Journal of Data Science **12**(2014), 35-52

A New Variable Selection Approach Inspired by Supersaturated Designs Given a Large-Dimensional Dataset

Christina Parpoula[1], Krystallenia Drosou[1],
Christos Koukouvinos[1*] and Kalliopi Mylona[2]
[1]*National Technical University of Athens and* [2]*University of Southampton*

Abstract: The problem of variable selection is fundamental to statistical modelling in diverse fields of sciences. In this paper, we study in particular the problem of selecting important variables in regression problems in the case where observations and labels of a real-world dataset are available. At first, we examine the performance of several existing statistical methods for analyzing a real large trauma dataset which consists of 7000 observations and 70 factors, that include demographic, transport and intrahospital data. The statistical methods employed in this work are the nonconcave penalized likelihood methods (SCAD, LASSO, and Hard), the generalized linear logistic regression, and the best subset variable selection (with AIC and BIC), used to detect possible risk factors of death. Supersaturated designs (SSDs) are a large class of factorial designs which can be used for screening out the important factors from a large set of potentially active variables. This paper presents a new variable selection approach inspired by supersaturated designs given a dataset of observations. The merits and the effectiveness of this approach for identifying important variables in observational studies are evaluated by considering several two-levels supersaturated designs, and a variety of different statistical models with respect to the combinations of factors and the number of observations. The derived results are encouraging since the alternative approach using supersaturated designs provided specific information that are logical and consistent with the medical experience, which may also assist as guidelines for trauma management.

Key words: Generalized linear model, penalized likelihood, supersaturated design, trauma, variable selection.

1. Introduction

Extensive research into variable selection has been carried out over the past four decades, (see [23] and [24]), and many studies are related to medicine and

*Corresponding author.

biology, such as [8], [9] and [34]. Factor screening in large-dimensional problems is an essential activity in which the main goal is to identify correctly and parsimoniously the factors that have an important influence on the measured response. In screening studies many of the effects are negligible. This is known as the sparsity-of-effects principle [3], which is very important for making the analysis feasible. To enhance predictability and obtain the "best" model derived from the screening procedure, viz., the model with the smallest residual sum of squares, the traditional variable selection techniques used are stepwise deletion and subset selection.

Although these selection procedures are practically useful, they lack of theoretic properties and ignore stochastic errors inherited in the stages of variable selection. The logistic regression model is assessing association between an antecedent characteristic and a quantal outcome statistically adjusting for potential confounding effects of other covariates, but does not fit the data accurately if the observed odds ratios deviate from its assumptions [18]. Furthermore the best subset variable selection suffers from several drawbacks, the most severe of which is that it lacks of stability [4], and can be computationally time-consuming when multiple predictors are considered. Fan and Li [7] proposed a class of variable selection procedures via nonconcave penalized likelihood which are different from traditional approaches of variable selection in that they delete insignificant variables by estimating their coefficients as 0. Recent related studies include [10], [26] and [38].

This work focuses on the variable selection issue, and specifically on the problem of selecting important variables in regression problems in the case where observations and labels of a real-world dataset are available. In particular, we deal with a large-dimensional problem of statistical modelling by providing a comparative study concerning various variable selection techniques, and considering an alternative approach using supersaturated designs (SSDs).

The rest of this paper is organized as follows. In Section 2, we describe the variable selection methods employed in this work. In Section 3, we discuss the use of SSDs for variable selection given a dataset of observations. In Section 4, we describe the proposed method using SSDs, apply all the above procedures to trauma annual data collected in Greece, and the merits of the alternative approach are also evaluated. Finally, in Section 5, the obtained results are discussed and some concluding remarks are made.

2. Variable Selection Methods

2.1 Generalized Linear Models

The generalized linear model (GLM) was developed to allow us to fit regres-

sion models for univariate response data that follow the exponential family which includes among others the binomial distribution, which describes the distribution of the errors, and will be the one upon which the analysis is based, when a logistic regression model is considered. The logistic regression model is often used to analyze data arising in medical studies, where it is often the case of a binary response variable, taking two possible values 1 or 0 for "success" or "failure", respectively. In this paper, we assume some basic familiarity with logistic regression concepts. The concepts necessary for a description of the generalized linear logistic regression can be found in [16], [23], [25], and in [36].

2.2 Best Subset Variable Selection

The procedure of subset selection is used with certain criteria which combine statistical measures with penalties for increasing number of predictors in the model. Reviews can be found in [14], [24], and [29]. Basically, these criteria are classified into four categories: (1) Prediction criteria; (2) Information-based criteria; (3) Data-reuse and Data-driven procedures; (4) Bayesian variable selection.

In this work, we focus on the information-based criteria which are related to likelihood or divergence measures. The most popular criteria of this class include Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) which are derived from distinct perspectives. AIC aims at minimizing the Kullback-Leibler divergence between the true distribution and the estimated from a candidate model, and BIC aims at selecting a model that maximizes the posterior model probability [37]. AIC was proposed by Akaike [1], and it selects the model that minimizes $\mathrm{AIC} = -2l + 2q$, where l is the log-likelihood of the model and q is the number of predictors or the dimension of the covariate vector X in the model. BIC was proposed by Schwarz [33] and has a similar form to AIC except that the log-likelihood is penalized by $q\log(n)$ instead of $2q$, selecting the model that minimizes $\mathrm{BIC} = -2l + q\log(n)$, where n is the number of observations.

All subsets approach is an exhaustive search, since it searches through all possible subsets and selects the subset with the smallest residual sum of squares. However, a drawback of this method is that it is very time consuming when the number of predictors is large. This computational difficulty prevents the all subsets algorithm from being widely used when there are a large number of predictors in practical problems.

2.3 Nonconcave Penalized Likelihood Methods

Although the aforementioned techniques are useful for exploratory investiga-

tions, in situations where the number of predictor variables of interest is large, several drawbacks are introduced. Fan and Li [7] proposed a class of variable selection procedures via nonconcave penalized likelihood. Li and Lin [20] introduced an extension of this method, i.e., the nonconcave penalized likelihood approaches extend to least squares, namely nonconvex penalized least squares, and focus on the situation in which the design matrix is not full rank. The proposed methods are different from traditional approaches of variable selection in that they delete insignificant variables by estimating their coefficients as 0.

Assume that the data (x_i, Y_i) are collected independently. Conditioning on x_i, Y_i has a density $f_i(g(x_i^T\beta), y_i)$, where g is a known link function and β the d-dimensional vector of unknown coefficients. In Fan and Li [7], a form of the penalized likelihood is defined as

$$Q(\beta) \equiv \sum_{i=1}^{n} l_i(g(x_i^T\beta), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{1}$$

where $l_i = \log f_i$ denote the conditional log-likelihood of Y_i, $p_\lambda(\cdot)$ is a penalty function, and λ is an unknown thresholding parameter, which can be chosen by data-driven approaches, such as cross-validation (CV) and generalized cross-validation (GCV, Craven and Wahba [5]). In general, C_p, C_L, CV, and GCV are useful techniques for selecting a good estimate from a proposed class of linear estimates and it is argued that CV and GCV can be viewed as some special ways of applying C_L (see [19]). Maximizing the penalized likelihood function is equivalent to minimizing

$$-\sum_{i=1}^{n} l_i(g(x_i^T\beta), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2}$$

with respect to β. To obtain a penalized maximum likelihood estimator of β we minimize (2) with respect to β, for some thresholding parameter λ.

In this paper, to obtain a penalized maximum likelihood estimator of β which is the vector of the unknown coefficients in the model, several penalty functions were considered. In particular we implement the L_1 penalty $p_\lambda(|\beta|) = \lambda|\beta|$ which results in the Least Absolute Shrinkage and Selection Operator method (LASSO, [35]), the Hard thresholding penalty $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$ (see [2]), where $I(\cdot)$ is an indicator function, and the Smoothly Clipped Absolute Deviation (SCAD) penalty the first derivative of which is defined by $p'_\lambda(|\beta|) = \lambda\{I(|\beta| \leq \lambda) + ((\alpha\lambda - \beta)_+/(\alpha - 1)\lambda)I(|\beta| > \lambda)\}$, for some $\beta > 0$ and $\alpha > 2$, with $p_\lambda(0) = 0$ ([7]). For the choice of α, according to the relevant literature (see [8]), the value $\alpha \approx 3.7$ appears to perform quite satisfactorily in numerous variable selection problems.

Fan and Li ([7]) proposed a unified algorithm for the minimization of (2) by local quadratic approximations. Given an initial value $\beta^{(0)}$ that is close to the true value of β, when $\beta_j^{(0)}$ is very close to 0 we set $\hat{\beta}_j = 0$ and when $\beta_j^{(0)}$ is not very close to 0, the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by a quadratic function as $[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j$, when $\beta_j \neq 0$. In other words, $p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)^2})/2$, for $\beta_j \approx \beta_j^{(0)}$. If the first term of (2) is regarded as a loss function of β denoted by $l(\beta)$, then (2) can be written in the unified form

$$l(\beta) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3}$$

Under the assumption that the first two derivatives of the loglikelihood are continuous, the first term of (3) can be locally approximated by a quadratic function. With the local quadratic approximation, the solution can be found by iteratively computing the following expression with an initial value

$$\beta^{(0)}\colon\ \beta^{(1)} = \beta^{(0)} - \{\nabla^2 l(\beta^{(0)}) + n\textstyle\sum_\lambda(\beta^{(0)})\}^{-1}\{\nabla l(\beta^{(0)}) + nU_\lambda(\beta^{(0)})\},$$

where

$$\nabla l(\beta^{(0)}) = \frac{\partial\, l(\beta^{(0)})}{\partial\beta}, \quad \nabla^2 l(\beta^{(0)}) = \frac{\partial^2\, l(\beta^{(0)})}{\partial\beta\partial\beta^T}, \quad U_\lambda(\beta^{(0)}) = \sum_\lambda(\beta^{(0)})\beta^{(0)},$$

$$\sum_\lambda(\beta^{(0)}) = \mathrm{diag}\{p'_\lambda(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \cdots, p'_\lambda(|\beta_d^{(0)}|)/|\beta_d^{(0)}|\}.$$

When the algorithm converges, the estimator satisfies the condition $\partial l(\hat{\beta}^{(0)})/\partial\beta_j + np'_\lambda(|\hat{\beta}_j^{(0)}|)\mathrm{sgn}(\hat{\beta}_j^{(0)}) = 0$, the penalized likelihood equation, for nonzero elements of $\hat{\beta}^{(0)}$. The standard errors for the estimated parameters can be obtained directly because the parameters and selecting variables are estimated at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula can be used as an estimator for the covariance of the estimates $\hat{\beta}^{(1)}$, the nonvanishing component of $\hat{\beta}$. That is, $\widehat{\mathrm{cov}}(\hat{\beta}^{(1)}) = \{\nabla^2 l(\hat{\beta}^{(1)}) + n\sum_\lambda(\hat{\beta}^{(1)})\}^{-1}\widehat{\mathrm{cov}}\{\nabla l(\widehat{\beta}^{(1)})\}\{\nabla^2 l(\hat{\beta}^{(1)}) + n\sum_\lambda(\hat{\beta}^{(1)})\}^{-1}$. Interested reader may refer to Fan and Li [7] for more details.

3. Supersaturated Designs

Supersaturated designs (SSDs) are fractional factorial designs in which the number of factors to be estimated exceeds the number of experimental runs. SSDs are widely used in experimentations in which the main goal is to identify

the important effects efficiently in terms of minimal computational cost and time. SSDs can be generally described as designs with m factors and n runs where $n \leq m$. The idea of SSDs was initiated by Satterthwaite [31]. Even though the construction methods of SSDs have been studied extensively (see, for example, the recent reviews [12] and [17]), the analysis of SSDs still remains a very challenging task. Some new approaches for analyzing SSDs have been developed in recent years. The interested reader may refer to Li and Lin (2003) [21], Gupta and Kohli (2008) [13] for analysis methods of SSDs up until 2007, and to Georgiou (2012) [11] for a detailed review later than 2008. We do not present here more details on construction and analysis methods of SSDs, since this paper focuses on the idea of using SSDs for variable selection.

The use of experimental designs for variable selection in problems of observational data has been introduced by Pumplün *et al.* (2005) ([27] and [28]). Schiffner and Weihs (2009) [32] extended the simulation study of [27] in order to verify the results and as basis for further research in this field; the appropriateness of D-optimal plans for training classification methods was additionally investigated. Rüping and Weihs (2009) [30] proposed an algorithm inspired by statistical design of experiments and kernel methods to deal with the problem of variable selection given a database of observations. Unlike observations resulting from experimental designs, massive data sets sometimes become available without predefined purposes. Usually, it is preferable to find some interesting features in the data sets that will provide valuable information to support decision making [22]. For experimental situations where there really is no prior knowledge of the effects of factors, but a strong belief in factor sparsity, and where the aim is to find out if there are any dominant factors and to identify them, experimenters should seriously consider using SSDs as suggested in [12]. In the early literature, there are several research papers regarding the practical use of SSDs in real life problems, for example see [12] and [15]. More research is needed for the best practical usage of SSDs, since the situations in which SSDs are really promising and essentially ready for use in practice are limited. Good designs (may not optimal) are already available for many sizes of an experiment [12], but their practical usage still remains a difficult and challenging task. In our paper, we used several SSDs combined with several existing statistical analysis methods in order to deal with the problem of variable selection in a large-dimensional dataset.

4. Proposed Methodology

In this section, we examine the performance of several existing variable selection methods, as well as the alternative approach considering SSDs, for analyzing a real large annual trauma dataset. Here is a brief summary. The data were collected in an annual registry conducted by the Hellenic Trauma and Emer-

gency Surgery Society involving 30 General Hospitals in Greece. The study was designed to assess the effects of differing prognostic factors on the outcome of injured persons. Note here that 70 covariates were thought to be possible risk factors, and after medical advice, all of the factors were treated equally during the variable selection approach, meaning that there was no factor that should be always maintained in the model. Altogether $N = 7000$ patients were recorded, and for each of them the binary response variable y (death: 1, otherwise: 0) was reported.

Penalized partial likelihood approach, with the SCAD, L_1, and Hard penalty, was applied to this dataset. In order to implement these methods, we estimated the thresholding parameter λ based on data via minimizing an approximate generalized cross-validation (GCV) statistic. The constant α in the SCAD was taken as 3.7 according to suggestions in the relevant literature (see [7]). The standard error formula described in Subsection 2.3 was also computed. We considered likelihood based generalized linear models as well. From now on we shall assume that the design matrix $X = (x_{ij})$ is standardized so that each column has mean 0 and variance 1. The best subset variable selection procedure with AIC and BIC was also conducted. The analysis was organized in two subsections, in which numerical comparisons among variable selection methods are illustrated. In Subsection 4.2 we considered different SSDs for the desired analysis. All models were conducted using MATLAB codes.

4.1 Sequential Variable Selection

In this subsection, we perform sequentially several variable selection methods until we achieve the model's convergence. Model I is the initial model, viz., the whole trauma dataset consisting of $N = 7000$ patients and 70 possible risk factors. The names of the 70 available prognostic factors of the initial Model I are listed in the Appendix. Due to the fact that best subset selection procedures are extremely computationally time-expensive, their application was not possible in this stage of the medical study.

Model II was obtained from the first execution of each of the LR, SCAD, LASSO, and Hard methods, and 29 variables were identified as statistically significant (ss). These 29 ss variables (x_9, x_{11}, x_{12}, x_{13}, x_{14}, x_{16}, x_{17}, x_{18}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{26}, x_{27}, x_{29}, x_{35}, x_{36}, x_{37}, x_{38}, x_{39}, x_{40}, x_{41}, x_{42}, x_{43}, x_{44}, x_{45}, x_{46}) were selected from at least one of these methods, were kept and used for further analysis.

Model III was obtained from the second execution of each of the LR, SCAD, LASSO, and Hard methods, and 14 variables were identified as ss. These 14 ss variables were selected from at least one of these methods, were kept and used for further analysis. Note here that when the above four methods were

applied for third time, we noticed that the values of the estimated coefficients were maintained different from zero, resulting to the model's convergence. Hence, after following two-step variable selection procedures (LR, SCAD, LASSO, and Hard) we identified Model III which consists of 14 ss variables listed in Table 1. The estimated coefficients and standard errors for Model III are also reported in Table 1. The unknown parameter λ was estimated to be 0.0008, 0.0001, and 0.0008 for the SCAD, L_1 and Hard thresholding penalties respectively. With the estimated value of λ, the penalized likelihood estimator was obtained at the 2nd, 3rd and 2nd step iterations for the penalized likelihood with the SCAD, L_1 and Hard thresholding penalties respectively.

Table 1: Estimated coefficients and standard errors for Model III

Method	Logistic	SCAD	LASSO	Hard	Best Subset (BIC)	Best Subset (AIC)
Intercept	-4.08 (0.11)	-4.08 (0.11)	-4.06 (0.11)	-4.08 (0.11)	-4.06 (0.11)	-4.08 (0.11)
x_9	0.33 (0.07)	0.33 (0.07)	0.33 (0.06)	0.33 (0.07)	0.32 (0.06)	0.33 (0.07)
x_{11}	-0.47 (0.10)	-0.47 (0.10)	-0.46 (0.10)	-0.47 (0.10)	-0.47 (0.09)	-0.47 (0.10)
x_{13}	0.24 (0.06)	0.24 (0.06)	0.24 (0.06)	0.24 (0.06)	0.23 (0.06)	0.24 (0.06)
x_{16}	-0.20 (0.06)	-0.20 (0.06)	-0.20 (0.06)	-0.20 (0.06)	-0.26 (0.06)	-0.20 (0.06)
x_{20}	-0.24 (0.10)	-0.24 (0.10)	-0.23 (0.10)	-0.24 (0.10)	0 (-)	-0.24 (0.10)
x_{21}	-0.67 (0.12)	-0.67 (0.12)	-0.66 (0.12)	-0.67 (0.12)	-0.69 (0.12)	-0.67 (0.12)
x_{24}	0.65 (0.07)	0.65 (0.07)	0.65 (0.07)	0.65 (0.07)	0.63 (0.07)	0.65 (0.07)
x_{26}	-0.18 (0.08)	-0.18 (0.07)	-0.17 (0.07)	-0.18 (0.07)	0 (-)	-0.18 (0.07)
x_{29}	-0.80 (0.12)	-0.80 (0.12)	-0.79 (0.12)	-0.80 (0.12)	-0.82 (0.12)	-0.80 (0.12)
x_{37}	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)	0.14 (0.03)
x_{38}	-0.71 (0.06)	-0.71 (0.06)	-0.71 (0.06)	-0.71 (0.06)	-0.74 (0.06)	-0.71 (0.06)
x_{39}	-0.34 (0.09)	-0.34 (0.09)	-0.34 (0.08)	-0.34 (0.09)	-0.42 (0.08)	-0.34 (0.09)
x_{42}	-0.21 (0.09)	-0.21 (0.09)	-0.20 (0.09)	-0.21 (0.09)	0 (-)	-0.21 (0.09)
x_{43}	-0.46 (0.09)	-0.46 (0.09)	-0.45 (0.09)	-0.46 (0.09)	-0.45 (0.09)	-0.46 (0.09)

The application of best subset selection procedures was not possible until this stage of the experimental study due to computational complexity. We thus applied best subset variable selection methods (with AIC and BIC) to the 14 ss variables of Model III derived from penalized likelihood methodology. We observe from Table 1 that SCAD, Hard, LR, and best subset with AIC score methods yield the same model, while estimates of LASSO are systematically lower. Best subset procedure via minimizing the BIC score rejects factors which may be statistically significant. For example, the estimated coefficient for covariate "spinal column x-ray" (x_{42}) is set equal to zero whereas is selected in all other models considered in this article. Note here that the estimated standard errors for the L_1 penalized likelihood estimator are consistently smaller than LR, SCAD and Hard methods. This implies that the biases in the LASSO estimators are larger.

As far as execution times are concerned, we observe that Hard method was much faster with 67 sec versus to SCAD which needed almost 147 sec, and to LASSO which needed 490 sec. The best subset procedures via minimizing the BIC and AIC scores needed 275331 and 276422 sec, respectively. In general, Hard method was the quickest compared to all considered methods.

4.2 Implementing Supersaturated Designs for Variable Selection

In this subsection, we study the use of SSDs for variable selection given a dataset of observations. We approach the problem of variable selection by considering SSDs, and conduct several statistical analysis methods. The measure used to determine the considered SSDs is the measure of non-orthogonality between two columns, i.e., the inner products of two columns in the two level designs does not always vanish to zero.

During our experiments, only main effects models were taken into consideration. We now present the proposed procedure used in order to identify SSDs as appropriate as possible for analyzing a real large dataset.

1. Given the initial dataset consisting of m input predictor variables $\{x_{i1}, \cdots, x_{im}\}$, and N observations, $i = 1, 2, \cdots, N$, define the SSDs properties.
   • Desired number of rows (n): $n = 1/100$ of N available runs
   • Desired number of columns (m): $m + 1 > n$.

2. Choose randomly from the initial dataset 100 $(n \times m)$ plans according to the uniform distribution. We assume that each of the initial m factors has two levels, ± 1. Generate the SSD matrix $X = [1, x_1, x_2, \cdots, x_m]$ for each plan, which is the $n \times (m + 1)$ model matrix. The first column of X is $1_n = [1, \cdots, 1]^T$, with the column j corresponding to the levels of the $(j - 1)$-th factor for $j = 2, \cdots, m + 1$. From columns $2, \cdots, m + 1$ of X, m columns were assigned to each plan at random. From $1, \cdots, N$ rows, $n = N/100$ rows were assigned to each plan at random.

3. Search and exclude the plans with the worst non-orthogonality property (the highest value of degree of non-orthogonality between columns is 1).

4. Identify the SSD plan(s) with the best possible non-orthogonality property, viz., the SSD plan(s) with the minimum correlation level (coefficient) between factors.

We applied the above procedure to our initial trauma data set, and 100 ($n = 70 \times m = 70$) plans were randomly chosen according to the uniform distribution. In other words, 100 SSDs have been constructed based on random samples of

observations of the trauma dataset, viz., by random selection of $n = 70$ runs from the set of $N = 7000$ available runs. We identified 5 SSDs with the minimum correlation coefficient. Hence, we considered five SSDs each with $m = 70$ factors and $n = 70$ runs. The disadvantage of these SSDs is that because of the nature of the problem it was not possible to be balanced. An SSD for two level factors is said to be balanced if every column of the design has levels $+1$ and -1 appearing equally often. Since the factors depend on the characteristics of the studied patients and cannot be pre-specified, it was impossible for us to take measures according to the balanced optimal two-level supersaturated designs which would protect us more from the unavoidable confound of the involved statistical analysis.

We perform sequentially several variable selection methods to the constructed 5 datasets, each of them consisting of $n = 70$ patients and $m = 70$ possible risk factors, until we achieve the model's convergence. $\text{Model}_{\text{SSDs}}$ was obtained from the first execution of each of the LR, SCAD, LASSO, and Hard methods, and 13 variables were selected as ss for all the 5 SSDs considered. These 13 ss factors (x_1, x_2, x_3, x_{10}, x_{11}, x_{16}, x_{19}, x_{38}, x_{42}, x_{43}, x_{47}, x_{48}, x_{62}) were selected from at least one of these methods, were kept and used for further analysis. The application of best subset selection procedures was possible in this stage of the experiment. We thus applied LR, SCAD, LASSO, Hard, and best subset variable selection methods (with AIC and BIC) to the 13 ss factors of $\text{Model}_{\text{SSDs}}$, and identified a new model consisting of 5 ss factors. Note here that when the above methods were applied once again, we noticed that the values of the estimated coefficients were maintained different from zero, resulting to this model's convergence. Hence, we identified the "best" possible model (Final $\text{Model}_{\text{SSDs}}$) consisting of 5 ss variables, appeared in Table 2. We derived the coefficients of this model by concerning all 7000 observations, so the same abbreviations with Model III are also used here. The unknown parameter λ was estimated to be 0.0007, 0.0001, and 0.0007 for the SCAD, L_1 and Hard thresholding penalties respectively. The needed steps to obtain the penalized likelihood estimator were 2, 3, and 2 for SCAD, L_1, and Hard penalty respectively.

Note here that even if the constructed SSDs may not be the best possible, however, these SSDs achieved to recognize the 5 ss factors that were included in Model III by using only the 1/100 of available runs. We also observe a stability in the results obtained from the statistical analysis of all 5 SSDs. In medical terms, a model with the less possible important factors was desirable in our case. We observe from Table 2 that all six methods gave the same results, and hence medical community should seriously take into account these five factors in order to decrease the death rate from trauma in Greece. Indeed, the 5 common factors should be considered as statistically significant, and all the others could be maintained for further investigation. The first and third factor are referred

to actions at the side of the accident. The second one concerns the decision a doctor should take after the patient was examined in the emergency room, and the last two factors concern actions taken in the emergency room.

Table 2: Estimated coefficients and standard errors for Final $\text{Model}_{\text{SSDs}}$

Method	Logistic	SCAD	LASSO	Hard	Best Subset (BIC)	Best Subset (AIC)
Intercept	-3.93 (0.10)	-3.93 (0.10)	-3.92 (0.10)	-3.93 (0.10)	-3.93 (0.10)	-3.93 (0.10)
x_{11}	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)	0.19 (0.05)
x_{16}	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)	-1.02 (0.06)
x_{38}	-0.65 (0.05)	-0.65 (0.05)	-0.65 (0.05)	-0.65 (0.05)	-0.65 (0.05)	-0.65 (0.05)
x_{42}	-0.49 (0.09)	-0.49 (0.09)	-0.49 (0.08)	-0.49 (0.09)	-0.49 (0.09)	-0.49 (0.09)
x_{43}	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)	-0.34 (0.07)

4.3 Comparative Results of the Proposed Variable Selection Procedures

Initially, we performed sequentially several variable selection methods to the whole dataset until we achieve the model's convergence. Hence, we identified the final model, viz., Model III which consists of 14 ss variables listed in Table 1. These selected 14 ss variables are x_9, x_{11}, x_{13}, x_{16}, x_{20}, x_{21}, x_{24}, x_{26}, x_{29}, x_{37}, x_{38}, x_{39}, x_{42}, x_{43}.

We then approached the problem of variable selection by considering several two-levels supersaturated designs (SSDs). After implementing the five constructed SSDs, we conducted several statistical analysis methods until we achieve the model's convergence. Hence, we identified the "best" possible model consisting of 5 ss variables, appeared in Table 2. These selected 5 ss variables are x_{11}, x_{16}, x_{38}, x_{42}, x_{43}.

There are four significant points worth mentioning here.

1. The Final $\text{Model}_{\text{SSDs}}$ consists of 5 variables which were identified as ss for all the five constructed SSDs.

2. The 5 ss variables of Final $\text{Model}_{\text{SSDs}}$ (see Table 2) are also identified as ss and are included in Model III (see Table 1). Note here that comparing with the results of the ss variables given in Tables 1 and 2, there only exist five accordant variables (x_{11}, x_{16}, x_{38}, x_{42}, x_{43}) obtained from the two proposed variable selection procedures. These accordant variables are the final sig-

nificant variables (Final $\text{Model}_{\text{SSDs}}$) identified by implementing the SSDs methodology and the assistance of the penalized likelihood methodologies.

3. The proposed SSDs methodology achieves to recognize these 5 ss factors (Final $\text{Model}_{\text{SSDs}}$) by using only the 1/100 of available runs.

4. Best subset procedure via minimizing the BIC score rejects factors which may be ss when performing only sequential variable selection. The estimated coefficient for "spinal column x-ray" (x_{42}) is set equal to zero whereas x_{42} is selected as ss from all other considered methods (see Table 1). The proposed method considering SSDs achieved to recognize x_{42} as ss even for the best subset procedure via minimizing the BIC (see Table 2). This fact implies that the proposed SSDs methodology tends to reduce the biases for the estimators of coefficients and standard errors.

5. Concluding Remarks

Recent proliferation of large-dimensional databases makes variable selection, or dimension reduction crucial in model building and challenging due to their complicated structure. Not only does judicious variable selection improves the model's predictive ability, but it generally provides a better understanding of the underlying concept that generates the data. In the literature, little work has been done in this area via penalized likelihood methods. This article comes to present an extensive application of SCAD, LASSO and Hard methods in combination with the usage of supersaturated designs for variable selection given a dataset of observations. We observed that the logistic regression compared to the two traditional variable selection methods, fails to eliminate the possible insignificant factors, and the best subset method (with AIC or BIC) is very time consuming, in some cases even impossible. SCAD, LASSO and Hard methods outperform the abovementioned methods as they are much faster, and more effective because they select important variables via optimizing a penalized likelihood simultaneously, and hence the standard errors of estimated parameters can be estimated accurately. Therefore, we believe that since these methods are easily and quickly implemented, even in a such large-dimensional problem, they should be all applied during a statistical analysis. In this way, factors that are selected from all three penalized likelihood methods should be considered as statistically significant, and the others that are chosen from at least one of these methods could be maintained for further investigation. The effective use of supersaturated designs in the statistical analysis of our medical data is very important, since it allowed us to obtain a parsimonious and meaningful model that identifies the significant prognostic factors affecting death from trauma, using only few patients (only the 1/100 of available runs).

Appendix

Trauma Study

- (Model I) (levels ± 1):
  x_1: gender
  x_2: arterial hypertension
  x_3: coronary disease
  x_4: heart failure
  x_5: arrhythmia
  x_6: asthma
  x_7: chronic obstructive pulmonary disease (COPD)
  x_8: chronic kidney disease
  x_9: car accident
  x_{10}: none safety measures at the side of the accident
  x_{11}: collar soft at the side of the accident
  x_{12}: RL at the side of the accident
  x_{13}: unconscious at the side of the accident
  x_{14}: life belt
  x_{15}: airbags
  x_{16}: transfer to a clinic, of the hospital
  x_{17}: transfer to surgery
  x_{18}: transfer to general surgery
  x_{19}: transfer to orthopedic clinic
  x_{20}: expected big temporary handicap
  x_{21}: expected small temporary handicap
  x_{22}: recovery
  x_{23}: transport by ambulance (in other hospital)
  x_{24}: transport by ambulance in emergency room
  x_{25}: transport by car in emergency room
  x_{26}: checking in the emergency room by general surgeon
  x_{27}: checking in the emergency room by neurosurgeon
  x_{28}: ICP Monitoring
  x_{29}: DW
  x_{30}: US Triplex
  x_{31}: evacuation
  x_{32}: A.T.L.S
  x_{33}: capillary refill
  x_{34}: peritoneum points
  x_{35}: sweating
  x_{36}: peripatetic

x_{37}: central cyanosis
x_{38}: peripheral vein
x_{39}: head x-ray
x_{40}: upper part of spinal column x-ray
x_{41}: pelvis x-ray
x_{42}: spinal column x-ray
x_{43}: extremities x-ray
x_{44}: CT abdomen
x_{45}: CT extremities
x_{46}: resident on charge
x_{47}: CT thorax
x_{48}: thorax x-ray
x_{49}: nasogastric tube
x_{50}: fluids
x_{51}: chest drainage
x_{52}: catheter
x_{53}: pericardiocentesis
x_{54}: blood
x_{55}: thoracotomy
x_{56}: angiography
x_{57}: diagnostic peritoneal lavage (DPL)
x_{58}: embolism
x_{59}: toxicology testing
x_{60}: ultrasound (US)
x_{61}: urea testing
x_{62}: mild trauma
x_{63}: Radiograph E.R.
x_{64}: immobility of limbs
x_{65}: face injury
x_{66}: head injury
x_{67}: breast injury
x_{68}: spinal column injury
x_{69}: upper limbs injury
x_{70}: lower limbs injury

Acknowledgements

The authors would like to thank Professors Dennis K. J. Lin and Runze Li for sending the Matlab code for the procedures proposed in their papers, and the First Propedeutic Surgical Clinic, in Hippocratio Hospital for giving the real medical data. This made the analysis feasible, the comparisons easier, and saved us a lot of

time. The research of the first author was financially supported by a scholarship awarded by the Secretariat of the Research Committee of National Technical University of Athens. The authors would also like to thank the Associate Editor and the referees for their constructive and useful suggestions which resulted in improving the quality of this manuscript.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716-723.

[2] Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion). *Journal of the Italian Statistical Society* **6**, 97-144.

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

[4] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350-2383.

[5] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377-403.

[6] Fan, J. (1997). Comments on "Wavelets in statistics: a review" by A. Antoniadis. *Journal of the Italian Statistical Society* **6**, 131-138.

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

[8] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74-99.

[9] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians* (Edited by M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), Volume III, 595-622.

[10] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**, 928-961.

[11] Georgiou, S. D. (2012). Supersaturated designs: a review of their construction and analysis. To appear in *Journal of Statistical Planning and Inference*. http://dx.doi.org/10.1016/j.jspi.2012.09.014.

[12] Gilmour, S. G. (2006). Factor screening via supersaturated designs. In *Screening Methods for Experimentation in Industry, Drug Discovery, and Genetics* (Edited by A. Dean and S. Lewis), 169-190. Springer, New York.

[13] Gupta, S. and Kohli, P. (2008). Analysis of supersaturated designs: a review. *Journal of Indian Society of Agricultural Statistics* **62**, 156-168.

[14] Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1-49.

[15] Holcomb, D. R., Montgomery, D. C. and Carlyle, W. M. (2007). The use of supersaturated experiments in turbine engine development. *Quality Engineering* **19**, 17-27.

[16] Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley, New York.

[17] Kole, B., Gangwani, J., Gupta, V. K. and Parsad, R. (2010). Two level supersaturated designs: a review. *Journal of Statistical Theory and Practice* **4**, 598-608.

[18] Lee, J. (1986). An insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence. *International Journal of Epidemiology* **15**, 22-29.

[19] Li, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics* **15**, 958-975.

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

[21] Li, R. and Lin, D. K. J. (2003). Analysis methods for supersaturated designs: some comparisons. *Journal of Data Science* **1**, 249-260.

[22] Li, R., Lin, D. K. J. and Li, B. (2012). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 1-11.

[23] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman & Hall, London.

[24] Miller, A. J. (2002). *Subset Selection in Regression.* Chapman & Hall, Boca Raton.

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

[26] Park, M. Y. and Hastie, T. (2007). L_1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* **69**, 659-677.

[27] Pumplün, C., Rüping, S., Morik, K. and Weihs, C. (2005a). *D-optimal Plans in Observational Studies.* Technical Report. Sonderforschungsbereich 475, Complexity Reduction in Multivariate Data Structures, No. 44. Technische Universität Dortmund, 44221 Dortmund, Germany. http://www.statistik.tudortmund.de/sfb-tr2005.html.

[28] Pumplün, C., Weihs, C. and Preusser, A. (2005b). Experimental design for variable selection in data bases. In *Classification - The Ubiquitous Challenge* (Edited by C. Weihs and W. Gaul), 192-199. Springer, Berlin.

[29] Rao, C. R. and Wu, Y. (2001). On model selection (with discussion). In *Institute of Mathematical Statistical Lecture Notes - Monograph Series* (Edited by P. Lahiri), Volume 38, 1-64. IMS, Beachwood, Ohio.

[30] Rüping, S. and Weihs, C. (2009). *Kernelized Design of Experiments.* Technical Report. Sonderforschungsbereich 475, Complexity Reduction in Multivariate Data Structures, No. 02. Technische Universität Dortmund, 44221 Dortmund, Germany. http://hdl.handle.net/10419/36602.

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

[32] Schiffner, J. and Weihs, C. (2009). *D-optimal Plans for Variable Selection in Data Bases.* Technical Report. Sonderforschungsbereich 475, Complexity Reduction in Multivariate Data Structures, No. 14. Technische Universität Dortmund, 44221 Dortmund, Germany. http://hdl.handle.net/10419/41052.

[33] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

[34] Svrakic, N. M., Nesic, O., Dasu, M. R. K., Herndon, D. and Perez-Polo, J. R. (2003). Statistical approach to DNA chip analysis. *Recent Progress in Hormone Research* **58**, 75-93.

[35] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

[36] Vittinghoff, E., Glidden, D. V., Shiboski, S. C. and McCulloch, C. E. (2005). *Regression Methods in Biostatistics.* Springer, New York.

[37] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937-950.

[38] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the LASSO. *Annals of Statistics* **35**, 2173-2192.

Received December 19, 2012; accepted June 21, 2013.

Christina Parpoula
Department of Mathematics
National Technical University of Athens
15773 Zografou, Athens, Greece
parpoula.ch@gmail.com

Krystallenia Drosou
Department of Mathematics
National Technical University of Athens
15773 Zografou, Athens, Greece
drosou.kr@gmail.com

Christos Koukouvinos
Department of Mathematics
National Technical University of Athens
15773 Zografou, Athens, Greece
ckoukouv@math.ntua.gr

Kalliopi Mylona
Southampton Statistical Sciences
Research Institute
University of Southampton
Southampston SO17 1BJ, UK
k.mylona@soton.ac.uk