Some Variable Selection Techniques for Finding Active Factors in Non-Geometric Plackett-Burman Designs

Oddgeir Samset
Department of Mathematical Sciences
The Norwegian University of Science and Technology
N-7034 Trondheim, Norway

February 11, 1998

Abstract

The non-geometric Plackett-Burman designs open for an economical screening of many factors in few runs. Traditionally these designs have been used as main effect plans assuming interactions to be negligible. However, if interactions are present, the complex alias patterns of these designs often entail a more thorough analysis of the experiments to identify the active factors. In this paper we present some variable selection techniques that account for active two-factor interactions. As an alternative to best subsets regression we propose a method where the principle of effect heredity guides the inclusion of terms. We also consider use of models where one includes all main effects and two-factor interactions for a given number of potentially active factors. It is demonstrated that forward regression on these models is a powerful tool for an extensive and efficient search for the active factors.

Key words: Effect heredity; Factor sparsity; Regression analysis; Screening.

1 Introduction

Typically a large number of factors, that the experimenter assumes might act upon the system under study, are included into the investigation in the initial stage of experimentation. The main objective of a screening experiment is then to identify those factors that actually cause variation in the responses of interest. Thus, by focusing on these active factors one can perform a refined subsequent experimentation much more efficient.

As a first approximation the response is often assumed a linear function of main effects and interactions between factors. One usually experiences however that only a few of the factors regarded initially are truly affecting the response. This is known as the *Pareto principle* or the *principle of factor sparsity* (Box and Meyer, 1986). In addition, based on previous experiences it is known that interactions of order three and higher often can be neglected. Thus, in such situations only a fraction of a full factorial design may be appropriate for identifying the active factors, and can thereby give large savings due to reduced number of runs.

The most popular screening designs are non-replicated 2^{k-p} fractional factorial designs (Finney, 1945; Box and Hunter, 1961) and Plackett-Burman designs (Plackett and Burman, 1946). When the number of runs n is a power of 2 the Plackett-Burman (PB) designs coincide with the 2^{k-p} fractional factorial designs, and are often called *geometric* PB designs. The other PB designs, with n not a power of 2, are referred to as *non-geometric* PB designs. All these designs have nice projection properties in the sense that small subsets of the factors often project into new designs with higher resolution. For example will a 2^{8-4}_{IV} fractional factorial design for studying 8 factors in 16 runs project into a replicated 2^3 full factorial design for any three factors. Thus, under factor sparsity these designs are regarded well suited for a systematic search for the active factors.

Traditionally experiments from these designs have been analysed using normal or half-normal plot (Daniel, 1959; 1976) or Bayes plot (Box and Meyer, 1986). However, active two-factor interactions may often cause problems identifying the active factors from such plots. For 2^{k-p} fractional factorial designs, of which resolution III or IV typically are used in screening situations, effects can be fully confounded. Then a number of models that equally well explain the variation in the response can be specified. Although the number of hypothesis usually is considerably reduced, the only way to distinguish between such models is to perform additional experiments. For non-geometric PB designs active two-factor interactions may seriously distort the traditional interpretation of the normal and half-normal plots due to their complex alias patterns. Hence, active factors may wrongly appear as inert and vice versa.

There has recently been an increased attention for non-geometric PB designs and how to analyse experiments from these designs. There are three main reasons for this. First, PB designs give the opportunity to study $n-1$ factors in n runs where n is a multiple of four. Use of non-geometric PB designs therefore open for an economical screening of many factors in few runs. Secondly, computer search made by Bisgaard in 1987 (Box and Bisgaard, 1993) and by Lin and Draper (1992) showed that saturated non-geometric PB designs may have better projection properties than saturated 2^{k-p} fractional factorial designs. See also Box and Tyssedal (1996) and Tyssedal and Samset (1997). Thirdly, as pointed out by Hamada and Wu (1991), non-geometric PB designs and other designs with partially confounded effects give under factor sparsity the ability of simultaneously estimating both main effects and interactions. Thus, unlike to 2^{k-p} fractional factorial designs where fully confounded effects may result in many models, non-geometric PB designs have in certain situations potential for separating out one plausible model that explains the variation in the response.

Section 2 gives a short overview of existing methods for analysing experiments from non-geometric PB designs. In Section 3 we present and compare some variable selection techniques that we have found useful for analysing such experiments. Here only main effects and two-factor interactions are taken into account. However, the methods described can easily be extended to also include interactions of order three or higher. In Section 4 some of the non-geometric PB designs are characterised with respect to the different methods presented in Section 3. Further use of the methods for identifying up to five active factors in simulated data are presented in Section 5.

2 Overview of Existing Methods

The non-geometric PB designs have traditionally been used as main effect plans when interactions between factors are assumed negligible. However, a few attempts have been accomplished analysing experiments from these designs when interactions are taken into account.

Hamada and Wu (1992) propose a method based on iterative use of forward selection. They start by including all main effect contrasts and all interaction contrasts orthogonal with these into the analysis. In step 1 all contrasts that are too large to be explained by noise alone are identified using a traditional method such as half-normal plot. In step 2 one includes all contrasts from step 1, all two-factor interaction contrasts containing at least one component factor found in step 1, and other contrasts suggested by the experimenter. Forward regression is then used to identify the significant contrasts among these. In step 3 they include all main effect contrasts and all contrasts from step 2, and use forward regression to identify additional main effect contrasts possibly missed in step 1 the first time. Step 2 and step 3 are repeated until the model stabilises.

Box and Meyer (1993) present a Bayesian approach that based on the principle of factor sparsity estimates the marginal posterior probabilities that the individual factors are active. By assuming only a small fraction of the factors being active a set of models is stated. For combination of potential active factors, models here include all main effects and interactions to a specified order. The posterior probabilities of the models given the observed response are computed. Finally, the marginal probabilities that a factor is active are found by summing over all models containing the factor.

Tyssedal and Samset (1997) use an alternative interpretation of the half-normal plot where the alias pattern of the applied design is taken into account. Half-normal plot of main effect contrasts and separate half-normal plots of two-factor interaction contrasts corresponding to each factor are generated. The motivation is that active main effects and two-factor interactions give rise to characteristic patterns in these plots, which may be explained based on the alias structure of the design. This simple method can be very informative when only a few active two-factor interactions are present. Tyssedal and Samset also use the projection properties of the 12 run PB design to identify up to four active factors.

Chipman, Hamada and Wu (1997) apply a Bayesian approach using the SSVS (stochastic search variable selection) algorithm of George and McCulloch (1993) and Gibbs sampling. Their method gives opportunity to model relations, such as effects heredity, between main effects and terms of higher order by specifying hierarchical priors for interactions.

3 Variable Selection Techniques

In this section we present some variable selection techniques for finding the active factors in non-geometric PB designs. To our experience these methods serve as useful supplement to the existing methods described above.

As the individual methods are presented in the following, their performance will be illustrated using example 4 of Hamada and Wu (1992). The response is simulated from the model $y = x_1 + 2x_1x_2 + 2x_1x_3 + \epsilon$ with $\epsilon \sim N(0, 0.25^2)$ using the 12 run PB design given in Table 1. In all situations we use the same realization from this model, see Table 1, so that the different methods can be compared to each other. Figure 1 shows the half-normal plot of the estimated main effect contrasts for this response. Using a traditional interpretation of the half-normal plot, the four largest estimated contrasts seem to be due to the inert factors x_7, x_8, x_9 and x_{10}. However, it is not difficult to verify that these large estimated contrasts are "produced" by the complex alias pattern of the design, see Tyssedal and Samset (1997). Thus, in this and similar situations the active factors can easily be overlooked.

All methods described in the following are implemented in MATLAB Version 4.2c.1, and

Table 1: The 12 run Plackett-Burman design.

run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	y
1	+	-	+	-	-	-	+	+	+	-	+	1.08
2	+	+	-	+	-	-	-	+	+	+	-	0.87
3	-	+	+	-	+	-	-	-	+	+	+	-5.00
4	+	-	+	+	-	+	-	-	-	+	+	1.40
5	+	+	-	+	+	-	+	-	-	-	+	1.21
6	+	+	+	-	+	+	-	+	-	-	-	5.07
7	-	+	+	+	-	+	+	-	+	-	-	-5.23
8	-	-	+	+	+	-	+	+	-	+	-	-1.02
9	-	-	-	+	+	+	-	+	+	-	+	3.04
10	+	-	-	-	+	+	+	-	+	+	-	-3.14
11	-	+	-	-	-	+	+	+	-	+	+	-1.08
12	-	-	-	-	-	-	-	-	-	-	-	3.10

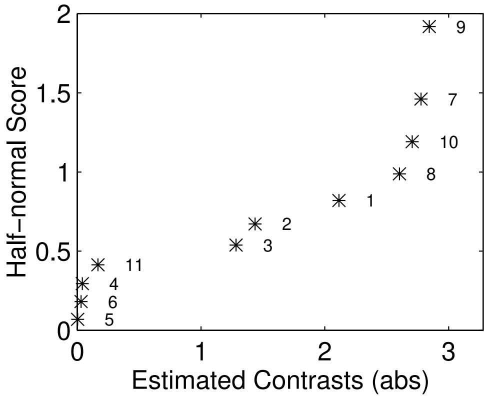

Figure 1: Half-normal plot of estimated main effect contrasts from the model $y = x_1 + 2x_1x_2 + 2x_1x_3 + \epsilon$ with $\epsilon \sim N(0, 0.25^2)$ using the 12 run PB design.

have been run on a Silicon Graphics Origin 200 with two 180MHz R10000 CPUs and 512Mb memory. The elapsed times for the different methods are presented to give an impression of the time it takes to perform the different analyses. For the time being, however, little effort is used to make the methods run as efficient as possible.

3.1 Best Subsets Regression

A popular method for variable selection in regression analysis is the *best subsets regression*. Here all models containing one regressor, two regressors, and so on are computed. Then the 'best' candidate models for given numbers of regressors are selected based on some predefined criteria such as the multiple coefficient of determination R^2, R^2_{adj}, Mallow's C_p or the residual means square error s^2.

Best subsets regression can also be used to analyse experiments from non-geometric PB designs. However, typically a relatively large number of factors are considered. Since both main effects and two-factor interactions are of interest, the number of models to examine may therefore grow very large as the number of terms to include in the models is increased. In

Table 2: The number of models to examine using best subsets regression.

No. of terms	PB(12)	PB(24)	PB(48)
1	66	276	1 128
2	2 145	37 950	635 628
3	45 760	3.47 e7	2.38 e8
4	720 720	2.37 e8	6.71 e10
5	8.94 e6	1.29 e10	1.51 e13

```
 Number of runs in design     :      12
 Number of factors in design  :      11
 Max number of terms          :       4
 Number of models examined    : 768692
 Number of rank warnings      :       0
 "Best model" criteria        : Residual MSe

Terms   R_sq  R_sqa      s    Terms in model
--------------------------------------------
    0   0.00   0.00   3.22
    1  45.27  39.80   2.50    1*3
    1  42.83  37.12   2.56    1*2
    1  32.95  26.24   2.77    6*7
    2  88.10  85.46   1.23    1*2 1*3
    2  74.37  68.67   1.80    8 4*11
    2  74.24  68.51   1.81    9 5*6
    3  99.82  99.76   0.16    1 1*2 1*3
    3  90.59  87.05   1.16    1*2 1*3 4*9
    3  90.43  86.85   1.17    1*2 1*3 8*11
    4  99.98  99.97   0.05    1 1*2 1*3 3*10
    4  99.92  99.87   0.12    1 1*2 1*3 4*9
    4  99.91  99.85   0.12    1 1*2 1*3 5*6
elapsed_time = 1738 seconds
```

Figure 2: Best subsets regression for the model $y = x_1 + 2x_1x_2 + 2x_1x_3 + \epsilon$ with $\epsilon \sim N(0, 0.25^2)$ using the 12 run PB design.

Table 2 we have listed the number of models to examine for a given number of terms (main effects and/or two-factor interactions in addition to the intercept) for the saturated 12, 24 and 48 run PB designs. As the table shows, the practical importance of this method as a useful tool for analysing such experiments is rather limited.

Since best subsets regression examines all possible combinations of main effects and two-factor interactions, this method will inevitable identify the 'best' model if it is distinguishable from the rest according to some criterion. If there are more models that equally well explain the variation in the response, all of these models will be identified. Best subsets regression is therefore included in this section as a comparison to the performance of the other methods presented. Mallow's C_p is often used as a measure for evaluating the different models. However, to use this criterion one needs a reliable estimate of σ^2. This is a problem here since the true model is (assumed) unknown and the design is not replicated. Often the residual mean square error s^2 for the full-term model is used as an estimate of σ^2, but the number of terms in the full-term models is too large in the situations considered here. We therefore choose to not include this measure.

In Figure 2 the three 'best' models according to residual mean square error are listed for zero to four terms (in addition to the intercept) for the response in Table 1 corresponding to example 4 of Hamada and Wu (1992). 768692 models were examined. All of the matrices of regressor variables including the constant column have full rank. When three terms are included, one of the models clearly separates out from the rest with respect to all measures. The 'best' of the models with four terms has slightly smaller value of s, but the increase in R^2 and R^2_{adj} is moderate. Furthermore, all three 'best' models with four terms contain the three-term model found above. This indicates that the model with terms x_1, x_1x_2 and x_1x_3 is the most plausible for describing the variation in the response.

Use of best subsets regression in this situation shows that the 12 run PB design, in addition to enable simultaneously estimation of both main effects and two-factor interactions, also under ideal conditions is able to separate out one single model and thereby the true active factors. In contrast a 2^{k-p} fractional factorial design will, due to fully confounded effects, require resolution V to be able to distinguish between two-factor interactions.

3.2 Submodel Regression

In this section we consider use of the "factor based" models proposed by Box and Meyer (1993). Instead of adopting the Bayesian approach we rather apply linear regression to fit all the proposed models to the response, and present the results as in best subsets regression.

The models are of the following types:
No active factor:

$$E(y) = \beta_0$$

One active factor:

$$E(y) = \beta_0 + \beta_1 x_i \quad {\scriptstyle i=1,\ldots,p}$$

Two active factors:

$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j + \beta_{12} x_i x_j \quad {\scriptstyle i=1,\ldots,p-1;\ j>i}$$

Three active factors:

$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j + \beta_3 x_k + \beta_{12} x_i x_j + \beta_{13} x_i x_k + \beta_{23} x_j x_k \quad {\scriptstyle i=1,\ldots,p-2;\ j>i;\ k>j}$$

We will in the following refer to these models as *submodels*. The motivation of using submodels is the principle of factor sparsity. By assuming no active factors, one active factor, two active factors and so on, we have the opportunity to perform a systematic search for every subset of potentially active factors. While Box and Meyer (1993) allow interactions to a specified order we here neglect three-factor and higher order interactions.

Computationally this "factor based" method is attractive due to the relatively few models that has to be examined compared to a "term based" method as best subsets regression. In Table 3 we have listed the number of submodels to examine for a given number of potentially active factors. However, this method also has some obvious limitations. Since the submodels contain all main effects and two-factor interactions for a given number of potentially active factors, the number of factors to examine is to some extent limited by the number of runs in the design. Generalised inverse can be used to find a solution to the normal equations if the matrix of regressor variables for a submodel does not have full rank. The active factors may

Table 3: The number of models to examine using submodel regression.

No. of factors	No. of terms	PB(12)	PB(24)	PB(48)
1	1	11	23	47
2	3	55	253	1 081
3	6	165	1 771	16 215
4	10	330	8 855	178 365
5	15	462	33 649	1.53e6
6	21	462	100 947	1.07e7

therefore be identified even in such situations. Nevertheless, as the number of terms get close to the number of runs in the design one might expect that also submodels containing inert factors will fit or even over-fit the data.

For the 12 run PB design all matrices of regressor variables for submodels with four active factors have full rank, and the 11 regression coefficients can be estimated with rather good precision, see Wang and Wu (1995). However, two of the runs in each matrix of regressor variables are always replicates of each other. For the first four columns in Table 1 we see that run 2 and 5 are replicates of each other. Thus, all submodels give perfect fit to the data except for the replicated runs where the fitted model goes through their mean. The judgement of what is the 'best' model is therefore based entirely on the two replicated runs. In addition, for each choice of four columns there is always one extra column having the same signs for the replicated runs. For example there are minus signs for both run 2 and 5 in column 6 in Table 1. Thus, at least $\binom{5}{4} = 5$ submodels have the same replicated runs and will give exactly the same residual mean square error. If only two of the 12 response values are similar to each other we will identify five submodels giving the same 'best' fit according to R^2, R^2_{adj} and s^2. The number of hypothesis is nevertheless considerably reduced since these five submodels together only contain five factors. If there are more response values rather similar to each other, there may be many submodels fitting the data almost equally well. In such situations it seems difficult to identify the active factors using submodel regression.

When searching for five active factors in a 12 run PB design the matrices of regressor variables for the submodels have 16 columns including the constant column. Hence generalised inverse has to be used. There are two types of projections into five columns for this design, see Lin and Draper (1992), and the matrices have rank equal to either 11 (one replicated run) or 12 (no replicated runs) depending on type of projection. Thus, all of the submodels having matrices of rank 12, where there are no replicated runs, will give perfect fit to the data. It therefore seems impossible to identify five active factors in the 12 run PB design using submodel regression. For PB designs where the number of runs is large relative to the number of terms in the matrices of regressor variables, use of submodels can nevertheless be very informative even if the matrices do not have full rank. We will illustrate this in Section 5 when searching for five active factors in a 24 run PB design.

In Figure 3 the three 'best' models according to the residual mean square error are listed for zero to four potentially active factors for the response in Table 1 corresponding to example 4 of Hamada and Wu (1992). Now only 562 models were examined. Also here the true active factors clearly are identified when three factors are included into the model. Adding three extra inert terms into this model seems to have little effect on the different criteria. Note

```
 Number of runs in design     :   12
 Number of factors in design  :   11
 Max number of active factors :    4
 Number of models examined    :  562
 Number of rank warnings      :    0
 "Best model" criteria        : Residual MSe

Active  R_sq  R_sqa      s    Terms in model
------------------------------------------
    0   0.00   0.00   3.22
    1  21.22  13.34   3.00    9
    1  20.23  12.26   3.02    7
    1  19.23  11.15   3.04    10
    2  71.46  60.75   2.02    8 9 8*9
    2  61.31  46.80   2.35    1 3 1*3
    2  59.97  44.96   2.39    1 2 1*2
    3  99.86  99.68   0.18    1 2 3 1*2 1*3 2*3
    3  93.92  86.62   1.18    8 9 10 8*9 8*10 9*10
    3  93.29  85.23   1.24    4 5 10 4*5 4*10 5*10
    4 100.00  99.98   0.04    7 8 9 10 7*8 7*9 7*10 8*9 8*10 9*10
    4 100.00  99.98   0.04    2 3 7 10 2*3 2*7 2*10 3*7 3*10 7*10
    4 100.00  99.98   0.04    1 8 9 10 1*8 1*9 1*10 8*9 8*10 9*10
elapsed_time = 2 seconds
```

Figure 3: Submodel regression for the model $y = x_1 + 2x_1x_2 + 2x_1x_3 + \epsilon$ with $\epsilon \sim N(0, 0.25^2)$ using the 12 run PB design.

however the problem with over-fitting when models with four factors are considered. There are actually ten models with four factors that experiences a value of $s = 0.04$. Only two of these models contain all three active factors. From Table 1 we see that the response values closest to each other are run 8 and 11 ($y = -1.02$ and $y = -1.08$) and run 9 and 12 ($y = 3.04$ and $y = 3.10$). The five columns having run 8 and 11 as repeated runs are columns 1,7,8,9 and 10, and the five columns having run 9 and 12 as repeated runs are columns 1,2,3,7 and 10. Thus there are $2 \times \binom{5}{4} = 10$ models having the same $s = 0.04$. Among these 10 models all three active factors will only appear in the two models containing the factors (x_1, x_2, x_3, x_7) and (x_1, x_2, x_3, x_{10}).

3.3 Heredity Least Term Regression

An alternative approach to best subsets regression is to guide the inclusion of terms into the models according to the *principle of effect heredity.* This principle states that in order to include an interaction at least one of the corresponding main effects also should be in the model. The notion 'effect heredity' was introduced by Hamada and Wu (1992). This principle has however been used for a long time in the analysis of experiments from fractional factorial designs to indicate the most rational models that explain the data, see for example the *Injection molding example* in Box, Hunter and Hunter (1978, p. 399).

We then use models of the type:
No term:

$$E(y) = \beta_0$$

One term:

$$E(y) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, p$$

Two terms:

$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j \quad i=1,\dots,p-1;\ j>i$$
$$E(y) = \beta_0 + \beta_1 x_i + \beta_{12} x_i x_j \quad i=1,\dots,p-1;\ j\neq i$$

Three terms:

$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j + \beta_3 x_k \quad i=1,\dots,p-2;\ j>i;\ k>j$$
$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j + \beta_{13} x_i x_k \quad i=1,\dots,p-1;\ j>i;\ k\neq i$$
$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j + \beta_{23} x_j x_k \quad i=1,\dots,p-1;\ j>i;\ k\neq (i,j)$$
$$E(y) = \beta_0 + \beta_1 x_i + \beta_{12} x_i x_j + \beta_{13} x_i x_k \quad i=1,\dots,p-1;\ j\neq i;\ k\neq (i,j)$$

Hamada and Wu (1992) argue for the principle of effect heredity by noting that it is difficult to provide a good physical interpretation of the x_1x_2 interaction if none of the main effects x_1 or x_2 are in the model. However, as pointed out by Meyer, Steinberg and Box (1996), effect heredity is not a principle that always can be relied on, especially if qualitative factors are present. Thus, effect heredity should not be forced into the models. It is therefore important to notice that the principle of effect heredity only guides the inclusion of terms into the models in the method described here. Anyhow it seems reasonable to give higher priority to the most rational models. Models that do not follow this principle can be found by including non-significant main effects into the model. For example is the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_{23} x_2 x_3$ not found among two-term models but rather among the three-term models as $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{23} x_2 x_3$ and $E(y) = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_{23} x_2 x_3$ with $\hat{\beta}_2$ and $\hat{\beta}_3$ close to zero. However, such models need extra terms, and this means that more models have to be examined compared to if effect heredity holds.

In Table 4 we have listed the number of models to examine for a given number of terms (in addition to the intercept). As is shown, the number of models to examine will be significantly reduced compared to best subsets regression. To our experience heredity least term regression performs very well in identifying the true models. Due to the reduced number of models to examine compared to best subsets regression this method may be a good alternative if an exhaustive search is preferred and there is a moderate number of factors.

In Figure 4 the three 'best' models according to the residual mean square error are listed for zero to four terms for the response in Table 1 corresponding to example 4 of Hamada and Wu (1992). Now only 17392 models were examined compared to the 768692 models examined by best subsets regression. Also here the true active factors clearly are identified when three terms are included into the model. As an illustration of identification of a model that does not follow the principle of effect heredity we can use the 'best' four-term model from best subsets

Table 4: The number of models to examine using heredity least term regression.

No. of terms	PB(12)	PB(24)	PB(48)
1	11	23	47
2	165	759	3 243
3	1 705	17 963	163 231
4	15 510	384 307	7.51 e6
5	125 202	7.51 e6	3.19 e8
6	902 649	1.35 e8	1.26 e10

```
 Number of runs in design     :     12
 Number of factors in design  :     11
 Max number of terms included :      4
 Number of models examined    : 17392
 Number of rank warnings      :      0
 "Best model" criteria        : Residual MSe

Terms   R_sq  R_sqa      s    Terms in model
--------------------------------------------
    0   0.00   0.00   3.22
    1  21.22  13.34   3.00    9
    1  20.23  12.26   3.02    7
    1  19.23  11.15   3.04    10
    2  56.99  47.43   2.34    1 1*3
    2  54.55  44.45   2.40    1 1*2
    2  53.67  43.38   2.43    9 9*8
    3  99.82  99.76   0.16    1 1*2 1*3
    3  82.93  76.52   1.56    7 7*6 7*11
    3  81.76  74.92   1.61    10 10*4 10*5
    4  99.88  99.81   0.14    1 11 1*2 1*3
    4  99.88  99.81   0.14    1 8 1*2 1*3
    4  99.87  99.79   0.15    1 1*2 1*3 1*11
elapsed_time = 36 seconds
```

Figure 4: Heredity least term regression for the model $y = x_1 + 2x_1x_2 + 2x_1x_3 + \epsilon$ with $\epsilon \sim N(0, 0.25^2)$ using the 12 run PB design.

regression. The model $\hat{y} = 0.02 + 1.10x_1 + 1.97x_1x_2 + 2.08x_1x_3 + 0.14x_3x_{10}$ with $s = 0.05$ is here found as the 'best' five-term model $\hat{y} = 0.02 + 1.10x_1 + 0.02x_3 + 1.98x_1x_2 + 2.08x_1x_3 + 0.14x_3x_{10}$ also with $s = 0.05$.

3.4 Forward Submodel Regression

We will here consider a method that to some extent combines the "factor based" strategy proposed by Box and Meyer (1993) and the "term based" strategy using forward regression proposed by Hamada and Wu (1992).

Both main effects and interactions are of interest in analysing experiments from non-geometric PB designs. The number of terms is then typically too large to use forward regression for choosing between all these terms simultaneously. Thus, it seems reasonable to somehow restrict the number of terms for which we perform forward regression. Hamada and Wu (1992) address this problem by first using a traditional analysis such as half-normal plot to identify the large main effect contrasts. Then forward regression is used to choose among these main effect contrasts and all two-factor interaction contrasts containing at least one factor corresponding to the main effects. To our experience however, there are two problems with this approach.

First, if two-factor interactions are active, inert main effect contrasts may wrongly appear as active and vice versa due to the complex alias pattern when using a traditional analysis method such as half-normal plot. Thus, there is a danger that focus initially is placed on inert terms, and the true model is therefore not found. Their example 4, see Figure 1, is an illustration of this problem. Secondly, considering all two-factor interactions containing at least one factor corresponding to the large main effects may easily result in a relatively large number of terms. As an example, with three large main effect contrasts, 27 additional two-factor interaction contrasts are to be considered by the forward procedure for the saturated 12 run PB design. As the number of terms grows large, the danger of having inert terms highly correlated with a 'part' of the response is increasing. As an illustration, let the true response be given by the model $E(y) = 2x_1+4x_3+2x_2x_3-2x_3x_4$ which is example 5 of Hamada and Wu (1992). Also assume that all main effect and two-factor interaction contrasts corresponding to the first four factors are considered by the forward procedure. The terms x_3 and x_3x_4 are first included into the model. However, now the correlation between the contrast column x_4x_5 and the signal left in the response $2x_1 + 2x_2x_3$, $\rho(x_4x_5, 2x_1 + 2x_2x_3)$, is equal to -0.58. This correlation is of the same size as $\rho(x_1, 2x_1 + 2x_2x_3) = 0.58$ and $\rho(x_2x_3, 2x_1 + 2x_2x_3) = 0.58$. Thus, considering too many terms when using forward regression increases the danger of choosing inert terms highly correlated with the response. This may also easily lead us to the wrong model.

We propose to use forward regression on each of the submodels used by Box and Meyer (1993). By using forward regression on each submodel we will get a systematic search for every subset of potentially active factors as in submodel regression described earlier. Opposed to 'traditional' forward regression, where the selection is among all possible terms and which results in only one promising model, we now get a number of models that later can be examined in more detail. In addition, if the assumption of factor sparsity holds we will always examine one submodel that contains all the active terms by specifying the number of factors to be included large enough.

Compared to the method by Hamada and Wu (1992) we avoid the problem of choosing terms to initially be considered by the forward procedure. Furthermore, within each submodel there are relatively few terms. In most of the non-geometric PB designs the correlations between contrast columns corresponding to main effects and two-factor interactions in a given submodel are rather small, see Table 7 to Table 10 in Appendix A. Thus, for the submodel containing the true active factors the danger of having inert terms that are highly correlated with the response is reduced. Compared to submodel regression we can also examine models with more potentially active factors since we do not fit the fully expanded models to the data. As a result, the problem with over-fitting will be reduced. Even though we here only consider main effects and two-factor interactions this method may also be well suited for examining models that allow inclusion of higher order interactions.

Compared to best subsets and heredity least term regression the number of models to examine will be reduced substantially. In Table 5 we have listed the maximum number of models that may have to be computed, i.e. the number of submodels multiplied by the number of terms in a submodel. In most situations the forward procedure terminates long before all terms in each submodel are included. Thus the number of models to examine is usually much smaller than the values listed.

Table 5: The maximum number of models to examine using forward submodel regression.

No. of factors	No. of terms	PB(12)	PB(24)	PB(48)
1	1	11	23	47
2	3	165	759	3 243
3	6	990	10 626	97 290
4	10	3 300	88 550	1.78e6
5	15	6 930	504 735	2.30e7
6	21	9 702	2.12e6	2.25e8

```
 Number of runs in design     :     12
 Number of factors in design  :     11
 Max number of active factors :      4
 F-enter, F-remove            :  4.00 0.00
 Increase in R_sq             :  0.00 0.00
 Number of submodels examined :   562
 Number of models examined    :  1072
 "Best model" criteria        : Residual MSe

Active   R_sq  R_sqa      s  Terms in model
-------------------------------------------
     0   0.00   0.00   3.22
     1  21.22  13.34   3.00  9
     1  20.23  12.26   3.02  7
     1  19.23  11.15   3.04  10
     2  71.46  60.75   2.02  8 9 8*9
     2  45.27  39.80   2.50  1*3
     2  42.83  37.12   2.56  1*2
     3  99.82  99.76   0.16  1 1*2 1*3
     3  82.93  76.52   1.56  7 6*7 7*11
     3  81.76  74.92   1.61  10 4*10 5*10
     4  99.98  99.97   0.05  1 1*2 1*3 3*10
     4  99.90  99.84   0.13  1 1*2 1*3 2*5
     4  99.82  99.76   0.16  1 1*2 1*3
elapsed_time = 7 seconds
```

Figure 5: Forward submodel regression for the model $y = x_1 + 2x_1x_2 + 2x_1x_3 + \epsilon$ with $\epsilon \sim N(0, 0.25^2)$ using the 12 run PB design. $F_{\text{ENTER}} = 4.0$.

Let

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_R + SS_E \\ SS_R &= SS(\beta_1|\beta_0) + SS(\beta_2|\beta_1, \beta_0) + \ldots \end{aligned}$$

Using standard forward regression, x_1 with largest $SS(\beta_1|\beta_0)$ is included into the model if

$$F_{\text{IN},1} = \frac{SS(\beta_1|\beta_0)/1}{SS_E(\beta_1, \beta_0)/(n-2)} \geq F_{\text{ENTER}}$$

and, given that x_1 is in the model, x_2 with largest $SS(\beta_2|\beta_1, \beta_0)$ is included into the model if

$$F_{\text{IN},2} = \frac{SS(\beta_2|\beta_1, \beta_0)/1}{SS_E(\beta_2, \beta_1, \beta_0)/(n-3)} \geq F_{\text{ENTER}}$$

In Figure 5 we have listed the four models with the smallest residual mean square error for zero to four active factors for the response in Table 1 corresponding to example 4 of Hamada and Wu (1992). A value of $F_{\text{ENTER}} = 4.0$ will here separate out the true model from the rest. Only 1072 models are now examined compared to 768692 and 17392 models for best subsets regression and heredity least term regression, respectively.

To our experience, use of $F_{\text{IN},i}$ will perform well in many situations and identifies the correct model. However, in some situations there seem to be problems with terminating the inclusion of variables at the correct model. This can result in a model with either too many or too few terms. We will demonstrate this in an example where use of $F_{\text{IN},i}$ has problem identifying the true model.

Table 6: Behaviour of the forward procedure for the model $y = x_1 + x_3 + x_2x_3 + x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.7^2)$ using the 12 run PB design.

Variables entered	s	R^2	$F_{\mathrm{IN},i}$	$\chi^2_{\mathrm{IN},i}$	ΔR_i^2
x_3	1.63	45.53	8.36	45.57	45.53
x_2x_3	1.53	57.27	2.47	11.74	11.73
x_1	1.24	74.74	5.53	17.49	17.47
x_3x_4	0.63	94.37	24.41	19.65	19.63
x_1x_4	0.49	97.02	5.35	2.66	2.65
x_1x_2	0.45	97.98	2.37	0.96	0.96
x_1x_3	0.48	98.14	0.34	0.16	0.16
x_2	0.48	98.57	0.91	0.43	0.43
x_2x_4	0.52	98.88	0.54	0.30	0.30
x_4	0.66	99.12	0.27	0.24	0.24

Let the response be simulated from the model $y = x_1 + x_3 + x_2x_3 + x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.7^2)$ using the 12 run PB design. A realization of this model is given by

$$y = (0.82\ \text{-}1.56\ 0.05\ 2.25\ \text{-}2.49\ 3.19\ 2.04\ 1.26\ \text{-}1.82\ 2.61\ \text{-}3.01\ \text{-}0.49)'.$$

Table 6 shows the performance of the forward procedure using $F_{\mathrm{IN},i}$ for the submodel containing the four true active factors. The term x_3 is first entered into the model with a value of $F_{\mathrm{IN},1} = 8.36$. This model accounts for 45.53% of the total variation, and the estimated variance is $s^2 = (1.63)^2$. In the next step the term x_2x_3 is included with $F_{\mathrm{IN},2} = 2.47$. The model with two terms accounts for 57.25% of the total variation, and the estimated variance is $s^2 = (1.53)^2$. Here F_{ENTER} has to be less or equal to 2.47 in order to include x_2x_3 and thereby to obtain the true model. However, then also the inert term x_1x_4 is included into the model before the forward procedure terminates.

Figure 6 shows the three models with the smallest residual mean square error for zero to four active factors using $F_{\mathrm{ENTER}} = 2.47$. Due to the 'low' value of F_{ENTER}, other models with relatively many terms are also found that seemingly fit the data well for four potentially active factors. The true model is here rated as number 16 among the four-factor models. In this situation it therefore seems natural to increase the value of F_{ENTER}. Consequently only x_3 is included for the submodel containing the four active factors, and the true model will not be identified.

Assume now that σ^2 is known. Then the inclusion of variables is determined by

$$\begin{aligned}
\chi^2_{\mathrm{IN},1} &= \frac{SS(\beta_1|\beta_0)/1}{\sigma^2} = \frac{SS_T}{\sigma^2}\frac{SS(\beta_1|\beta_0)/1}{SS_T} = CR_1^2 \\
\chi^2_{\mathrm{IN},2} &= \frac{SS(\beta_2|\beta_1,\beta_0)/1}{\sigma^2} = \frac{SS_T}{\sigma^2}\frac{SS(\beta_2|\beta_1,\beta_0)/1}{SS_T} = C(R_2^2 - R_1^2) \\
\chi^2_{\mathrm{IN},3} &= \frac{SS(\beta_3|\beta_2,\beta_1,\beta_0)/1}{\sigma^2} = \frac{SS_T}{\sigma^2}\frac{SS(\beta_3|\beta_2,\beta_1,\beta_0)/1}{SS_T} = C(R_3^2 - R_2^2)
\end{aligned}$$

which all have a χ^2-distribution with 1 degree of freedom. Note that the proportionality constant C only depends on SS_T and σ^2, and is therefore independent of which of the submodels we are examining. Since σ^2 is unknown we can not find C, and we can therefore not perform a formal test of significance. However, the increase in the coefficient of multiple determination $\Delta R_i^2 = R_i^2 - R_{i-1}^2$ will still be very helpful in identifying the true model. Our main intention is here to compare a number of models and identify those models that are 'better' than the others according to some criterion. Even if σ^2 was known the significance level would be varied to 'filter' out the most promising models. Hence, by varying the level for $\Delta R^2_{\mathrm{ENTER}}$ we achieve exactly the same filtering as if σ^2 was known.

Use of the coefficient of multiple determination is a well-known tool for variable selection. If there are k terms, $\binom{k}{i}$ values of R_i^2 are computed for every number of terms i in the model, and $\max R_i^2$ for every i is plotted versus i. This plot typically shows a curve that initially increases rapidly and then remains flat. The number of terms i to include into the model is then subjectively identified to be where the 'knee' occurs. However, use of the increase in the coefficient of multiple determination as a criterion in forward regression is to our knowledge not known.

In Table 6 we have also listed the values of $\chi^2_{\mathrm{IN},i}$ that we would get if it was known in advance that $\sigma^2 = 0.7^2$. Note the large differences between $F_{\mathrm{IN},i}$ and $\chi^2_{\mathrm{IN},i}$ for the first variables entered into this model. This is due to over-estimation of σ^2 when too few terms are entered into the model. $F_{\mathrm{IN},i}$ is therefore less than what we would get if a reliable independent estimate of σ^2 was available. F_{ENTER} must be specified 'small' in order to include x_2x_3 the first time. When the true model is found there is a danger that F_{ENTER} is specified too 'small', and the inclusion of variables does therefore not terminate at the correct model. In addition, when F_{ENTER} is specified too 'small' there is a danger that other models with many terms seem to fit the response better than the true model. The values of ΔR_i^2 are also listed in Table 6. These values are in general only proportional to $\chi^2_{\mathrm{IN},i}$. The close correspondence between $\chi^2_{\mathrm{IN},i}$ and ΔR_i^2 for this specific response comes from the fact that $SS_T = 49.04$ which is approximately $100\sigma^2$.

Figure 7 shows the three models with the smallest residual mean square error for zero to four active factors. Instead of using $F_{\mathrm{IN},i}$ as in Figure 6 we here rather use ΔR_i^2 and a required increase in R^2, $\Delta R^2_{\mathrm{ENTER}}$, of 10%. Then the true model, with the correct number of terms, is rated as number 2 for four active factors. In both these models all terms are significant at 1.0% level. Both models follow the principle of effects heredity. However, one of the models has two extra terms compared to the other. If we use $\Delta R^2_{\mathrm{ENTER}} = 5$ there are 22 models with lower residual mean square error than the true model. Most of these models have 6 or 7 terms. There are two other models having only four terms, but none of them follow the principle of effect heredity. If we use $\Delta R^2_{\mathrm{ENTER}} = 7.5$ there are 5 models with lower residual mean square error than the true model.

We have here focused on forward regression. Stepwise regression can also be used, and the implemented program has this possibility. To our experience, there are very few if any of the terms at all that is removed from the models at least for a moderate number of potentially active factors for the 12 and 24 run PB designs. This is probably due to the low correlation between the different contrast columns corresponding to main effects and two-factor interactions in the submodels of these designs. One advantage by using forward compared to stepwise regression is of course the reduced computing time. For forward regression there is no need for recomputing any of the results even if another choice of F_{ENTER} or $\Delta R^2_{\mathrm{ENTER}}$ is varied. For stepwise regression one has to recompute most of the results.

```
 Number of runs in design     :    12
 Number of factors in design  :    11
 Max number of active factors :     4
 F-enter, F-remove            :  2.47  0.00
 Increase in R_sq             :  0.00  0.00
 Number of submodels examined :   562
 Number of models examined    :  1773
 "Best model" criteria        : Residual MSe

Active   R_sq  R_sqa      s  Terms in model
-------------------------------------------
     0   0.00   0.00   2.11
     1  45.53  40.09   1.63  3
     1  21.50  13.66   1.96  11
     1  10.00   1.00   2.10  6
     2  67.04  59.71   1.34  3 11
     2  54.99  50.49   1.49  7*9
     2  59.33  50.29   1.49  3 3*9
     3  89.35  85.36   0.81  3 3*9 7*9
     3  91.84  85.05   0.82  2 3 2*3 2*10 3*10
     3  86.58  78.91   0.97  3 6 11 3*11
     4  99.89  99.69   0.12  3 8 9 3*8 3*9 7*9 8*9
     4  99.94  99.64   0.13  3 5 6 8 3*5 3*6 3*8 5*6 6*8
     4  99.96  99.51   0.15  3 4 5 6 3*4 3*5 3*6 4*5 4*6 5*6
elapsed_time = 12 seconds
```

Figure 6: Forward submodel regression for the model $y = x_1 + x_3 + x_2x_3 + x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.7^2)$ using the 12 run PB design. $F_{\text{ENTER}} = 2.47$.

```
 Number of runs in design     :    12
 Number of factors in design  :    11
 Max number of active factors :     4
 F-enter, F-remove            :  0.00  0.00
 Increase in R_sq             : 10.00  0.00
 Number of submodels examined :   562
 Number of models examined    :  1694
 "Best model" criteria        : Residual MSe

Active   R_sq  R_sqa      s  Terms in model
-------------------------------------------
     0   0.00   0.00   2.11
     1  45.53  40.09   1.63  3
     1  21.50  13.66   1.96  11
     1  10.00   1.00   2.10  6
     2  67.04  59.71   1.34  3 11
     2  54.99  50.49   1.49  7*9
     2  59.33  50.29   1.49  3 3*9
     3  89.35  85.36   0.81  3 3*9 7*9
     3  82.89  76.47   1.02  3 11 4*11
     3  78.43  70.34   1.15  3 11 3*5
     4  97.12  93.66   0.53  4 6 7 4*7 6*7 6*8
     4  94.37  91.15   0.63  1 3 2*3 3*4
     4  89.35  85.36   0.81  3 3*9 7*9
elapsed_time = 11 seconds
```

Figure 7: Forward submodel regression for the model $y = x_1 + x_3 + x_2x_3 + x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.7^2)$ using the 12 run PB design. $\Delta R^2_{\text{ENTER}} = 10$.

4 Characterisations of the Non-Geometric PB Designs

Due to the complex alias patterns of the non-geometric PB designs it seems difficult to give a short and complete description of how good these designs are for identifying the active factors in given situations. A few attempts are, however, done characterising these designs. Lin and Draper use the projection properties (Lin and Draper, 1992; 1995) and repeat and mirror-image runs (Draper and Lin, 1995) as a general characterisation of the designs. As a more specific description of the designs ability to simultaneously estimate all main effects and a given number of two-factor interactions Wang and Wu (1995) use the D and D_s efficiency. These measures are defined as $D = |\boldsymbol{Z}'\boldsymbol{Z}|^{1/k}$ and $D_s = [\boldsymbol{z}_i'\boldsymbol{z}_i - \boldsymbol{z}_i'\boldsymbol{Z}_{(i)}(\boldsymbol{Z}_{(i)}'\boldsymbol{Z}_{(i)})^{-1}\boldsymbol{Z}_{(i)}'\boldsymbol{z}_i]/\boldsymbol{z}_i'\boldsymbol{z}_i$ where $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k]$ is a matrix of regressor variables with the columns scaled to unit length, and where $\boldsymbol{Z}_{(i)}$ is $\boldsymbol{Z}$ with column $\boldsymbol{z}_i$ deleted.

In this chapter we try to characterise some of the non-geometric PB designs with special focus on the variable selection techniques presented in Section 3. All tables referred are given in Appendix A. In Table 7 we first give an overview over the size and percentage of correlations that appear in saturated non-geometric PB design with 12 to 60 runs. Here the % cases of all correlations, $|\frac{1}{n}\sum_k x_{ik}x_{jk}|$, between main effect and two-factor interaction columns and between two-factor interaction columns are listed. Thus, for the 12 run PB design 18.18% of the $11 \times 55 = 605$ correlations between main effects and two-factor interactions and 33.33% of the $\binom{55}{2} = 1485$ correlations between two-factor interactions equal 0.0.

Table 7 does not give any good characterisation of the designs when using submodels under the principle of factor sparsity described in Section 3.2 and 3.4. In Table 8 to Table 10 we therefore focus on the use of submodels with 2 to 7 potentially active factors for the 12, 20 and 24 run saturated PB designs, respectively. For a given number of potentially active factors the matrices of regressor variables $\boldsymbol{X}$ (including the constant column, all main effect columns and the corresponding two-factor interaction columns) are generated for all submodels. For each submodel we first find the correlation, $|\frac{1}{n}\sum_k x_{ik}x_{jk}|$, between all columns in $\boldsymbol{X}$. If $\boldsymbol{X}$ has full rank the *variance inflation factors* (Marquardt, 1970) related to the individual regression coefficients are computed. Let $\boldsymbol{Z}$ be a matrix of regressor variables with all columns scaled to unit length. Then the variance inflation factor related to the j-th regression coefficient is given by $VIF_j = C_{jj}$ where $\boldsymbol{C} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Thus, VIF is a measure of loss in precision in the estimation of the regression coefficients due to non-orthogonality in the design. Since $C_{jj} = 1/(1-R_j^2)$ where R_j^2 is the multiple coefficient of determination when regressing column j on the other columns, see Montgomery and Peck (1982, p. 292), the VIF measure is here equivalent to the D_s efficiency measure used by Wang and Wu (1995); i.e. $VIF = 1/D_s$. Montgomery and Peck (1982) point out that if any of the VIF-measures in a matrix of regressor variables exceeds 5 or 10 this should be taken as an indication of poor estimation of the coefficients due to multicollinearity. As pointed out by Draper (1985) different choice of factors may have projections of different types. For example, there are 2 types of projections into both 5 and 6 factors for the 12 run PB design. Table 8 to Table 10 give a summary of the above mentioned measures over all submodels irrespectively type of projection. Thus, for the 12 run PB design there are 165 submodels with potentially 3 active factors. Of the $165 \times \binom{7}{2} = 3465$ correlations between the 7 columns in the matrices of regressor variables 85.7% equal 0.0. All of the submodels have full rank, and the maximum VIF found in any of the submodels is 1.12.

From Table 8 to Table 10 we see that for both three and four active factors all submodels have matrices of regressor variables with full rank. In addition, all coefficients in the submodels can be estimated with rather good efficiency compared to VIF=1.0 for an orthogonal design. Full rank and low values of VIF gives however no guaranty of identifying the active factors. Hynén (1996) presented an example where the response is simulated from the model $E(y) = x_1 + x_2 + x_3 + 0.5(x_1x_2 + x_1x_3 + x_2x_3)$ using a 12 run PB design. If no noise is present there are four submodels with three factors, together containing 9 of the 11 factors, that fit this response equally well, see Tyssedal and Samset (1997). Thus, the 12 run PB design does not necessarily enable identification of the true model and thereby the active factors even if VIF=1.12 and the parameters hence are estimated with good efficiency. In this situation additional experiments are required. Contrary, when searching for five active factors in the 20 and 24 run PB designs there are 32.4% and 9.8% of the possible submodels, respectively, that do not have full rank. As mentioned before the use of generalised inverse can be helpful for identifying the active factors even if the matrices of regressor variables do not have full rank and thus have VIF equal to infinity.

As for submodel regression we try in Table 11 to Table 13 to give a characterisation of the 12, 20 and 24 run saturated PB designs when using heredity least term regression. Note that for models including up to six terms all possible $\boldsymbol{X}$-matrices have full rank, and that in most situations there are only moderate loss of efficiency in the estimation of the parameters due to non-orthogonality.

5 Further Examples

5.1 Case 1

In this section we will consider example 5 from Hamada and Wu (1992). The response is simulated from the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design. We use a realization from this model given by Hynén (1996)

$$y = (6.27 \text{ -3.21 } 5.51 \text{ } 1.94 \text{ -2.48 } 10.35 \text{ } 2.47 \text{ -2.84 -2.93 -1.19 -10.09 -5.98})'.$$

Both Hamada and Wu using their proposed forward selection technique, and Hynén using the Bayesian method of Box and Meyer (1993), fail to identify the active factors in this example, and indicate that additional experiments are needed.

In Figure 8 to Figure 11 we have listed the three 'best' models with the smallest residual mean square error for zero to four terms or potentially active factors for the four variable selection techniques presented earlier. From best subsets regression we see that this model is clearly identified using four terms. Thus, additional runs are not required to identify this model when we are using the 12 run PB design. Also heredity least term and forward submodel regression identify the true model using considerably less computer time than best subsets regression. Only submodel regression has problem identifying the true model. There are five submodels with four potentially active factors that seemingly gives an almost perfect fit to the data with $s = 0.06$. The model that contains the true active factors are listed as number 35 among the four-term models with $s = 0.52$.

```
 Number of runs in design     :     12
 Number of factors in design  :     11
 Max number of terms          :      4
 Number of models examined    : 768692
 Number of rank warnings      :      0
 "Best model" criteria        : Residual MSe

Terms   R_sq  R_sqa      s    Terms in model
-------------------------------------------------------
    0   0.00   0.00   5.71
    1  57.16  52.88   3.92    3
    1  45.10  39.61   4.44    6*11
    1  41.18  35.30   4.59    8*10
    2  77.63  72.66   2.98    5*9 8*10
    2  76.99  71.88   3.03    3 3*4
    2  76.97  71.85   3.03    3 6*11
    3  93.68  91.31   1.68    5*9 8*10 10*11
    3  92.95  90.31   1.78    3 3*4 7*9
    3  92.09  89.13   1.88    3 7 6*11
    4  99.57  99.32   0.47    1 3 2*3 3*4
    4  97.01  95.30   1.24    8 5*9 8*10 10*11
    4  96.99  95.27   1.24    1*3 5*9 8*10 10*11
elapsed_time = 1711 seconds
```

Figure 8: Best subsets regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design.

```
 Number of runs in design     :   12
 Number of factors in design  :   11
 Max number of active factors :    4
 Number of models examined    :  562
 Number of rank warnings      :    0
 "Best model" criteria        : Residual MSe

Active  R_sq  R_sqa      s    Terms in model
-------------------------------------------------------
    0   0.00   0.00   5.71
    1  57.16  52.88   3.92    3
    1  15.17   6.69   5.51    1
    1   7.19   0.00   5.77    10
    2  80.30  72.91   2.97    3 4 3*4
    2  72.37  62.01   3.52    1 3 1*3
    2  67.36  55.12   3.82    3 10 3*10
    3  94.30  87.46   2.02    3 4 5 3*4 3*5 4*5
    3  93.75  86.26   2.12    1 2 3 1*2 1*3 2*3
    3  92.75  84.04   2.28    3 4 11 3*4 3*11 4*11
    4 100.00  99.99   0.06    2 4 5 8 2*4 2*5 2*8 4*5 4*8 5*8
    4 100.00  99.99   0.06    1 4 5 8 1*4 1*5 1*8 4*5 4*8 5*8
    4 100.00  99.99   0.06    1 2 5 8 1*2 1*5 1*8 2*5 2*8 5*8
elapsed_time = 2 seconds
```

Figure 9: Submodel regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design.

```
 Number of runs in design     :    12
 Number of factors in design  :    11
 Max number of terms included :     4
 Number of models examined    : 17392
 Number of rank warnings      :     0
 "Best model" criteria        : Residual MSe

Terms   R_sq  R_sqa      s    Terms in model
---------------------------------------------------------
    0   0.00   0.00   5.71
    1  57.16  52.88   3.92    3
    1  15.17   6.69   5.51    1
    1   7.19   0.00   5.77    10
    2  76.99  71.88   3.03    3 3*4
    2  72.33  66.18   3.32    1 3
    2  65.24  57.52   3.72    3 3*2
    3  91.62  88.48   1.94    1 3 3*2
    3  85.07  79.47   2.59    3 3*2 3*4
    3  83.53  77.35   2.72    1 3 3*4
    4  99.57  99.32   0.47    1 3 3*2 3*4
    4  94.24  90.95   1.72    1 3 1*10 3*2
    4  94.06  90.66   1.74    1 3 1*6 3*2
elapsed_time = 40 seconds
```

Figure 10: Heredity least term regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design.

```
 Number of runs in design     :    12
 Number of factors in design  :    11
 Max number of active factors :     4
 F-enter, F-remove            :  0.00  0.00
 Increase in R_sq             :  5.00  0.00
 Number of submodels examined :   562
 Number of models examined    :  2565
 "Best model" criteria        : Residual MSe

Active   R_sq  R_sqa      s  Terms in model
---------------------------------------------------
     0   0.00   0.00   5.71
     1  57.16  52.88   3.92  3
     1  15.17   6.69   5.51  1
     1   7.19   0.00   5.77  10
     2  76.99  71.88   3.03  3 3*4
     2  72.33  66.18   3.32  1 3
     2  65.24  57.52   3.72  3 2*3
     3  91.62  88.48   1.94  1 3 2*3
     3  85.59  80.18   2.54  3 3*4 4*11
     3  85.07  79.47   2.59  3 2*3 3*4
     4  99.57  99.32   0.47  1 3 2*3 3*4
     4  99.57  99.06   0.55  1 8 1*4 1*8 2*4 2*8
     4  98.44  96.58   1.06  3 5 10 3*10 3*11 5*10
elapsed_time = 15 seconds
```

Figure 11: Forward submodel regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design. $\Delta R^2_{\mathrm{ENTER}} = 5$

5.2 Case 2

In this section we reconsider example 5 from Hamada and Wu (1992) but expand this model by one extra term so that there now are five active factors. The response is simulated from the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 - 2x_3x_5 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design. The following realization from this model is used

$$y = (\ 8.58\ \text{-}3.69\ 4.04\ 4.18\ \text{-}0.35\ 8.85\ 4.03\ \text{-}3.10\ 0.13\ 0.44\ \text{-}12.72\ \text{-}8.35)'.$$

In Figure 12 to Figure 14 we show the result from three of the variable selection techniques. Submodel regression gives very little information in this example and is therefore not included. For best subsets regression there are 16 models with five terms having smaller value of s than the true model. This will of course make it more difficult to identify the active factors. However, 8 of these 16 models contain only two-factor interactions. The other 8 models contain from 6 to 8 factors and do not follow the principle of heredity. Both heredity least term regression and forward submodel regression identify the true model as one of the most plausible using considerably less computer time than best subsets regression.

```
 Number of runs in design      :      12
 Number of factors in design   :      11
 Max number of terms           :       5
 Number of models examined     : 9705620
 Number of rank warnings       :       0
 "Best model" criteria         : Residual MSe

Terms   R_sq  R_sqa      s    Terms in model
--------------------------------------------------------
    0   0.00   0.00   6.44
    1  47.68  42.45   4.89    3
    1  46.35  40.98   4.95    8*10
    1  44.06  38.47   5.05    6*11
    2  85.01  81.68   2.76    5*9 8*10
    2  76.43  71.19   3.46    2*11 6*11
    2  71.16  64.75   3.83    3 7*8
    3  95.30  93.54   1.64    8 5*9 8*10
    3  91.51  88.33   2.20    4*6 5*9 8*10
    3  91.40  88.17   2.22    1*9 5*9 8*10
    4  98.14  97.08   1.10    8 4*6 5*9 8*10
    4  97.51  96.08   1.28    1 8 5*9 8*10
    4  97.33  95.80   1.32    3 8 5*9 8*10
    5  99.91  99.84   0.26    3*6 3*9 5*9 7*11 8*10
    5  99.91  99.83   0.27    2*7 2*9 5*9 6*11 8*10
    5  99.86  99.75   0.32    5*8 5*9 6*11 7*10 8*10
elapsed_time = 23013 seconds
```

Figure 12: Best subsets regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 - 2x_3x_5 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design.

```
 Number of runs in design      :      12
 Number of factors in design   :      11
 Max number of terms included :       5
 Number of models examined     : 142594
 Number of rank warnings       :       0
 "Best model" criteria         : Residual MSe

Terms   R_sq  R_sqa      s    Terms in model
-------------------------------------------------------
    0   0.00   0.00   6.44
    1  47.68  42.45   4.89    3
    1  21.07  13.18   6.00    1
    1  11.42   2.56   6.36    9
    2  68.75  61.81   3.98    1 3
    2  67.65  60.46   4.05    3 3*4
    2  66.34  58.86   4.13    3 3*5
    3  86.31  81.18   2.80    3 3*4 3*5
    3  84.35  78.48   2.99    1 3 3*2
    3  80.81  73.62   3.31    3 10 10*8
    4  95.89  93.55   1.64    5 8 5*9 8*10
    4  95.37  92.72   1.74    8 9 8*10 9*5
    4  92.32  87.93   2.24    3 11 3*4 3*5
    5  99.70  99.46   0.47    1 3 3*2 3*4 3*5
    5  98.82  97.84   0.95    1 5 8 5*9 8*10
    5  98.40  97.06   1.10    1 3 10 1*3 10*8
elapsed_time = 339 seconds
```

Figure 13: Heredity least term regression for the model $y = 2x_1+4x_3+2x_2x_3-2x_3x_4-2x_3x_5+\epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design.

```
 Number of runs in design      :    12
 Number of factors in design   :    11
 Max number of active factors :      5
 F-enter, F-remove             :  0.00  0.00
 Increase in R_sq              :  6.00  0.00
 Number of submodels examined :  1024
 Number of models examined     :  4749
 "Best model" criteria         : Residual MSe

Active   R_sq  R_sqa      s  Terms in model
---------------------------------------------------
     0   0.00   0.00   6.44
     1  47.68  42.45   4.89  3
     1  21.07  13.18   6.00  1
     1  11.42   2.56   6.36  9
     2  68.75  61.81   3.98  1 3
     2  67.65  60.46   4.05  3 3*4
     2  66.34  58.86   4.13  3 3*5
     3  90.73  85.44   2.46  1 3 9 1*9
     3  86.31  81.18   2.80  3 3*4 3*5
     3  84.35  78.48   2.99  1 3 2*3
     4  98.40  97.06   1.10  1 3 10 1*3 8*10
     4  98.31  96.27   1.24  3 6 8 3*4 3*8 6*8
     4  98.03  95.67   1.34  8 4*6 4*11 6*8 6*11 8*11
     5  99.93  99.80   0.29  1 4 8 11 2*4 2*8 2*11
     5  99.70  99.46   0.47  1 3 2*3 3*4 3*5
     5  98.40  97.06   1.10  1 3 10 1*3 8*10
elapsed_time = 30 seconds
```

Figure 14: Forward submodel regression for the model $y = 2x_1+4x_3+2x_2x_3-2x_3x_4-2x_3x_5+\epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 12 run PB design. $\Delta R^2_{\mathrm{ENTER}} = 6$

5.3 Case 3

We here reconsider the example with five active factors given in the previous section using a 24 run instead of the 12 run PB design. The PB design is constructed by choosing the generator given by Plackett and Burman (1946)

$$+ + + + + - + - + + - - + + - - + - + - - - -$$

as the $n - 1$ first rows in the first column. Cycling downward, and finally adding a last row of minus signs give the other columns. The following realization of the response is used

$$y = (\text{-3.42 -7.69 12.04 8.18 3.65 0.85 0.03 -3.10 4.13 -3.56 3.28 -0.35} \\ \text{4.62 -4.32 8.29 -0.18 3.93 -8.67 7.36 -7.51 -0.02 -4.40 -4.38 -7.57})'.$$

Both best subsets regression and heredity least terms regression require here too many models to be examined, and we therefore only include the results from submodel and forward submodel regression, see Figure 15 and Figure 16. For all matrices of regressor variables that do not have full rank we have used the Moore-Penrose pseudo-inverse to find a solution to the normal equations in submodel regression. None of the models listed in Figure 15 have any rank problems. Again the true model is identified as the most plausible within reasonable computer time.

The factors (1,2,3,6 and 7) is one choice of five factors where the matrix of regression variables for the submodel does not have full rank. We have also generated data from the same model as before but with factors 4 and 5 now replaced with 6 and 7, i.e. $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_6 - 2x_3x_7 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 24 run PB design. The realization of the response is given by

$$y = (\text{-0.51 -3.84 12.76 12.37 11.75 4.44 -0.12 -4.36 -0.22 -0.31 3.90 0.28} \\ \text{-0.53 -3.82 0.57 0.08 -3.65 -8.03 1.01 -7.54 3.09 -3.98 -4.90 -7.49})'.$$

The results from submodel regression is given in Figure 17. Only one of the models listed in the figure has rank problem. This model is marked with (rp). Also in this situation the use of submodel regression will identify the active factors.

```
 Number of runs in design     :   24
 Number of factors in design  :   23
 Max number of active factors :    5
 Number of models examined    : 44552
 Number of rank warnings      : 3289
 "Best model" criteria        : Residual MSe

Active  R_sq  R_sqa      s    Terms in model
-------------------------------------------------------
    0   0.00   0.00   5.72
    1  48.22  45.86   4.21    3
    1  13.13   9.19   5.45    1
    1   6.24   1.98   5.66    13
    2  61.52  55.75   3.81    3 4 3*4
    2  61.50  55.72   3.81    2 3 2*3
    2  61.37  55.57   3.81    1 3 1*3
    3  86.14  81.24   2.48    3 13 21 3*13 3*21 13*21
    3  80.56  73.70   2.93    3 4 5 3*4 3*5 4*5
    3  79.79  72.66   2.99    2 3 5 2*3 2*5 3*5
    4  94.08  89.52   1.85    3 13 19 21 3*13 3*19 3*21 13*19 13*21 19*21
    4  93.32  88.18   1.97    2 3 13 21 2*3 2*13 2*21 3*13 3*21 13*21
    4  93.11  87.82   2.00    3 13 15 21 3*13 3*15 3*21 13*15 13*21 15*21
    5  99.84  99.55   0.38    1 2 3 4 5 ...
    5  97.85  93.81   1.42    3 7 13 15 21 ...
    5  97.80  93.67   1.44    3 13 19 21 23 ...
elapsed_time = 400
```

Figure 15: Submodel regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 - 2x_3x_5 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 24 run PB design.

```
 Number of runs in design     :     24
 Number of factors in design  :     23
 Max number of active factors :      5
 F-enter, F-remove            :   0.00  0.00
 Increase in R_sq             :  10.00  0.00
 Number of submodels examined :  44552
 Number of models examined    : 117180
 "Best model" criteria        : Residual MSe

Active   R_sq  R_sqa      s  Terms in model
---------------------------------------------------
     0   0.00   0.00   5.72
     1  48.21  45.85   4.21  3
     1  13.15   9.20   5.45  1
     1   6.24   1.98   5.66  13
     2  61.47  57.80   3.72  3 2*3
     2  61.46  57.79   3.72  3 3*4
     2  61.35  57.67   3.72  1 3
     3  74.75  70.96   3.08  3 3*4 4*9
     3  74.73  70.93   3.08  3 2*3 3*4
     3  74.62  70.81   3.09  1 3 2*3
     4  87.87  85.32   2.19  1 3 2*3 3*4
     4  86.26  83.36   2.33  3 2*3 3*4 3*5
     4  86.15  83.23   2.34  1 3 2*3 3*5
     5  99.40  99.24   0.50  1 3 2*3 3*4 3*5
     5  87.87  85.32   2.19  1 3 2*3 3*4
     5  86.26  83.36   2.33  3 2*3 3*4 3*5
elapsed_time = 821 seconds
```

Figure 16: Forward submodel regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_4 - 2x_3x_5 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 24 run PB design.$\Delta R^2_{\mathrm{ENTER}} = 10$

```
 Number of runs in design     :   24
 Number of factors in design  :   23
 Max number of active factors :    5
 Number of models examined    : 44552
 Number of rank warnings      : 3289
 "Best model" criteria        : Residual MSe

Active  R_sq  R_sqa      s    Terms in model
--------------------------------------------------------
    0   0.00   0.00   5.82
    1  52.26  50.09   4.11    3
    1  12.75   8.78   5.56    1
    1   6.49   2.23   5.76    2
    2  71.32  67.02   3.34    2 3 2*3
    2  69.30  64.69   3.46    3 7 3*7
    2  68.20  63.43   3.52    3 6 3*6
    3  89.23  85.43   2.22    1 2 3 1*2 1*3 2*3
    3  87.65  83.30   2.38    1 3 7 1*3 1*7 3*7
    3  87.37  82.91   2.41    1 3 6 1*3 1*6 3*6
    4  95.41  91.88   1.66    1 2 3 12 1*2 1*3 1*12 2*3 2*12 3*12
    4  95.14  91.40   1.71    1 2 3 18 1*2 1*3 1*18 2*3 2*18 3*18
    4  94.86  90.91   1.76    1 2 3 19 1*2 1*3 1*19 2*3 2*19 3*19
    5  99.71  99.25   0.51    1 2 3 6 7 ... (rp)
    5  98.90  96.85   1.03    1 2 3 12 16 ...
    5  98.56  95.87   1.18    1 3 5 6 18 ...
elapsed_time = 389
```

Figure 17: Submodel regression for the model $y = 2x_1 + 4x_3 + 2x_2x_3 - 2x_3x_6 - 2x_3x_7 + \epsilon$ with $\epsilon \sim N(0, 0.5^2)$ using the 24 run PB design.

6 Summary and Conclusions

In this paper we present and compare some variable selection techniques that we have found useful in analysing experiments from non-geometric Plackett-Burman designs in the presence of active two-factor interactions. The techniques are *heredity least term regression*, *submodel regression* and *forward submodel regression*.

If interactions are present, the complex alias patterns of the non-geometric Plackett-Burman designs often entail a more thorough analysis of the experiments in order to identify the active factors. However, when both main effects and two-factor interactions are taken into account, use of best subsets regression may require a large number of models to be examined. By guiding the inclusion of terms according to the principle of effect heredity one gives higher priority to models that often are encountered to give more plausible explanations of the variation in the response. Thus, models that follow this principle can be found with considerably less effort than using best subsets regression. Models that not follow this principle may also be found at the expense of including more terms and thereby increased computer time.

Use of the 'factor based' approach proposed by Box and Meyer (1993) gives a systematically and efficient search for the active factors. For each model we only include all main effects and two-factor interactions for a given number of potentially active factors. Instead of the Bayesian approach by Box and Meyer we rather regress each submodel to the response and present the results as in best subsets regression. This technique will in many situations be helpful identifying the active factors, and due to the limited computer time compared to the other techniques one should consider using this method preliminary to the others. There are situations however, when many of the models from submodel regression seem to over-fit the data due to too many terms, and therefore makes it difficult to separate out a few plausible models. In such situations we propose to use forward regression on each of the submodels. Instead of using the traditionally F_{IN} we rather use the increase in the multiple coefficient of determination R^2 to 'filter' out the most plausible models. One thereby achieves the same filtering as if the standard deviation of the error was known.

We have used the different techniques with promising results for identifying from three to five active factors in the 12 run Plackett-Burman design, and for identifying five active factors in the 24 run Plackett-Burman design in simulated examples. In most of the examples the performances of the techniques are compared to the performance of *best subsets regression*. These examples also illustrate the large potential of the non-geometric Plackett-Burman designs as screening designs.

A Characterisation of Some Non-Geometric PB Designs

Table 7: Correlation structure of some non-geometric PB designs.

Run	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ between main and two-factor (% cases) two-factor and two-factor (% cases)				
12	0.00	0.33			
	(18.18)	(81.82)			
	(33.33)	(66.67)			
20	0.00	0.20	0.60		
	(10.53)	(84.21)	(5.26)		
	(20.00)	(75.29)	(4.71)		
24	0.00	0.33			
	(60.87)	(39.13)			
	(64.29)	(35.71)			
28	0.00	0.14	0.43		
	(7.41)	(81.48)	(11.11)		
	(14.29)	(75.43)	(10.29)		
36	0.00	0.11	0.33		
	(5.71)	(77.15)	(17.14)		
	(11.11)	(72.73)	(16.16)		
40	0.00	0.20	0.60	1.00	
	(59.57)	(37.87)	(2.37)	(0.20)	
	(61.64)	(35.93)	(2.25)	(0.19)	
44	0.00	0.09	0.27		
	(4.65)	(72.09)	(23.26)		
	(9.09)	(68.74)	(22.17)		
48	0.00	0.17	0.33		
	(42.55)	(51.06)	(6.38)		
	(45.00)	(48.89)	(6.11)		
52	0.00	0.08	0.23	0.38	0.85
	(3.92)	(75.66)	(16.60)	(3.70)	(0.12)
	(7.69)	(72.69)	(15.96)	(3.55)	(0.11)
56	0.00	0.14	0.43	1.00	
	(56.93)	(37.82)	(5.16)	(0.10)	
	(58.49)	(36.44)	(4.97)	(0.10)	
60	0.00	0.07	0.20		
	(3.39)	(61.02)	(35.59)		
	(6.67)	(58.95)	(34.38)		

Table 8: Correlation structure in submodels for the 12 run PB design.

No. active	No. submodels	Rank problems	max VIF	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ 0.00	0.33
2	55	0	1.00	100.00	
3	165	0	1.12	85.71	14.29
4	330	0	1.63	72.73	27.27
5	462	462	-	62.50	37.50

Table 9: Correlation structure in submodels for the 20 run PB design.

No. active	No. submodels	Rank problems	max VIF	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ 0.00	0.20	0.60
2	171	0	1.00	100.00		
3	969	0	1.56	85.71	13.45	0.84
4	3876	0	1.89	72.73	25.67	1.60
5	11628	3762	5.03	62.50	35.51	1.99
6	27132	27132	-	54.55	43.57	1.88

Table 10: Correlation structure in submodels for the 24 run PB design.

No. active	No. submodels	Rank problems	max VIF	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ 0.00	0.33
2	253	0	1.00	100.00	
3	1771	0	1.13	93.88	6.12
4	8855	0	1.50	89.87	10.13
5	33649	3289	5.10	87.50	12.50
6	100947	51359	241.25	80.52	19.48
7	245157	245157	-	77.09	22.91

Table 11: Correlation structure in heredity models for the 12 run PB design.

No. terms	No. active	No. models	Rank problems	max VIF	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ 0.00	0.33
2	2	165	0	1.00	100.00	
3	2	55	0	1.00	100.00	
	3	1650	0	1.12	90.00	10.00
4	3	1980	0	1.12	90.00	10.00
	4	13530	0	1.31	82.44	17.56
5	3	990	0	1.12	86.67	13.33
	4	33660	0	1.53	86.67	13.33
	5	90552	0	1.91	80.61	19.39
6	3	165	0	1.12	85.71	14.29
	4	39939	0	1.57	80.95	19.05
	5	374220	0	2.91	85.71	14.29
	6	488334	0	4.91	81.07	18.93

Table 12: Correlation structure in heredity models for the 20 run PB design.

No. terms	No. active	No. models	Rank problems	max VIF	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ 0.00	0.20	0.60
2	2	513	0	1.00	100.00		
3	2	171	0	1.00	100.00		
	3	9690	0	1.56	90.00	10.00	
4	3	11628	0	1.56	90.00	9.41	0.59
	4	158916	0	1.71	82.12	16.68	1.21
5	3	5814	0	1.56	86.67	12.55	0.78
	4	395352	0	1.83	85.44	13.41	1.15
	5	2279088	0	2.86	79.74	18.45	1.81
6	3	969	0	1.56	85.71	13.45	0.84
	4	468996	0	1.89	80.17	18.57	1.26
	5	9418680	0	3.63	84.86	13.56	1.57

Table 13: Correlation structure in heredity models for the 24 run PB design.

No. terms	No. active	No. models	Rank problems	max VIF	Correlation $\lvert\frac{1}{n}\sum_k x_{ik}x_{jk}\rvert$ 0.00	0.33
2	2	759	0	1.00	100.00	
3	2	253	0	1.00	100.00	
	3	17710	0	1.13	95.71	4.29
4	3	21252	0	1.13	95.71	4.29
	4	363055	0	1.31	92.43	7.57
5	3	10626	0	1.13	94.29	5.71
	4	903210	0	1.50	91.82	8.18
	5	6595204	0	1.91	89.26	10.74
6	3	1771	0	1.13	93.88	6.12
	4	1071455	0	1.50	92.89	7.11

References

Box, G. E. P. and Bisgaard, S. (1993). What Can You Find Out from 12 Experimental Runs, *Quality Engineering* **5**, 663–668.

Box, G. E. P. and Hunter, J. S. (1961). The 2^{k-p} Fractional Factorial Designs, *Technometrics* **3**, 311–351, 449–458.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters*, John Wiley & Sons, New York.

Box, G. E. P. and Meyer, R. D. (1986). An Analysis for Unreplicated Fractional Factorials, *Technometrics* **28**, 11–18.

Box, G. E. P. and Meyer, R. D. (1993). Finding the Active Factors in Fractionated Screening Experiments, *Journal of Quality Technology* **25**, 94–105.

Box, G. E. P. and Tyssedal, J. (1996). Projective Properties of Certain Orthogonal Arrays, *Biometrica* **83**, 950–955.

Chipman, H., Hamada, M. and Wu, C. F. J. (1997). A Bayesian Variable-Selection Approach for Analyzing Designed Experiments With Complex Aliasing, *Technometrics* **39**, 372–381.

Daniel, C. (1959). Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments, *Technometrics* **1**, 311–341.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, John Wiley & Sons, New York.

Draper, N. R. (1985). Small Composite Designs, *Technometrics* **27**, 173–180.

Draper, N. R. and Lin, D. K. J. (1995). Characterizing Projected Designs: Repeat and Mirror-Image Runs, *Commun. Statist. Theory Meth.* **24**, 775–795.

Finney, D. J. (1945). The Fractional Replication of Factorial Arrangements, *Annals of Eugenics* **12**, 291–301.

George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling, *Journal of the American Statistical Association* **88**, 881–889.

Hamada, M. and Wu, C. F. J. (1991). Analysis of Censored Data From Highly Fractionated Experiments, *Technometrics* **33**, 25–38.

Hamada, M. and Wu, C. F. J. (1992). Analysis of Designed Experiments with Complex Aliasing, *Journal of Quality Technology* **24**, 130–137.

Hynén, A. (1996). Screening for Main and Interaction Effects with Plackett and Burman's 12 Run Design, *Technical Report 1*, Division of Quality Technology and Management. Lindköping University, Sweden.

Lin, D. K. J. and Draper, N. R. (1992). Projection Properties of Plackett and Burman Designs, *Technometrics* **34**, 423–428.

Lin, D. K. J. and Draper, N. R. (1995). Screening Properties of Certain Two-Level Designs, *Metrika* **42**, 99–118.

Marquardt, D. W. (1970). Generalized inverses, Ridge Regression, Biased Linear Estimation, and Non-Linear Estimation, *Technometrics* **12**, 591–612.

Meyer, R. D., Steinberg, D. M. and Box, G. E. P. (1996). Follow-Up Designs to Resolve Confounding in Multifactor Experiments (with discussion), *Technometrics* **38**, 303–332.

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

Plackett, R. L. and Burman, J. P. (1946). The Design of Optimum Multifactorial Experiments, *Biometrica* **33**, 305–325.

Tyssedal, J. and Samset, O. (1997). Analysis of the 12 Run Plackett-Burman Design, *Technical Report 8*, Department of Mathematical Sciences. The Norwegian University of Science and Technology, Norway.

Wang, J. C. and Wu, C. F. J. (1995). A Hidden Projection Property of Plackett-Burman and Related Designs, *Statistica Sinica* **5**, 235–250.