Journal of Data Science **10**(2012), 653-672

Variable Selection and Computation of the Prior Probability of a Model via ROC Curves Methodology

Christos Koukouvinos* and Christina Parpoula
National Technical University of Athens

Abstract: Nowadays, extensive amounts of data are stored which require the development of specialized methods for data analysis in an understandable way. In medical data analysis many potential factors are usually introduced to determine an outcome response variable. The main objective of variable selection is enhancing the prediction performance of the predictor variables and identifying correctly and parsimoniously the faster and more cost-effective predictors that have an important influence on the response. Various variable selection techniques are used to improve predictability and obtain the "best" model derived from a screening procedure. In our study, we propose a variable subset selection method which extends to the classification case the idea of selecting variables and combines a nonparametric criterion with a likelihood based criterion. In this work, the Area Under the ROC Curve (AUC) criterion is used from another viewpoint in order to determine more directly the important factors. The proposed method revealed a modification (ϵBIC) of the modified Bayesian Information Criterion (mBIC). The comparison of the introduced ϵBIC to existing variable selection methods is performed by some simulating experiments and the Type I and Type II error rates are calculated. Additionally, the proposed method is applied successfully to a high-dimensional Trauma data analysis, and its good predictive properties are confirmed.

Key words: AIC, AUC, best subset, BIC, logistic regression, modified BIC (mBIC), prior, ROC, trauma, variable/feature selection.

1. Introduction

Variable and feature selection have become the focus of much research to high-dimensional statistical modeling in diverse fields of sciences. Many studies to variable selection are related to medicine and biology, such as Fan and Li (2001; 2002; 2006), Sierra *et al.* (2001), Svrakic *et al.* (2003), Fan *et al.* (2005), and Genkin *et al.* (2007). The main problem in any model-building situation is to

*Corresponding author.

choose from a large set of covariates those that should be included in the "best" model. "Best" predictive modeling is the process by which a model is created or chosen to try to best predict the probability of an outcome Seymour (1993). In many cases the best model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input training data.

There are two basic ways to address the problem of variable selection, i.e., variable ranking and variable subset selection. Many variable selection algorithms include a simple, scalable criterion for ranking variables, and are widely used in medical studies to discriminate between healthy and diseased patients, survival or not. Ranking criteria include correlation criteria, which enforce a ranking according to goodness of linear fit of individual variables (correlation criteria can detect linear dependencies between variable and target; a simple way of lifting this restriction is to make a non-linear fit of the target with single variables and rank according to the goodness of fit), and information theoretic criteria which rely on empirical estimates of the mutual information between each variable and the target. On the other hand, several approaches to the variable selection problem illustrate the usefulness of selecting subsets of variables that together have good predictive power, as opposed to ranking variables according to their individual predictive power Guyon and Elisseeff (2003).

The prediction issue is very important in medical problems. When given patient-specific covariate information, the predictive model that is used should be able to predict accurately the response-outcome (survival or not, the presence/absence of a disease). A right decision to keep a variable in the model might be based not solely on the clinical significance but also on statistical significance. This study works on the problem of high-dimensional statistical modeling through the analysis of the trauma annual data in Greece for the year 2005. The dataset selected to be examined, which has been properly divided into training and test sets, deals with a binary response variable (death or not). In addition, the performance of the proposed variable subset selection procedure is examined by simulating multiple tests, i.e., different types of predictor variables with a binary outcome ($y = 1$ or $y = 0$) are considered. The simulation study assumes a binary logistic regression model (LR) which is the appropriate method to present the relationship between the dichotomous response's measurements and its predictor factors which are of any type. The interested reader for LR is referred to Myers *et al.* (2002) and Montgomery *et al.* (2006) for more details.

This paper focuses mainly on selecting subsets of features that are useful to build a good predictor and excluding simultaneously many redundant variables. We propose a variable subset selection method which extends to the classification case the idea of selecting variables. The proposed technique is compared to existing variable selection methods, viz., best subset (AIC), best subset (BIC) and best subset (mBIC). For more details for best subset in regression see Miller (2002).

The rest of this paper is organized as follows. In Section 2, a review of several variable selection criteria is presented. In Section 3, we describe the proposed variable subset selection method. In Section 4, we perform some simulation experiments to evaluate the merits of the proposed method. Additionally, the proposed method is applied to analyze real medical data and is compared to other variable selection techniques. Finally, in Section 5, the obtained results are discussed and some concluding remarks are made.

2. Variable Selection Criteria

The procedure of subset selection is used with certain criteria which combine statistical measures with penalties for increasing the number of predictors in the model. Basically, the criteria used in the best subset variable selection procedure are classified into four categories: (1) Prediction criteria; (2) Information-based criteria; (3) Data-reuse and Data-driven procedures; (4) Bayesian variable selection. In our study, we use only the information-based criteria which are related to likelihood or divergence measures.

The most popular criteria include AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria). AIC was proposed by Akaike (1974) and it selects the model that minimizes AIC $= -2l + 2q$. BIC was proposed by Schwarz (1978) 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 BIC $= -2l + q\log(n)$, where l is the log-likelihood of the model, q is the number of parameters in the model and n is the number of observations.

Consider the situation in which a large database needs to be analyzed and we expect that only a few explanatory variables influence the response. In this case classical criteria, such as the AIC or the BIC information criteria usually overestimate the number of regressors. Bogdan *et al.* (2008) developed the modified version of BIC (mBIC), which enables the incorporation of prior knowledge on a number of regressors preventing this phenomenon of overestimation. In particular, mBIC is adapted using the binomial prior and recommends choosing the model for which

$$-2l + q\log(n) + 2q\log\left(\frac{1-p}{p}\right)$$

attains its minimum value, where p denotes the prior probability that a randomly chosen regressor influences the response-outcome.

Variable subset selection techniques choose subsets of variables that together have good predictive power. However, a variable that is useless when taken with others can be useful by itself. Therefore, we propose a method which extends to the classification case the idea of variable subset selection, taking into account the individual predictive power of each variable. We use as criterion the performance of a classifier built with a single variable performing Receiver Operating Characteristic (ROC) curves analysis (see Pepe, 2000a; 2000b), using in particular the criterion of the Area Under the ROC Curve (AUC) (for more details see Hanley and McNeil, 1982 and Bradley, 1997). Precisely, using the AUC metric, we estimate the probability that for a pair of patients of which one has survived and the other has not, the surviving patient is given a greater probability of survival. This probability was estimated from the test data.

3. The Proposed Best Subset Variable Selection Method

1. Separate data into training and test sets. Given the $n \times m$ model matrix $D = [x_1, x_2, \cdots, x_m]$, where x_j, $j = 1, 2, \cdots, m$, is a column of the matrix, as well as an $n \times 1$ vector Y, which is the common response vector, compute the AUC measure of every variable x_j with respect to the Y.

2. Sort out the factors with respect to their AUC measures, sorting in descending order the AUC vector. Hence, the vector of AUC values is $\text{AUC} = (\text{AUC}_1, \text{AUC}_2, \cdots, \text{AUC}_m)$, where AUC_j, for $j = 1, 2, \cdots, m$, is the corresponding value of the AUC measure for the j-th variable.

3. Retain the factors of the model that have corresponding AUC_j larger than $\theta = 0.5$. These factors together constitute a better than a random guess model. The threshold $\theta = 0.5$ is set on the value of the response variable, i.e., at the mid-point between the center of gravity of the two classes ($y = 0$ or $y = 1$).

4. Compute the prior probability that a randomly chosen regressor influences Y, that is, the $p = r/q$, where r denotes the number of the identified significant factors following the previous steps and q denotes the number of parameters in the model, viz., $x_1, \cdots, x_m$.

5. Finally, ϵBIC recommends choosing the model for which

$$-2l + q\log(n) + 2\epsilon q \log\left(\frac{1-p}{p}\right) \tag{1}$$

attains its minimum value, for some predefined $0 < \epsilon < 1$, where l is the log-likelihood of the model and n is the number of observations.

There are two things to note here:

First of all, separating data into training and testing sets is an important part of evaluating models. Typically, when a data set is partitioned into a training set and testing set, most of the data is used for training (75%), and a smaller portion of the data is used for testing (25%). This ratio is often used in data mining, but this ratio varies according to the requirements of an experimental study. Also, note here that setting of the threshold value θ, which determines the number of significant factors is very crucial. The decision about θ is made according to the fact that θ varies over the possible values of a variable, and to be informative, the entire ROC curve should lie above the 45° line where sensitivity$(\theta) = 1 -$ specificity(θ) DeLong *et al.* (1988). The default cutoff value in two class classifiers is 0.5. Some recent papers (e.g., Lachiche and Flach, 2003 and Rosset, 2004) discuss that 0.5 may not be the optimal threshold and the fact that AUC is oblivious to that threshold is an illustration of the "bias" in using AUC for model selection. However, our interest here is purely in comparing training and test sets and their performance. The threshold is a point chosen based on the ROC curve of the training data set. If this threshold remains optimal on the test data, the AUC remains unbiased. So, even if our threshold may be suboptimal for classification, it has no bearing on the validity of our comparisons as long as the same threshold is used for training and testing evaluation, since we retain the factors of the model that have corresponding AUC larger than 0.5 for both training and test sets. The classification threshold in this view is a part of the model specification, not a separate parameter to be optimized. We thus assume here equal probabilities for the two classes ($y = 0$ or $y = 1$) and equal misclassification costs. Hence, the threshold value θ is set on 0.5.

4. Comparative Study

4.1 Criteria for Performance Evaluation

In the comparative study that follows, we compare the performance of four best subset methods: the best subset (AIC), best subset (BIC), best subset (mBIC) and the proposed best subset variable selection method (ϵBIC).

When testing a single hypothesis, one is usually concerned with testing the null hypothesis $H_0 : \beta_j = 0$ versus an alternative $H_1 : \beta_j \neq 0$, for $j = 1, 2, \cdots, m$. A Type I error occurs when declaring an inactive factor to be active, but H_0 is really true; a Type II error occurs when declaring an active effect to be inactive, but H_1 is really true. Recently, we have seen a recent increase in the size of data sets available. It is now often up to the statistician to find as many interesting features in a data set as possible rather than test a very specific hypothesis on one item. In our simulation study, we deal with a problem of multiple testing since hundreds of parameters are tested simultaneously (1000 iterations). We thus compute the average Type I (reject the true null hypothesis) and Type II (accept the false null hypothesis) error rate over 1000 iterations. Here, we test the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$. This is a test of the null that none of the independent variables have predictive power. This implies the alternative hypothesis $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\cdots \beta_m \neq 0$.

For our real data analysis, we use a multiple testing error measure, viz., the false discovery rate (FDR). The FDR is especially appropriate for analyses based on real data in which one is interested in finding several significant results among many tests. Benjamini and Hochberg (2000) suggest that the False Discovery Rate (FDR) is the appropriate error rate to control in many applied multiple testing problems and that the experimenter should be concerned with controlling the FDR at a desired level, while maintaining the power (1-Type II error) of each test as much as possible. At first, we compute the number of false positives FP, i.e., the number of chosen variables that do not appear in the true model and the number of false negatives FN, i.e., the number of true regressors which were not detected. These numbers are then used to compute the values of the following characteristics:

1. Power (1-Type II error), defined as $(k-\mathrm{FN})/k$, where k denotes the number of explanatory variables with non-zero regression coefficients. The power is not defined when $k = 0$, since the cases for which $k = 0$ are excluded from this analysis.

2. False Discovery Rate (FDR), defined as $\mathrm{FDR} = \mathrm{FP}/(\mathrm{FP}+k-\mathrm{FN})$. If there are no discoveries, then $\mathrm{FP} + k - \mathrm{FN} = 0$ yielding $\mathrm{FDR} = 0$.

3. Number of misclassified regressors, defined as $\mathrm{MR} = \mathrm{FP} + \mathrm{FN}$.

4. Prediction error d which is defined as $d = \sum_{i=1}^{n} |Y_i - \hat{Y}_i|/n$, where n is the number of observations. Prediction error computes the mean absolute deviation between these observations and their predicted values. Mean prediction error d is calculated for both training and test sets.

4.2 Simulation Study

For our simulation experiments, we develop logistic regression models with $n = 100$ experimental runs, and $m = 100$ explanatory variables, with coefficients taking random values from the vector β. Searching the space of possible models, note that if a variable does not improve the appropriate penalized likelihood is deleted as insignificant by estimating its β coefficient as 0. All simulations were conducted using MATLAB software. 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. During our simulation experiments, only main effects models were taken into consideration.

For each of 1000 iterations in our simulation experiments we used the following simulation protocol:

1. Suppose that the model has the form
$$y_i = P(x_i) + \varepsilon,$$
where the quantity $P(x_i)$ is defined as
$$P(x_i) = 1/(1 + e^{-x_i'\beta}),$$
where the term $x_i'\beta$ is equal to $\beta_1 x_{i1} + \cdots + \beta_m x_{im}$, $i = 1, 2, \cdots, n$, and the response variable y_i takes on the value either 0 or 1.

2. The true active variables were selected randomly from the set of $\{1, \cdots, m\}$ potentially active factors. From columns $1, \cdots, m$ of X, k columns were assigned to active factors according to the following procedure: the true active variables were selected randomly from the set of $\{1, \cdots, m\}$ potentially active factors and only main effects were taken into consideration. For each experiment, we estimated all the main effects and labeled each effect as either active or inactive. Then, by calculating the percentage of the estimated active effects within each experiment, we estimated the number of active effects that can be identified. Note here that for Case I $(0 < p < 0.5)$ we estimated that the number of active effects that can be identified does not exceed $m/2$, where m is the number of columns of the design matrix.

3. To obtain the coefficients for the active factors, a sample of size k was drawn from a $N(4, 0.2)$, and ± 1 signs randomly allocated to each number. For the non-active variables, in the true model, their coefficients were obtained as a random draw from a $N(0, 0.2)$.

4. 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 $P(x_i)[1 - P(x_i)]$.

5. The design matrix is $X = [x_1, x_2, \cdots, x_{100}]$ where $x_1, \cdots, x_{100}$ are a mix of dichotomous, ordinal and continuous variables, taking randomly different values as described in Table 1.

The results of the proposed method (ϵBIC) for several ϵ, are compared to best subset (AIC), best subset (BIC) and best subset (mBIC). It is worthy to note that ϵBIC is equivalent to BIC if we take $\epsilon = 0$ and equivalent to mBIC if we take $\epsilon = 1$. The final decision for ϵ is made according to the value which is

Table 1: The chosen random structure for x_j

x_j	Values
x_j	±1 equally distributed
x_j	±1 randomly distributed
x_j	generated randomly using rating scales based on four (0-3) ordinal levels
x_j	generated from a Normal distribution $x_i \sim N(10, 5)$
x_j	generated from a Normal distribution $x_i \sim N(0, 0.5)$
x_j	generated from a uniform distribution $x_i \sim U(a = 5, b = 15)$

observed to be good choice from simulation trials. The optimal ϵ value is chosen so that ϵ succeeds the smallest Type I and Type II error values simultaneously for the following key reason. Type I and Type II error rates are important and should be kept as low as possible. The low Type I error rates are important since the ability to exclude unnecessary factors reduces the cost of additional experiments and low Type II error rates are especially desirable since the main goal is to find out the important factors that influence most the response.

In our simulation study, several different ϵ values (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95) were examined in order to provide a more generalized suggestion about how to choose the optimal ϵ value, and to cover all possible cases since variable selection may have different purposes in different cases. We have thus considered the following two cases for p, viz., the prior probability that a randomly chosen regressor influences Y:

1. Case I: If $0 < p < 0.5$, the penalty term $2\epsilon q \log((1-p)/p)$ is positive. In this case, the greater ϵ is, the more penalized term will be.

2. Case II: If $0.5 < p < 1$, the penalty term $2\epsilon q \log((1-p)/p)$ is negative. In this case, the greater ϵ is, the less penalized term will be.

Note here that the case of $p = 0.5$ corresponds to the usual BIC.

We performed the simulations 1000 times for each Case I and Case II, and the obtained results are summarized in Table 2 and Table 3 respectively. In these tables, the first column lists the used criterion. The two columns named "Type I" and "Type II" present the Type I error rate and Type II error rate, respectively, averaged over 1000 iterations for each best subset method.

We observe from Table 2 (Case I: $0 < p < 0.5$) that AIC and BIC have a strong tendency to overestimate the number of regressors, since the Type I error rate equals 0.3275 for AIC and 0.1650 for BIC, respectively. The highest average probability of the Type I error is for AIC, since AIC typically includes more regressors (for a sample size of $n \geq 8$ the penalty on model dimension using AIC is smaller than the penalty using BIC). The standard version of mBIC helps to control the overall Type I error reducing its value to 0.14. As demonstrated by

Table 2: Empirical performance of best subset variable selection methods for Case I

Criterion	Type I	Type II
AIC	0.3275	0.0667
BIC	0.1650	0.1067
mBIC	0.1400	0.1150
ϵBIC ($\epsilon = 0.1$)	0.1550	0.1083
ϵBIC ($\epsilon = 0.15$)	0.1575	0.1000
ϵBIC ($\epsilon = 0.2$)	0.1500	0.1100
ϵBIC ($\epsilon = 0.25$)	0.1550	0.1000
ϵBIC ($\epsilon = 0.3$)	0.1500	0.1117
ϵBIC ($\epsilon = 0.35$)	0.1550	0.1017
ϵBIC ($\epsilon = 0.4$)	0.1450	0.1133
ϵBIC ($\epsilon = 0.45$)	0.1500	0.1050
ϵBIC ($\epsilon = 0.5$)	0.1450	0.1133
ϵBIC ($\epsilon = 0.55$)	0.1500	0.1067
ϵBIC ($\epsilon = 0.6$)	0.1425	0.1133
ϵBIC ($\epsilon = 0.65$)	0.1475	0.1067
ϵBIC ($\epsilon = 0.7$)	0.1425	0.1133
ϵBIC ($\epsilon = 0.75$)	0.1450	0.1134
ϵBIC ($\epsilon = 0.8$)	0.1400	0.1133
ϵBIC ($\epsilon = 0.85$)	0.1375	0.1134
ϵBIC ($\epsilon = 0.9$)	0.1400	0.1133
ϵBIC ($\epsilon = 0.95$)	0.1350	0.1150

Table 2, the ϵBIC prevents overestimation and performs better than AIC and BIC, since the proposed method has smaller values for Type I errors for all examined ϵ. The proposed method also succeeds lower Type II error values for almost all considered ϵ compared to mBIC, The proposed method also succeeds lower Type II error values for almost all considered ϵ compared to mBIC, and these correspond to cases where the majority of the active effects are detected correctly. We also observe that ϵBIC succeeded lower Type I and Type II error values simultaneously for $\epsilon = 0.85$ compared to mBIC.

We observe from Table 3 (Case II: $0.5 < p < 1$) that the highest average probability of the Type I error is for AIC, as expected. In this case, mBIC does not prevent overestimation and BIC performs better than mBIC reducing the Type I error value to 0.1625. As demonstrated by Table 3, the ϵBIC succeeds to control the overall Type I error compared to AIC and mBIC, since the proposed method has smaller values for Type I errors for all examined ϵ. The proposed method also succeeds lower Type II error values for all considered ϵ compared to BIC. We also observe that ϵBIC performs better than BIC for $\epsilon = 0.1$.

Table 3: Empirical performance of best subset variable selection methods for Case II

Criterion	Type I	Type II
AIC	0.3275	0.0533
BIC	0.1625	0.0933
mBIC	0.2000	0.0817
ϵBIC ($\epsilon = 0.1$)	0.1625	0.0917
ϵBIC ($\epsilon = 0.15$)	0.1775	0.0917
ϵBIC ($\epsilon = 0.2$)	0.1800	0.0900
ϵBIC ($\epsilon = 0.25$)	0.1775	0.0833
ϵBIC ($\epsilon = 0.3$)	0.1850	0.0867
ϵBIC ($\epsilon = 0.35$)	0.1850	0.0833
ϵBIC ($\epsilon = 0.4$)	0.1850	0.0867
ϵBIC ($\epsilon = 0.45$)	0.1875	0.0917
ϵBIC ($\epsilon = 0.5$)	0.1850	0.0867
ϵBIC ($\epsilon = 0.55$)	0.1900	0.0917
ϵBIC ($\epsilon = 0.6$)	0.1850	0.0867
ϵBIC ($\epsilon = 0.65$)	0.1925	0.0900
ϵBIC ($\epsilon = 0.7$)	0.1850	0.0867
ϵBIC ($\epsilon = 0.75$)	0.1975	0.0900
ϵBIC ($\epsilon = 0.8$)	0.1925	0.0850
ϵBIC ($\epsilon = 0.85$)	0.1975	0.0850
ϵBIC ($\epsilon = 0.9$)	0.1950	0.0817
ϵBIC ($\epsilon = 0.95$)	0.1975	0.0817

We shall now give a more generalized suggestion about which one among AIC, BIC, mBIC or ϵBIC is better according to the case under consideration, since model selection may have different purposes in different cases.

According to the results in Table 2, if $0 < p < 0.5$, we suggest the use of AIC, BIC, or ϵBIC for $0 < \epsilon < 0.5$ as a model selection criterion, if overfitting is acceptable in order to ensure that all possible influent regressors can be included. The experimenter can choose which one among AIC, BIC, or ϵBIC for $0 < \epsilon < 0.5$ is better, according to what extent of overfitting is acceptable under consideration. On the other hand, if choosing influent regressors is crucial, we suggest the use of ϵBIC for $0.5 < \epsilon < 1$ as a model selection criterion.

According to the results in Table 3, if $0.5 < p < 1$, we suggest the use of AIC, mBIC, or ϵBIC for $0.5 < \epsilon < 1$ as a model selection criterion, if overfitting is acceptable so that all possible regressors will be included. The experimenter can choose which one among AIC, mBIC, or ϵBIC for $0.5 < \epsilon < 1$ is better, according to what extent of overfitting is acceptable under consideration. On the other hand, if choosing one more additional regressor included is serious, we suggest the use of BIC or ϵBIC for $0 < \epsilon < 0.5$ as a model selection criterion.

Table 4 displays the average execution times (sec) for the best subset methods with AIC, BIC and mBIC in both considered cases. Table 5 displays the average execution times (sec) for the proposed best subset method with ϵBIC (for several ϵ in both considered cases).

Table 4: Execution times (sec)

Method	Best Subset (AIC)	Best Subset (BIC)	Best Subset (mBIC)
	$1.7784 * 10^4$	$1.8096 * 10^4$	$1.8174 * 10^4$

Table 5: Execution times (sec) for ϵBIC

ϵBIC	$\epsilon = 0.1, 0.15$	$\epsilon = 0.2, 0.25$	$\epsilon = 0.3, 0.35$	$\epsilon = 0.4, 0.45$	$\epsilon = 0.5$
	$3.6114 * 10^4$	$5.4366 * 10^4$	$7.2462 * 10^4$	$9.0871 * 10^4$	$10.8889 * 10^4$
ϵBIC	$\epsilon = 0.55$	$\epsilon = 0.6, 0.65$	$\epsilon = 0.7, 0.75$	$\epsilon = 0.8, 0.85$	$\epsilon = 0.9, 0.95$
	$10.8889 * 10^4$	$12.7063 * 10^4$	$14.5315 * 10^4$	$16.3489 * 10^4$	$18.1585 * 10^4$

In a nutshell, this simulation study demonstrates 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. Since the aim is mainly to find out the important factors that influence most the response and should be considered for further investigation, the low Type II error rates are especially desirable, even though both Type I and Type II error rates are important and should be kept as low as possible. The proposed method achieves this task successfully. The fact that ϵBIC is slower compared to the other best subset methods which perform almost similarly (see Table 4 and Table 5) does not influence these desirable predictive properties of the proposed method. Note here that the mBIC criterion incorporates the prior distribution for the number of effects assuming the binomial prior, while when prior knowledge on the number of regressors is not available, Bogdan *et al.* (2004) proposed choosing a constant c in such a way that the family-wise error rate (FWER, the probability of detecting at least one false positive) for the sample size $n \geq 200$ is controlled at the level below 10%. This modified version of mBIC (ϵBIC) can be used in any experiment, allowing the incorporation of prior knowledge by calculating p, viz., the probability that a randomly chosen regressor influences Y, through the proposed AUC-based variable selection method.

4.3 Analysis Based on Trauma Data

The proposed method presented in the previous section is now applied to a real medical dataset and is compared to other traditional variable selection techniques. The data derive from an annual registry conducted during the period

01/01/2005−31/12/2005 by the Hellenic Trauma and Emergency Surgery Society involving 30 General Hospitals in Greece. The study is designed to assess the effects of differing prognostic factors on the outcome of injured persons. For each patient the main outcome of interest, which is vital status at the last follow-up, dead ($y = 1$) versus alive ($y = 0$), is reported. The dataset consists of 8862 observations and 92 factors that include demographic, transport and intrahospital data used to detect possible risk factors of death. According to medical advices, all the prognostic factors should be treated equally during the statistical analysis and there is no factor that should be always maintained in the model. The Trauma data set which is used to compare the results of variable selection methods is presented in Table 6.

Table 6: Trauma study

Continuous covariates
$x1$: weight, kg
$x2$: age, years
$x3$: Glasgow Coma Score, score
$x4$: pulse, N/min
$x6$: systolic arterial blood pressure, mmHg
$x7$: diastolic arterial blood pressure, mmHg
$x8$: Hematocrit (Ht), %
$x9$: haemoglobin (Hb), g/dl
$x11$: white cell count, /ml
$x12$: platelet, /ml
$x14$: potassium, /ml
$x15$: glucose, mg %
$x16$: creatinine, mg %
$x17$: urea, mg %
$x18$: amylase, score
$x20$: Injury Severity Score, score
$x21$: Revised Trauma Score, score

In the experiment, the data set is split into training (75%) and test (25%) sets for data analysis in order to evaluate the prediction performance on new data. Firstly, we consider one predictor at a time to see how well each predictor alone predicts the target variable ($y = 0$ or $y = 1$). This first step is added to the analytical process because it allows the variable set to be reduced in size, creating a more manageable set of attributes for modeling. The variables are ranked according to a specified criterion depending on the measurement levels of the predictors. A common technique used in data mining is ranking the attributes based on the measure of importance which is defined as $(1 - p)$, where p is the

Table 6: (continued) Trauma Study

Categorical covariates
$x19$: evaluation of disability (0 = expected permanent big, 1 = expected permanent small, 2 = expected impermanent big, 3 = expected impermanent small, 4 = recovery)
$x23$: cause of injury (0 = fall, 1 = trochee accident, 2 = athletic, 3 = industrial, 4 = crime, 5 = other)
$x24$: means of transportation (0 = airplane, 1 = ambulance, 2 = car, 4 = on foot)
$x25$: Ambulance (0 = no, 1 = yes)
$x26$: hospital of records
$x27$: substructure of hospital (0 = orthopaedic, 1 = CT, 2 = vascular surgeon, 3 = neurosurgeon, 4 = Intensive Care Unit)
$x28$: comorbidities (0 = no, 1 = yes)
$x31$: sex (0 = female, 1 = male)
$x35$: doctor's speciality (0 = angiochirurgeon, 1 = non specialist, 2 = general doctor 3 = general surgeon, 4 = jawbonesurgeon, 5 = gynaecologist, 6 = thoraxsurgeon, 7 = neurosurgeon, 8 = orthopaedic, 9 = urologist, 10 = paediatrician, 11 = children surgeon, 12 = plastic surgeon)
$x36$: major doctor (0 = no, 1 = yes)
$x41$: dysphoria (0 = no, 1 = yes)
$x52$: collar (0 = no, 1 = yes)
$x55$: immobility of limbs (0 = no, 1 = yes)
$x56$: fluids (0 = no, 1 = yes)
$x64$: Radiograph E.R. (0 = no, 1 = yes)
$x66$: US (0 = no, 1 = yes)
$x67$: urea test (0 = no, 1 = yes)
$x71$: destination after the emergency room (0 = other hospital, 1 = clinic, 2 = unit of high care, 3 = intensive care unit I.C.U, 4 = operating room)
$x72$: surgical intervention (0 = no, 1 = yes)
$x86$: arrival at emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 = 16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)
$x87$: exit from emergency room (0 = 00:00-04:00, 1 = 04:01-08:00, 2 = 08:01-12:00, 3 = 12:01-16:00, 4 = 16:01-18:00, 5 = 18:01-20:00, 6 = 20:01-24:00)
$x101$: head injury (0 = none, 1 = AIS $\leqslant 2$, 2 = AIS > 2)
$x102$: face injury (0 = none, 1 = AIS $\leqslant 2$, 2 = AIS > 2)
$x104$: breast injury (0 = none, 1 = AIS $\leqslant 2$, 2 = AIS > 2)
$x106$: spinal column injury (0 = none, 1 = AIS $\leqslant 2$, 2 = AIS > 2)
$x107$: upper limbs injury (0 = none, 1 = AIS $\leqslant 2$, 2 = AIS > 2)
$x108$: lower limbs injury (0 = none, 1 = AIS $\leqslant 2$, 2 = AIS > 2)

p-value of a chosen statistical test of association between the candidate predictor and the target variable.

In our study some predictors are continuous and some are categorical. The criterion used for continuous predictors is the p-value based on an one-way ANOVA F test, while the criterion for categorical predictors is restricted to the p-value based on Pearson chi-square test Pearson (1983). These p-values are compared and therefore are used to rank the predictors. In the initial data set, we had 92 explicative variables and after following the procedure of feature selection, we execute and detect the most statistically significant of them, for significance level $\alpha = 5\%$. Table 7 presents the importance values for the 44 significant variables of the feature selection generated model. In fact, feature selection method allowed us to minimize the set of regressor variables from 92 to 44 otherwise best subset methods would not be applicable due to computational complexity ($n = 8862$ records).

Table 7: The most important fields of the data set

Fields	$x71$	$x3$	$x21$	$x101$	$x20$	$x64$	$x56$	$x19$	
Importance value $(1-p)$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
Fields	$x8$	$x6$	$x26$	$x7$	$x35$	$x9$	$x36$	$x11$	$x66$
Importance value $(1-p)$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Fields	$x108$	$x72$	$x55$	$x27$	$x67$	$x41$	$x107$	$x52$	
Importance value $(1-p)$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
Fields	$x2$	$x16$	$x18$	$x104$	$x28$	$x12$	$x17$		
Importance value $(1-p)$	1.0	1.0	1.0	0.999	0.999	0.998	0.997		
Fields	$x31$	$x1$	$x4$	$x24$	$x15$	$x102$	$x106$	$x25$	
Importance value $(1-p)$	0.996	0.994	0.988	0.988	0.988	0.985	0.978	0.965	
Fields	$x86$	$x87$	$x14$	$x5$					
Importance value $(1-p)$	0.964	0.962	0.961	0.952					

Regarding the proposed method, we sort out the factors with respect to their AUC measures. We retain the factors of the model that have corresponding AUC$_j$ larger than $\theta = 0.5$. These factors are then declared to be significant ($r = 6$). Let p denote the probability that a randomly chosen regressor influences Y. In our study, $p = r/q = 6/44$. We set ϵ to be 0.85 for ϵBIC, since this value was found to be the optimal value from the simulation study for Case I ($0 < p < 0.5$). The results of the proposed method (ϵBIC) are compared to best subset (AIC), best subset (BIC) and best subset (mBIC). The estimated β coefficients and standard errors of the selected model for each examined best subset method are listed in Table 8. The insignificant variables are deleted by estimating their coefficients as 0. The execution times of all best subset methods are reported in Table 9.

Table 8: Estimated β coefficients and standard errors (in parentheses) for best subset methods

Method	MLE	Best Subset (AIC)	Best Subset (BIC)	Best Subset (mBIC)	Best Subset (ϵBIC)
Intercept	-5.97 (0.21)	-5.97 (0.21)	-5.96 (0.21)	-5.96 (0.21)	-5.96 (0.21)
x_2	0.72 (0.09)	0.71 (0.09)	0.73 (0.09)	0.73 (0.09)	0.73 (0.09)
x_{11}	0.25 (0.07)	0.25 (0.07)	0.28 (0.07)	0.28 (0.07)	0.28 (0.07)
x_{16}	0.09 (0.05)	0.09 (0.05)	0 (-)	0 (-)	0 (-)
x_{20}	0.56 (0.05)	0.56 (0.05)	0.55 (0.05)	0.55 (0.05)	0.55 (0.05)
x_{23}	0.06 (0.10)	0 (-)	0 (-)	0 (-)	0 (-)
x_{25}	1.02 (0.15)	1.00 (0.14)	1.02 (0.14)	1.02 (0.14)	1.02 (0.14)
x_{27}	-0.16 (0.09)	-0.16 (0.09)	0 (-)	0 (-)	0 (-)
x_{71}	1.46 (0.07)	1.47 (0.07)	1.43 (0.07)	1.43 (0.07)	1.43 (0.07)
x_{101}	1.30 (0.08)	1.30 (0.08)	1.29 (0.08)	1.29 (0.08)	1.29 (0.08)

Table 9: Execution times (sec)

Method	Best Subset (AIC)	Best Subset (BIC)	Best Subset (mBIC)	Best Subset (ϵBIC)
	$8.535 * 10^4$	$8.475 * 10^4$	$8.548 * 10^4$	$1.709 * 10^5$

We observe from Table 8 that BIC, mBIC and ϵBIC perform similarly and have identical values for both β and standard errors. This fact does not come as a surprise taking into account the properties of BIC, since in the case where m is fixed ($m = 44$ regressors) and n increases ($n = 8862$ records) or n goes to infinity, mBIC and ϵBIC criteria asymptotically approach BIC. AIC fails to exclude two redundant variables (x_{16}, x_{27}) compared to BIC, mBIC and ϵBIC, as expected, since AIC in many practical applications typically includes more regressors, as discussed in Bogdan *et al.* (2008). The proposed AUC-based best subset variable selection method succeeds to identify the same subset of significant variables affecting death from Trauma, compared to the well-known BIC and mBIC criteria, which includes x_2, x_{11}, x_{25}, x_{20}, x_{71} and x_{101}.

In our trauma study, the training set is used to select the significant regressors and to estimate their regression coefficients. After the model has been determined by using the training set, we test the model by making predictions using the test set. Because the data in the test set already contains known values for the response variable we want to predict, it is easy to examine whether the model's guesses are correct. The test set is used to compute the number of FP and FN, and these numbers are then used to compute the performance criteria values presented in Table 10. The results reported in Table 10 are averaged over 1000 different random partitions.

Table 10: Results over 1000 different random partitions

	Best Subset (AIC)	Best Subset (BIC)	Best Subset (mBIC)	Best Subset (ϵBIC)
FP	15.4	12.7	12.5	12.4
FN	1.92	3.45	2.95	2.48
Power	0.808	0.655	0.705	0.752
FDR	0.655	0.659	0.639	0.622
MR	17.32	16.15	15.45	14.88
d test	0.325	0.213	0.204	0.198
d training	0.432	0.397	0.381	0.376

As demonstrated in Table 10, most of the signals detected are false positive for all information criteria. The number of false positives detected by AIC is 15.4, and is considerably higher than the number of false positives detected by BIC, mBIC and ϵBIC which perform similarly. The tendency of AIC to include many false positives is also reflected in the higher value of MR. Note here that if the cost of a false positive is the same as the cost of a false negative, then MR is proportional to the total cost of the experiment. This strong tendency has a relatively small influence on the prediction error using AIC. The small values of the prediction error d for both training and test sets for AIC, demonstrate that for large sample sizes the overestimation of the number of regressors does not substantially deteriorate the predictive properties of AIC. Furthermore, the FDR of the proposed method is 0.622, smaller than the corresponding value for AIC, BIC and mBIC. The proposed method seems to perform better than best subset methods with AIC, BIC and mBIC criteria, since it has also smaller misclassification rate and prediction error for both training and test sets. Additionally, the proposed method has the highest power value compared to BIC and mBIC.

Conclusively, the trauma study illustrates the merits of the proposed best subset variable selection method. Despite having such high dimensional data set, the proposed method achieved the main goal, i.e., to identify a rather small subset of features which is sufficient for modeling. Further, the revealed prognostic model for the outcome for trauma patients includes only the factors of highest importance for prediction. As a result, the outcome prediction model is now plausible and provides specific information which may assist as guidelines for trauma management and may help trauma personnel to focus mostly on features that are observed to be the most relevant for prediction.

5. Concluding Remarks

Best subset variable selection remains as a challenging and promising area of research. In this paper, we have proposed a method for selecting subsets of features that are useful to build a good predictor. Since the use of a variable selection method is mainly to screen the factors that should not be considered for further investigation, the proposed method achieves this task successfully excluding the redundant variables and maintaining only the factors of highest importance. A drawback of the proposed best subset variable selection method is the computational time. Generally, best subset methods are very time consuming because all subsets approach is an exhaustive search, since it searches through all possible subsets to find the optimal one. This computational difficulty prevents the best subset methods from being widely used when there are a large number of predictors in practical problems. However, the proposed method performed satisfactorily on the problem of high-dimensional statistical modeling, identifying the significant prognostic factors affecting death from trauma.

The simulation study demonstrated that the proposed method succeeded very low values of Type II error rate which is very crucial in order not to omit important factors of the model. Additionally, the Type I error rate was also maintained at low level for several ϵ values. 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. Furthermore, the real data analysis demonstrated that the proposed method has very good properties with respect to controlling the false discovery rate, minimizing the number of misclassified regressors and the prediction error.

The innovation of the proposed method places in the the combination of a nonparametric technique (AUC measure) with a likelihood based method (ϵBIC information-based criterion) which is uncommon type of variable selection to the best of our knowledge. In the proposed method, only main effects models were considered. We are currently looking into problems involving models with interactions. We aim to develop techniques that take into account the correlated nature of the data and deal with intra-subject (between two measurements on the same subject) correlation.

Acknowledgements

The authors would like to thank the First Propedeutic Surgical Clinic, in Hippocratio Hospital for giving the real medical data. The authors would also like to thank the Associate Editor and referees for constructive and useful suggestions that improved the quality of the manuscript.

References

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

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60-83.

Bogdan, M., Ghosh, J. K. and Doerge, R. W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**, 989-999.

Bogdan, M., Ghosh, J. K. and Żak-Szatkowska, M. (2008). Selecting explanatory variables with the modified version of Bayesian information criterion. *Quality and Reliability Engineering International* **24**, 627-641.

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

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845.

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.

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

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), Vol. **III**, 595-622. European Mathematical Society, Zürich.

Fan, J., Chen, Y., Chan, H. M., Tam, P. K. H. and Ren, Y. (2005). Removing intensity effects and identifying significant genes for Affymetrix arrays in macrophage migration inhibitory factor-suppressed neuroblastoma cells. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17751-17756.

Genkin, A., Lewis, D. D. and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**, 291-304.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3**, 1157-1182.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36.

Lachiche, N. and Flach, P. (2003). Improving accuracy and cost of the two-class and multi-class probabilistic classifiers using ROC curves. In *Proceedings of the 20th International Conference on Machine Learning* (Edited by T. Fawcett and N. Mishra), 416-423. AAAI Press.

Miller, A. J. (2002). *Subset Selection in Regression*, 2nd edition. Chapman and Hall, New York.

Myers, R. H., Montgomery, D. C. and Vining, G. G. (2002). *Generalized Linear Models*: *With Applications Engineering and the Sciences*, 1st edition. Wiley, New York.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2006). *Introduction to Linear Regression Analysis*, 4th edtion. Wiley, Hoboken, New Jersey.

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

Pepe, M. S. (2000a). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95**, 308-311.

Pepe, M. S. (2000b). An interpretation for ROC curve and inference using GLM procedures. *Biometrics* **56**, 352-359.

Rosset, S. (2004). Model selection via the AUC. In *Proceedings of the 21st International Conference on Machine Learning* (Edited by R. Greiner and D. Schuurmans), 89. ACM Press, Banff, Alberta, Canada.

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

Seymour, G. (1993). *Predictive Inference*: *An Introduction*, 1st edition. Chapman and Hall, New York.

Sierra, B., Lazkano, E., Inza, I., Merino, M., Larrañaga, P. Y. and Quiroga, J. (2001). Prototype selection and feature subset selection by estimation of distribution algorithms: a case study in the survival of cirrhotic patients treated with TIPS. In *Lecture Notes in Artificial Intelligence* (Edited by A. L. Rector *et al.*), Vol. **2101**, 20-29. Springer, Berlin .

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.

Received June 27, 2011; accepted May 10, 2012.

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

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