AN INFORMATION THEORETICAL METHOD FOR ANALYZING UNREPLICATED DESIGNS WITH BINARY RESPONSE

Authors: Krystallenia Drosou
– Department of Mathematics, National Technical University of Athens, Athens, Greece (drosou.kr@gmail.com)

Christos Koukouvinos
– Department of Mathematics, National Technical University of Athens, Athens, Greece (ckoukouv@math.ntua.gr)

Abstract:

- The analysis of unreplicated factorial designs constitutes a challenging but difficult issue since there are no degrees of freedom so as to estimate the error variance. In the present paper we propose a method for screening active effects in such designs, assuming Bernoulli distributed data rather than linear; something that hasn't received much attention yet. Specifically, we develop an innovating algorithm based on an information theoretical measure, the well-known symmetrical uncertainty, so that it can measure the relation between the response variable and each factor separately. The powerfulness of the proposed method is revealed via both, a thorough simulation study and a real data set analysis.

Key-Words:

- *Two-level factorial designs; Unreplicated experiments; Generalized linear models; Symmetrical Uncertainty.*

AMS Subject Classification:

- 62-07; 62K15; 62J12.

1. INTRODUCTION

Factorial designs constitute a powerful tool especially in screening experiments where the goal is to identify the factors with a significant impact on the response of interest. Although two-level factorial designs are commonly used as experimental plans, the number of runs grows exponentially as the number of factors increases; thus, in case when the replication of the experiment is prohibitive due to economical or technical issues, unreplicated designs constitute an appropriate choice. Such designs are saturated; means that the number of examined factors d equals to $n-1$, where n is the number of runs. As a result, the experimenter can estimate all the d main and interaction effects, but there are no degrees of freedom to estimate the error; therefore, the conventional analysis of variance (ANOVA) techniques cannot be applied.

Many methods, either theoretical or graphical ones, have been proposed to overcome the aforementioned problem. The standard method for identifying active effects in unreplicated designs is the probability plot of the effects, proposed by Daniel [7]. This approach consists of plotting the factor estimates on a normal or half-normal probability plot, where the inactive effects fall along a straight line while the active ones tend to fall off the line. The subjective nature of that method motivated many authors to provide more objective procedures. For a detailed review article, we refer the interested reader to Hamada and Balakrishnan [10]. Some important works include: Box and Meyer [5], Lenth [11], Dong [8], Chen and Kunert [6], Aboukalam [1], Miller [14], Voss and Wang [22], Angelopoulos and Koukouvinos [2], and Angelopoulos et al. [3,4].

Although many methods have been proposed for analyzing unreplicated designs for a normal response, it is evident the lack of research papers for non-normally distributed responses. This fact prompted us to develop a methodology for screening out the important effects assuming that the response of interest is a binary one; therefore, we developed a generalized linear model, say a logistic model. Our approach for analyzing unreplicated designs constitutes a statistical method inspired by some information theoretical measures, the main of which was the symmetrical uncertainty (SU). To the best of our knowledge, this is the first time such an algorithm is modified and appropriately used for variable selection in unreplicated designs. The merits of our study is encouraging enough.

The rest of the paper is organized as follows. In Section 2, we briefly discuss the basic concepts of the information theoretical measures, the formulation of the problem as well as our new SU algorithm. In Section 3, we carry out an empirical study comparing our method with two well-known feature selection algorithms, the CMIM and the mRMR. Finally, in the last Section 4, we summarize the merits of our study providing some concluding remarks.

2. A METHOD FOR SEARCHING ACTIVE EFFECTS IN UNREPLICATED DESIGNS WITH BINARY RESPONSE

Generalized linear models (Nelder and Wedderburn [17], McCullagh and Nelder [13] and Myers et al. [16]) were developed to allow the fit of regression models for response data that follow a distribution belonging to the exponential family. This family includes not only the exponential but also the normal, binomial, Poisson, geometric, negative binomial, gamma and the inverse normal distributions. All these models have a common property: the mean (or expected) response at each data point and the variance of the response are related.

Consider a two-level full factorial unreplicated design where one wants to estimate the main and interaction effects in d factors with n runs. Let X be the corresponding $n \times d$ design matrix where at the i_{th} data point, $i = 1, \ldots, n$ the response is a Bernoulli random variable y_i, that takes only two possible values, 0 and 1, representing "failure" or "success", respectively. It is well known that $\mu_i = E(y_i) = P_i = P(\mathbf{x}_i)$, where P_i is the probability of success in a Bernoulli process, $\mathbf{x}_i$ is a $d-$dimensional vector of the predictor variables and $Var(y_i) = P_i(1 - P_i)$ is the variance of the response. It is obvious that the variance is a function of the mean. The probability of success, $P(\mathbf{x}_i)$, in case of the logistic regression model is given as follows

$$P(\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}}}, \tag{2.1}$$

where the term $\mathbf{x}_i^T \boldsymbol{\beta}$ is said to be the linear predictor. For more details on logistic regression model, we refer the interested reader to Montgomery et al. (2006). In accordance with this scenario, we perform our simulation study by generating logistic models that has the form

$$y_i = P(\mathbf{x}_i) + \varepsilon, \tag{2.2}$$

where ε has a distribution with zero mean and variance $P(x_i)[1 - P(x_i)]$. More precisely, ε takes two possible values: $\varepsilon = 1 - P(\mathbf{x}_i)$ with probability $P(\mathbf{x}_i)$ if $y = 1$, and $\varepsilon = -P(\mathbf{x}_i)$ with probability $1 - P(\mathbf{x}_i)$ if $y = 0$. Consequently, the conditional distribution of the outcome variable has a Bernoulli distribution with success probability $P(\mathbf{x}_i)$.

2.1. Information measures

Information theory provides useful tools to quantify the uncertainty of random variables. Our method is inspired from the information theory field with the aim of identifying those effects that carry as much information as possible. This section provides some information measures which constitutes the theoretical basis of our methodology.

Let U and V be two discrete random variables. One of the most fundamental concept in information theory is that of entropy measure which was introduced by Shannon [21] and it is defined as

$$H(U) = -\sum_{u \in \mathcal{U}} p(u) \log_2(p(u)). \tag{2.3}$$

The entropy quantifies the uncertainty of U, where $p(u)$ is the prior probability for all values of U. It is a measure of the amount of information required on average to describe the random variable. The information entropy of a Bernoulli trial used in our study is defined as

$$H(Y) = -p(y)log_2p(y) - (1 - p(y))log_2(1 - p(y)), \tag{2.4}$$

where $p(y)$ is the prior probability for all values of Y.

In case of two variables we could define the mutual information (MI) which is a quantity that measures the mutual dependence of these variables. It is also called information gain (Quinlan [18]) and it is defined as

$$I(U|V) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y). \tag{2.5}$$

Note that the MI of a random variable with itself, is its entropy. MI can be used for feature selection with the aim to select a small subset of features that carries as much information as possible (Fleuret [9], Peng et al. [19]). Information gain is a symmetrical measure for two random variables. Symmetry is an appealing property for a measure of correlations between factors, but information gain is biased in favor of factors with more values. Symmetrical uncertainty (Press et al. [20]) counterbalances the bias of information gain towards factors with more values, and normalizes its value to the range $[0, 1]$. The definition of Symmetrical Uncertainty is given as

$$SU(U,V) = 2 \times \left[\frac{I(U|V)}{H(U) + H(V)}\right]. \tag{2.6}$$

2.2. Symmetrical uncertainty algorithm

The proposed method is a modification of a feature selection algorithm, known as Fast Correlation Based Filter (FCBF, Yu and Liu [23]). More precisely, it actually performs a typical variable selection using the SU coefficient so as to determine the significant effects. The algorithm can be described as follows:

Algorithm

a) Given a $n \times d$ unreplicated design matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d]$, where $\mathbf{x}_l$, $l = 1, 2, \ldots, d$, is the l_{th} column of the matrix, as well as a $n \times 1$ Bernoulli distributed vector $\mathbf{y}$, which is the response vector, compute the entropy and the conditional entropy with respect to the response variable.

b) Compute the vector entropy values and the conditional entropy values for each variable as: $H(X) = (H(\mathbf{x}_1), H(\mathbf{x}_2), \ldots, H(\mathbf{x}_d))$ and $H(\mathbf{X}|\mathbf{Y}) = (H(\mathbf{x}_1|\mathbf{y}),$
$H(\mathbf{x}_2|\mathbf{y}), \ldots, H(\mathbf{x}_d|\mathbf{y}))$, where $H(\mathbf{x}_j)$ is the corresponding value of the entropy measure and $H(\mathbf{x}_j|\mathbf{y})$ is the corresponding value of the conditional entropy for the j-th, $j = 1, \ldots, d$ variable, respectively.

c) Compute the vector of information gain values as: $I(X|Y) = (I(\mathbf{x}_1|\mathbf{y}), I(\mathbf{x}_2|\mathbf{y}),$ $\ldots, I(\mathbf{x}_d|\mathbf{y}))$, where $I(\mathbf{x}_j|\mathbf{y})$ is the information gain value for each variable with respect to the response variable.

d) Compute the symmetrical uncertainty measure, $SU = (su_1, su_2, \ldots, su_d)$, where

$$su_j = 2 \times \left[\frac{I(\mathbf{x}_j|\mathbf{y})}{H(\mathbf{x}_j) + H(\mathbf{y})} \right],$$

for $j = 1, \ldots, d$, represents the value of SU for the j-th variable with respect to the response variable.

e) The last step is to identify and maintain the significant effects by retaining only those with scores greater than the predefined threshold value of the SU vector values.

2.3. Performance Criteria

The performance of the proposed methodology is evaluated using the two most known criteria, the Type I and Type II error rates. In screening designs, there are two, the probability of declaring an inactive factor to be active (Type I error), and the probability of declaring an active factor to be inactive (Type II error). Type II errors are troublesome, as addressed in Lin [12], as well as Type I errors, since they can result in unnecessary cost in follow-up experiments. Type I errors are very likely in situations of effect sparsity. Undoubtedly, Type II error rates are of highly importance and we have considered that importance during the creation and implementation of our algorithm.

3. EXPERIMENTAL RESULTS

This section presents a simulation study examining the performance of our algorithm. To assess the performance of the proposed method, we applied simulations for a wide range of underlying models. Our information-theoretic method is compared with two feature selection algorithms which are widely used in many fields of science: the Conditional Mutual Information Maximization (CMIM) algorithm proposed by Fleuret [9] and the minimal-redundancy-maximal-relevance

feature selection (mRMR) algorithm proposed by Peng et al. [19]. These algorithms were selected to be compared with SU-algorithm since they were made based on information measures. More precisely, CMIM constitutes a feature selection technique based on conditional mutual information and it iteratively picks features which maximize their mutual information with the class to predict, conditional to any feature has already picked. MRMR algorithm performs feature selection by maximizing the mutual information between the selected features and the desired output (relevance), as well as by minimizing the mutual information between the selected features (redundancy).

3.1. Simulation scheme

Two unreplicated factorial designs served as the design matrices in our simulations experiments: a 2^4 and a 2^5 full factorial design. We used these designs since they are commonly used in a wide range of problems; thus, our results can be comparable to other existing methods and problems. For the examined designs, the true active variables were selected using two different scenarios. For each design and each number of the active factors, we randomly generated 1000 Bernoulli distributed response vectors $\mathbf{y} \sim \text{Bernoulli}(P(\mathbf{X}^T\boldsymbol{\beta}))$, where $P(u) = \frac{1}{1+e^{-u}}$. All simulations were conducted using MATLAB codes.

Scenario A: We developed logistic models with coefficients taking predefined values. The coefficients of inactive effects are set equal to zero. However, in order to examine the sensitivity of the results in terms of the selection and the number of active factors, we changed the order of columns of the active factors, using different values of β as well as different number of active factors for each unreplicated design. As a result, we considered several models that were different in this regard. We considered the cases for $p = 1, 2, 3, 4, 5, 6, 7, 8$ active effects involved in a 2^4 factorial design and for $p = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ active effects involved in a 2^5 factorial design.

Scenario B: We developed logistic models with coefficients taking randomly selected values from the range -5 to 5. When a generated coefficient was "almost zero", it was replaced by 50% of the maximum coefficient. Concerning the true active variables, they were also selected randomly, according to the uniform distribution, using the set of $\{1, \ldots, d\}$ potentially active factors and with respect to the number of active factors of the design matrix. The coefficients of the non-active variables in the true model, were set equal to zero. The number of true active variables was set at most $d/2$, based on the sparsity of effects principle (Box and Meyer [5]). This principle states that, in contrast with the initial large number of potentially active factors, only few of them are dominant, meaning that their multitude hardly exceeds $1/2$ of the total number of factors.

3.2. Simulation results

The simulation results listed in the following Tables and Figures, contain the application of the SU method along with that of CMIM and mRMR. Before performing the simulation experiments, we should set the threshold value which determines whether a factor is significant or not. Several different threshold values (0.001, 0.01, 0.05, 0.1, 0.15, 0.2, median (SU)) were examined in order to find the optimal one for the proposed method. We finally selected the median(SU) as a threshold value, since it acquires the best results. Not to mention the fact that median(SU) is based on the estimated values of the SU vector and it seems to be a reasonable choice. The following Tables summarize the results concerning scenario A of simulation study. Specifically, in Tables 1 and 3, we present the examined models for designs with four and five factors, respectively. The first column represents the number corresponding to each model with predefined values for the coefficients depicted in the second column.

Model	**Predefined values of coefficients**
1	$[0,0,0,0,3,0,0,0,0,0,0,0,0,0,0]^T$
2	$[0,0,0,0,0,0,0,0,0,0,0,2,0,0,3]^T$
3	$[0,0,-7,0,0,0,0,-8,0,0,0,0,0,0,-6]^T$
4	$[0,0,-9,0,4,0,0,0,0,-2,0,0,0,0,10]^T$
5	$[6,0,0,0,0,7,0,0,-5,-5,0,-7,0,0,0]^T$
6	$[0,7,0,-2,0,5,2,0,4,0,0,0,-8,0,0]^T$
7	$[0,0,-9,2,0,0,0,4,5,8,0,0,-5,-7,0]^T$
8	$[5,0,-6,8,0,-5,6,0,0,7,0,0,-7,0,1]^T$

Table 1: Models considered in the simulation study for a 2^4 unreplicated design (Scenario A).

	Type I Error			**Type II Error**		
Model	**SU**	**CMIM**	**mRMR**	**SU**	**CMIM**	**mRMR**
1	0.00	0.00	0.03	0.00	0.00	0.00
2	0.15	0.04	0.08	0.17	0.31	0.45
3	0.08	0.08	0.11	0.00	0.33	0.33
4	0.09	0.10	0.11	0.24	0.27	0.28
5	0.09	0.10	0.13	0.08	0.23	0.23
6	0.00	0.11	0.22	0.33	0.33	0.33
7	0.12	0.34	0.35	0.28	0.39	0.40
8	0.00	0.13	0.18	0.12	0.12	0.16
Average	**0.07**	**0.11**	**0.15**	**0.15**	**0.25**	**0.27**

Table 2: 2^4 unreplicated design: Performance of the proposed method for models $1-8$, using 1000 simulations (Scenario A).

The obtained results are summarized in Tables 2 and 4 for four and five

factors, respectively. More precisely, the first column in both Tables contains the number that corresponds to each model. The remaining columns present the results for Type I and Type II error rates correspond to each method separately. Table 2 clearly shows that SU algorithm outperforms all the others in terms of both Type I and Type II error rates. Especially, the average values of Type II error is comparatively smaller; with SU equals to 0.15 compared to 0.25 and 0.27 of CMIM and mRMR, respectively. This fact is extremely important in factorial designs since low Type II means low probability of declaring an active factor to be inactive.

Model	**Predefined values of coefficients**
1	[0,5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-4,0,0,0,0,0,0,0,0,0,0,0,0,6,0]^T
2	[20,0,-17,0,0,0,0,0,0,0,0,0,12,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]^T
3	[7,0,5,0,0,3,0,0,0,0,0,0,0,-5,0,0,0,0,0,0,0,0,0,0,0,0,-7,0,0,0]^T
4	[0,0,0,17,0,0,0,0,0,0,0,8,0,0,0,-7,0,12,0,0,0,0,0,,0,3,0,0,0,0,-8,0]^T
5	[0,0,-9,-5,0,-9,0,0,0,0,0,0,0,0,0,0,-2,0,0,5,0,0,0,0,,4,0,0,0,0,0,0,8]^T
6	[0,0,0,0,0,0,5,0,0,7,0,0,0,0,0,7,0,0,0,5,0,0,0,0,0,0,0,5,9,9,0,0]^T
7	[0,2,4,0,0,0,0,0,0,0,0,0,2,0,3,0,0,-2,0,2,0,0,0,,0,0,0,3,2,0,0,0]^T
8	[5,0,4,5,0,0,0,0,0,0,0,0,0,0,0,0,0,9,5,0,4,5,0,0,0,0,0,0,0,0,0,9,6]^T
9	[0,0,2,-4,-3,0,0,0,0,0,-4,3,0,0,0,0,-4,0,0,0,0,0,0,0,0,0,4,3,0,2,-1]^T
10	[0,-5,0,0,0,0,-9,0,-7,0,0,0,-4,0,0,0,-5,-7,0,0,0,-2,-9,0,0,0,0,-3,-8,0,-5]^T
11	[0,7,9,9,0,0,0,0,0,17,0,0,0,10,7,19,0,0,0,0,10,0,0,14,13,0,0,0,3,-10,0]^T
12	[0,3,0,-2,0,0,0,0,1,0,0,-4,-3,2,0,0,0,-5,1,4,0,-3,0,2,0,3,0,2,2,4,0]^T

Table 3: Models considered in the simulation study for a 2^5 unreplicated design (Scenario A).

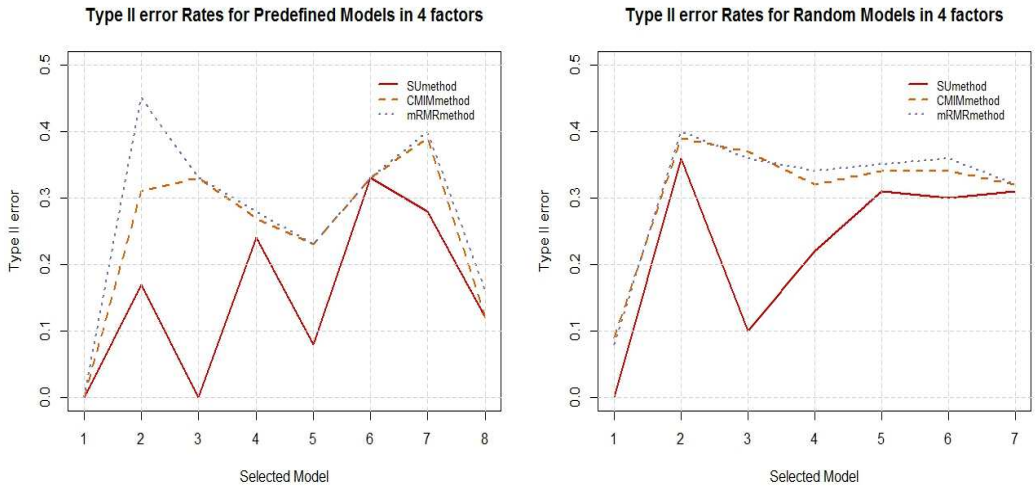

Figure 1: Comparisons of Type II error rates for Scenario A(left panel) and Scenario B (right panel) in case of four factor model. SU algorithm vs CMIM and mRMR (Scenario A).

Table 4 shows that SU algorithm achieves the lowest error rates outperforming the other two methods. More precisely, SU gathers extremely low average value of Type II while keeping low values of Type I error. Figure 1 illustrates the performance of the proposed methods considering Type II errors for scenario A and scenario B at the left and right panel, respectively, considering the design with the four factors. As depicted in this Figure, SU reveals extremely better results compared to CMIM and mRMR in all the considered cases, establishing its effectiveness.

	Type I Error			**Type II Error**		
Model	**SU**	**CMIM**	**mRMR**	**SU**	**CMIM**	**mRMR**
1	0.04	0.04	0.03	0.00	0.33	0.26
2	0.04	0.04	0.04	0.00	0.33	0.33
3	0.19	0.08	0.04	0.00	0.39	0.21
4	0.15	0.06	0.10	0.00	0.26	0.42
5	0.17	0.08	0.08	0.14	0.29	0.28
6	0.08	0.07	0.09	0.14	0.24	0.32
7	0.13	0.09	0.10	0.00	0.26	0.30
8	0.22	0.18	0.15	0.00	0.43	0.36
9	0.14	0.08	0.13	0.09	0.17	0.27
10	0.18	0.20	0.21	0.27	0.36	0.39
11	0.21	0.19	0.22	0.08	0.29	0.34
12	0.04	0.24	0.17	0.13	0.23	0.18
Average	**0.13**	**0.11**	**0.12**	**0.07**	**0.30**	**0.31**

Table 4: 2^5 unreplicated design: Performance of the proposed method for models 1-12 (Scenario A).

Tables 5 and 6 are referred to scenario B. According to the simulation scheme, first column shows the number of true active effects in the simulated models which were selected randomly, and the next columns are referred to the average values of the Type I and Type II error rates for the examined approaches. A four factor unreplicated design is considered and seven different active factors from 1 to 7 were taken. Observing Table 5 we could confirm that the SU algorithm achieves an excellent performance since it has the lowest percentages of both Type I and Type II errors, say 0.08 and 0.23, while CMIM and mRMR achieve almost similar results with average values equal to 0.13 and 0.32 for Type I and Type II errors. Lastly, Table 6 aggregates the results for a five-factor unreplicated design considering different number of active factors, varying from 1 to 15. The average values of Type I error, show that SU overall outperforms the other algorithms. It is obvious that in terms of Type II error rates, SU revealed much better performance. Figure 2 illustrates a comparison of Type II error rates for scenario A (left panel) and scenario B (right panel), considering the design with the five factors. The horizontal axes show the active factors that we ex-

amined each time while the vertical axes the percentage of the Type II error. It should be noted that in cases of 12 active factors and above, the performance is relatively smaller. This fact justified by the assumption of effect sparsity which holds in the present experiment.

	Type I Error			**Type II Error**		
Active effects	**SU**	**CMIM**	**mRMR**	**SU**	**CMIM**	**mRMR**
1	0.00	0.01	0.01	0.00	0.09	0.08
2	0.02	0.06	0.15	0.36	0.39	0.40
3	0.12	0.08	0.10	0.10	0.37	0.36
4	0.13	0.12	0.12	0.22	0.32	0.34
5	0.10	0.17	0.17	0.31	0.34	0.35
6	0.09	0.21	0.23	0.30	0.34	0.36
7	0.08	0.27	0.28	0.31	0.32	0.32
Average	**0.08**	**0.13**	**0.15**	**0.23**	**0.31**	**0.32**

Table 5: 2^4 unreplicated design: Performance of the examined methods for random model coefficients (Scenario B).

	Type I Error			**Type II Error**		
Active effects	**SU**	**CMIM**	**mRMR**	**SU**	**CMIM**	**mRMR**
1	0.00	0.02	0.01	0.00	0.02	0.04
2	0.01	0.03	0.03	0.30	0.38	0.39
3	0.05	0.04	0.04	0.10	0.34	0.30
4	0.16	0.05	0.05	0.07	0.30	0.29
5	0.17	0.09	0.06	0.09	0.32	0.32
6	0.18	0.07	0.07	0.12	0.31	0.31
7	0.17	0.11	0.10	0.18	0.31	0.32
8	0.17	0.12	0.11	0.21	0.32	0.32
9	0.16	0.15	0.14	0.25	0.32	0.32
10	0.15	0.16	0.16	0.28	0.33	0.33
11	0.15	0.20	0.19	0.31	0.33	0.33
12	0.14	0.21	0.21	0.37	0.33	0.34
13	0.14	0.23	0.25	0.38	0.32	0.33
14	0.12	0.26	0.26	0.41	0.32	0.32
15	0.13	0.29	0.29	0.44	0.31	0.31
Average	**0.13**	**0.14**	**0.13**	**0.23**	**0.30**	**0.30**

Table 6: 2^5 unreplicated design: Performance of the examined methods for random model coefficients (Scenario B).

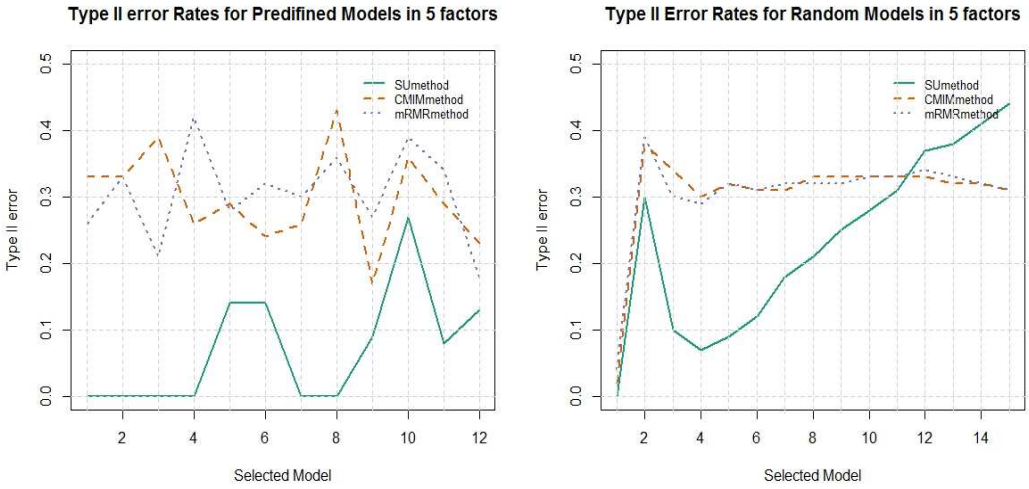

Figure 2: Comparisons of Type II error rates for Scenario A(left panel) and Scenario B (right panel) in case of five factor model. SU algorithm vs CMIM and mRMR.

3.3. Real experiment

In this subsection, we examine how the proposed screening methodology performs in the presence of real data. More precisely, we examined a real medical dataset that was collected in an annual registry conducted during the period 01/01/2005−31/12/2005 by the Hellenic Trauma and Emergency Surgery Society and which involves 30 General Hospitals in Greece. Each week, there was selected two data sets, each forms a factorial design with four and five factors, respectively, according to medical advice. There was the necessity of finding significant factors and their interactions without using an extremely large number of patients. For each patient a corresponding response variable, y, was reported which takes only two possible outcomes, denoted as 0 for survival and 1 for death. Taking all the interactions among factors a factorial design without replicates was formed.

Variable	Description
x_{55}	immobility of limbs (0 = no, 1 = yes)
x_{56}	fluids (0 = no, 1 = yes)
x_{64}	Radiograph E.R. (0 = no, 1 = yes)
x_{72}	surgical intervention (0 = no, 1 = yes)

Table 7: Description of variables for a 2^4 experiment.

This experiment helped us to confirm the effectiveness of our method to identify the significant factors in real life problems. This case study is of particular

interest since one can identify the most significant variables and their interactions with respect to a certain effect (survival or death). The main purpose of the present real case study is to validate the practical use of our approach and to give some insights into how the proposed screening procedure contributes in real life scenarios. First of all, we present the analysis of the real data in the presence of four factors. Table 7 gives a description of the variables used in our study. Table 8 presents the merits of this experiment. We denote variable x_{55} as factor A, variable x_{56} as factor B, variable x_{64} as factor C and x_{72} as the D factor. According to this notation we present the second order interactions of variables x_{55} and x_{56} as AB, of variables x_{55} and x_{64} as AC and so on. In this way we acquired the third and fourth order interactions of factors presented in Table 8. As we can conclude, all the applied methods recognize exactly the same significant variables something that confirms the efficiency of our algorithm to correctly identify significant factors.

Method	A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	ABCD
SU	•	•	○	○	•	○	○	○	•	○	○	○	○	○	•
CMIM	•	•	○	○	•	○	○	○	•	○	○	○	○	○	•
mRMR	•	•	○	○	•	○	○	○	•	○	○	○	○	○	•

Table 8: Significant variables for the real medical dataset using a 2^4 experiment.

Variable	**Description**
x_{36}	x36: major doctor (0 = no, 1 = yes)
x_{55}	immobility of limbs (0 = no, 1 = yes)
x_{56}	fluids (0 = no, 1 = yes)
x_{64}	Radiograph E.R. (0 = no, 1 = yes)
x_{72}	surgical intervention (0 = no, 1 = yes)

Table 9: Description of variables for a 2^4 experiment.

The second stage of this real experiment regards to the full factorial unreplicated design with five factors. Table 9 summarizes the description of the five variables used for this case study. In the same way as that of the four-factor case, variable x_{36} is denoted as factor A, variable x_{55} as factor B, variable x_{56} as factor C, variable x_{64} as factor D and x_{72} as the E factor. According to this notation, we present the second order interactions of variables x_{36} and x_{55} as AB, of variables x_{36} and x_{56} as AC and so on. In this way we acquired the third, fourth and fifth order interactions of the factors presented in Tables 10 and 11. Observing Tables 10 and 11 there are some interesting results that should be highlighted. First and foremost, SU seems to identify only the most significant variables and do not add additional and possibly unnecessary information in the final model. This fact leads to low levels of Type I error rates something that proved in the previous section through the simulation study. The aforementioned fact is also confirmed through the results of the other applied algorithms. We should state

that all the other methods identify an additional significant factor which is different for each method; for instance, CMIM remarks ACDE as a significant one, and mRMR factor A. It should be noted that applying mRMR algorithm requires the number of significant factors as an input. As a consequence we present both: the first nine (9 sig.) and the first ten significant (10 sig.) factors, respectively. When nine factors were requested, the results were exactly the same as SU. This fact confirms that the effects identified by SU algorithm are the most significant ones.

Method	A	B	AB	C	AC	BC	ABC	D	AD	BD	ABD	CD	ACD	BCD	ABCD	E
SU	○	•	•	○	○	•	•	•	○	○	○	○	○	○	•	○
CMIM	○	•	•	○	○	•	•	•	○	○	○	○	○	○	•	○
mRMR (9 sig.)	○	•	•	○	○	•	•	•	○	○	○	○	○	○	•	○
mRMR(10 sig.)	•	•	•	○	○	•	•	•	○	○	○	○	○	○	•	○

Table 10: Significant variables for real medical dataset using a 2^5 experiment.

Method	AE	BE	ABE	CE	ACE	BCE	ABCE	DE	ADE	BDE	ABDE	CDE	ACDE	BCDE	ABCDE
SU	○	○	•	○	○	•	○	○	○	○	○	○	○	○	•
CMIM	○	○	•	○	○	•	○	○	○	○	○	○	•	○	•
mRMR (9 sig.)	○	○	•	○	○	•	○	○	○	○	○	○	○	○	•
mRMR (10 sig.)	○	○	•	○	○	•	○	○	○	○	○	○	○	○	•

Table 11: Significant variables for real medical dataset using a 2^5 experiment (continue).

4. CONCLUDING REMARKS

Unreplicated experiments can be conducted in various improvement processes due to their economic run size and structure. However, the analysis of unreplicated designs doesn't constitute an easy issue since there are no degrees of freedom to estimate the experimental error. This fact makes the analysis of variance of such designs infeasible. An additional hindrance is that of dealing with a non-normal response, fon instance a binary one. In this work, we propose a method for selecting the active effects in unreplicated designs, assuming a logistic regression model. We take advantage of the simpleness and the effectiveness of the SU measure so as to introduce a new method for analyzing unreplicated factorials. The novelty of the proposed method is contained on the usage of information gain and symmetrical uncertainty for analyzing unreplicated designs with a binary response. The simulation study of section 3 shows that the proposed method tends to declare at the highest rate inactive effects to be active and at the lowest rate active effects to be inactive. Compared with CMIM and mRMR, our approach has an almost similar performance concerning Type I error rates; however, Type II error is notably higher for both CMIM and mRMR leading to an unstable performance compared to SU. This fact, simultaneously leads to a very satisfactory power, that is 1-(Type II error rate), of the algorithm, something that constitutes an extremely characteristic for a screening procedure, such as the analysis of unreplicated designs. In conclusion, SU achieves a general stable

performance and yields significantly low Type II errors, while it keeps Type I at a low level as well. It should be highlighted that there are problems, especially in real life, where one needs to perform an economic experiment with the smallest possible error. SU method gave the best average results in case of a real medical analysis not only by identifying the significant factors but also by keeping low Type I error rates. The empirical performance of the proposed algorithm reveals that this new approach constitutes a very efficient way of tackling the problem of unreplicated factorial designs while it opens new research opportunities for the application of information-theoretic methods in experimental designs where there are no degrees of freedom to estimate the experimental error.

ACKNOWLEDGMENTS

The research of the first author (K.D.) was financially supported by a scholarship awarded by the Secretariat of the Research Committee of National Technical University of Athens.

REFERENCES

[1] Aboukalam, M.A.F. (2005). Quick, easy and powerful analysis of unreplicated factorial designs, *Communications in Statistics-Theory and Methods*, **34**, 1169–1175.

[2] Angelopoulos, P. and Koukouvinos, C. (2008). Detecting active effects in unreplicated designs, *Journal of Applied Statistics*, **35**, 277–281.

[3] Angelopoulos, P., Evangelaras, H. and Koukouvinos, C. (2010). Analyzing unreplicated 2^k factorial designs by examining their projections into $k-1$ factors, *Quality and Reliability Engineering International*, **26**, 223–233.

[4] Angelopoulos, P., Koukouvinos, C. and Skountzou, A. (2013). Clustering effects in unreplicated factorial experiments, *Communications in Statistics-Simulation and Computation*, **42**, 1998–2007.

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

[6] Chen, Y. and Kunert, J. (2004). A new quantitative method for analysing unreplicated factorial designs, *Biometrical Journal*, **46**, 125-140.

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

[8] Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials, *Statistica Sinica*, **3**, 209–217.

[9] Fleuret, F. (2004). Fast binary feature selection with conditional mutual information, *Journal of Machine Learning Research*, **5**, 1531–1555.

[10] Hamada, M. and Balakrishnan, N. (1998). Analysing unreplicated factorial experiments: A review with some new proposals, *Statistica Sinica*, **8**, 1–41.

[11] Lenth, R.V. (1989). Quick and easy analysis of unreplicated factorial, *Technometrics*, **31**, 469-473.

[12] Lin, D.K.J. (1995). Generating systematic supersaturated designs, *Technometrics*, **37**, 213−225.

[13] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

[14] Miller, A. (2005). The analysis of unreplicated factorial experiments using all possible comparisons, *Technometrics*, **47**, 51-63.

[15] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2006). *Introduction to Linear Regression Analysis*, 4th ed. New York: Wiley.

[16] Myers, R.H; Montgomery, D.C. and Vining G.G. (2002). *Generalized Linear Models. With applications in Engineering and the Sciences*, John Wiley and Sons, New York.

[17] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models, *Journal of the Royal Statistical Society Series A*, **135**, 370–384.

[18] Quinlan, J.R. (1986). Induction of decision trees, *Machine Learning*, **1**, 81–106.

[19] Peng, H., Long, F. and Ding, C. (2005). Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1238.

[20] Press, W.H.; Flannery, B.P. and Teukolsky, S.A. and Vetterling, W. T. (1988). *Numerical Recipes*, Cambridge University Press, Cambridge.

[21] Shannon, C.E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, **27**, 379–423 and 623–656.

[22] Voss, D.T. and Wang, W. (2006). On adaptive testing in orthogonal saturated designs, *Statistica Sinica*, **16**, 227–234.

[23] Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the Twentieth International Conference on Machine Learning* (ICML-2003), Washington DC, pp. 856–863.