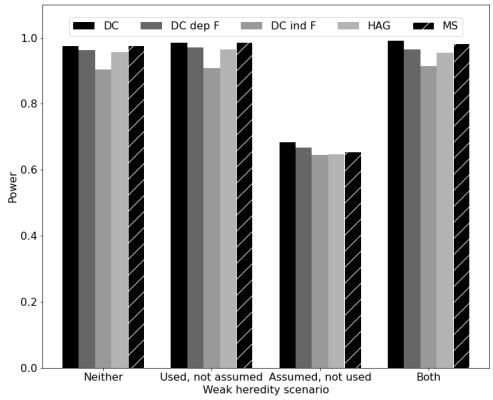
heredity assumption and see whether the residuals improve. This is less straight-forward for the MS and HAG method, for which patterns in the residuals from the model including second-order effects can also be caused by unidentified odd effects.



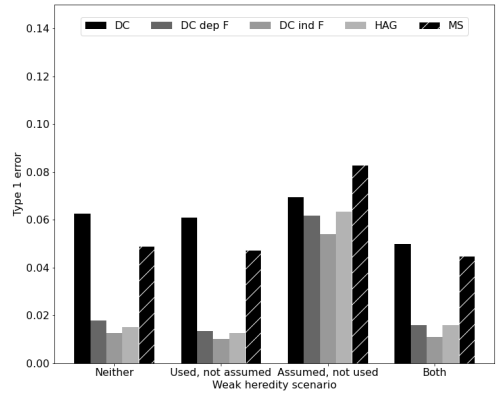
(a) Power, using a baseline of 0.5.



(b) Type 1 error, using a baseline of 0.5.



(c) Power, using a baseline of 1.

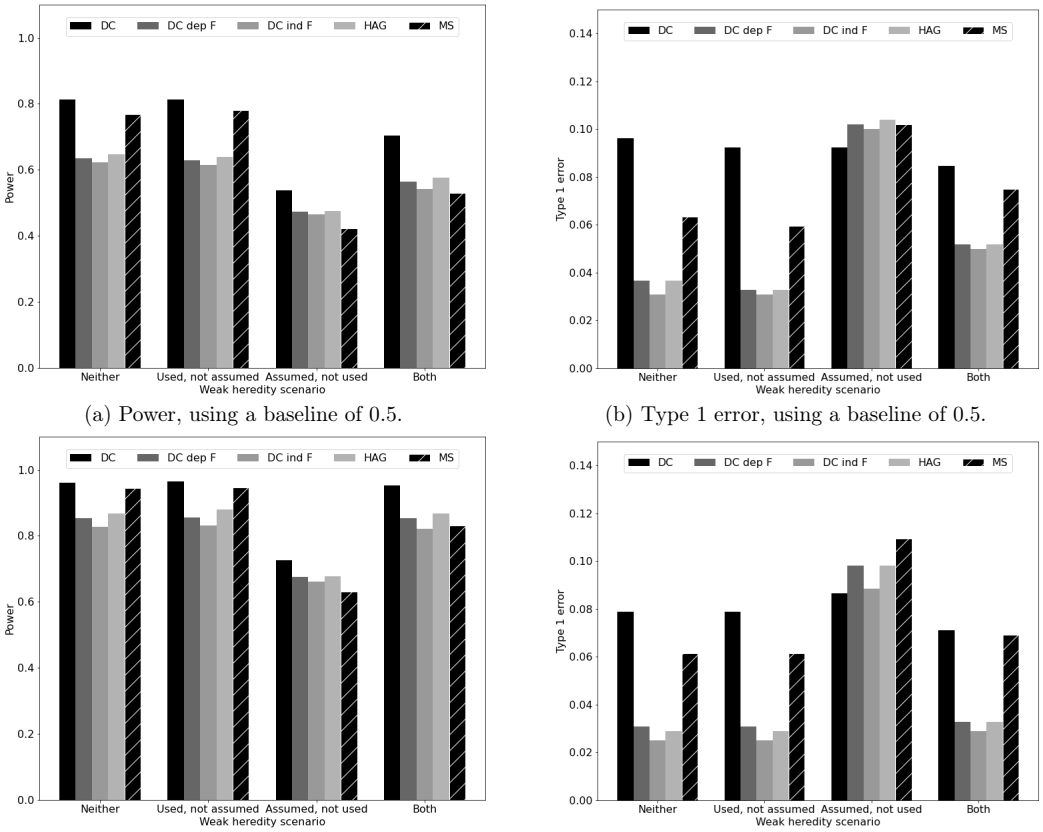


(d) Type 1 error, using a baseline of 1.

Figure 13: Power and type 1 error for selecting two-factor interactions when a PB_{12+12} design with 8 factors was used and 2 two-factor interactions were active.

Comparing the plots for the PB_{12+12} design when using 6 and 8 factors respectively, the patterns are similar, but the powers are larger when there are only 6 factors in the design. When having 8 factors instead of 6, the number of two-factor interaction candidates increases from 15 to 28, making it harder to select the correct ones. Note that type 1 error is relative to the number of available 2FIs, thus for a given number of inactive 2FI chosen, the type 1 error will be smaller in the 8-factor case. An interesting observation is that for 6 factors, the DC method with a dependent F-test performs very similarly to HAG, but in the 8-factor case, the total results seem slightly better as the achieved powers are higher and type 1 errors rather equal or lower.

5.3 Results for two-factor interactions, using an OMARS design



(a) Power, using a baseline of 0.5.

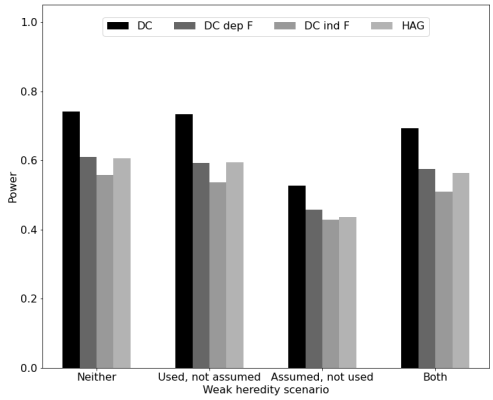
(b) Type 1 error, using a baseline of 0.5.

(c) Power, using a baseline of 1.

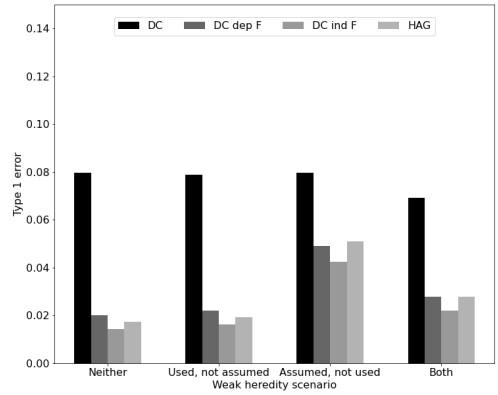
(d) Type 1 error, using a baseline of 1.

Figure 14: Power and type 1 error for selecting two-factor interactions when a 27-run OMARS design with 6 factors was used and 2 two-factor interactions were active.

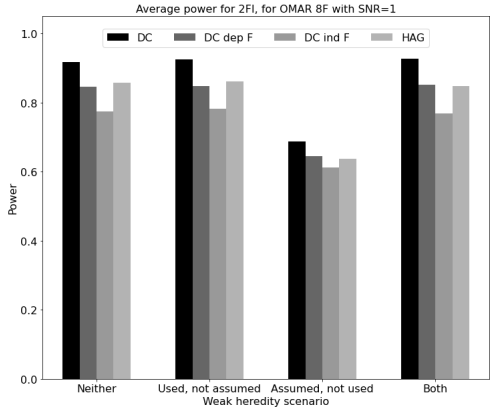
For the OMARS design, plots with results for the second-order effects can be found in Figure 14 and 15. For this design, the DC method performed better than all the other methods with regards to power. The MS method clearly suffers from often failing to select the correct main effects, which is especially severe when weak heredity is assumed. The HAG method gets a slight increase in power compared to the DC methods with F-tests when using the OMARS design. Also in this case, using the DC method results in a higher type 1 error than F-test based methods. The DC method with a dependent F-test and the HAG method yielded very similar results, while the DC method with an independent F-test combines low power and low type 1 error, seeming like an inferior alternative.



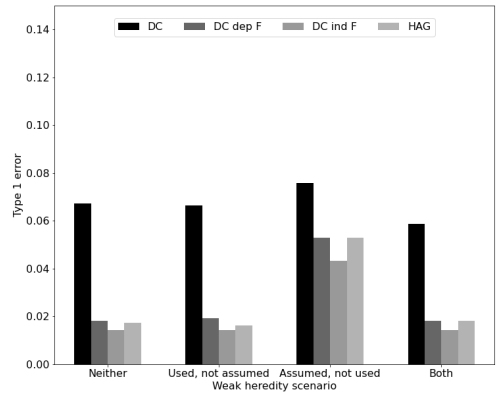
(a) Power, using a baseline of 0.5.



(b) Type 1 error, using a baseline of 0.5.



(c) Power, using a baseline of 1.



(d) Type 1 error, using a baseline of 1.

Figure 15: Power and type 1 error for selecting two-factor interactions when a 27-run OMARS design with 8 factors was used and 2 two-factor interactions were active.

5.4 Results for quadratic effects, using an OMARS designs

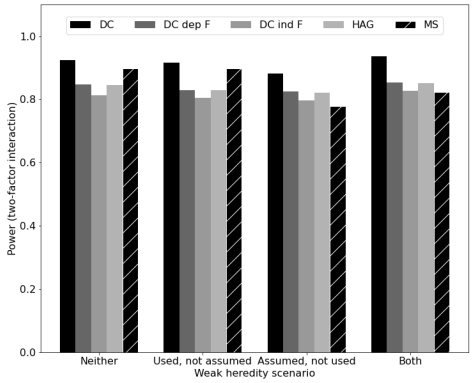
The above simulations only included two-factor interactions as second-order effects, facilitating comparison between the PB_{12+12} design and the 27-run OMARS design. To demonstrate the methods' ability to detect quadratic effects, complementary simulations were conducted. Unlike the intercept column and the two-factor interaction columns, the intercept column and the quadratic effect columns are not orthogonal. Using the DC method, the intercept, two-factor interactions and quadratic effects are handled together in step 2. The procedure was therefore conducted as before, always searching for models where the intercept is included. Using the HAG method, the non-orthogonality was handled by centering the second-order columns, as suggested in Jones and Nachtsheim (2017).

To enable comparison of all methods, the design used was the 6 first factor columns of the 27-run OMARS design. The drawn models now had 1 two-factor interaction and 1 or 2 quadratic effects, and quadratic effects were included among the second-order effect

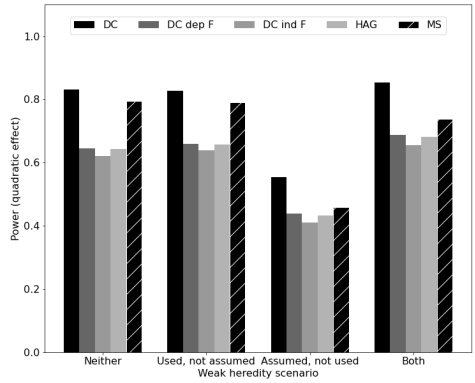
candidates when searching for the correct model. To still be able to assess different heredity scenarios, the number of active factors was reduced from 6 to 5. Except for that, the same settings were used as in the previous sections, applying a baseline of 1 when drawing the coefficients and including 4 active main effects. The results can be found in Figure 16.

When not assuming heredity for the candidate models, the power for identifying two-factor interactions was slightly lower than the power achieved for two-factor interactions in the corresponding scenarios when quadratic effects were not present (shown in Figure 14). This can be due to the larger number of candidate effects. When heredity was assumed for the candidate models, but not explicitly required in the drawn models, the results were in this case better, since the two-factor interactions had to fulfill the heredity assumptions in all cases (if not, there would have been 6 active factors). Only the MS method yielded substantially worse results when heredity was assumed than when not assuming heredity, as it is the only method that does not identify the correct active main effects in nearly all cases when the baseline is 1, as shown in Figure 11. For the quadratic effects, the power was notably lower when heredity was falsely assumed, as it was possible to include a quadratic effect not fulfilling the heredity assumption, but still have 5 active factors. The power for detecting quadratic effects was clearly lower than for two-factor interactions, but the behavior of the methods was similar, with the DC method always yielding the highest power for detecting the correct effects. In fact, the difference in power for the DC method and the other methods was larger for the quadratic effects than for the two-factor interactions. As before, the DC method with a dependent F-test and the HAG method behaved very similarly.

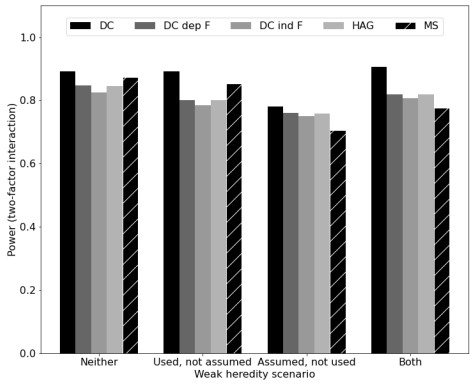
The powers were in most cases slightly lower if 2 quadratic effects were included in the true models instead of 1. The type 1 errors can however not be compared for the two scenarios, as the maximum number of included second-order effects was 4, thus the type 1 error is likely to be lower the more active effects the true model contains. Therefore, only type 1 error for the situation with 1 two-factor interaction and 1 quadratic effect is included, as it can be compared to the type 1 error in Figure 14. Note that the number of inactive two-factor interactions are used in the denominator for the type 1 error for the two-factor interactions, and likewise for the quadratic effects. As before, the high power for the DC method is achieved at the cost of a higher type 1 error than the other methods. For all methods, the type 1 error for the two-factor interactions was of similar magnitude as when quadratic effects were not included, and although heredity was always fulfilled, the type 1 error was higher when heredity was falsely assumed. The increase in type 1 error from falsely assuming heredity was paradoxically more prominent for the two-factor interactions than for the quadratic effects, although only the quadratic effects could fail to fulfill the heredity assumption. This indicates that the methods are more prone to include two-factor interactions than quadratic effects to compensate when the correct effect(s) cannot be chosen due to heredity assumptions. As before, the type 1 errors of the F-test based methods are more strongly affected by false heredity assumptions than the type 1 errors of the DC method and the MS method.



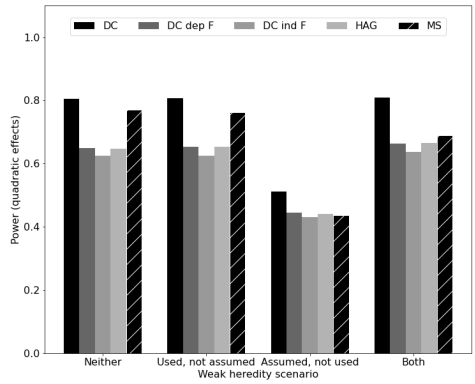
(a) Power for two-factor interactions when including 1 quadratic effect and 1 two-factor interaction in the drawn models.



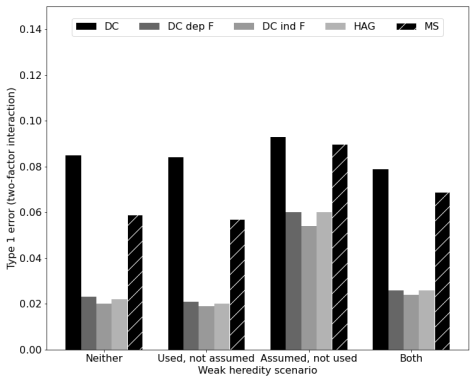
(b) Power for quadratic effects when including 1 quadratic effect and 1 two-factor interaction in the drawn models.



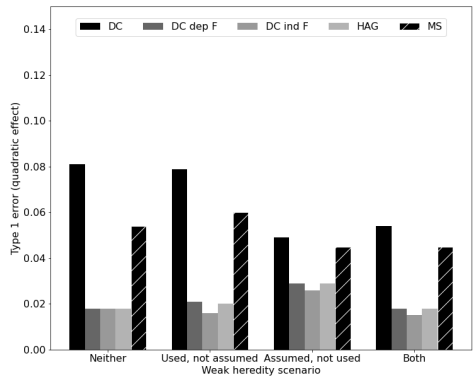
(c) Power for two-factor interactions when including 2 quadratic effects and 1 two-factor interaction in the drawn models.



(d) Power for quadratic effects when including 2 quadratic effects and 1 two-factor interaction in the drawn models.



(e) Type 1 error for two-factor interactions when including 1 quadratic effect and 1 two-factor interaction in the drawn models.



(f) Type 1 error for quadratic effects when including 1 quadratic effect and 1 two-factor interaction in the drawn models.

Figure 16: Power and type 1 error for selecting second-order effects when a 27-run OMARS design with 6 factors was used. There were 5 active factors, distributed between 4 main effects, 1 two-factor interaction and 1 or 2 quadratic effects. A baseline of 1 was used for all drawn models.

168

5.5 Summary of simulation results

The DC method demonstrates a strong ability to select the correct active effects in a variety of situations. It achieves a higher power than the F-test based methods in all cases and outperforms the MS method when there are few unassigned columns available for variance estimation. It performs particularly well compared to the other methods when weak heredity was a precept when drawing models or assumed when assessing second-order effects. However, assuming weak heredity when analyzing the designs yielded inferior results for all methods even if the drawn models obeyed the principle.

The DC method in most cases yield a higher type 1 error compared to the F-test based methods. This may not be considered a problem when the design is used for screening. If one for some reason wishes to reduce the type 1 error but keep the possibility of assessing the main effects model (for instance inspect the residuals), an alternative is to use the DC method with an F-test in step 2. The DC method with a dependent F-test and the HAG method seem to yield similar results. Note that for some OMARS designs, the HAG method can utilize more degrees of freedom for the variance estimate than the other methods.

To ease the comparison, the simulation study only covered models which were compatible with all analysis methods. Had three-factor interactions been included, the DC method is the only method with a formal way of testing for presence and a procedure for including them. To enable comparison of results for the PB_{12+12} design and the OMARS design, the main focus was on models including main effects and two-factor interactions only, but complementary simulations with quadratic effects included in the drawn models were conducted to demonstrate the methods' ability to detect those when using an OMARS design. The results showed that quadratic effects are harder to detect than two-factor interactions. The difference in power for the DC method and the other methods is even larger for quadratic effects than for two-factor interactions, and in addition it is less affected by false heredity assumptions. When falsely assuming weak heredity in the presence of quadratic effects, all methods seem to be prone to compensate by including more two-factor interactions rather than quadratic effects. As the results are dependent on the structure of the drawn models, assessing the performance of the methods for more combinations of model structures would be an interesting path of further work.

6 Concluding remarks

The new proposed DC method for analyzing foldover designs is based on decoupling the response values into two parts, one part for finding odd effects and one part for finding even effects. Unlike the methods proposed in Miller and Sitter (2001, 2005), Jones and Nachtsheim (2017) and Hameed et al. (2023), it yields the opportunity to search for odd and even effects separately in a two-step procedure, in which choices made in one step do not affect the results in the other. In addition, common procedures for variable selection and model checking can be used in each step, rather than just for the final model. This

makes the method less prone to error. Furthermore, it is possible to assess whether effects of an order greater than two should be included.

The performance of the method was assessed using real data from a foldover of a 12-run PB design, and further investigated and compared to the methods in Miller and Sitter (2001) and Hameed et al. (2023) in a simulation study where first only main effects and two-factor interactions were active (using a PB_{12+12} design and an OMARS design), followed by simulations where also quadratic effects were included (using an OMARS design). The DC method showed an overall superior performance in identifying both main effects and second-order effects. It performed well also when weak heredity was assumed, but we do not recommend using that assumption unless there are very good reasons to do so. Higher power is attained without the weak heredity assumption, even if it is correct.

Despite its good and stable performance compared to other methods, we believe that the greatest advantage of the DC method is the possibility to use standard statistical procedures for variable selection and model assessment, and that these can be performed in steps which are unaffected of each other. Using several criteria and model assessment methods to check whether the models seem reasonable is a benefit in real life settings to obtain trust in the models. The independent steps make it easier to find the cause of an eventual lack of fit.

Acknowledgements

The authors would like to thank two anonymous referees for constructive feedback and interesting ideas which greatly helped clarify and improve the paper.

References

Ares, J.N. and P. Goos (2020), "Enumeration and multicriteria selection of orthogonal minimally aliased response surface designs." *Technometrics*, 62, 21–36.

Banks, H. T. and M.L. Joyner (2017), "AIC under the framework of least squares estimation." *Applied Mathematics Letters*, 74, 33–45.

Bertsimas, D., A. King, and R. Mazumder (2016), "Best subset selection via a modern optimization lens." *The Annals of Statistics*, 44, URL `http://dx.doi.org/10.1214/15-aos1388`.

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

Box, G. E. P, W. G Hunter, and J.S. Hunter (1978), *Statistics for Experimenters*. Wiley.

Box, G.E.P and J. S. Tyssedal (1996), "Projective properties of certain orthogonal arrays." *Biometrica*, 83, 950–955.

Box, G.E.P. and K.P. Wilson (1951), "On the experimental attainment of optimum conditions." *Journal of the Royal Statistical Society*, 13, 1–45.

Burnham, K. P. and D.R. Anderson (2004), "Multimodel inference: Understanding AIC and BIC in model selection." *Sociological Methods & Research*, 33, 261–304.

Candes, E. and T. Tao (2007), "The Dantzig selector: Statistical estimation when p is much larger than n." *Annals of Statistics*, 35, 2313–1351.

Cheng, C.S. (1995), "Some projection properties of orthogonal arrays." *The Annals of Statistics*, 23, 1223 – 1233.

Cheng, C.S. (1998), "Some hidden projection properties of orthogonal arrays with strength three." *Biometrika*, 85, 491–495.

Daniel, C. (1959), "Use of half-normal plots in interpreting factorial two-level experiments." *Technometrics*, 1, 311–341.

Draper, N.R. and H. Smith (1998), *Applied Regression Analysis, Third Edition*. Wiley Series in Probability and Statistics, Wiley.

Evangelaras, H. and C. Koukouvinos (2004), "On generalized projectivity of two-level screening designs." *Statistiscs & Probability Letters*, 68, 429–434.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013), *Regression*. Springer.

Garzon, I. (2000), "Optimisation for product and process improvement: Investigation of Taguchi tools and genetic algorithms. PhD dissertation, ISRU." *University of Newcastle Upon Tyne*.

Hameed, M. S. I., J. Nunez, and P. Goos (2023), "Analysis of data from ortogonal minimally aliased response surface designs." *Journal of Quality Technology*, 55, 366–384.

Hamre, Y. and J. S. Tyssedal (2022), "On the identification of active factors in nonregular two-level designs with a small number of runs." *Quality and Reliability Engineering International*, 38, 4099–4121.

Jones, B. and C. J. Nachtsheim (2017), "Effective design-based model selection for definite screening designs." *Technometrics*, 59, 319–323.

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

Kulachi, M. and G. E. P. Box (2003), "Catalysis of discovery and development in engineering and industry." *Quality Engineering*, 15, 513–517.

Miller, A. and R.R. Sitter (2001), "Using the folded-over 12-run Plackett-Burman design to consider interactions." *Technometrics*, 43, 44–55.

Miller, A. and R.R. Sitter (2005), "Using folded-over nonorthogonal designs." *Technometrics*, 47, 502–513.

Montgomery, D. C. and E. A. Peck (1982), *Introduction to Linear Regression Analysis*. Wiley, New York.

Mønnes, E. (2012), "Data transformations with a full 2^6 experimental design - a metal-cutting case study." *Quality Engineering*, 24, 37–48.

Mønnes, E., M.J. Linsley, and I. E. Garzon (2007), "Comparing different fractions of a factorial design: a metal cutting case study." *Applied Stochastic Models in Business and Industry*, 23, 117–128.

Plackett, R.L. and J.P. Burman (1946), "The design of optimum multifactorial experiments." *Biometrika*, 33, 305–325.

Tyssedal, J. and S. Hussain (2016), "Factor screening in nonregular two-level designs based on projection-based variable selection." *Journal of Applied Statistics*, 43, 490–508, URL `https://doi.org/10.1080%2F02664763.2015.1070805`.

Tyssedal, J. and O. Samset (1997), "Analysis of the 12 run Plackett-Burman design." *Technical Report no. 8*.

Tyssedal, J. S. and O. Samset (1999), "Two-level designs with good projection properties." *Preprint, StatisticsNo.12/1999, NTNU,Norway*.

Tyssedal, J.S. (2008), "Plackett-Burman designs." In *Encyclopedia of Statistics in Quality and Reliability* (F. Ruggeri, R.S. Kenett, and F.W. Faltin, eds.), 1361–1365, Wiley, New York.

Vazquez, A. R., E. D. Schoen, and P. Goos (2020), "A mixed integer optimization approach for model selection in screening experiments." *Journal of Quality Technology*, 53, 243–266, URL `http://dx.doi.org/10.1080/00224065.2020.1712275`.

Walpole, E. R., R. H. Myers, S. L. Myers, and K. Ye (2012), *Probability Statistics for Engineeers and Scientists*, 9. edition edition. Pearson, Boston USA.

Wang, J.C. and C. F. J. Wu (1995), "A hidden projection property of Plackett-Burman and related designs." *Statistica Sinica*, 5, 235–250.

Webb, S. (1968), "Non-orthogonal designs of even resolution." *Technometrics*, 10, 291–299.

7 Appendix: Data for example

Table 4: A PB_{12+12} design for 6 factors, data from Mønnes (2012).

Run	A	B	C	D	E	F	y
1	-1	-1	1	-1	-1	-1	0.358355
2	-1	1	1	1	-1	-1	1.102892
3	-1	-1	-1	1	-1	1	1.095865
4	-1	-1	-1	-1	-1	-1	0.644945
5	-1	1	1	1	-1	1	1.151139
6	-1	1	-1	-1	-1	1	0.203328
7	-1	-1	1	1	1	-1	1.023245
8	1	-1	1	1	-1	1	1.285193
9	-1	1	1	-1	1	1	0.963643
10	1	1	1	-1	-1	-1	0.417663
11	1	1	-1	1	-1	-1	1.188810
12	-1	1	-1	1	1	-1	0.958609
13	1	1	-1	1	1	1	0.904882
14	1	-1	-1	-1	1	1	0.762895
15	1	1	1	-1	1	-1	0.962983
16	1	1	1	1	1	1	1.055025
17	1	-1	-1	-1	1	-1	1.066483
18	1	-1	1	1	1	-1	0.926640
19	1	1	-1	-1	-1	1	0.073068
20	-1	1	-1	-1	1	-1	0.962608
21	1	-1	-1	1	-1	-1	1.159382
22	-1	-1	-1	1	1	1	0.888587
23	-1	-1	1	-1	1	1	1.057513
24	1	-1	1	-1	-1	1	0.048384

ISBN 978-82-326-8032-0 (printed ver.)
ISBN 978-82-326-8031-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU

Norwegian University of
Science and Technology