Sequential selection of variables using short permutation procedures and multiple adjustments: An application to genomic data

Article *in* Statistical Methods in Medical Research · January 2015

DOI: 10.1177/0962280214566262 · Source: PubMed

René Natowicz

University of Paris-Est

Antônio Pádua Braga

Federal University of Minas Gerais

Statistical Methods in Medical Research
000(00):1–36
©The Author(s) 2010
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI:doi number
http://mms.sagepub.com

Sequential Selection of Variables using Short Permutation Procedures and Multiple Adjustments: An Application to Genomic Data

Marcelo Azevedo Costa*
Department of Industrial Engineering, Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil

Thiago de Souza Rodrigues
Computer Department, Centro Federal de Educação Tecnológica Minas Gerais

André Gabriel F. C. da Costa
Department of Statistics, Universidade Federal de Minas Gerais

René Natowicz
University of Paris - ESIEE/Paris, Computer Sciences Department.

Antônio Pádua Braga
Graduate Program in Electrical Engineering, Universidade Federal de Minas Gerais

Abstract

This work proposes a sequential methodology for selecting variables in classification problems in which the number of predictors is much larger than the sample size. The methodology includes a Monte Carlo permutation procedure that conditionally tests the null hypothesis of no association among the outcomes and the available predictors. In order to improve computing aspects, we propose a new parametric distribution, the Truncated and Zero Inflated Gumbel Distribution. The final application is to find compact classification models with improved performance for genomic data. Results using real data sets show that the proposed methodology selects compact models with optimized classification performances.

Keywords
classification models, variable selection, extreme distributions

* Corresponding author; e-mail: macosta@ufmg.br

1. Introduction

Consider the problem of evaluating whether a particular random variable X can predict the expected value of a dependent Bernoulli random variable $Y, Y \in \{0, 1\}$. Specifically, let $Y \sim Bernoulli(\pi)$, represent a Bernoulli random variable with the mean parameter π, $0 \leq \pi \leq 1$. In this particular case, the logistic function can link a linear predictor, $\eta = \beta_0 + \beta_1 \cdot x$ to the mean parameter π, thus:

$$P(Y = 1|X = x) = \exp(\beta_0 + \beta_1 \cdot x)/\left[1 + \exp(\beta_0 + \beta_1 \cdot x)\right] \tag{1}$$

Therefore, given the statistical model defined in Equation 1, testing the hypothesis of association between the random variables Y and X means testing whether the parameter β_1 is zero ($H_0 : \beta_1 = 0$, means no association) or not ($H_1 : \beta_1 \neq 0$, means association). For the Generalized Linear Models (MacCullagh and Nelder, 1989), the estimate of β_1 given a large sample size is approximately normally distributed. Therefore, in this case, a single t-test can be applied.

Now, consider a large number of random variables, X_1, X_2, ..., X_p and the sample size n. Assume that $p \gg n$, and that n is not large enough to assume asymptotic normality for the predictors. This is quite common in genomic data in which a large number of covariates are available and a small sample size is given. In this case, it is of great interest to identify a small group of variables which are statistically correlated to the mean value of Y. Nevertheless, finding the optimal group of variables is not a trivial task. Briefly, there are 2^p possible combinations of variables and, due to the small sample size, problems such as multicolinearity will compromise the individual t-test for each coefficient. For instance, one can think that, in this case, the null hypothesis should be written as

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \tag{2}$$

The alternative hypothesis then represents the $2^p - 1$ possible combinations of variables and, therefore, rejecting the null hypothesis does not solve the problem of finding the proper alternative hypothesis.

One alternative to select covariates is to minimize the error function or the likelihood function subject to a constraint function that provides selection of coefficients. The lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996) applies the sum of the absolute values of the coefficients as the constraining function, and therefore it forces some of the coefficients towards zero. Furthermore, it allows a proper fit of the data. Faster lasso algorithms have been successfully applied in variable selection problems (Friedman et al., 2012) including gene selection (Zare et al., 2013). Nevertheless, the degree of shrinkage of the coefficients is related to the constraining function. Zou and Hastie (2005) propose mixing ridge (Hoerl and Kennard, 1970) and lasso penalties to enable lasso to handle multicollinearity.

In practice, it is quite common to explore the data by first testing univariate models, one for each variable. Those significant variables can be later aggregated into the same model. However, variables which rejected the null hypothesis for the univariate analysis might be strongly correlated, and those variables which failed to reject the null using the univariate analysis may be powerful predictors, as long as they are combined with other variables. In conclusion, evaluating over the 2^p possible combinations of variables is still very attractive, but not always realistic.

Since the main interest generally relies on the evaluation of different combinations of variables, and not on the statistical significance of each coefficient in the model, then hypothesis testing can be driven towards the model performance. In this case, the test statistic can represent a measure of the model fitness to the data. For example, for the normal regression problem the sum of squared errors, or simply the error statistic, can be applied (Montgomery et al., 2012). For the Generalized Linear Models, the log-likelihood ratio statistic or the Deviance statistic is widely used as a measure of fitness (MacCullagh and Nelder, 1989; Dobson, 2008).

Given a test statistic that measures the model fit to the data, then hypothesis testing can be applied to each model to test whether the variables included in the models are predictors of the mean of Y. In other words, now we aim to test the following null hypothesis,

$$H_0 : \mathbf{b}_k = \mathbf{0} \tag{3}$$

where $\mathbf{b}_k$ is the vector of coefficients of model k, $k = 1, 2, ..., 2^p$. In this case, the test statistic is defined as a measure of the model fitness. One can think about the error statistic ($error_k$), but a new problem arises which is the unknown distribution of the error statistics under the null. Furthermore, it does not solve the problem of multiple testing.

Finally, in the face of such a huge amount of candidate models, each one of them having different observed error statistic, then looking at the model which has the best fitness further simplifies the problem. In this case, the new test statistic is written as $T = \min_k\{error_k\}$. Its distribution under the null hypothesis although not known in advance can be generated by means of permutation procedures.

In this paper we address the problem of model selection in which the number of predictors are much higher than the sample size. We propose a sequential analysis in which first univariate models are adjusted. Then, more variables are evaluated and sequentially aggregated into the model. The procedure requires intensive simulations in order to run conditional hypothesis testing. In order to save computing time without loss in performance, we propose a new parametric distribution: the truncated and zero inflated Gumbel distribution. Previous empirical analysis (Abrams et al., 2010) has shown that the Gumbel distribution properly fits maximum values of log likelihood ratio statistics.

The paper is organized as follows: section 2 presents the proposed sequential procedure, and the truncated and zero inflated Gumbel distribution. Section 3 introduces the two evaluated data sets. Section 4 presents the simulation study. Section 5 presents the results. Discussion and conclusion are presented in section 6.

2. Methodology

The proposed methodology for variable selection in classification problems is a variation of the spatial-based cluster method presented by Kulldorff (1997). In spatial clustering the random variable of interest can be represented by the number of cases in small areas of a larger studied region. This random variable can be modeled as independent binomial random variables: $c_i \sim Binomial(n_i, p)$, where n_i is the population in each area i, $i = 1, \ldots, S$ and p is the probability of an individual in area i being a case. Under the null hypothesis of spatial randomness, the maximum likelihood estimate of p is $\hat{p} = \frac{C}{N}$, where C is the total number of cases and N is the total population in the region. The alternative hypothesis assumes that there is one spatial cluster at an unknown location. In order to find one candidate, the method considers all candidates of a set Z as potential clusters. Due to the number of areas, the set Z can reach up to hundreds or thousands of cluster candidates. To avoid the problem of multiple comparison, the method calculates likelihood ratio test statistics for each cluster z ($z \in Z$) and stores the candidate with the highest likelihood ratio. Further details can be found in Kulldorff (1997); Costa et al. (2012).

Kulldorff (1997) uses a process of Monte Carlo simulation (Dwass, 1957) to generate an empirical distribution of a likelihood ratio test statistic under the null. The Monte Carlo procedure consists of randomly distributing the observed cases among the areas based on their population at risk and then finding the new cluster with the maximum likelihood ratio, i.e., the test statistic. This procedure is repeated many times to generate the empirical distribution. Finally, the observed test statistic of the original candidate cluster is compared to the null distribution. If the test statistic is higher than the $100(1-\alpha)\%$ percentile from the empirical distribution then the null hypothesis is rejected and the candidate cluster is assumed to be the most critical region.

The empirical distribution can also be used to assess the significance of other geographical regions known as secondary clusters. Secondary clusters also have a test statistic greater than the estimated threshold, i.e., they reject the null hypothesis. It is worth noting that for each permutation procedure, the cluster with the highest test statistic is stored. Therefore, the null distribution represents the test statistic behavior of the most likely cluster under the null hypothesis.

In a similar way, the permutation method presented in this paper aims at identifying a set of variables or predictors which achieve maximum discrimination for two possible outcomes. Kulldorff (1997) proposed a method to identify among a large set of cluster candidates a small group of clusters which are statistically correlated to the occurrence of a high rates of diseases. We extend the purely spatial approach by replacing the set of spatial clusters by an extensive collection of variables, hereafter called predictors. In a first stage, the method selects variables based on univariate Bernoulli models. Further significant variables are sequentially incorporated into the univariate models using a conditional Monte Carlo simulation. The conditional Monte Carlo simulation stops the aggregation of new variables if no statistically significant variables are found.

2.1. Statistical model and test statistic

For the classification regression model, we apply a logistic regression model in which the response variable follows a Bernoulli distribution, $P(Y = y) = \mu^y(1-\mu)^{1-y}, y \in \{0, 1\}$ and $0 \leq \mu \leq 1$. The outcome of the model or the dependent variable is related to a linear predictor by means of a logistic function, $\mu = \exp(\beta_0 + \beta_1 x)/[1 + \exp(\beta_0 + \beta_1 x)]$, where μ is the probability of the event $Y = 1$. The parameters of the model (β's) are obtained applying the Fisher Score algorithm (MacCullagh and Nelder, 1989). The test statistic chosen for the logistic model is the log-likelihood ratio statistic or the Deviance statistic. The Deviance statistic is the difference of the log-likelihood functions between the saturated and the adjusted model, and it is written as:

$$D = 2 \cdot \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i) \log\left(\frac{1 - y_i}{1 - \hat{\mu}_i}\right) \tag{4}$$

The interpretation of the Deviance statistic is similar to the error statistic for the Gaussian regression model. The closer the model prediction to the observed responses, the smaller is the value of the Deviance.

In our methodology the Deviance statistic is first calculated for each predictor using univariate logistic models. Then, the predictors are ranked by increasing order of their Deviances. The predictors with smaller Deviance fit better the data. There is special interest to create a cutting point to the Deviance statistic so that predictors with Deviances below the threshold can be regarded as being statistically significant predictors. This can be achieved using permutation procedures under the assumption that there is no association among the predictors and the response, as it will be shown next.

2.2. The Permutation Procedure

Permutation procedures have been extensively used to address statistical significance of models and variables (Dudoit et al., 2002; Wald and Wolfowitz, 1944). In our approach, the permutation method aims at finding a threshold for the values of the Deviance statistic. Predictors with values below this threshold reject the null hypothesis of no association between the predictor variable and the response variable, at α-level. The procedure is as follows: let D_k be the Deviance value associated with the univariate logistic model of the k-th predictor, $k = 1, ..., p$, and $\mathbf{x}_k^T = [x_{k1}, ..., x_{kn}]$ is the associated vector of samples with dimension n, i.e., the sample size. Initially, the elements of vector $\mathbf{x}_k^T$ are randomly shuffled and the new value of the Deviance is calculated. The test statistic is defined as the minimum value among all values of the Deviance calculated for each shuffled vector $\mathbf{x}_k^T$, $T = min\{D_k\}$. Since the analytical distribution of the T statistic under the null is unknown, the previous procedure is repeated S times. Standard values for S are 1,000 or 10,000. The threshold is defined as the 100α percentile of the simulated test statistics.

As opposed to the Monte Carlo permutation procedure presented by Kulldorff (1997), our approach looks for the minimum value of Deviance amongst all univariate models. This differs from the original approach that seeks for the maximum value of the likelihood ratio test. As a consequence, the threshold of Deviance is defined as the 100α percentile and not the $100(1 - \alpha)$ percentile.

The elements of the permutation algorithm are shown below:

1. Adjust univariate logistic models for each dependent variable k.
2. Sort the values of Deviance in ascending order.
3. Generate S simulated values of the test statistic T as follows:
   (a) Randomly shuffle the elements of each vector $\mathbf{x}_k^T$.
   (b) Calculate the new Deviance, D_{ik}, where i is the i^{th} simulation of S.
   (c) Make $T_i = min\{D_{ik}\}$
4. Get the threshold as the 100α percentile from the simulated T_i's.
5. Select the probes whose Deviances are smaller than the threshold.

An important fact in this procedure is that the empirical distribution is estimated from the minimum value of Deviance among all predictors. Thus, the distribution represents the simulated performance of the best predictor under the null. Consequently, the alternative hypothesis is that the best predictor is strongly associated with the observed response. Furthermore, more than one predictor can reject the null hypothesis, allowing this procedure to identify a smaller subset of predictors. Nevertheless, the procedure does not take into account the correlation between the significant predictors and therefore it may find a small group of strongly correlated predictors. In order to overcome this problem we propose a multivariate approach next.

2.3. The Multivariate Approach

In order to adjust the correlation between the predictors and build multivariate models, we propose a sequential procedure. Initially, an univariate logistic model is adjusted using the predictor whose value of Deviance rejected the most the null hypothesis, using the procedure shown in section 2.2. Hereafter, this predictor is named as the seed predictor. Having found the univariate model, the $n-1$ remaining predictors are evaluated as the potential second predictor in the logistic model. The second predictor, which provides the smallest value of the Deviance, is definitely aggregated into the model and a new scan starts for the $n-2$ remaining probes. The procedure is repeated until the model reaches K variables, defined by the user, or may reach a smaller number of variables based on a conditional hypothesis testing.

This procedure ensures the selection of models with a maximum of K variables with optimized classification fit starting from the seed probe. Inference is evaluated for each value of j, where j is the number of probes in the multivariate models, $j = 1, 2, ..., K$.

Suppose T_1 is the test statistics for the univariate model, then section 2.2 describes the permutation procedure used to evaluate the null hypothesis. As described before, if the null hypothesis is rejected then a second scan searches for the predictor that best fits the first one. Let T_2 be the test statistic for the multivariate model with two variables. The second variable is chosen so that the model fitness is improved, i.e. $T_2 < T_1$. Statistical inference tests whether the addition of the new variable rejects the null hypothesis. In this case, the new p-value is defined as $P(T_2 < \hat{T}_2|\hat{T}_1, H_0)$, where $\hat{T}_2$ is the observed Deviance statistic for the model with two variables. In order to run a conditional permutation procedure we simply keep the previously selected predictor in the model, with its previously estimated parameter, and we apply the permutation procedure to the remaining variables. This procedure can be further extended if more than two variables are fixed in the model, that is, their estimated parameters are kept in the model and the permutation is run only with the remaining variables. By doing so, the method requires that for each new variable aggregated into the model, a new permutation procedure has to be run.

Estimated p-values using permutation procedures are calculated as

$$\left[1 + \sum_i I(T_k^{(i)} < \hat{T}_k)\right] / (N_p + 1) \tag{5}$$

where N_p is the total number of permutations, usually chosen as 999 or $9,999$, $T_k^{(i)}$ is the simulated test statistic at the i permutation step, for the model with k variables, and $\hat{T}_k$ is the observed test statistic.

It is important to mention that the proposed variable selection procedure is stepwise based but at each step the previously selected covariates and the estimated coefficients are kept constant in the model as an offset. Therefore, in the current step only the candidate covariate that provides the best fit is evaluated, i.e., the previously estimated coefficients do not change. Furthermore, the Monte Carlo simulation analysis is also conditional. It evaluates whether the change of the deviance related to the best variable is statistically significant, given the covariates which are already fixed into the model. In fact, the proposed analysis represents a statistical sequential procedures in which at each step the chance of rejecting the null hypothesis is α. The α value must be chosen by the user, standard values are 0.05 ou 0.10. If the null hypothesis is true then the average number of variables that are aggregated into the model is $1/(1-\alpha)$. For example, if $\alpha = 0.05$ then, under the null hypothesis, the average number of variables in the model is 1.05. Even if $\alpha = 0.10$ the average number of variables under the null is also small, say $1/(1-0.10) = 1.11$. Therefore, our approach controls the expected number of variables under the null hypothesis. As a consequence, our proposal selects models with smaller order. Finally, the estimated p-value provides empirical evidence against the null.

The elements of the multivariate permutation algorithm, starting with the previous best univariate variable $x^{(1)}$, are shown below:

1. Let offset $= \mathbf{x}_{(1)}^T\hat{\beta}_1$ be the linear predictor of the best univariate logistic model.
2. For the remaining predictors adjust logistic models with the following linear predictor:

$$\eta_j = \beta_0 + \text{offset} + \mathbf{x}_k^T\beta_k$$

3. Sort the values of Deviance in ascending order and select the predictor with the minimum value of the Deviance, say $\mathbf{x}_i^T\hat{\beta}_i$.
4. Generate S simulated values of the test statistic T as follows:
   (a) Randomly shuffle the elements of each vector $\mathbf{x}_k^T$.
   (b) Calculate the new Deviance, D_{ik}, where i is the i^{th} simulation of S.
   (c) Make $T_i = min\{D_{ik}\}$
5. Calculate the p-value using Equation 5 and the simulated test statistics: $T_1, \ldots, T_S$.
6. If the p-value is greater than α then stop. Otherwise let

$$\text{offset} \leftarrow \text{offset} + \mathbf{x}_i^T\hat{\beta}_i$$

   and exclude the i^{th} predictor from the remaining predictors.
7. If the number of variables in the offset is equal to K then stop, otherwise go to step 2.

As can be seen in the multivariate algorithm, for each new variable in the multivariate analysis a new sequence of S simulations is required. Therefore, the Monte Carlo simulation represents a major computing burden. Abrams et al. (2010) propose and evaluate estimated p-value for spatial scan statistics by first running a limited number of permutation procedures, then fitting a parametric distribution to the simulated sample and then estimating the p-value from the fitted distribution, as if the parametric distribution were the real one, under the null. The normal, log-normal, gamma, and Gumbel distributions were evaluated and the Gumbel distribution achieved the best results. In general, the estimated p-values proposed by Abrams et al. (2010) provides proper accuracy and requires a reduced number of permutation procedures.

The Gumbel distribution can save computing time by running just a few permutations and then fitting the distribution. However, our procedure shares an uncommon behavior. The more predictors enters the model, the smaller is its deviance, and eventually the simulated deviances might reach zero. This means that, under the null, for a large number of predictors and a small sample size, the more predictors enters the model the more likely is the model to eventually overfitting the data, and therefore reaching a null deviance. Thus, after a certain number of predictors, the distribution of the simulated test statistic becomes inflated with zeros. Furthermore, the test statistic is positive only and the more variables enter the model the closer is the distribution to zero. Thus, in addition to the zeros, the distribution becomes a truncated distribution.

As a consequence of the model overfit to the data, the number of predictors in the multivariate model might be less than K. Furthermore, statistical inference using the standard Gumbel distribution is not reliable. Therefore, we propose a new Gumbel distribution which accounts for the increased number of zeros and the truncation of the distribution: The Truncated and Zero Inflated Gumbel Distribution (TZIGD), shown next.

2.4. The Truncated and Zero Inflated Gumbel Distribution

If we initially consider the univariate model, then the T_1 statistic is the minimum value of the deviance among all univariate models. If more variables are evaluated, then the T_k statistics are also extreme values of the deviances. In these cases, it is natural to consider one of the extreme value distributions as a possible candidate to approximate the simulated distribution. Following Abrams et al. (2010), we first introduce the Gumbel distribution, shown in Equation 6.

$$F(X = x) = \exp\{-\exp[-(x - \eta)/\tau]\} \tag{6}$$

where η ($\eta \in \Re$) is the location parameter and τ ($\tau > 0$) is the scale parameter.

Abrams et al. (2010) applies the method of moments to estimate the parameters as: $\hat{\tau} = s \cdot \sqrt{6}/\pi$, and $\eta = \overline{x} - \gamma\hat{\tau}$ where s is the sample standard deviation and γ is the Euler's constant. ($\gamma = 0.577215665$). From Equation 6, the estimated p-value using the Gumbel approximation is written as

$$\text{p-value} = 1 - \exp\{-\exp[-(x_{obs} - \hat{\eta})/\hat{\tau}]\} \tag{7}$$

where x_{obs} is the observed test statistic.

The proposed Truncated and Zero Inflated Gumbel density distribution is given by:

$$f(x) = \begin{cases} (1-\pi)f_G(x)/F_0, & \text{if } x < 0 \\ \pi, & \text{if } x = 0 \end{cases} \tag{8}$$

where $f_G(x) = \tau^{-1}(\exp[-(x - \eta)/\tau]) \cdot (\exp\{-\exp[-(x - \eta)/\tau]\})$ is the standard Gumbel density distribution, and $F_0 = P(X > 0) = \exp\{-\exp[\eta/\tau]\}$ is the truncated component.

The cumulative density distribution (cdf) of the truncated and zero inflated Gumbel distribution is given as:

$$F(x) = \begin{cases} (1-\pi)F_G(x)/F_0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases} \tag{9}$$

where $F_G(x) = \exp\{-\exp[-(x - \eta)/\tau]\}$.

Figures 1 and 2 show the Gumbel density and the cumulative Gumbel distribution adjusted to a simulated data of minimum Deviances. As will be shown, the Gumbel provides a proper fit and can be used to reduce the number of Monte Carlo simulations.

In order to estimate the parameters η, τ and π of the proposed distribution we apply the Newton-Raphson algorithm (Ypma, 1995). Therefore, the log-Likelihood equation using the TZIG distribution is given by:

$$
\begin{aligned}
l(x_1, ..., x_n; \eta, \tau, \pi) = n_0 \log \pi + (n - n_0) \log(1 - \pi) - (n - n_0) \log F_0 \\
+ \sum_{j=n_0+1}^{n} \log f_G(x_j)
\end{aligned} \tag{10}
$$

where n is the sample size, n_0 is the number of samples in which $x = 0$ and $(n - n_0)$ is the number of samples in which $x < 0$.

The maximum likelihood estimate for the parameter π, defined as $\frac{\partial l(.)}{\partial \pi} = 0$, is given by $\hat{\pi} = n_0/n$. The elements of the Score vector (Equations 11 and 12) and the Hessian matrix (Equations 13, 14 and 15) used in the Newton Raphson algorithm are given as:

$$
\frac{\partial l(.)}{\partial \eta} = (n - n_0) \exp(\eta/\tau)\tau^{-1} + (n - n_0)\tau^{-1} - \tau^{-1} \sum_j c_j \tag{11}
$$

$$
\begin{aligned}
\frac{\partial l(.)}{\partial \tau} =& - (n - n_0) \exp(\eta/\tau)\eta\tau^{-2} - (n - n_0)\tau^{-1} \\
&+ \sum_j (x_j - \eta)\tau^{-2} - c_j(x_j - \eta)\tau^{-2}
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
\frac{\partial^2 l(.)}{\partial \tau \partial \eta} =& - (n - n_0) \exp(\eta/\tau)\tau^{-2}[\eta\tau^{-1} + 1] - (n - n_0)\tau^{-2} \\
&+ \tau^{-2} \left\{ \sum_j c_j \right\} - \tau^{-3} \sum_j c_j(x_j - \eta)
\end{aligned} \tag{13}
$$

$$
\frac{\partial^2 l(.)}{\partial \eta \partial \eta} = (n - n_0) \exp(\eta/\tau)\tau^{-2} - \tau^{-2} \sum_j c_j \tag{14}
$$

$$
\begin{aligned}
\frac{\partial^2 l(.)}{\partial \tau \partial \tau} =& (n - n_0)\eta \exp(\eta/\tau)\tau^{-2} \left[\eta\tau^{-2} + 2\tau^{-1}\right] + (n - n_0)\tau^{-2} \\
&- 2\tau^{-3} \sum_j (x_j - \eta)(1 - c_j) - \tau^{-4} \left\{ \sum_j (x_j - \eta)^2 c_j \right\}
\end{aligned} \tag{15}
$$

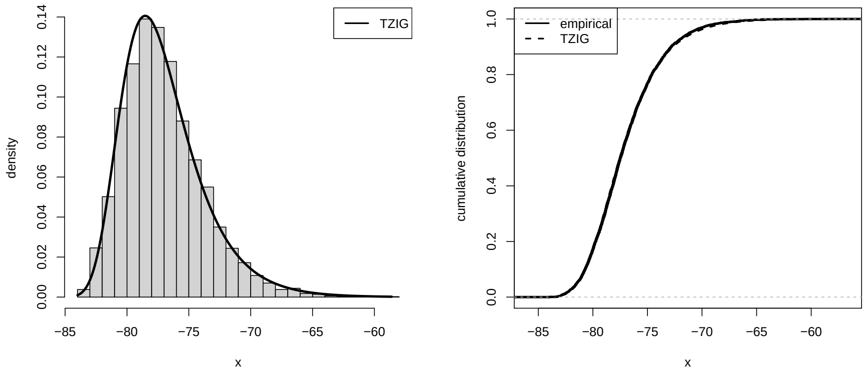
where $c_j = \exp[-(x_j - \eta)/\tau]$ and $\frac{\partial^2 l(.)}{\partial \eta \partial \tau} = \frac{\partial^2 l(.)}{\partial \tau \partial \eta}$.

Initial values for η and τ are generated using only the negative values of x ($x < 0$) and applying the method of moments, previously presented. Then, the Newton-Raphson algorithm is applied until convergence.

It is worth noting that the Truncated and Zero Inflated Gumbel distribution proposed in this work assumes a negative random variable ($x \leq 0$) which is not the case of the Deviance statistic. In fact, the Deviance statistic is a positive value. Therefore, we apply the Gumbel distribution to the negative values of Deviance, that is, $x_i = -T_i$. We do so in order to be consistent with the original methodology proposed by Abrams et al. (2010).

The estimated p-value after fitting the TZIGD and using the negative values of the Deviance is given as:

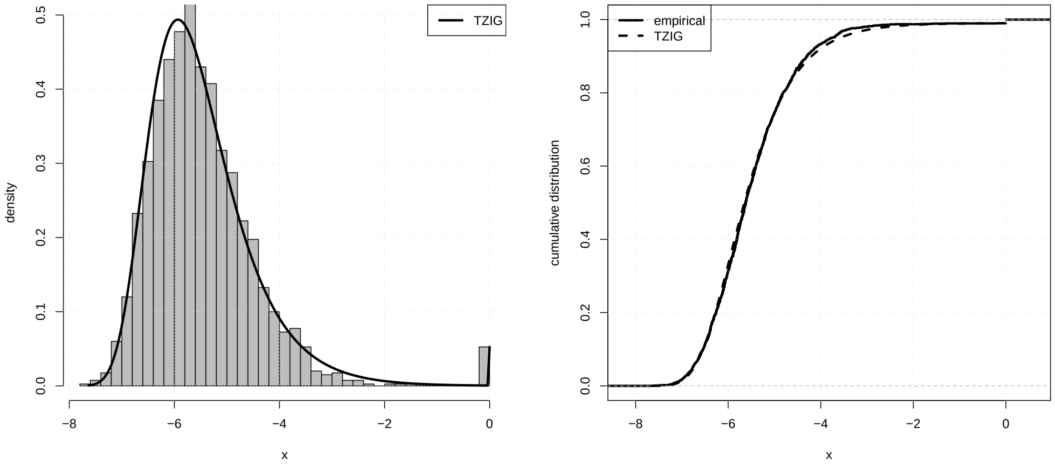
$$
\text{p-value} = 1 - \hat{F}(-T_{obs}) \tag{16}
$$



(a) Density function

(b) Cumulative distribution

Fig. 1. Density function and cumulative distribution for the proposed Truncated and Zero Inflated Gumbel Distribution, and their comparison with the empirical estimates. In this case the data does not present a truncated nor zero inflated behavior.



(a) Density function

(b) Cumulative distribution

Fig. 2. Density function and cumulative distribution for the proposed Truncated and Zero Inflated Gumbel Distribution, and their comparison with the empirical estimates. In this case the data presents truncated and zero inflated behavior.

where $\hat{F}$ is the estimated cumulative distribution, using the estimated parameters, given in Equation 9, and T_{obs} is the observed test statistic.

2.5. Least absolute shrinkage and selection operator method (lasso), and variants

In addition to our proposed method, we investigated the performance of different variable selection methods such as lasso, elastic-net, sparsenet, Bolasso, and stepAIC.

The lasso method (Tibshirani, 1996) aims at minimizing the Deviance function (Equation 4), subject to the sum of the absolute value of the coefficients being less than a constant, $\sum_j |\beta_j| \leq s$, where $s \geq 0$. This constraint also can be written as a penalty function of form $\lambda \sum_j |\beta_j|$, where λ ($\lambda \geq 0$) is a monotonic function of s, $\lambda_{(s)}$. Values of λ can produce coefficients that are exactly zero, therefore, the lasso minimizes the objective function and selects variables, simultaneously. Recently, Friedman et al. (2012) proposed fast algorithms to select generalized linear models with convex penalties. These algorithms are available in the R package glmnet(Friedman et al., 2010), which implements the coordinate descent algorithm, and the elastic-net model. The elastic-net method is a variant of lasso in which the penalty function is a linear combination of both lasso and ridge penalty functions: $\lambda[\gamma \sum_j |\beta_j| + (1-\gamma) \sum_j \beta_j^2]$, where $\gamma \in [0,1]$ defines the relative emphasis on lasso and ridge penalty functions and λ controls the overall elastic-net penalty. It can be seen that lasso is a special case of elastic-net if $\gamma = 1$. In general, it selects more variables than the standard lasso. Furthermore, cross-validation procedures are normally applied to estimate the penalty parameters λ and γ.

Alternatively, Mazumder et al. (2011) applies a non-convex penalty function of form $\lambda P(|\beta|; \lambda; \gamma)$ which defines a family of penalty functions concave in $|\beta|$, where λ and γ are the parameters which control the degrees of regularization and concavity of the penalty, respectively. Known as sparsenet, it induces less bias than lasso and elastic-net. In our simulation study, we applied the minimax concave (MC+) penalty (Zhang, 2010) which assumes the following form: $\lambda P(t; \lambda; \gamma) = \lambda \int_0^{|t|} (1 - x/(\gamma\lambda))_+ \, dx$, where $(u)_+ = max(u, 0)$. This method is also available in the R package sparsenet (Mazumder et al., 2011).

The bootstrap lasso, known as Bolasso, applies the standard lasso method to B bootstrap samples, and then selects the variables which are consistently selected in each sample. According to Bach (2008), it leads to consistent model selection. In practice, it selects a smaller number of variables as compared to lasso.

It is worth noting that lasso, selected variants, and our proposed method aim at selecting variables, which is the real issue in genomic data. Furthermore, some of lasso variants such as elastic net and sparsenet, also handles multicollinearity. We explore the ability of each method to select variables using a simulation study presented in section 4.

3. The Application

Since the advent of high throughput genomic technologies a decade ago, massive information at the genomic level is available which can be exploited in medical studies. Microarrays allow the simultaneous measurement of the expression levels of many thousands of genes. Relying on these data, it has become possible to design efficient predictors that significantly outperform the previous clinic-pathologic ones. Particularly, this work is focused on prediction of response to pre-operative chemotherapy treatments for breast cancer.

In the process of designing predictors of the outcomes of chemotherapy treatments, the main issues are that of identifying the genes actually involved in the response to the chemotherapy treatment; combining the expression levels of these genes for predicting the response to treatment; and assessing the robustness of predictors, i.e. the statistical independence of the method - gene selection and computational model - related to the learning set of patient cases.

In its essence, designing a predictor is a supervised learning process (Haykin, 2008) which aims at recognizing the class of patient cases who are responders to the treatment, and that of the non-responders. The novelty comes from the nature of the high throughput genomic data and clinical trials. The data are very noisy due to the present state of DNA microarrays technology, and also because we are dealing with biological data (Klebanov and Yakovlev, 2007). Hence, selecting a very small subset of relevant genes is crucial for designing robust predictors.

3.1. The datasets

Dataset 1: Our dataset comes from clinical trials which were conducted at MD Anderson Cancer Center, Houston-Texas (USA) in 2004 and 2005 (Hess et al., 2006). For each patient case, the data are the outcome of the pre-operative chemotherapy treatment, either complete responder (PCR) or non complete responder (residual disease or no pathologic complete response - noPCR), and the expression level of 22,283 probe sets, measured on the tumor tissues. The gene expression profiling was performed using oligonucleotide microarrays (Affymetrix U133A). The data set has previously been used by Natowicz et al. (2008a,b,c).

Training and validation tests represent two groups of women from two different nationalities, respectively. Training is a sample of 82 North-American women and validation is a sample of 51 French women.

An important feature of our method is that it does not require data cross-validation to select probes. As will be shown, after selecting the multivariate models for the training set, the validation set is used only to provide a classification matrix. Therefore, the validation set also works as the test set.

Both training and validation datasets were pre-processed by applying the logarithm function with base 10 to the expression levels of the genes, followed by the quantile normalization process (Bolstad et al., 2003). We set a significance level of $\alpha = 0.10$ (10%) for both univariate and multivariate selection procedures. The maximum number of probes in the multivariate model was set at $K = 5$.

Dataset 2: The second data set is also known as the van't Veer data set (van 't Veer et al., 2002). The data has 24,481 gene expression signatures from 78 young patients diagnosed with breast cancer. The patients were grouped into poor prognosis and good prognosis groups, with 34 and 44 patients, respectively. A validation set of 19 patients consisting of 7 patients with poor diagnosis and 12 patients with good diagnosis is also provided.

Originally, van 't Veer et al. (2002) evaluated gene expression levels as potential predictors of cancer prognosis. Succinctly, approximately 5,000 genes were primarily selected amongst the 24,481 genes on the microarray. From the pre-selected genes, 231 genes were statistically associated with the disease outcome. A combination of unsupervised clustering methods and supervised classification methods were applied to select significant genes.

Similar to the previous data set, both training and validation datasets were pre-processed by applying the logarithm function with base 10 to the expression levels of the genes. The quantile normalization process was not applied. We set a significance level of $\alpha = 0.10$ (10%) for both univariate and multivariate selection procedures. The maximum number of probes in the multivariate model was set at $K = 5$.

4. A simulation study

We propose a simulation study to compare our proposed method, named Permutation, the standard AIC stepwise selection procedure (Sakamoto et al., 1999), named stepAIC, the lasso (using glmnet package), the elastic-net (using glmnet package), Bolasso (Bach, 2008) and MC+ (using sparsenet package). Following Zou and Hastie (2005), we simulate data from three scenarios named A, B and C. The first two scenarios were used in the original lasso paper (Tibshirani, 1996). We simulate data from the true model:

$$\eta = \mathbf{X}\beta$$

where β is the parameter vector of size 40, and the columns of the prediction matrix $\mathbf{X}$ were generated from standard normal distributions, $\mathbf{x}_i \sim N(0, 1)$ with distinct covariance matrices.

For each scenario, we simulate a total of 50 training and 50 validation sets. The training and validation data sets have 100 observations each. For each simulation, the parameters of lasso, elastic-net, Bolasso and sparsenet are selected using a leave-one-out cross validation applied to the training set. We use the following grid of values for γ, $\gamma \in \{0.1, 0.2, ..., \ldots, 0.9\}$, for elastic-net.

In scenario A, we let $\beta_1 = 3$, $\beta_2 = 1.5$ and $\beta_5 = 2$, and the remaining elements of the vector β were set as zero. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ was set to be $corr(i, j) = 0.5^{|i-j|}$. In scenario B, we let $\beta_i = 2$ for $i \in \{11, 12, ..., 20, 31, 32, ..., 40\}$, the remaining elements were set as zero, and $corr(i, j) = 0.5$ for all i and j. In scenario C, we let $\beta_i = 3$ for $i \in \{1, 2, ..., 15\}$, the remaining elements were set as zero, and the $\mathbf{x}_i$ are independent identically distributed.

A logistic function was applied to the outcome η to define a two class outcome variable y with $Pr(y = 1) = 1/(1 + exp(-\eta))$, $Pr(y = 0) = 1 - Pr(y = 1)$, as proposed in Friedman et al. (2012).

Each method was investigated with respect to their performance to predict both training and validation data sets, i.e., the training and validation accuracy. The accuracy was calculated as the average proportion of predicted outputs which were matched to the observed output. That is, if the predictive value is equal or larger than 0.5 we assume that the predicted output is $y = 1$, and $y = 0$ otherwise. We evaluated the average number of detected variables by each method. Furthermore, among the detected variables we calculated the average number of variables which belong the real variables, as described in the simulations scenarios. We also included the average number of variables which were falsely detected. These variables do no belong to the real variables. Finally, we evaluated the average proportion of the detected variables which belong to the real set of variables, this measure is known as the positive predictive value (Costa and Assunção, 2005; Tango and Takahashi, 2005).

5. Results

5.1. The simulation study

Table 1 presents the results of the simulated scenarios. The number of real variables in scenarios A, B and C are 3, 20 and 15, respectively. The stepAIC method presented the best training accuracy values (100%) for all scenarios, which indicates overfit of the training data . The Bolasso method presented the worst training accuracy values. The elastic-net method achieved the best validation results, and the Bolasso presented the worst values of validation accuracy. For scenario A the stepAIC method detected the largest number of variables, on average. For scenarios B and C the elastic-net detected the largest number of variables, on average. On the contrary, the Bolasso method detected the smallest number of variables for all scenarios, on average. As a consequence, the Bolasso method presented the smallest values of falsely detected variables. It worth noting that the Bolasso method can detect models with size of zero. These results affect the number of falsely detected and the number of real detected variables. Furthermore, if the detected model is of size zero, the positive predictive value can not be calculated, as shown in the results for the Bolasso method for scenarios B and C. Our proposed method detects models with smaller sizes as compared to lasso, elastic-net, sparsenet and stepAIC; and it presents the best positive predictive values. Thus, the simulation study indicates that our proposed approach detects smaller models in which most variables belong to the real variables, on average. The lasso and elastic-net detect largest models with a large number of detected variables which belong to the real variables, but with a smaller positive predictive value, as compared to our proposal. In conclusion, the simulation study indicates that our proposed model detects compact models in which most of the detected variables belong to the real variables, on average.

Table 1. Simulation study results for data sets A, B and C

Simulated data A						
Statistics/Methods	Permutation	lasso	elastic-net	Bolasso	Sparsenet	stepAIC
Training accuracy	88.5% (3%)	89.7% (4%)	90.1% (4%)	49.2% (5%)	89.2% (3%)	100% (0%)
Validation accuracy	86.0% (5%)	86.9% (4%)	86.9% (4%)	83.6% (5%)	86.1% (5%)	80.6% (6%)
Numb. detected variables	2.82 (0.48)	5.16 (1.86)	5.68 (2.55)	1.84 (0.71)	4.48 (3.45)	11.02 (3.21)
Falsely detected var.	0.1 (0.30)	2.18 (1.83)	2.7 (2.53)	0 (0)	1.84 (3.28)	8.24 (3.35)
Real detected variables	2.72 (0.50)	2.98 (0.14)	2.98 (0.14)	1.84 (0.71)	2.64 (0.53)	2.78 (0.46)
Positive predictive value	96.7% (4.6%)	65.0% (22%)	60.4% (21%)	100% (0%)	75.1% (27%)	27.6% (10%)
Simulated data B						
Statistics/Methods	Permutation	lasso	elastic-net	Bolasso	Sparsenet	stepAIC
Training accuracy	97.0% (2%)	99.3% (1%)	99.7% (1%)	50.8% (5%)	96.5% (3%)	100% (0%)
Validation accuracy	87.1% (4%)	92.0% (3%)	93.9% (3%)	54.0% (5%)	88.5% (5%)	86.2% (4%)
Numb. detected variables	5.02 (1.12)	16.58 (2.15)	28.6 (7.1)	0.16 (0.42)	11.86 (6.29)	5.82 (1.17)
Falsely detected variables	1.16 (1.06)	5.22 (1.82)	12 (4.7)	0.02 (0.14)	4.34 (3.29)	1.62 (1.34)
Real detected variables	3.86 (1.11)	11.36 (2.03)	16.6 (2.97)	0.14 (0.40)	7.52 (3.75)	4.2 (1.21)
Positive predictive value	78.0% (18%)	68.6% (10%)	59.6% (8%)	NA	66.6% (16%)	73.5% (21%)
Simulated data C						
Statistics/Methods	Permutation	lasso	elastic-net	Bolasso	Sparsenet	stepAIC
Training accuracy	74.7% (8%)	95.8% (3%)	96.8% (3%)	49.4% (5%)	95.2% (3%)	100% (0%)
Validation accuracy	61.2% (5%)	81.3% (6%)	82.4% (5%)	57.5% (6%)	80.1% (6%)	80.1% (6%)
Numb. detected variables	2.88 (1.49)	20.94 (4.27)	25.4 (5.9)	1.42 (1.01)	19.8 (7.68)	13.06 (1.20)
Falsely detected variables	0.04 (0.20)	6.88 (3.51)	10.9 (5.47)	0.00 (0.00)	6.82 (6.13)	1.44 (1.39)
Real detected variables	2.84 (1.48)	14.06 (1.15)	14.5 (0.8)	1.42 (1.01)	12.98 (2.32)	11.62 (1.59)
Positive predictive value	98.9% (5%)	69.3% (11%)	59.7% (12%)	NA	71.3% (17%)	89.0% (10%)

Table 2. Simulation study results for data sets B and C, using the Permutation method for different values of the parameter α.

Simulated data B			
Statistics/Methods	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
Training accuracy	96.2% (3%)	97% (3%)	98.1% (2%)
Validation accuracy	85.9% (4%)	87.1% (4%)	87.8% (3%)
Numb. detected variables	4.32 (1.08)	5.02 (1.11)	6.06 (1.43)
Falsely detected var.	0.94 (0.94)	1.16 (1.06)	1.46 (1.25)
Real detected variables	3.38 (1.16)	3.86 (1.11)	4.6 (1.36)
Positive predictive value	78.86% (20%)	77.98% (18%)	76.88% (18%)
Simulated data C			
Statistics/Methods	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
Training accuracy	70.5% (8%)	74.7% (8%)	79.6% (7%)
Validation accuracy	59.5% (5%)	61.2% (5%)	64.5% 6%)
Numb. detected variables	1.98 (1.24)	2.88 (1.49)	4.82 (2.17)
Falsely detected var.	0.02 (0.14)	0.04 (0.20)	0.2 (0.4)
Real detected variables	1.96 (1.19)	2.84 (1.48)	4.62 (2.05)
Positive predictive value	99.6% (3%)	98.9% (5%)	96.5% (8%)

Table 2 shows the performance of the Permutation method for different values of α, for data sets B and C. We did not include results for data set A because it has a small number of real variables (only three). In general, the larger the value of α, the larger is the number detected variables. As a consequence, the larger the number of falsely detected variables, real detected variables, and the higher the training and validation accuracies. Nevertheless, the positive predictive value statistic decreases for larger values of α.

In addition, computing times (in seconds) using data sets with different numbers of variables: 100, 1,000, 10,000, 20,000 and 25,000 variables; and for different values for the parameter K (maximum number of variables in the model), are presented in Figure 3. In general, there is a linear correlation between the time required to run the Permutation procedure, the number of variables and the parameter K. Computing times were estimated using simulated data sets with 200 observations and $K \times 100$ Monte Carlo simulations, i.e., 100 Monte Carlo simulations for each potential variable in the model.

5.2. Case study 1:

Initially, we investigate the use of the Gumbel distribution as the proper model for p-value estimates in a situation with limited samples. Therefore, we designed the following experiment: we first generate through Monte Carlo simulations a sample of size 5,000 of the test statistic. Next, we choose the following order statistics from the set of simulated samples: the 90^{th} order statistic, the 95^{th} order statistic and the 99^{th} order statistic. If these values are chosen as the observed statistic then the associated p-values, also named as the true p-values, are 0.10, 0.05 and 0.01, respectively. Finally, we randomly select small samples of size 100 from the original simulated sample and estimate p-values using the standard Monte Carlo approach (Equation 5) and the Gumbel approximation (Equation 16). We repeat this procedure 2,000 times in order to generate empirical distributions for the p-values for both standard and Gumbel methods.

Table 3 shows the descriptive statistics for the simulated p-values. In general, the statistical means and medians for the Gumbel approximations are closer to the real p-value than the standard Monte Carlo approach. The quantile distances, as well as the distance between the maximum and the minimum values, are also smaller for the Gumbel approximation. From these results we can conclude that there is empirical evidence that the Gumbel approach provides more accurate p-values, and it requires a smaller set of simulated test statistics than the standard Monte Carlo approach. These findings are applied only to data set 1.

Table 4 shows the univariate model and also the models with two and three probes; both multivariate models were selected using the multivariate approach. Based on the p-value, the model with two probes is statistically significant whereas the model with three probes is not statistically significant. It is worth noting that, in general, the sequential procedure generates

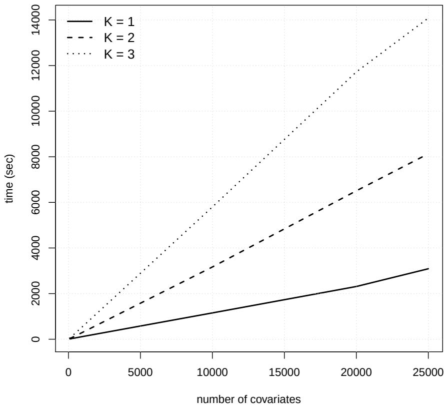


Fig. 3. Computing time (in seconds) required to run the Permutation method using data sets with different number of variables, and different values for the maximum number of variables in the model (parameter K).

Table 3. Descriptive statistics for the estimated p-values using the standard Monte Carlo approach and the Gumbel approximation, using a sample size of 100 and 2,000 simulated values.

real	type	minimum	1st Quant.	median	mean	3rd Quant	maximum
0.10	standard	0.0198	0.0891	0.1090	0.1083	0.1287	0.2475
	Gumbel	0.0446	0.0837	0.0971	0.0986	0.1123	0.1757
0.05	standard	0.0099	0.0396	0.0594	0.0594	0.0693	0.1485
	Gumbel	0.0149	0.0403	0.0488	0.0499	0.0584	0.1033
0.01	standard	0.0099	0.0099	0.0198	0.0197	0.0297	0.0594
	Gumbel	0.0022	0.0086	0.0112	0.0119	0.0146	0.0319

smaller AIC statistics as more probes are included. This is due to the multivariate selection criteria. That is, the multivariate analysis chooses the next probe as the best probe from among more than 22,000 candidates. As a consequence of the large number of candidates, the best probe provides, in general, a substantial reduction of the Deviance and the AIC statistics. Therefore, the AIC statistic can not be used as the selection criterion in our approach. Our findings also suggest using the p-value as the model selection criterion. We apply a similar procedure based on the $stepwise$ procedure, a standard variable selection approach in linear regression analysis (Montgomery et al., 2012). The $stepwise$ approach generally applies p-values of 0.15 or 0.20 as stopping criteria in sequential variable selection procedures. Therefore, if we choose a maximum p-value of 0.20 then the final model is the one with two probes: $205548\_s\_at + 217016\_x\_at$.

Table 5 presents the validation results. As mentioned previously, our approach does not require any validation set to select the proper number of variables. Therefore, Table 5 shows predictive results for independent data. Our selected model with two probes has an accuracy measure of 72.5%.

Results using lasso are presented in Table 6. The final model selected by lasso has 3 probes. It is worth noting that the probe 205548_s_at was also selected in our approach as the best univariate probe. In addition, the lasso model achieved a validation accuracy measure of 76.47%, which is 5.47% higher that the validation accuracy achieved by our approach.

Table 4. Selected probes applying the training dataset for univariate and multivariate models. The p-values under the null were obtained conditioned on the number of probes in the models. Both Gumbel and Monte Carlo p-values are shown. The AIC statistic is also displayed.

Seed probe	2nd probe	3rd probe	Deviance	pvalue	Gumbel	AIC
205548_s_at			66.25	0.0055	0.0097	70.25
205548_s_at	217016_x_at		47.57	0.1954	0.1906	52.95
205548_s_at	217016_x_at	205219_s_at	33.12	0.5557	0.5498	37.94

Table 5. Classification matrix for training and validation sets, and accuracy statistics.

Training set			Validation set				
			Classification matrix				Accuracy
Seed probe	2nd probe	3rd probe	TP	FP	FN	TN	(%)
205548_s_at			3	10	3	35	74.5
205548_s_at	217016_x_at		4	9	5	33	72.5
205548_s_at	217016_x_at	205219_s_at	5	8	6	32	72.5

Table 6. Selected probes using the lasso algorithm

Training set		Validation set				
Selected probes	Deviance	Classification matrix				Accuracy
		TP	FP	FN	TN	(%)
201976_s_at+204825_at+205548_s_at	43.86	1	12	0	38	76.47

The elastic-net model selected a larger model with 151 probes, which includes those selected by lasso, and achieved an improved validation accuracy of 84.3%.

Secondary models The univariate approach can be applied to select more than one seed probe. For each probe selected, a test statistic is assigned; therefore, probes with smaller test statistics may reject the null hypothesis as long as their p-values are smaller than the α-level. These statistically significant seed probes, or secondary seed probes, can be used to build multivariate models. This approach is appropriate for exploring potential candidate models which were not selected as the best model, and therefore may include different subsets of variables. These secondary models may present similar training and validation results as compared to the best model.

Table 7 shows multivariate models which were built based on secondary probes. Only probes 216295_s_at and 212207_at were found to be statistically significant at the α-level for the univariate analysis. From these probes, the multivariate analysis provided two models with two probes each. It is worth mentioning that probes 218671_s_at and 202261_at were not statistically significant based on the univariate analysis. Nevertheless, these probes are significant if they share the model with secondary seed probes. It is also interesting that the secondary multivariate models and the best multivariate model achieved similar validation results.

Finally, Figure 4 shows the computing time required to run the methods using an Intel Core i7, 1.73GHz with 8GB (RAM) with Windows 7. The fastest method is stepAIC, which required 368.77 seconds (or 6.15 minutes). This was followed by lasso, which required 396.04 seconds (or 6.6 minutes); sparsenet, which required 581.7 seconds (or 9.7 minutes), and elastic-net, which required 795.5 seconds (or 13.3 minutes). The Bolasso and Permutation methods required 5,290.5 seconds (1.46 hours) and 6,436.65 (or 1.79 hours), respectively. It is worth noting that the lasso method is 16.25 times faster than the Permutation procedure.

5.3. Case study 2:

Initially, we replicated the previous experiment which investigates the use of the Gumbel distribution as the proper model for p-value estimates. Table 8 shows the descriptive statistics for the simulated p-values. In general, the statistical means and medians for the Gumbel approximations are closer to the real p-value than the standard Monte Carlo approach. Nevertheless, if the true p-value gets smaller, for instance if p-value$=0.01$, the Gumbel approximation presents some bias, i.e., the median

Table 7. Secondary models built using statistically significant seed probes, which also rejected the null hypothesis for the univariate analysis.

Model	Deviance	Gumbel pvalue	MCMC pvalue	validation accuracy (%)
216295_s_at+218671_s_at	50.45	0.1885	0.2178	74.51
212207_at+202261_at	50.10	0.0450	0.0495	72.55

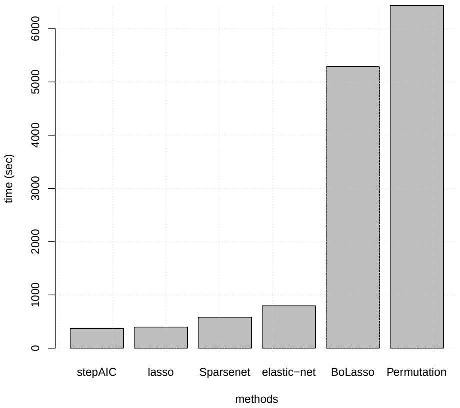


Fig. 4. Computing time (in seconds) required do run glmnet, Permutation, Sparsenet, Bolasso, and StepAIC.

Table 8. Descriptive statistics for estimated p-values using the standard Monte Carlo approach and the Gumbel approximation using the van't Veer data set.

real	type	minimum	1rst Quant.	median	mean	3rd Quant	maximum
0.10	standard	0.0198	0.0891	0.1089	0.1083	0.1287	0.2079
	Gumbel	0.0486	0.0917	0.1064	0.1068	0.1203	0.1862
0.05	standard	0.0099	0.0396	0.0594	0.0598	0.0693	0.1485
	Gumbel	0.0202	0.0486	0.0584	0.0592	0.0690	0.1152
0.01	standard	0.0099	0.0099	0.0198	0.0199	0.0297	0.0693
	Gumbel	0.0030	0.0115	0.0148	0.0154	0.0188	0.0435
sample size of 500 and 2,000 simulated values							
0.01	standard	0.0020	0.0080	0.0120	0.0118	0.0140	0.0279
	Gumbel	0.0090	0.0138	0.0152	0.0153	0.0168	0.0242

and mean estimates are slightly greater than the true or the standard results. In order to investigate this behavior, the sample size was changed to 500 and the differences among the Gumbel and standard mean and medians were found to be more evident than using the sample size of 100. This suggests that if the Gumbel p-value estimate is small, the true unknown p-value might be slightly higher. Nevertheless, this behavior may not affect main results since we use an α-level of 0.10 (10%) for both univariate and multivariate analyses.

Table 9 shows the univariate model, with probe $AK001044$, and the multivariate model which includes probes $M94046$ and $Contig30388$. Table 10 presents the validation results. It can be seen from Table 10 that the multivariate model achieved higher accuracy measures than the univariate model. The accuracy measure of the multivariate model is compared to the lasso approach in Table 11. Different from the previous data set, the lasso did not find a short subset of probes. In this case, it found a model with 30 probes whose validation accuracy was much smaller than our multivariate approach. The elastic-net model selected a larger model with 277 probes, and achieved a validation accuracy of 63.1%, which also is smaller than our proposed multivariate model (68.42%). Therefore, there is evidence that our approach is able to generate smaller subsets of probes with improved validation accuracy measure.

Table 9. Selected probes applying the training dataset for univariate and multivariate models. The p-values under the null were obtained conditioned on the number of probes in the models. Both Gumbel and Monte Carlo p-values are shown. The AIC statistic is also displayed.

Seed probe	2nd probe	3rd probe	Deviance	pvalue	Gumbel	AIC
AK001044			53.02	0.001	3.2e-11	57.02
AK001044	M94046	Contig30388	13.61	0.050	0.063	21.61

Table 10. Classification matrix for training and validation sets, and accuracy statistics.

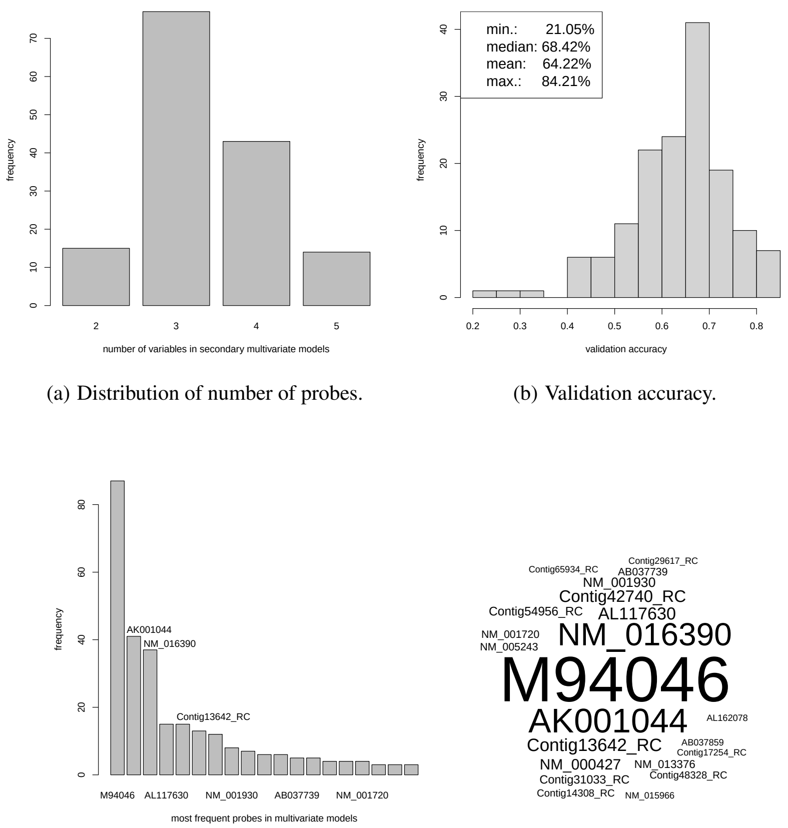
Training set			Validation set				
			Classification matrix				Accuracy
Seed probe	2nd probe	3rd probe	TP	FP	FN	TN	(%)
AK001044			4	3	7	5	47.36
AK001044	M94046	Contig30388	2	5	1	11	68.42

Table 11. Selected probes using the lasso approach for Van't Veer data set

Training set	Validation set					
Selected probes	Deviance	Classification matrix				Accuracy
		TP	FP	FN	TN	(%)
Contig50122_RC+NM_000978+M94046+NM_001720+Contig45670_RC+NM_004090+NM_001923+NM_001930+NM_001964+NM_002848+Contig726_RC+AL117630+NM_005243+NM_003992+NM_014262+Contig48328_RC+Contig42740_RC+Contig54956_RC+NM_006423+AF055033+NM_007359+Contig47544_RC+Contig37063_RC+AB037863+Contig1025_RC+Contig63102_RC+Contig17360_RC+AL050353+Contig45953_RC+Contig51070_RC+Contig21655_RC	2.14	3	4	5	7	52.63

Secondary models In addition to the multivariate model, we also explore secondary seed probes. As mentioned earlier, secondary seed probes do not have the minimum deviance statistic but their deviance statistic also rejects the null hypothesis. Remarkably, we found 149 secondary seed probes. For each one of these 149 secondary seed probes we applied the multivariate approach. Therefore, we found 149 statistically significant multivariate models. For each multivariate model we calculate the validation accuracy measure.

By applying the multivariate approach to each seed probe, the final model may have more than one predictor. Figure 5 presents the distribution of the number of predictors in the multivariate models and their validation performance. Figure 5 (a) shows the distribution of the number of probes in the secondary multivariate models. In this case, the average number of predictors is 3.4, the maximum number of predictors is 5, and the minimum number of predictors in the multivariate model is 2. Interestingly, the multivariate approach is applied to each secondary seed probe and it evaluates all remaining probes as potential predictor candidates. Therefore, the final multivariate models may share common probes as predictors. The frequency of these probes is presented in Figure 5 (c) and its associated word cloud is presented in Figure 5 (d). It can be seen that probe $M94046$, also selected in the best multivariate model shown in Table 9, is the most frequent probe chosen in the multivariate models, generated using secondary seed probes. The best univariate probe, $AK00144$, is also repeatedly chosen as an important predictor in the secondary seed probe models. The third most frequent probe is the NM_016390 probe, which has not appeared in any previous model. We conclude that, in this particular data set, the secondary models provides evidence of important probes which were not detected in the multivariate analysis using the best univariate probe, or in the best secondary univariate probe set. Furthermore, validation results for multivariate secondary models are presented in Figure 5 (b). It can be seen that, on average, the validation accuracy is 68.42% and the maximum validation accuracy is 84.21%. It is also worth noting that, on average, the accuracy of secondary multivariate models is superior to the results provided using the lasso and elastic-net algorithms.



(a) Distribution of number of probes.

(b) Validation accuracy.

(c) Frequency distribution of selected probes.

(d) Word cloud of selected probes.

Fig. 5. Distribution of selected probes in the multivariate secondary models. Results does not consider the seed probes.

6. Discussion and Conclusion

In this work we propose and evaluate a multivariate procedure for variable selection, applied to genomic data. Given the commonly large number of variables and the smaller sample size of genomic data, our proposal relies on a sequential procedure which first selects statistically significant univariate probes and then evaluates whether additional variables would still generate statistically significant models. Although the procedure searches for the single model with the best performance, the proposed statistical inference procedure can successfully select secondary models. By searching for best candidates, the procedures generate, on average, smaller models with improved performance. Therefore, the procedure is suitable for choosing small subsets of statistically significant probes.

Our method does differ from standard variable selection methods such as lasso, elastic-net and step AIC in which the estimates of coefficients change based on the objective function and penalty parameter. Our method relies on a sequential analysis (Wald, 1973) like a decision tree rather than an optimization search. For instance, if a variable is found earlier in the procedure as a significant variable, then our proposal looks for the next variable most correlated to the residuals of the previous model. This approach deals with the problem of multicollinearity, reduces the space of search, and further improves computational performance since the next step of the algorithm can be seen as an univariate model. Furthermore, the statistical significance is conditional and is related to the expected number of variables in the model under the null, which is controlled by the user. Further information on sequential analysis can be found in Wald (1973); Siegmund (1985).

Finally, we comment on another recent method for variable selection proposed by Zare et al. (2013), named as FeaLect, which is similar to Bolasso. FeaLect calculates a score statistic for each feature using Bootstrap samples, but without replacement. For each Bootstrap sample, it applied the LARS (Least Angle Regression) algorithm (Efron et al., 2004), a more general algorithm for variable selection that includes lasso. The LARS penalty parameter is chosen so that k features are selected. The parameter k is defined by the user. From the k selected features for each Bootstrap sample, a score statistic is calculated for each feature in the data set. The score statistic is sorted in ascending order and a three-segment spline is adjusted in the log-scale. The features in the top non-linear bending part of the curve are informative feature candidates. We believe that our approach is simpler than FeaLect and more intuitive, since we do not need any curve fitting analysis.

Regarding the results in the two genomic data sets, in the first one, about the two probes model with smaller AIC statistic, the probe 205548_s_at is associated with negative regulation of cell proliferation and was also selected as a relevant probe by Natowicz et al. (2008b) and Hess et al. (2006). The probe 217016_x_at is associated with Lung Cancer and can be a target to be investigated in breast cancer disease (Cheng et al., 2012). For the probes selected by lasso algorithm, 205548_s_at and 204825_at (positive regulation of apoptotic process) and 213134_x_at (negative regulation of cell proliferation) was also selected in Natowicz et al. (2008b) and Hess et al. (2006), and the gene 201976_s_at is involved in lymphatic metastasis of breast cancer (Pandit et al., 2009).

In the second genomic data set, the gene $M94046$, selected by the multivariate model and by the lasso algorithm, is involved in regulation of transcription and drives tumor-specific expression in breast cancer cells (Wang et al., 2008). About the lasso algorithm result, the NM_000978 is a candidate oncogene in breast cancer (Clark et al., 2002), the gene NM_001930, involved in positive regulation of cell proliferation, is a component of the protein translation apparatus, over-expressed in tumors (Ramaswamy et al., 2002), the gene NM_001964, involved in regulation of transcription, is up-regulated in inflammatory breast cancer (Bièche et al., 2004), the gene NM_002848 is involved in tamoxifen sensitivity for breast cancer (Ramaswamy et al., 2009), the genes $AL117630$, NM_005243 and NM_003992 are over-expressed in breast cancer (Henry et al., 1993; Xie et al., 2002; Sloan et al., 2004) and the gene $AF055033$ is involved in regulation of cell growth in breast cancer (McCaig et al., 2002). In both data sets the majority of selected genes is involved with cancer or with biological function associated with the disease, showing coherency with the problem involved in the genomic data sets.

The main drawback of our proposal is the computing burden. Since the method relies on a Monte Carlo procedure to evaluate the inclusion of new variables in the model, we propose the use of a new parametric distribution and a fixed number of simulations to reduce the computing time. By doing so, we may include bias in the p-value estimates. In our study cases, we found some bias which did not compromise the main findings. Furthermore, results show that the lasso method, one of the fastest methods, is 16 times faster than our method. Nevertheless, our computing time can be reduced using parallel processing, i.e., the Monte Carlo simulations can be executed independently and simultaneously. For instance, if we run 100 Monte Carlo simulations using 8 core processors simultaneously the time can be reduced by a factor of 8. In this case, the lasso would be only 2 times faster than our procedure.

Not only the number of simulations, but the classification model and the programming language also affect the computing time. For instance, the logistic model was selected as the classification model due to its fitting algorithm, which is reasonably fast. Other classification models may require more computing time. Furthermore, due to the small sample size, a classification model based on a linear predictor seems appropriate, which is the case of the logistic model.

A major advantage of our approach is the fact that it selects a very small subset of relevant predictors. In addition, it does not require any validation set, or any cross-validation procedure. Thus, the full sample size is used to estimate the parameters of the model.

Acknowledgements

This research was funded by CAPES/COFECUB [grant number 591/08], FAPEMIG and CNPq.

References

Abrams, A. M., Kleinman, K., Kulldorff, M., 2010. Gumbel based p-value approximations for spatial scan statistics. International Journal of Health Geographics 9 (61).

Bach, F. R., 2008. Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th international conference on Machine learning. ACM, pp. 33–40.

Bièche, I., Lerebours, F., Tozlu, S., Espie, M., Marty, M., Lidereau, R., 2004. Molecular profiling of inflammatory breast cancer identification of a poor-prognosis gene expression signature. Clinical Cancer Research 10 (20), 6789–6795.

Cheng, P., Cheng, Y., Li, Y., Zhao, Z., Gao, H., Li, D., Li, H., Zhang, T., 2012. Comparison of the gene expression profiles between smokers with and without lung cancer using rna-seq. Asian Pacific Journal of Cancer Prevention 13, 3605–3609.

Clark, J., Edwards, S., John, M., Flohr, P., Gordon, T., Maillard, K., Giddings, I., Brown, C., Bagherzadeh, A., Campbell, C., et al., 2002. Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cdna clones. Genes, Chromosomes and Cancer 34 (1), 104–114.

Costa, M. A., Assunção, R. M., 2005. A fair comparison between the spatial scan and the besag–newell disease clustering tests. Environmental and Ecological Statistics 12 (3), 301–319.

Costa, M. A., Assunção, R. M., Kulldorff, M., 2012. Constrained spanning tree algorithms for irregularly-shaped spatial clustering. Computational Statistics & Data Analysis 56 (6), 1771–1783.

Dobson, A. J., 2008. An Introduction to Generalized Linear Models. Chapman and Hall.

Dudoit, S., Yang, Y., Callow, M., Speed, T., 2002. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Statistica sinica 12 (1), 111–140.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. The Annals of Mathematical Statistics 28, 181—-187.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al., 2004. Least angle regression. The Annals of statistics 32 (2), 407–499.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33 (1), 1.

Friedman, J. H., Hastie, T., Tibshirani, R., 2012. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (1), 1–22.

Haykin, S., 2008. Neural Networks and Learning Machines, 3rd Edition. Prentice Hall.

Henry, J. L., Coggin, D. L., King, C. R., 1993. High-level expression of the ribosomal protein l19 in human breast tumors that overexpress erbb-2. Cancer research 53 (6), 1403–1408.

Hess, K., Anderson, K., Symmans, W., Valero, V., Ibrahim, N., Mejia, J., Booser, D., Theriault, R., Buzdar, A., Dempsey, P., R., R., Sneige, N., Ross, J., Vidaurre, T., Gomez, H., Hortobagyi, G., Pusztai, L., 2006. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. Journal of Clinical Oncology 24 (26), 4236–4244.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Klebanov, L., Yakovlev, A., 2007. How high is the level of technical noise in microarray data. Biol Direct 2 (9).

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics: Theory and Methods 26, 1481–1496.

MacCullagh, P., Nelder, J. A., 1989. Generalized Linear Models. Monographs on Statistics and Applied Probability. Chapman and Hall.

Mazumder, R., Friedman, J. H., Hastie, T., 2011. Sparsenet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association 106 (495).

McCaig, C., Perks, C. M., Holly, J. M., 2002. Intrinsic actions of igfbp-3 and igfbp-5 on hs578t breast cancer epithelial cells: inhibition or accentuation of attachment and survival is dependent upon the presence of fibronectin. Journal of cell science 115 (22), 4293–4303.

Montgomery, D. C., Peck, E. A., Vining, G. G., 2012. Introduction to Linear Regression Analysis, 5th Edition. Wiley.

Natowicz, R., Braga, A., Incitti, R., Horta, E., Rouzier, R., Rodrigues, T., Costa, M., 2008a. A new method of dna probes selection and its use with multi-objective neural network for predicting the outcome of breast cancer preoperative chemotherapy. In: ESANN. pp. 71–76.

Natowicz, R., Incitti, R., Horta, E. G., Charles, B., Guinot, P., Yan, K., Coutant, C., Andre, F., Pusztai, L., Rouzier, R., 2008b. Prediction of the outcome of preoperative chemotherapy in breast cancer by dna probes that convey information on both complete and non complete responses. BMC Bioinformatics 9 (149).

Natowicz, R., Incitti, R., Rouzier, R., Cela, A., Braga, A. P., Horta, E. G., Rodrigues, T., Costa, M., 2008c. Downsizing Multigenic Predictors of the Response to Preoperative Chemotherapy in Breast Cancer. Lecture Notes in Computer Science. Springer, pp. 157–164.

Pandit, T. S., Kennette, W., MacKenzie, L., Zhang, G., Al-Katib, W., Andrews, J., Vantyghem, S. A., Ormond, D. G., Allan, A. L., Rodenhiser, D. I., et al., 2009. Lymphatic metastasis of breast cancer cells is associated with differential gene expression profiles that predict cancer stem cell-like properties and the ability to survive, establish and grow in a foreign environment. International journal of oncology 35 (2), 297.

Ramaswamy, B., Majumder, S., Roy, S., Ghoshal, K., Kutay, H., Datta, J., Younes, M., Shapiro, C. L., Motiwala, T., Jacob, S. T., 2009. Estrogen-mediated suppression of the gene encoding protein tyrosine phosphatase ptpro in human breast cancer: mechanism and role in tamoxifen sensitivity. Molecular Endocrinology 23 (2), 176–187.

Ramaswamy, S., Ross, K. N., Lander, E. S., Golub, T. R., 2002. A molecular signature of metastasis in primary solid tumors. Nature genetics 33 (1), 49–54.

Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1999. Akaike Information Criterion Statistics. Springer.

Siegmund, D., 1985. Sequential analysis: tests and confidence intervals. Springer.

Sloan, D. D., Nicholson, B., Urquidi, V., Goodison, S., 2004. Detection of differentially expressed genes in an isogenic breast metastasis model using rna arbitrarily primed-polymerase chain reaction coupled with array hybridization (rap-array). The American journal of pathology 164 (1), 315–323.

Tango, T., Takahashi, K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. International journal of health geographics 4 (1), 11.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H., Jan. 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415 (6871), 530–536.

Wald, A., 1973. Sequential analysis. Courier Corporation.

Wald, A., Wolfowitz, J., 1944. Statistical tests based on permutations of the observations. The Annals of Mathematical Statistics, 358–372.

Wang, X., Southard, R. C., Allred, C. D., Talbert, D. R., Wilson, M. E., Kilgore, M. W., 2008. Maz drives tumor-specific expression of ppar gamma 1 in breast cancer cells. Breast cancer research and treatment 111 (1), 103–111.

Xie, D., Jauch, A., Miller, C. W., Bartram, C. R., Koeffler, H. P., 2002. Discovery of over-expressed genes and genetic alterations in breast cancer cells using a combination of suppression subtractive hybridization, multiplex fish and comparative genomic hybridization. International journal of oncology 21 (3), 499–507.

Ypma, T. J., 1995. Historical development of the Newton-–Raphson method. Society for Industrial and Applied Mathematics 37, 531–551.

Zare, H., Haffari, G., Gupta, A., Brinkman, R. R., 2013. Scoring relevancy of features based on combinatorial analysis of lasso with application to lymphoma diagnosis. BMC genomics 14 (Suppl 1), S14.

Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 894–942.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2), 301–320.