STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **8**, June 2020, pp 414–424.
Published online in International Academic Press (www.IAPress.org)

Active Effects Selection which Considers Heredity Principle in Multi-Factor Experiment Data Analysis

Bagus Sartono [1], Achmad Syaiful[2], Dian Ayuningtyas[3], Farit M. Afendi[1], Rahma Anisa[1], Agus Salim[4],*

[1]*Department of Statistics, IPB University, Indonesia*
[2]*Salim Group Enterprise Digital Technology Services, Indonesia*
[3]*PT Tokopedia, Indonesia*
[4]*Department of Mathematics and Statistics, La Trobe University, Australia*

Abstract The sparsity principle suggests that the number of effects that contribute significantly to the response variable of an experiment is small. It means that the researchers need an efficient selection procedure to identify those active effects. Most common procedures can be found in literature work by considering an effect as an individual entity so that selection process works on individual effect. Another principle we should consider in experimental data analysis is the heredity principle. This principle allows an interaction effect is included in the model only if the correspondence main effects are there in. This paper addresses the selection problem that takes into account the heredity principle as Yuan and Lin [23] did using least angle regression (LARS). Instead of selecting the effects individually, the proposed approach perform the selection process in groups. The advantage our proposed approach, using genetic algorithm, is on the opportunity to determine the number of desired effect, which the LARS approach cannot.

Keywords Factorial Experiments, Genetic Algorithm, Heredity Priciple, Variable Selection

AMS 2010 subject classifications 17D92, 90C31

DOI: 10.19139/soic-2310-5070-628

1. Introduction

Fractional-factorial (FF) experiments have been commonly used by researchers in a circumstance where several number of factors are involved but it is impossible to run all factor-level combinations. The FF experiments are usually implemented for a screening research and involving factors with two or three levels. Empirically, there would be only a small portion of effects which are active as stated by the sparsity concept. The active effects could be in form of main effects of factors, or interaction effects among factors. Wu and Hamada [21] discuss detail on designing regular FF experiments, while Schoen et. al. [15] provide methodology to generate non-regular designs. A more specific type of FF experiment is a saturated design where the number of runs is the number of factors plus one. There is also a super-saturated design whose number of factors exceeds the number of runs. Georgiou [7] discusses these designs extensively. The use of FF designs in experiments reduce experimental cost. However, a difficulty rises at the stage of data analysis since the number of runs is not sufficiently large to estimate all possible effects, even the subset of them.

Suppose we have an FF experiment with k factors, and each with two levels. Further, suppose that we are interested in the main effects and two-factor interaction effects only. In total, there are k main effects and

*Correspondence to: Bagus Sartono (Email: bagusco@apps.ipb.ac.id). Department of Statistics, IPB University, Jl. Meranti Wing 22 Level 4 Kampus IPB Darmaga, Bogor, West Java, Indonesia 16680.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2020 International Academic Press

Figure 1. Dependence structure among variables, (a) Group Lasso and (b) Desired dependence structure

$\binom{k}{2} = k(k-1)/2$ interaction effects that the researcher would like to examine. Due to the limited number of runs, the number of effects is likely to exeed the number of observations, especially k is large.

In the context of regression analysis, this kind of situation is known as high dimensional problem. Applying a linear model, it is impossible to include all possible effects in the model because the number of effects to be estimated is larger than the sample size. Therefore, selection process is needed to identify the the subset of active effects.

There are several techniques we can find in the discussion of high dimensional regressions that could be used as alternatives to select active or significant variables. They are included forward selection [14], best subset, LASSO [16], SCAD [6], and other penalized regression approaches. LASSO has attracted attention of many authors since the development of least angle regreesion (LARS) algorithm by Efron et. al. [5] which could reduce the computational effort significantly. Another approach is two-step procedure proposed by Kazemi and Arashi [9].

Those aforementioned approaches could not be directly implemented in the analysis of multi-factor experiment data because they select the predictor variable individually. We mean "individually" as a circumstance that the selection of a variable is performed to one by one variable.

There are several approaches could be found in literatures that work the variables selection in group, such as Group LASSO [24, 11] and Fused LASSO [17]. In the analysis of experiment data, however, Group LASSO could not properly meet the need of the modeling since for a certain group of variables, the algorithm would either include all variables in the group or exclude them all. Figure 1 (a) represents this concept of dependency. The dependence among variables is depicted by the arrow on the Figure. A variable that is pointed must be included in the model as long as the variable pointing it is in the model. Suppose that X_1 and X_2 belong to the same group. It means that if X_1 is included in the model then so is X_2, and the other way around if X_2 is included in the model then X_2 must be included also. Explicitly, we could also said that the variable group in Group LASSO has non-overlapping properties. Each pair of groups are mutually exclusive. This definition of a group is not appropriate in the modeling data of factorial experiments as explained below.

Suppose that an experiment includes factors A and B, so that there are three different effects: main effect of A, main effect of B, and interaction effects of AB. Following the strong heredity principle, if the interaction AB is in the model then the main effects A and B have to be also in the model. But not vice versa, if the main effects are in the model, the interaction is not required to be in. Figure 1 (b) depicts this kind of dependence structure between $X1$ and $X2$ where X_1 might represent an interaction effect and X_2 might represent a main effect. We adopt the way of representing the dependence structure from Yuan et. al. [23]. By this definition of variable group, the groups may have overlap. For example, there is a group containing both X_1 and X_2, but another group only consist of one variable X_2.

A variable selection approach that is suited to experimental data is the one proposed by Yuan et. al. [23]. This approach start with identification of groups by dependence structure as introduced previously and then implement the LARS algorithm to select a group to be included in each step. The group that is selected in each step is the one that has the largest multiple correlation with the residual of the model in the previous step. Since the group may contain one, two, three, or more number of effects, then this approach do not have good control in determining how many effects should be included in the model. The only way to control the complexity is by using the penalty coefficient.

This paper proposes an alternative algorithm to handle variable selection for experimental data. Our approach does not takes into account only the heredity principle in forming the groups, but it also allows the number of effects in the model to be controlled. A genetic algorithm (GA) would be utilized as the optimization technique

in order to reach the desired result. The use of this algorithm is similar to other metaheuristic algorithms such as firefly algorithm implemented by Algamal [2].

The paper is organised as follows. Section 2 provides a general discussion about GA and some notes on its usage in the context of variable selection. Section 3 gives detail on the proposed approach for two-level fractional factorial experiments. Section 4 explains how to deal with three-level quantitative factors. Some illustrations are given in those both sections, along with the comparison of the results of the proposed approach to the competing approaches. A discussion section concludes the paper where some possible extensions are discussed.

2. Genetic Algorithm

The genetic algorithm (GA) is a metaheuristic technique to solve a wide range of optimization problems. The stage of the algorithm imitates the evolutionary process of life beings with the main idea that only the best individuals will survive. In the algorithm, an individual is a point in the feasible region of solution and the best individual is the point which provides optimal solution.

The GA is commonly used in scheduling task to obtain optimum way in completing tasks [3]. The GA is also popularly used in logistic and transportation field to find the best route. Lesiak and Bojarczyk [10] provide some examples of logistic and transportation problem handled by the GA.

In general, GA starts with a group of initial solutions called a population that consists of random individuals. An individual solution is represented by a series of genes forming a chromosome. The value of a gene is defined as it is needed so that the chromosome can represent well a point in the feasible region. This group of individuals is known also as a generation. A generation then developed to new generations by a sequence of stages: selection, cross-over, and mutation.

The selection stage aims to discard ”bad” individuals/chromosomes. Only chromosomes whose best performance in term of a certain fitness/objective function would be kept and continue to cross-over and mutation stages. In the optimization term, this process can be seen as selecting the best solution among candidates.

Next, come the cross over stage. The main idea of this stage is to create a better solution by combining two survived chromosomes. We could choose an arbitrary way of combining two chromosomes as long as it is able to maintain good properties of the chromosomes. A single-point cross-over works by splitting a chromosome into two segments, the left and right segments. Next, it crosses the segments from an individual with the segments from other individual in a cross way. It means that the left segment of an individual is combined with the right segment of the other one. There are other strategies of crossing over such as two-point, multiple-point, and uniformly cross-over, as described by Umbarkar and Seth [18].

The last stage in an iteration is the mutation. Within this stage, a few chromosomes are slightly changed. The changes happen in the values of genes which were randomly picked with a very small probability.

The sequence of selection, cross-over, and mutation stages is running for several generations and it is expected that in every generation there is a gradual improvement of solution. The algorithm may stop whenever the improvement is negligible or if the number of generation exceeds a certain number that was previously set.

GA has been widely used in the context of variable selection. Yang and Honavar [22] used GA in selecting predictor for neural-network classifier. Vafaie and De Jong [19] and Zelenkov et. al. [25] elaborated the possibility of GA implementation to features selection in pattern recognition and machine learning. The selection was highly needed to decrease the processing time and GA was helpful in doing selection while maintain the prediction accuracy. In the field of chemometrics, Broadhurst et. al. [4] discussed how GA contributes in variable selection for regression model using spectometry data. Aalaei et. al. [1] and Vandewater et. al. [20] implemented GA for variable selection in the detection of breast cancer and alzheimer, respectively.

3. Proposed Approach for Experiments with All Two-Level Factors

Let us start our by focusing on experiments involving factors, all with two levels. This kind of experiments is commonly found in a screening research where there are many factors to be examined with the constraint of small

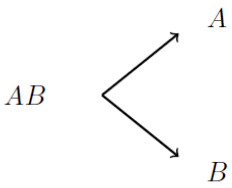

Figure 2. Dependence structure interaction and main effects of factors whose two levels, A and B

budget. The budget restriction requires researchers not to run all possible level combinations. A certain fractional factorial design might be implemented and an analysis to identify active effects should follow afterward. The common situation is that the number of observations would be less than the number of interested effects.

First, let k be the number of factors and all have two levels. Second, we are only interested in main effects and two-factor interaction (2fi) effects, so that in total we have $K = k + \binom{k}{2}$ effects to be estimated. The total number of runs in the experiment is n and it is assumed that $n < K$.

First we define the dependence structure among main effects and 2fi effect as depicted by Figure 2. As previously discussed this structure suggests that if AB is in the model, then the effects of A and B have to be included also. However, the existence of A or B does not imply the existence of AB. Therefore, we then have the possible groups of effects as follows:

- $A = \{A\}$
- $B = \{B\}$
- $AB = \{A, B, AB\}$

To implement the GA, we define a chromosome as an object that represents a certain model containing a set of effects. Every chromosome consists of a sequence of K binary-genes. Each of the first k genes represents a group containing a single main effect of k factors, while each other gene represents a group of effects containing a 2fi effect and two main effects of the factors contribute to the interaction effects.

The binary code of the genes is either 1 or 0, where 1 means that the effect is included in the model and 0 for the otherwise. Since the group of effects may overlap, the set of effects included in the model is the union of effects correspond to included genes. Suppose that p is the total effect in the set, then for some sense of effect selection, we should also add other condition that $p \leq t < (n-1)$ where n is the number of observation, and the smaller t implies the tighter selection process and less number of effects included in the model.

We illustrate the idea of this chromosome representation here. An 8-run experiment involving 5 (five) two-level factors so that there are in total 15 intersted effects (i.e. 5 main-effects and 10 two-factor interactions effects). A chromosome that represent a certain model therefore consists of 15 genes. The first five genes would represent 5 groups each of which consists a main effect A, B, C, D, and E, while each of the remaining ten genes represent the following series of groups of effects $\{AB, A, B\}$, $\{AC, A, C\}$, $\{AD, A, D\}$, $\{AE, A, E\}$, $\{BC, B, C\}$, $\{BD, B, D\}$, $\{BE, B, E\}$, $\{CD, C, D\}$, $\{CE, C, E\}$, and $\{DE, D, E\}$. Suppose there is a chromosome in form of

1	1	0	0	0	0	1	0	0	0	0	0	0	0	0

This chromosome would represent a model containing the following set of effects

$$\{A\} \cup \{B\} \cup \{AC, A, C\} = \{A, B, C, AC\}.$$

Other components of GA that we have to define is the fitness function to be optimized. We use PRESS and AIC (Akaike's Information Criterion) which consider both the fit of the prediction and the complexity of the model. The PRESS value is calculated by the following formula

$$PRESS = \sum (y_i - \hat{y}_{[i]})^2 \tag{1}$$

where y_i is the response value of the i-th observation and $\hat{y}_{[i]}$ is the predicted value for the i-th observation from a model estimated using the other $(n-1)$ observations. The AIC value is obtained using the formula of AIC = $-2L + 2p$, where L is the log-likelihood function and p is the number of parameter in the model.

The genetic algorithm to identify active effects was proposed as follows. Initially, a population consisting of M individual chromosomes is generated. The codes of genes for each chromosome are assigned randomly from a Bernoulli distribution having a certain probability parameter. In this paper we use the probability parameter of $0.6n/K$.

We now begin the selection process. The initial selection process is checking the estimability of the models represented by the chromosomes. It ensures that all represented model contains less than $(n-1)$ effects. A chromosome that does not meet the estimability constraint is immediately removed from the population. The PRESS (or AIC) values are then calculated for the remaining chromosome. After all chromosomes are assessed, up to s chromosomes/models with the lowest PRESS (or AIC) values are selected.

Those selected chromosomes are then combined among each other using a single-point cross-over process. The point location to segment the chromosomes is selected randomly over all possible point along the chromosomes. For each two parent-chromosomes we could generate two offspring-chromosomes which are formed by combining the first-part segment from a chromosome with a second-part segment of the other chromosome and in the other way around. By this cross-over technique, the offspring chromosomes represent a model containing effects that some are included in a parent model and some others are included in the other parent model. At the end of cross over process we have s parent chromosomes and $s(s-1)$ offspring chromosomes, representing s^2 candidates model for further selection.

A mutation stage comes next. In this stage, each gene value may change from 0 to 1 or form 1 to 0. It means that during the mutation stage, a certain effects might be removed from a model, or in contrast, the previously excluded effects would be included. For this paper, the probability of mutation is set to be 10^{-3}. This means that our initial selection needs to be very good.

The algorithm is then back to the selection stage, followed by cross-over and mutation stages. This cyclic procedure is repeated until one of some stopping criteria is satisfied such as the iteration reach the pre-determined maximum number and the improvement is very small.

Illustration #1

To illustrate how the approach works, we would use a 2^{9-5} experiment described by [12]. It involved nine two-level factors and only a fraction of 1/32 of the full-factorial design that was tried. The design along with the data are given in Table 1.

The algorithm started with defining a chromosome as a series of 45 genes, where first nine genes represent nine main effects and 36 other genes represent groups of three effects (a two-factor interaction effect and two main effects of the factors). Table 2 shows the effects which were selected by the proposed approach when the maximum number of effects was restricted to be 5 to 8. Coincidently, the use of both criteria AIC and PRESS resulted in identical effects for all different constraints of maximum effect numbers.

Interestingly, the selection result is exactly the same as reported by Yuan et. al. [23] using LARS algorithm, which is also identical as resulted by Raghavarao [12].

Illustration #2

Rais et. al. [13] describes an experiment of sulfated amides of fatty acids derived from olive pomace oil. The experiment employed a 18-run super saturated design involving 31 two-level factors, $U_1, \ldots, U_{31}$. Readers could find the detail description about the factors in the paper. The authors proposed a sequential method involving ridge regression, stepwise procedure, best subset, and final effect test based on the projected design. At the end of the analysis, they suggested that there are nine factors that should be considered in the follow-up experiments. Those selected factors are U_{13}, U_{18}, U_{19}, U_{20}, U_{24}, U_{27}, U_{28}, U_{29} and U_{30}.

While Rais et. al. [13] explicitly ignored the possibilities of interaction among factors, we believe that some degree of interactions exist. The used design implies that some interactions are confounded partially by the factors' main effects. It means that we might wrongly conclude that a certain factor is active due to the active interaction

Table 1. Illustration 1

Run	A	B	C	D	E	F	G	H	J	Y
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	136.475
2	1	1	−1	1	1	−1	−1	−1	−1	147.775
3	1	−1	−1	1	−1	−1	1	1	−1	142.425
4	1	−1	1	1	1	−1	1	1	1	141.800
5	1	1	1	1	−1	−1	−1	−1	1	136.675
6	−1	1	−1	−1	1	−1	1	1	−1	150.725
7	−1	1	1	−1	−1	−1	1	1	1	142.800
8	−1	−1	1	−1	1	−1	−1	−1	1	135.825
9	1	1	1	−1	−1	1	−1	1	−1	143.476
10	−1	1	−1	1	1	1	1	−1	1	145.150
11	1	−1	1	−1	1	1	1	−1	−1	142.600
12	−1	−1	−1	1	−1	1	−1	1	1	139.375
13	1	1	−1	−1	1	1	−1	1	1	139.650
14	1	−1	−1	−1	−1	1	1	−1	1	144.775
15	−1	−1	1	1	1	1	−1	1	−1	148.275
16	−1	1	1	1	−1	1	1	−1	−1	141.075

Table 2. The selected effects based on AIC and PRESS criteria

max number of effects	selected effects
5	E, G, J, EJ, GJ
6	E, G, H, J, EJ, GJ
7	E, G, H, J, EJ, GJ, HJ
8	$B, E, G, H, J, EJ, GJ, HJ$

effect confounded by the factor. By this reason, we implemented our approach to the data of Rais et. al. [13] with the constraint that the number of effects in the model is nine.

The result suggests that those nine effects are U_4, U_{19}, U_{21}, U_{27}, U_{28}, U_{29}, $U_4 \times U_{27}$, $U_{19} \times U_{28}$, and $U_{21} \times U_{29}$. The model which contains those nine effects has AIC value of 77.60 and PRESS value of 132.04. Based on both criteria, this model is far better compared to the result of Rais et. al. [13] which has AIC value of 106.19 and PRESS value of 502.02.

4. Dealing with Three-Level Quantitative Factors

Now, suppose that in an experiment, instead of involving factors with two levels, we found also some factors with three levels. For simplicity, let us first assume that those three-level factors are quantitative factors whose levels are equally spaced. In this circumstance, we might decompose the main effect of a certain factor into two orthogonal polynomial contrast as follows:

	level		
contrast	1	2	3
linear	−1	0	1
quadratic	−1	2	−1

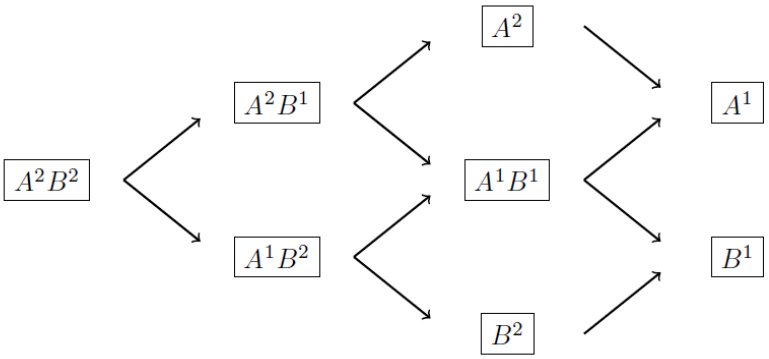

Figure 3. Dependence structure among two-factor interaction and main effects of two three-level factors, A and B

Further, suppose that A is the factor with three levels so that the main effects of A would be decomposed into a linear effect A^1 and a quadratic effect A^2. Similarly, for a three-level factor B whose effects of B^1 and B^2 for linear and quadratic effect. The two-factor interaction effect between A and B are then could be splitted into four components:

- linear-by-linear interaction effect, A^1B^1
- linear-by-quadratic interaction effect, A^1B^2
- quadratic-by-linear interaction effect, A^2B^1
- quadratic-by-quadratic interaction effect, A^2B^2

Following the heredity principle, Yuan et. al. [23] defined a dependence structure among two-factor interaction effect and main effects for three-level factors as shown in Figure 3. This structure and the heredity principles imply that if an effect is included in the model then so are the effects which are pointed by that effect. For example, if A^2B^1 is in the model then the model should also include A^2 and A^1B^1. Next, because A^2 is included then so is A^1, and the inclusion of A^1B^1 implies that B^1 should be in the model. Therefore, at the end, the selection of A^2B^1 is equivalent to the selection of a group of effects $\{A^1, A^2, B^1, A^1B^1, A^2B^1\}$. The full list of group effects for main and two-factor interactions generated by the structure of Figure 3 is as follow:

- $A^1 = \{A^1\}$
- $A^2 = \{A^1, A^2\}$
- $B^1 = \{B^1\}$
- $B^2 = \{B^1, B^2\}$
- $A^1B^1 = \{A^1, B^1, A^1B^1\}$
- $A^1B^2 = \{A^1, B^1, B^2, A^1B^1\}$
- $A^2B^1 = \{A^1, A^2, B^1, A^1B^1\}$
- $A^2B^2 = \{A^1, A^2, B^1, B^2, A^1B^1, A^1B^2, A^2B^1, A^2B^2\}$

The identification of groups of effects is important because it is needed in defining the chromosome when the GA is implemented. It is clear that each gene in a chromosome will represent a certain group of effect that may contain a single effect such as A^1 or B^1, or may contain two effects such as $A^2 = \{A^1, A^2\}$, three effects such as $A^1B^1 = \{A^1, B^1, A^1B^1\}$, and so on. If the chromosome and the genes inside of it are already appropriately defined, the GA might but submitted to select which effects should be considered to be active or significant.

Illustration #3

As an illustration of the implementation of the proposed approach, we use a blood glucose experiment reported by Hamada and Wu [8]. The experiment was performed in 18 runs and employed a two-level factor (named A) and

Table 3. Illustration 3

Run	A	B	C	D	E	F	G	H	Y
1	1	1	1	1	1	1	1	1	97.94
2	1	2	2	2	2	2	1	2	83.40
3	1	3	3	3	3	3	1	3	95.88
4	1	1	1	2	2	3	2	3	88.86
5	1	2	2	3	3	1	2	1	106.58
6	1	3	3	1	1	2	2	2	89.57
7	1	1	2	1	3	2	3	3	91.98
8	1	2	3	2	1	3	3	1	98.41
9	1	3	1	3	2	1	3	2	87.56
10	2	1	3	3	2	2	1	1	88.11
11	2	2	1	1	3	3	1	2	83.81
12	2	3	2	2	1	1	1	3	98.27
13	2	1	2	3	1	3	2	2	115.52
14	2	2	3	1	2	1	2	3	94.89
15	2	3	1	2	3	2	2	1	94.70
16	2	1	3	2	3	1	3	2	121.62
17	2	2	1	3	1	2	3	3	93.86
18	2	3	2	1	2	3	3	1	96.10

seven three-level factors (named B, C, D, E, F, G, and H), so that in total there are eight factors. Table 3 presents the design along with the response variable values for each run.

All three-level factors are quantitative and almost evenly-spaced so that we could decompose the main effect into linear and quadratic contrasts. Meanwhile, we used $\{-1, +1\}$ contrast for the two-level factor.

Using aforementioned setting, we have $1 + 7(2) = 15$ main effects, $7(1 \times 2) = 14$ two-factor interaction effects between A and the other factors, and $\binom{7}{2}(2 \times 2) = 84$ two-factor interaction effects among the three-level factors. It means that in total we should consider 113 effects to be examined and a chromosome in the GA procedure would consist of 113 genes.

Table 4 presents the effects selected by GA method when the number of effects in the model were restricted at most 1 to 13. Those results were produced by using AIC as the fitness value. The table also provides the effects obtained by Yuan et. al. [23] using the LARS/LASSO approach.

As previously mentioned in the introduction section, LARS in Yuan et. al. [23] works by entering a single group of effect on each of its iteration. Since the group sizes vary, we could not control the total number of effects which entered in the model. It explains why we do not have the result in most rows of Table 4 . From the table, we observe that when the number of effects was restricted to be 3, 4, and 5, the proposed approach produced identical results with what Yuan et. al. [23] had. The selected effects included main effects of E and F, and also liner-by-linear interaction of both factors. LARS methodology was only able to select all eight main effects and interaction effects of B and H at their next iteration so that in total there are 13 effects. It is obvious why LARS did not produce models with 6 to 12 effects.

When we executed the GA methodology with at most 13 effects, we ended up with quite different result. All effects of B and H are there, but the other effects differ from what Yuan et. al. [23] produced. In term of AIC and PRESS the linear model including effects that the GA produced is better. Respectively, it has AIC and PRESS of 43.29 and 150.98, while the model with effects resulted by LARS methodology has AIC = 89.43 and PRESS=510.22.

Table 4. Selected effects in blood glucose experiment

maximum number of effects	selected effect	result of [23]
1	G	not available
2	F, F^2	not available
3	E, F, E^1F^1	E, F, E^1F^1
4	E, F, E^2, E^1F^1	E, F, E^2, E^1F^1
5	E, F, E^2, F^2, E^1F^1	E, F, E^2, F^2, E^1F^1
6	$B, G, H, H^2, B^1H^1, B^1H^2$	not available
7	$B, G, H, G^2, H^2, B^1H^1, B^1H^2$	not available
8	$B, H, B^2, H^2, B^1H^1, B^1H^2, B^2H^1, B^2H^2$	not available
9	$B, F, H, B^2, H^2, B^1H^1, B^1H^2,$ B^2H^1, B^2H^2	not available
10	$B, F, H, B^2, H^2, B^1F^1, B^1H^1,$ B^1H^2, B^2H^1, B^2H^2	not available
11	$B, D, F, H, B^2, H^2, F^1H^1,$ $B^1H^1, B^1H^2, B^2H^1, B^2H^2$	not available
12	$B, C, F, H, B^2, H^2, B^1C^1,$ $B^1F^1, B^1H^1, B^1H^2, B^2H^1, B^2H^2$	not available
13	$A, B, C, F, H, B^2, H^2, B^1C^1,$ $B^1F^1, B^1H^1, B^1H^2, B^2H^1, B^2H^2$	$B, E, F, H, B^2, E^2, F^2, H^2,$ $E^1F^1, B^1H^1, B^1H^2, B^2H^1, B^2H^2$

5. Conclusion

This current paper discusses an alternative approach to identify active effects based on a fractional factorial experiment data. The existing of heredity principle insist us not to utilize ordinary variable selection approaches since they perform the selection of variable individually. By applying the dependence structure of effects explained in Yuan et. al. [23], we propose to use GA to find effects that optimize a certain model selection criterion. In this paper, AIC and PRESS were chosen because both are considering goodness of fit of the prediction as well as preventing models to be too complicated. The proposed approach offers some benefit. It could well handle the conformity to the heredity principle by carefully define the group of effects represented by each gene. Also, it gives an opportunity to determine the number of effects to be selected. Both advantages might be useful for most researchers and make this approach as a competitive method.

However, because the optimization algorithm employs the GA, we suggest that the users should run the algorithm several times and pick the best result among those trials. The different results for one try to other tries might happen because of the different random initial population. This possibility would commonly occur when the number of effects is very large.

We demonstrated how this approach works when the model have to satisfy the strong heredity principle. We aware that it is possible that some analyst may work with the weak heredity principle instead. A modification of dependence structure should be defined based on that and the GA method could work without any difference. Yuan et. al. [23] discussed the dependence structure for this principle and readers could adopt it literally. Even we only discussed the algorithm for two-level factors and quantitative three-level factors, this approach could also be implemented for other situations such as factors with three or more levels, no matter whether they are either quantitative or qualitative ones.

6. Acknowledgement

This work was supported by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia through the research grant of International Research Collaboration 2018/2019.

REFERENCES

[1] Aalaei, S., H. Shahraki, A. Rowhanimanesh, and S. Eslami (2016). Feature selection using genetic algorithm for breast cancer diagnosis experiment on three different datasets. *Iranian Journal of Basic Medical Sciences 19*, 476–482.

[2] Algamal, Z. Y. (2019). Variable selection in count data regression model based on firefly algorithm. *Stat. Optim. Inf. Comput. 7*, 520–529.

[3] Asadzadeh, L. and K. Zamanifar (2010). An agent-based parallel approach for the job shop scheduling problem with genetic algorithms. *Mathematical and Computer Modelling 52*, 1957–1965.

[4] Broadhurst, D., R. Goodacre, A. Jones, J. J. Rowland, and D. B. Kell (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta 348*, 71–86.

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

[6] 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.

[7] Georgiou, S. (2014). Supersaturated designs: A review of their construction and analysis. *Journal of Statistical Planning and Inference 144*, 92–109.

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

[9] Kazemi, M., S. D. and M. Arashi (2018). Variable selection and structure identification for ultrahigh-dimensional. *Stat. Optim. Inf. Comput. 6*, 373–382.

[10] Lesiak, P. and P. Bojarczyk (2015). Application of genetic algorithms in design of public transport network. *Logistics and Transport 52*, 75–81.

[11] Meier, L., S. van de Geer, and P. Buhlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B (methodological) 70*, 53–71.

[12] Raghavarao, D. and S. Altan (2003). A heuristic analysis of highly fractionated 2^n factorial experiments. *Metrika 156*, 185–191.

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

[14] Rawlings, J., S. Pantula, and D. A. Dickey (1998). *Applied Regression Analysis: A Research Tool, Second Edition*. Springer.

[15] Schoen, E. D., P. T. Eendebak, and M. V. M. Nguyen (2010). Complete enumeration of pure-level and mixed-level orthogonal arrays. *Journal of Combinatorial Designs 18*, 123–140.

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological) 58*, 267–288.

[17] Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B (methodological) 67*, 91–108.

[18] Umbarkar, A. and P. Sheth (2015). Crossover operators in genetic algorithms: a review. *ICTACT Journal on Soft Computing 6*, 1083–1092.

[19] Vafaie, H. and K. De Jong (1992, November). Genetic algorithms as a tool for feature selection in machine learning. In *Proceeding of the 4th International Conference on Tools with Artificial Intelligence*.

[20] Vandewater, L., V. Brusic, W. Wilson, L. Macaulay, and P. Zhang (2015). An adaptive genetic algorithm for selection of blood-based biomarkers for prediction of alzheimer's disease progression. *BMC Bioinformatics 16*, 1–10.

[21] Wu, C. F. J. and M. Hamada (2000). *Experiments: Planning, Analysis and Parameter Design Optimization*. New York: Wiley.

[22] Yang, J. and V. Honavar (1997). Feature subset selection using a genetic algorithm. *Computer Science Technical Reports 156*.

[23] Yuan, M., V. R. Joseph, and Y. Lin (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics 49*, 430–438.

[24] Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (methodological) 68*, 49–67.

[25] Zelenkov, Y., E. Fedorova, and D. Chekrizov (2017). Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Systems with Applications 88*, 393 – 401.