Analyzing Supersaturated Designs by means of an Information Based Criterion

Christos Koukouvinos, Christina Parpoula

Analyzing Supersaturated Designs by means of an Information Based Criterion

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2010-0271.R3
Manuscript Type:	Original Paper
Date Submitted by the Author:	30-Mar-2011
Complete List of Authors:	Koukouvinos, Christos; National Technical University of Athens Parpoula, Christina; National Technical University of Athens, Mathematics
Keywords:	Screening, Supersaturated design, Information theory, Symmetrical uncertainty, ROC
Abstract:	In this paper, we propose a method for analyzing data using a correlation-based measure, named as symmetrical uncertainty. This method combines measures from the information theory field and is used as the main idea of variable selection algorithms developed in data mining. In this work, the symmetrical uncertainty is used from another viewpoint in order to determine more directly the important factors.We evaluate our method by using some of the existing supersaturated designs, obtained according to methods proposed by Tang and Wu \cite{Tang1997} as well as by Koukouvinos et al. \cite{Simos2008}.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
CS-SC.tex	

Analyzing Supersaturated Designs by means of an Information Based Criterion

C. Koukouvinos and C. Parpoula

Department of Mathematics
National Technical University of Athens
15773 Zografou, Athens, Greece
e-mail: ckoukouv@math.ntua.gr, parpoula.ch@gmail.com

Abstract

The cost and time consumption of many industrial experimentations can be reduced using the class of supersaturated designs since this can be used for screening out the important factors from a large set of potentially active variables. A supersaturated design is a design for which there are fewer runs than effects to be estimated. Although there exists a wide study of construction methods for supersaturated designs, their analysis methods are yet in an early research stage. In this paper, we propose a method for analyzing data using a correlation-based measure, named as symmetrical uncertainty. This method combines measures from the information theory field and is used as the main idea of variable selection algorithms developed in data mining. In this work, the symmetrical uncertainty is used from another viewpoint in order to determine more directly the important factors. The specific method enables us to use supersaturated designs for analyzing data of generalized linear models for a Bernoulli response. We evaluate our method by using some of the existing supersaturated designs, obtained according to methods proposed by Tang and Wu [31] as well as by Koukouvinos et al. [18]. The comparison is performed by some simulating experiments and the Type I and Type II error rates are calculated. Additionally, Receiver Operating Characteristics (ROC) curves methodology is applied as an additional statistical tool for performance evaluation.

Key words and phrases: Screening, Supersaturated design, Information theory, Symmetrical uncertainty, Generalized linear models, Error rates, ROC.

AMS Subject Classification: 62K15, 62-07, 62J12

1 Introduction

Designed experiments are used in a very wide range of industry applications since they aim at evaluating the performance of a system, or optimizing its performance in terms of one or more output responses. Typically these experiments involve many potentially important factors, of which only a few are expected to be active. However, the effective factors are not known a priori. The situation where many effects are unimportant is called "effect

sparsity" and this phenomenon was studied by Box and Meyer in [7]. To save experimental cost and time, experimenters have to investigate methods for the reduction of the number of factors. The objective of screening experiments is to identify the factors of an experiment that mainly affect its performance and to investigate those that have dominant effects. In such conditions, the experimenter settles on only up to p active factors from the initial set of m factors involved in the experiment.

To address the challenges of these industrial experimentations, research in experimental design has lately focused on the class of supersaturated designs (SSDs) for their run-size economy and mathematical novelty. A supersaturated design is a design for which there are fewer runs than effects to be estimated, that is a design with m factors and n runs where $n \leq m$. The idea of SSDs was initiated by random balance experimentation, which was customary in industry in 1950s, and more specifically, SSDs were proposed by Satterthwaite in [29]. The basic principle of these experiments is that the combinations of the factor level are chosen randomly, but having the same number of runs at each level of each factor, and moreover the runs may be fewer than the included factors. In [6], Box also supported the latter idea and few years later, Booth and Cox in [5] presented the first supersaturated designs. For the next thirty years, no more papers were published in this discipline, but Lin in [20] and Wu in [34] initiated a new period of research about SSDs. Since then, many works which concern construction methods of SSDs have been published, for example, among others, [24], [10] and [9].

In contrast with the wide study of construction methods of SSDs, their analysis methods are yet in an early research stage, although many different approaches for analyzing SSDs are provided in the statistical literature. Hamada and Wu [15] used the Plackett-Burman designs in screening experiments for identifying important main effects. In this paper, an iterative guided stepwise regression strategy that entertains interactions in addition to main effects, is proposed. However, their strategy provides a restricted search in a rather large model space. A half fraction of Plackett-Burman designs were used by Lin [20] which suggested and performed the forward selection method using data based on Plackett-Burman designs. Lin's analysis was used by Wang [32] which applied this analysis on the other half fraction of Plackett-Burman design and observed that four of the five active factors found in one half fraction were not found in the other. However, when effect sparsity holds, Type I errors can easily occur. A Bayesian variable selection method for analyzing experiments with complex aliasing was proposed by Chipman et al. [11], and Westfall et al. [33] suggested an error control skill in forward selection. Abraham et al. [1] applied stepwise and all-models methods to investigate the active factors, since they have pointed out that the analysis of SSDs can give uncertain result independently of the method used. Beattie et al. [3] proposed a two-stage Bayesian model selection strategy for SSDs. According to this method it is feasible to keep all possible models under consideration while providing a level of robustness similar to Bayesian analysis incorporating noninformative priors. A variable selection approach for identifying the active effects, based on penalized least squares was proposed by Li and Lin [19]. Holcomb et al. [16] proposed contrast-based methods, while Zhang et al. [36] suggested a method based on partial least squares.

The rest of this paper is organized as follows. In Section 2, we describe the statistical method for analyzing supersaturated designs by using an information-based measure. In

Section 3, we describe the criteria used to evaluate the performance of the proposed method. In Section 4, we perform some simulation experiments to evaluate the merits of the proposed method. In Section 5, the proposed method is compared to other methods. Finally, in Section 6, the obtained results are discussed and some concluding remarks are made.

2 A method for analyzing supersaturated designs with an information-based measure

In this section, we present a method that can be applied to supersaturated designs for screening out the important factors from a large set of potentially active factors. As we have already mentioned, in screening experiments, the experimenter can assume that there are only up to p active factors from the total set of m factors involved (effect sparsity).

A logistic regression (LR) model, in which the response variable has only two possible outcomes, denoted by 0 and 1, has been used for the implementation of the proposed method. LR model is part of a category of statistical models called generalized linear models (GLMs) [2].

If y_i, i=1,2,...,n, represent the response values, then the GLM is given by

$$g(\mu_i) = g[E(y_i)] = \mathbf{x}_i'\boldsymbol{\beta},$$

where $\mathbf{x}_i$ is a vector of regressor variables or covariates for the i-th observation and $\boldsymbol{\beta}$ is the vector of parameters or regression coefficients. A GLM has three components - a response variable distribution, a linear predictor that involves the regressor variables or covariates, and a link function g that connects the linear predictor to the natural mean of the response variable. Consider the situation in which, at the i-th data point, the response is a Bernoulli random variable y_i, that takes on only two possible values, 0 and 1, where

$$E(y_i) = P_i = P(\mathbf{x}_i),\ i=1,2,...,n.$$

Here, P_i is the probability in a Bernoulli process,

$$Var(y_i) = P_i(1 - P_i),$$

and $\mathbf{x}_i$ is a vector of predictor variables. The parameter P_i, and consequently the variance, is a function of the regressors $\mathbf{x}_i$.

We now present the Logistic Regression Model used in screening experiments as well as in the simulation experiments in the present work.

The dependent variable in logistic regression is usually dichotomous, i.e., the dependent variable can take the value 1 with a probability of success q, or the value 0 with probability of failure *1-q*. Thus, it is a Bernoulli (or binary) variable. Suppose that there are n experimental runs with a binary response y (0 or 1) depending on a vector of regressor variables $\mathbf{x}_i$ for the i-th observation. If we arbitrarily denote y=*1* for a success and y=*0* for a failure, then we are truly modeling the mean response $P(x_i)$, where $P(x_i)$ is the success probability and x_i denotes the covariates or regressors at the i-th data point. The logistic model for $P(x_i)$ is then given by

$$P(x_i) = 1/(1 + e^{-x_i'\beta}) \quad (1)$$

where the term $x_i'\beta = \beta_0+\beta_1 x_{i1}+...+ \beta_k x_{ik}$ is said to be the linear predictor. We note that $0 \leq P(x_i) \leq 1$ which renders the logistic model quite natural [23]. For more details on logistic regression model, we refer the interested reader to [22].

For our simulation experiments, we develop logistic models with coefficients taking random values from the vector β and only main effects models were considered. To generate the experiments we used the following simulation protocol.

Suppose that the model has the form

$$y_i = P(x_i) + \varepsilon$$

where the quantity $P(x_i)$ is defined in Eq.(1) in the aforementioned section, and the response variable y_i takes on the value either 0 or 1. Here the quantity ε may assume one of two possible values. If y=1 then $\varepsilon = 1 - P(x_i)$ with probability $P(x_i)$, and if y=0 then $\varepsilon = -P(x_i)$ with probability $1 - P(x_i)$. Thus, ε has a distribution with mean zero and variance $P(x_i)[1 - P(x_i)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution with probability given by the conditional mean $P(x_i)$.

2.1 Variable selection via symmetrical uncertainty

The aim of factor screening will be to identify those factors which have non-zero effects. The proposed method is based on an information theoretic approach. More precisely, it uses the entropy of the model, and hence of the factors, which is the number of bits it would take, on average, to correct the output of the model [14]. Entropy is a measure of the uncertainty or unpredictability in a system. Suppose X is a random variable with probability distribution $p(X = x) = p(x)$ and Y is a random variable with probability distribution $p(Y = y) = p(y)$, the entropy of the variable X is defined as

$$H(X) = -\sum_{x \in X} p(x) log_2(p(x)),$$

where $p(x)$ is the prior probabilities for all values of X. The information entropy of the distribution over Y, or the information entropy of a Bernoulli trial in our study, is defined as

$$H(Y) = -p(y) log_2 p(y) - \{1 - p(y)\} log_2\{1 - p(y)\}.$$

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

Entropy of a variable X after observing values of another variable Y is defined as

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) log_2(p(x|y)).$$

where the conditional probability $p(x|y) = p(X = x|Y = y)$ is the posterior probabilities of X given the values of Y. The amount by which the entropy of X decreases reflects additional information X provided Y and is called *information gain* [28], or alternatively, *mutual information* [30]. Information gain is given by

$$I(X|Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(Y, X).$$

It is easy to prove that information gain is a symmetrical measure for two random variables X and Y. Symmetry is a desired property for a measure of correlations between factors, but information gain is biased in favor of factors with more values. Furthermore, the values have to be normalized to ensure they are comparable and have the same affect. *Symmetrical uncertainty* [27] compensates for information gain's bias towards factors with more values and normalizes its value to the range $[0, 1]$. Symmetrical uncertainty is defined by the form

$$SU(X, Y) = 2 \times \left[\frac{I(X|Y)}{H(X)+H(Y)}\right].$$

This measure has been used to the development of a feature selection algorithm, known as Fast Correlation Based Filter (FCBF), which is widely used in data mining. We must observe that the FCBF was selected among other feature selection algorithms because FCBF runs significantly faster (in degrees), a fact which verifies FCBF's superior computational efficiency. Additionally, FCBF is the most effective algorithm due to the fact that it can remove a large number of features that are redundant, [35], compared to other related feature selection algorithms, i.e., ReliefF [17], CorrSF [14], ConsSF [13]. Moreover, a new algorithm, the Pearson Redundancy Based Filter (PRBF) [4], after empirical comparisons with the aforementioned state-of-the-art features selection algorithms (FCBF, CorrSF, ReliefF and ConnSF), had very encouraging results but was not better than FCBF.

The proposed method, in this paper, is a modification of FCBF, since it does a typical variable selection using SU coefficient to determine the significant factors, setting some threshold value of the form $SU > \delta$. The proposed procedure can be described as follows:

1. Given a $n \times k$ supersaturated design matrix $X = [x_1, x_2, \ldots, x_k]$, where x_j, $j = 1, 2, \ldots, k$, is a column of the matrix, as well as an $n \times 1$ vector Y, which is the common response vector, compute the entropy and the conditional entropy with respect to the response variable.

2. Firstly, compute the vector of entropy values $H(X) = (H(x_1), H(x_2), \ldots, H(x_k))$, where $H(x_j)$, for $j = 1, 2, \ldots, k$, is the corresponding value of the entropy $H(X)$ measure for the j-th variable. Furthermore, compute the conditional entropy values $H(X|Y) = (H(x_1|y), H(x_2|y), \ldots, H(x_k|y))$, where $H(x_j|y)$, for $j = 1, 2, \ldots, k$, is the corresponding value of the conditional entropy $H(X|Y)$ measure for the j-th variable.

3. Compute the vector of information gain values $I(X|Y) = (I(x_1|y), I(x_2|y), \ldots, I(x_k|y))$, where $I(x_j|y) = H(x_j) - (H(x_j|y))$, for $j = 1, 2, \ldots, k$, is the corresponding value of the information gain $I(X|Y)$ measure for the j-th variable.

4. Finally, compute the symmetrical uncertainty measure $SU = (su_1, su_2, \ldots, su_k)$, where $su_j = 2 \times \left[\frac{I(x_j|y)}{H(x_j)+H(y)}\right]$, for $j = 1, 2, \ldots, k$, is the corresponding value of the SU measure for the j-th variable, with respect to the response variable.

5. If the SU measure of a variable is at least as large as a predefined threshold value, then the corresponding variable is considered to be significant, otherwise it is declared to be an unimportant one.

3 Criteria for Performance Evaluation

We used two criteria for evaluating the performance of the proposed method. Initially, the Type I and Type II error rates were used as first criterion. As with many decision problems, errors of various types must be balanced against costs. In screening designs, there is a cost of declaring an inactive factor to be active (Type I error), and also a cost of declaring an active effect to be inactive (Type II error). Type II errors are troublesome, as addressed in [21], as well as Type I errors, since they can result in unnecessary cost in follow-up experiments. Under situations of effect sparsity, Type I errors are very likely.

Given a classifier and an instance, there are four possible outcomes. If the instance is positive (P) and it is classified as positive, it is counted as a true positive (TP); if it is classified as negative (N), it is counted as a false negative (FN). If the instance is negative and it is classified as negative, it is counted as a true negative (TN); if it is classified as positive, it is counted as a false positive (FP). Given a classifier and a set of instances (the test set), a two-by-two confusion matrix (also called a contingency table) can be constructed representing the dispositions of the set of instances. The numbers along the major diagonal represent the correct decisions made, and the remaining numbers represent the errors, the confusion, between the various classes. Table 1 displays the confusion matrix. Finally, we define

$$\text{Type I error} = FP/(TN + FP)$$

and

$$\text{Type II error} = 1\text{-}[\text{TP} / (\text{TP} + \text{FN})]$$

Table 1: Confusion matrix

True Class	P	N
Hypothesized Class Y	TP	FP
Hypothesized Class N	FN	TN

The second criterion used is the average Receiver Operating Characteristic (ROC) [26] score of the models. We constructed the ROC curve of the method and computed the Area Under the ROC curve (AUC) [8]. ROC curve is a graphical representation of the trade-offs between True Positive Rate (TPR $= TP/(TP + FN)$ and False Positive Rate (FPR $= FP/(TN + FP)$). The AUC is useful in this way because it aggregates performance across the entire range of trade-offs. The higher the AUC, the better, with 0.50 indicating random performance and 1.00 denoting perfect performance.

4 Empirical Study

To assess the performance of the proposed method we apply simulations for a range of underlying models and supersaturated designs. In our simulations, the s-block-orthogonal

two-level $E(s^2)$-optimal supersaturated designs with n runs and $m = s(n-1)$ factors, constructed by Tang and Wu [31], are included. More specifically, the supersaturated design, obtained by this method, for $m = 22$ factors in $n = 12$ runs is included in our simulations. Apart from the above mentioned supersaturated designs, the $E(s^2)-$optimal and minimax-optimal cyclic supersaturated designs constructed by Koukouvinos et al. [18] are also used. According to this construction method, the obtained $E(s^2)-$optimal and minimax-optimal cyclic supersaturated designs have n runs and $m = q \cdot (n-1)$ factors, where q is even. In our simulations we used the $E(s^2)-$optimal and minimax-optimal cyclic supersaturated designs with the following (n, m) values: $(10, 18)$, $(12, 22)$, $(14, 26)$, $(16, 30)$, $(18, 34)$, $(20, 38)$, $(22, 42)$, $(6, 10)$ and also $(8, 14)$.

During our simulation experiments, the true active variables were selected randomly from the set of $\{1, \ldots, m\}$ potentially active factors and only main effects models were taken into consideration. The coefficients of the non-active variables, in the true model, were set equal to zero. The distribution of the contrasts is not affected by the magnitude of the coefficients, but it depends on the relative size of the coefficients. Since the conditions in practice are usually different from those under simulation and that the experimenter does not know how many factors may be active, in order to examine how sensitive the results are to the selection and the number of active columns, we changed the order of columns of the active factors, used different values of β and different number of active factors for each SSD in our experiments. For this reason, we considered several models that were different in this regard. For the non-active variables, in the true model, their coefficients were set to be zero. All simulations were conducted using MATLAB code.

The threshold values that determine whether a factor is significant or not, are very crucial. 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 values for the proposed method. The final decision was made according to the threshold values which were observed to be good choices from simulation trials. We thus selected three threshold values in order to study the results, the efficiency and the stability of our method. The first two thresholds were constant numbers while the third threshold was based on the estimated values of the SU. Hence, the three applied thresholds were 0.01, 0.05 and median(SU).

In our empirical study, we used all the models listed below.

1. SSD for $m = 22$ factors and $n = 12$ runs, from [31], and β = [1, 0, 0, -13, 0, 0, 4, -3, 0, 0, -6, 0, 0, -7, 0, -24, 0, -5, 0, -21, 0, 0]$'$

2. SSD for $m = 22$ factors and $n = 12$ runs, from [18], and β = [1, 0, 0, 0, -1, 0, 1, 0, 0, 0, 0, 0, -2, 0, 0, -4, 0, 0, 0, -1, 0, 0]$'$

3. SSD for $m = 18$ factors and $n = 10$ runs, from [18], and β = [0, 0, -5, 0, -3, 4, 0, 0, 7, 0, 0, 1, 2, 0, 0, -7, 0, 3]$'$

4. SSD for $m = 18$ factors and $n = 10$ runs, from [18], and β = [0, 0, -7, 0, 0, 6, 4, 0, 0, -3, 0, 0, 1, 1, 0, 0, 0, 1]$'$

5. SSD for $m = 18$ factors and $n = 10$ runs, from [18], and β = [0, -4, 1, 0, -2, 0, 3, -4, 0, 0, 0, -6, 0, 0, -2, 0, 2, 0]$'$

6. SSD for $m = 18$ factors and $n = 10$ runs, from [18], and β = [1, 0, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0]′

7. SSD for $m = 18$ factors and $n = 10$ runs, from [18], and β = [0, 0, 0, -13, 0, -24, 0, 2, 0, 0, 11, 0, 0, 21, 0, 0, -6, 0]′

8. SSD for $m = 26$ factors and $n = 14$ runs, from [18], and β = [-2, -1, 0, 0, 2, -3, 0, 0, -3, 0, 0, 2, 0, -4, 0, 0, 0, -6, 0, 0, -4, 0, 0, 0, -5, 0]′

9. SSD for $m = 26$ factors and $n = 14$ runs, from [18], and β = [1, 0, 1, 0, 2, -4, 0, 0, 5, 0, 0, -1, 0, 0, 2, 1, 0, 7, 3, 1, 0, 0, -2, 0, 1, 0]′

10. SSD for $m = 30$ factors and $n = 16$ runs, from [18], and β = [0, -1, 0, 0, 0, 0, 2, 0, 0, 0, -5, 0, -2, 0, -1, 0, 0, -7, 0, 0, -3, 0, 0, 0, 0, -7, 0, 0, 0, 0]′

11. SSD for $m = 30$ factors and $n = 16$ runs, from [18], and β =[-2, -1, 0, 0, 0, 0, 2, 0, 0, 2, 0, 0, 0, 0, -1, 0, 0, -7, 0, 0, -3, 0, 0, 0, 0, 0, 0, 0, -4, 0]′

12. SSD for $m = 34$ factors and $n = 18$ runs, from [18], and β = [0, 0, 0, 2, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0]′

13. SSD for $m = 34$ factors and $n = 18$ runs, from [18], and β = [2, 0, 0, 3, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]′

14. SSD for $m = 34$ factors and $n = 18$ runs, from [18], and β = [2, 3, 4, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]′

15. SSD for $m = 38$ factors and $n = 20$ runs, from [18], and β = [0, 0, 0, 0, -17, 0, 0 ,0, 0, 0, -23, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 26, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]′

16. SSD for $m = 38$ factors and $n = 20$ runs, from [18], and β = [0, -5, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0 ,2 ,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0]′

17. SSD for $m = 38$ factors and $n = 20$ runs, from [18], and β = [0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0]′

18. SSD for $m = 38$ factors and $n = 20$ runs, from [18], and β = [0, 0, 0, 0, -17, 0, 0, 0, 0, 0, -23, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 26, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -21, 0, 0, 0, 0, 0]′

19. SSD for $m = 42$ factors and $n = 22$ runs, from [18], and β = [0, 0, 0, 0, -2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, -4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -5, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0]′

20. SSD for $m = 42$ factors and $n = 22$ runs, from [18], and β = [0, 0, 3, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 12, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0]′

21. SSD for $m = 10$ factors and $n = 6$ runs, from [18], and β = [-7, 0, 0, 0, 0, -7, 0, 0, 2, 0]′

22. SSD for $m = 10$ factors and $n = 6$ runs, from [18], and β = [-17, -13, 0, 0, 0, -2, 0, 0, 6, 0]′

23. SSD for $m = 14$ factors and $n = 8$ runs, from [18], and β = [1, 0, -3 ,0 ,0, -1, 0, 0, 0, 0, 7, 0, 0, 0]$'$

24. SSD for $m = 14$ factors and $n = 8$ runs, from [18], and β=[-6, 0, 0 ,0 ,2, 0, 0, 4, 0, 0, 0, -8, 0, 0]$'$

For each of these models, 1000 datasets were generated, and the obtained results after the application of our method, are presented in Table 2 in accordance with the threshold values. More precisely, in the first row of Table 2, the selected threshold value is presented. In the first column, the number that corresponds to each used model is given. Columns named as "Type I" and "Type II", present the Type I and Type II error rates that correspond to every threshold value.

Table 2: Empirical performance of the proposed method for Models 1-24

Threshold values	0.01		0.05		median(**su**)	
Model	Type I	Type II	Type I	Type II	Type I	Type II
1	0.15	0.11	0.15	0.11	0.15	0.11
2	0.25	0.17	0.25	0.17	0.25	0.17
3	0.20	0.12	0.20	0.12	0.20	0.12
4	0.27	0.14	0.27	0.14	0.27	0.14
5	0.30	0.25	0.30	0.25	0.30	0.25
6	0.36	0.00	0.36	0.00	0.36	0.00
7	0.33	0.17	0.33	0.17	0.33	0.17
8	0.37	0.10	0.37	0.10	0.37	0.30
9	0.30	0.15	0.30	0.15	0.30	0.23
10	0.36	0.12	0.36	0.12	0.36	0.50
11	0.18	0.25	0.18	0.25	0.18	0.36
12	0.27	0.00	0.27	0.00	0.41	0.00
13	0.41	0.00	0.41	0.00	0.40	0.10
14	0.24	0.20	0.24	0.20	0.45	0.20
15	0.31	0.00	0.31	0.00	0.46	0.00
16	0.18	0.25	0.18	0.25	0.47	0.25
17	0.23	0.25	0.23	0.25	0.47	0.25
18	0.09	0.25	0.09	0.25	0.47	0.25
19	0.19	0.00	0.19	0.00	0.43	0.00
20	0.21	0.40	0.21	0.40	0.46	0.20
21	0.14	0.00	0.14	0.00	0.14	0.00
22	0.17	0.25	0.17	0.25	0.17	0.25
23	0.20	0.25	0.20	0.25	0.20	0.25
24	0.20	0.25	0.20	0.25	0.20	0.25

We observe from Table 2 that the proposed method has small values for both errors for almost all the models, and the Type II error rates are considerably less than the Type I error rates for the majority of the models (16/24 models). Table 2 shows that the proposed method tends to declare at a higher rate inactive effects to be active and at a much lower rate active effects to be inactive. Thus, the proposed method is indeed conservative in this sense.

We also observe in Table 2 that the proposed method succeeds identical Type I and Type II error rates for all the models, in cases where the threshold value equals to 0.01 or 0.05. The proposed method seems to have one of these error rates to be considerably higher when the threshold value is equal to median (SU) (13/24 models). Thus, we use the threshold values 0.01 and 0.05 for further comparative study. Moreover, we note that the error rates are also low for models in which the effect sparsity principle is weaker, that is, models in which the number of active factors p is greater than $\frac{n}{2}$.

5 Comparative Study

In order to examine the efficiency of the proposed method and its novelty, we performed two chi-square tests for two-way tables. Firstly, we performed Pearson's chi-square test which is a test of independence between X and Y that involves the difference between the observed and expected frequencies [25]. Then, we performed the Likelihood ratio chi-square test [12] which is a test of independence between X and Y that involves the ratio between the observed and expected frequencies. The expected frequencies are estimated under the null hypothesis of independence for both tests.

Firstly, we computed the *p-values* of the chi-square test. We sorted the predictors by *p-value* in the ascending order to rank the variables. Then, we selected the important ones based on the previous ranking and computed the Type I and Type II error rates which occurred. We followed these steps for each Pearson and Likelihood ratio chi-square statistic. Finally, we compared these methods to the proposed one, for rank ordering of the variables for two cut-offs, i.e., 0.01 and 0.05.

Table 3: Comparative empirical performance for both thresholds for models 1-24

Threshold values	Symmetrical Uncertainty (0.01 & 0.05)		Pearson & Likelihood ratio (0.01)		Pearson & Likelihood ratio (0.05)	
Model	Type I	Type II	Type I	Type II	Type I	Type II
1	0.15	0.11	0.19	0.28	0.21	0.26
2	0.25	0.17	0.38	0.31	0.42	0.28
3	0.20	0.12	0.22	0.17	0.25	0.14
4	0.27	0.14	0.28	0.21	0.30	0.19
5	0.30	0.25	0.34	0.32	0.37	0.28
6	0.36	0.00	0.34	0.34	0.38	0.28
7	0.33	0.17	0.35	0.25	0.36	0.22
8	0.37	0.10	0.45	0.18	0.47	0.14
9	0.30	0.15	0.43	0.29	0.45	0.27
10	0.36	0.12	0.46	0.15	0.48	0.12
11	0.18	0.25	0.26	0.30	0.31	0.28
12	0.27	0.00	0.35	0.08	0.38	0.05
13	0.41	0.00	0.58	0.06	0.59	0.02
14	0.24	0.20	0.33	0.25	0.36	0.23
15	0.31	0.00	0.37	0.04	0.41	0.02
16	0.18	0.25	0.22	0.28	0.25	0.26
17	0.23	0.25	0.26	0.33	0.27	0.28
18	0.09	0.25	0.15	0.27	0.19	0.25
19	0.19	0.00	0.22	0.09	0.24	0.04
20	0.21	0.40	0.28	0.67	0.31	0.66
21	0.14	0.00	0.25	0.07	0.27	0.04
22	0.17	0.25	0.38	0.31	0.39	0.28
23	0.20	0.25	0.24	0.34	0.27	0.31
24	0.20	0.25	0.25	0.32	0.26	0.30

We observe from Table 3 that the proposed method succeeds identical Type I and Type II error rates for all the models, in both cases where the threshold value equals to 0.01 or 0.05. Moreover, we observe that the Pearson and Likelihood ratio chi-square statistics have identical error rates, considering firstly the threshold value equal to 0.01 and then equal to 0.05. The proposed method has Type I and Type II error rates considerably less than the error rates of both chi-square statistics. As a result, we assume that the proposed method using symmetrical uncertainty outperforms the Pearson and Likelihood ratio chi-square tests for both threshold values, i.e., 0.01 and 0.05.

Moreover, we put the proposed method, the Pearson chi-square test, and the Likelihood ratio chi-square test on ROC curves and we calculated their AUCs for further comparison. Note that Pearson and Likelihood ratio chi-square tests perform exactly the same way. We

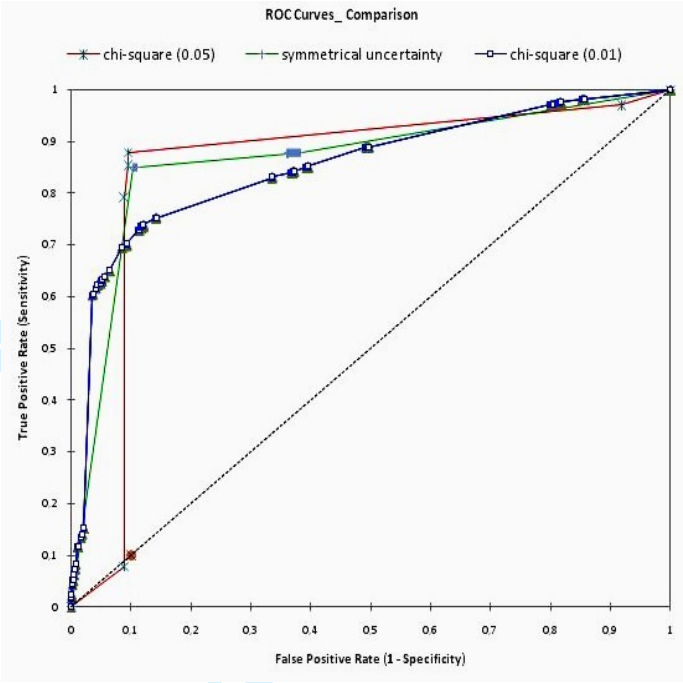


Figure 1: Comparison of ROC curves

have combined the results from the 24 different designs/models and constructed Fig. 1 which presents the comparison of the average ROC curves, generated by using symmetrical uncertainty and the chi-square test at the two different cutoffs, i.e., 0.01 and 0.05. The average ROC curve for symmetrical uncertainty is the same for both different cutoffs due to the fact that the proposed method succeeds identical Type I and Type II error rates for both 0.01 and 0.05.

A ROC curve that is located closer to the top left corner of the plot is evidently more accurate. We observe from Fig. 1 that the proposed method succeeds the highest AUC even though all ROC curves lie above the diagonal and are close to the (0,1) point, i.e., the top-left corner, indicating a good performance. The AUC (Symmetrical Uncertainty)= 0.865 > AUC (chi-square test 0.01)= 0.853 > AUC (chi-square test 0.05)= 0.851. Note that there is little difference in AUC value of the chi-square test at the two different cutoffs. The AUC is a measure of predictive power and estimates the probability that the predictions and the outcomes are concordant. The higher the AUC, the better. Precisely, the AUC of the proposed method equals to 0.865, i.e., is really close to 1.00 denoting perfect performance. Thus, ROC curves analysis revealed that the proposed method using symmetrical uncertainty is more efficient when compared to the chi-square method.

Finally, we assume that the used criteria for performance evaluation, i.e., the error rates and ROC curves analysis, both suggested that the proposed method outperforms Pearson and Likelihood ratio chi-square tests.

6 Concluding Remarks

In this paper, we have introduced a method for analyzing data in supersaturated designs using the symmetrical uncertainty measure. We approach the problem of variable selection combining fundamental measures of information theory that led to symmetrical uncertainty. Empirical performance of our method, based on simulations, is tested and the results are presented for different values of the threshold point. As with any method for analyzing supersaturated designs, high risk is expected for this method as well. Both Type I and Type II error rates are important and should be kept as low as possible. Our method succeeded low values of Type II error rates which is very crucial in order not to omit important factors of the model. Additionally, the Type I error rates were also maintained at low levels. A threshold point that implies a low Type I error rate has the ability to exclude unnecessary factors, so it can be helpful in reducing the cost of additional experiments based on the selected factors. From the simulations above, we notice that suitable for this purpose are 0.01 threshold value as well as 0.05. The symmetrical uncertainty has been used as component of data mining algorithms targeting on feature selection, but in this work it has been used in a different way. The innovation of the proposed method places in the analysis of generalized linear models for a Bernoulli response using supersaturated designs, which is currently uncommon and poorly developed type of analysis when supersaturated designs are used.

Acknowledgements

The authors would like to thank the Associate Editor and referees for constructive and useful suggestions that improved the quality of the manuscript.

References

[1] Abraham, B., Chipman, H., Vijayan, K., (1999). Some risks in the construction and analysis of supersaturated designs, *Technometrics*, **41**, 135-141.

[2] Agresti, A., (2007). *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley, New York.

[3] Beattie, S.D., Fong, D.K.F., Lin, D.K.J., (2002). A two-stage Bayesian model selection stategy for supersaturated designs, *Technometrics*, **44**, 55-63.

[4] Biesiada, J., Duch, W., (2007). Feature Selection for High-Dimensional Data: A Pearson Redundancy Based Filter, *Computer Recognitions systems 2, Kurzynski et al., Springer Berlin / Heidelberg*, **45**, 242-249.

[5] Booth, K.H.V., Cox, D.R., (1962). Some systematic supersaturated designs, *Technometrics*, **4**, 489-495.

[6] Box, G.E.P., (1959). Discussion of Satterthwaite and Budne papers, *Technometrics*, **1**, 174-180.

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

[8] Bradley, A.P., (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30** (7), 1145-1159.

[9] Bulutoglu, D.A., (2007). Cyclicly constructed $E(s^2)$-optimal supersaturated designs, *J.Statist.Plann.Inference*, **137**, 2413-2428.

[10] Bulutoglu, D.A., Cheng, C.S., (2004). Construction of $E(s^2)$-optimal supersaturated designs, *Ann.Statist.*, **32**, 1662-1678.

[11] Chipman, H., Hamada, M., Wu, C.F.J., (1997). A bayesian variable selection approach for analyzing designed experiments with complex aliasing, *Technometrics*, **39**, 372-381.

[12] Cox, D.R., Hinkley, D.V., (1974). *Theoretical Statistics*, Chapman and Hall.

[13] Dash, M., Liu, H., Motoda, H., (2000). Consistency based feature selection, *Proceedings of the Fourth Pacic Asia Conference on Knowledge Discovery and Data Mining. Springer-Verlag*, 98-109.

[14] Hall, M.A., (1999). *Correlation Based Feature Selection for Machine Learning*, Phd Thesis, Department of Computer Science, Waikato, Hamilton, New Zeland.

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

[16] Holcomb, D.R., Montgomery, D.C., Carlyle, W. M., (2003). Analysis of supersaturated designs, *Journal of Quality Technology*, **35**, 13-27.

[17] Kira, K., Rendell, L., (1992). The feature selection problem: Traditional methods and a new algorithm, *Proceedings of the Tenth National Conference on Artificial Intelligence. Menlo Park: AAAI Press/The MIT Press.*, 129-134.

[18] Koukouvinos, C., Mylona, K., Simos, D.E., (2008). $E(s^2)$-optimal and minimax-optimal cyclic supersaturated designs via multi-bjective simulated annealing, *Journal of Statistical Planning and Inference*, **138**, 1639-1646.

[19] Li, R., Lin, D.K.J., (2002). Data analysis in supersaturated designs, *Statistics and Probability Letters*, **59**, 135-144.

[20] Lin, D.K.J., (1993). A new class of supersaturated designs, *Technometrics*, **35**, 28-31.

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

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

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

[24] Nguyen, N.K., (1996). An algorithmic approach to constructing supersaturated designs, *Technometrics*, **38**, 69-73.

[25] Pearson, R.L., (1983). Karl Pearson and the chi-squared test, *International Statistical Review*, **51**, 59-72.

[26] Pepe, M.S., (2000). An interpretation for ROC curve and inference using GLM procedures, *Biometrics*, **56**, 3529.

[27] Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., (1988). *Numerical Recipes in C.*, Cambridge University Press, Cambridge.

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

[29] Satterthwaite, F.E., (1959). Random balance experimentation (with discussions), *Technometrics*, **1**, 111-137.

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

[31] Tang, B., Wu, C.F.J., (1997). A method for constructing supersaturated designs and its $E(s^2)$-optimality, *Canadian Journal of Statistics*, **25**, 191-201.

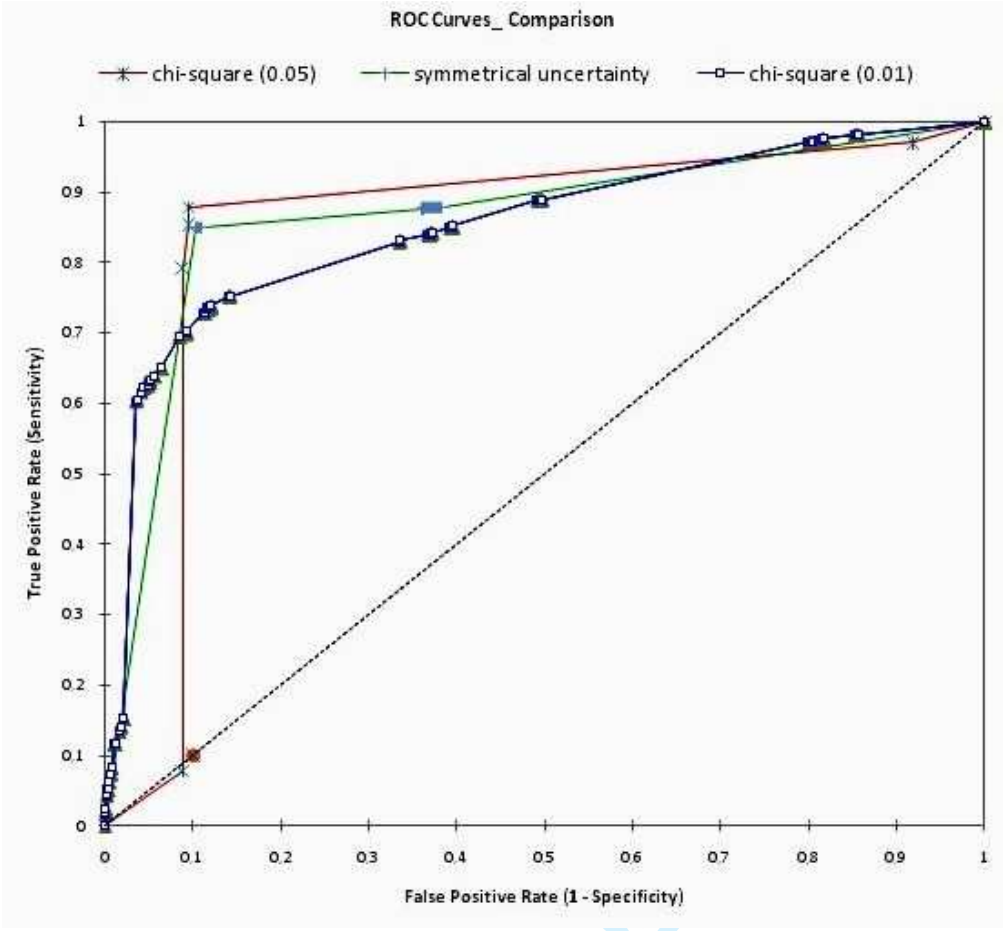
[32] Wang, P.C., (1995). Comments on Lin(1993), *Technometrics*, **37**, 358-359.

[33] Westfall, P.H., Young, S.S., Lin, D.K.J., (1998). Forward selection error control in the analysis of supersaturated designs, *Statistica Sinica*, **8**, 101-117.

[34] Wu, C.F.J., (1993). Construction of supersaturated designs through partially aliased interactions, *Biometrika*, **80**, 661-669.

[35] Yu, L., 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)*, 856-863, Washington DC.

[36] Zhang, Q.Z., Zhang, R.C, Liu, M.Q., (2007). A method for screening active effects in supersaturated designs, *Journal of Statistical Planning and Inference*, **137**, 235-248.



Letter to the Associate Editor

concerning the submission to Communications in Statistics - Simulation and Computation

Analyzing Supersaturated Designs by means of an Information Based Criterion

by

C. Koukouvinos and C.Parpoula

This letter concerns the revision of our submission entitled "Analyzing Supersaturated Designs by means of an Information Based Criterion" according to the Reviewer(s)' comments to Author. Taking into consideration the point you kindly mentioned at your report, we made the requested change.

1. "Please redo the figure. It is blurry. Please produce a clear version."

   Answer: Taking into consideration your suggestion, we have redone the figure and produced a more clear version. In addition to the jpg file which is used in the manuscript in Latex, we send you a doc file which includes the figure, so that you can edit the figure at the Journal needs, if it is necessary.