Pesquisa Operacional (2015) 35(2): 401-421

Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope
doi: 10.1590/0101-7438.2015.035.02.0401

# SELECTING PROFILES OF IN DEBT CLIENTS OF A BRAZILIAN TELEPHONE COMPANY: NEW LASSO AND ADAPTIVE LASSO ALGORITHMS IN THE COX MODEL

Marcelo A. Costa[1*], Enrico A. Colosimo[2] and Carolina G. Miranda[2]

Received January 24, 2013 / Accepted November 16, 2014

**ABSTRACT.** Variable selection plays an important rule in identifying possible factors that could predict the behavior of clients with respect to the bill payments. The Cox model is the standard approach for modeling the time until starting the lack of payments. Parsimony and capacity of predicting are some desirable characteristics of statistical models. This paper aims at proposing a new forward stagewise Lasso (least absolute shrinkage and selection operator) algorithm and applying it for variable selection in the Cox model. The algorithm can be easily extended to run the Adaptive Lasso (ALasso) approach.

**Keywords**: proportional hazards model, partial likelihood, lasso regression.

## 1 INTRODUCTION

The cell phone market has grown fast in recent years. Modern cell phones, advantageous plans and services provided by the telephone companies have attracted new clients. Also due to the variety of companies with competitive services, the clients may change their telephone provider. One of the greatest challenges faced by telephone companies, which were generated by such a huge expansion in the market, is to identify the characteristics of their new and current clients. It is in the company's interest to attract faithful clients, or in a different perspective, to identify the profile of in debt clients. This is imperative in order to define new politics which leads the company to increase and to offer its clients more appropriate and specific services such as prepaid services.

In debt clients are either clients who start the service and do not pay their bill or who stop payments after a certain time. In both cases, the time until the lack of payments is the response

*Corresponding author.
[1]Department of Production Engineering, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil.
E-mail: azevedo@est.ufmg.br
[2]Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil.
E-mails: enricoc@est.ufmg.br; carolina.miranda@oi.net.br

variable. The goal is then to identify possible factors related to the client's socioeconomic conditions, the service provided by the company, among others that could predict the behavior of the clients with respect to bill payments. In this context, survival analysis and variable selection techniques are the common tools to handle the problem.

Survival analysis involves the study of response times from a well defined point in time to a pre-specified outcome event and, in general, the interest focuses on the effect of covariates on survival. Statistical analysis is often complicated by the presence of incomplete or censored observations in data sets. That is, frequently, the failure event is not observed for some subjects by the end of follow-up or it is not observed exactly but, in such cases, it is known that it occurred in a certain period of time. Literature in survival analysis has been increasing enormously after the seminal work of Cox [1, 2] on the proportional hazards model. Classical texts include Cox & Oakes [3], Kalbfleish & Prentice [4] and Lawless [5].

The most popular expression of Cox regression model, for covariates not dependent on time, uses the exponential form for the relative hazard, so that the hazard function is given by:

$$\lambda(t) = \lambda_0(t) \exp(\beta^T x), \tag{1}$$

where $\lambda_0(t)$, the baseline hazard function, is an unknown non-negative function of time, $\beta^T = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of unknown parameters and $x = (x_1, ..., x_p)^T$ is a row vector of covariates. Inference for $\beta$ is usually based on the partial log-likelihood function, taken into consideration a sample of $n$ individuals, in which it is observed $k \leq n$ failures at times $t_1 \leq t_2 \leq \cdots \leq t_k$. In the absence of ties, for the model (1), this function is written as:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left[ \beta^T x_i - log \left( \sum_{j \varepsilon R_{(t_i)}} \exp(\beta^T x_j) \right) \right], \tag{2}$$

where $\delta_i$ is the failure indicator and $R_{(t_i)}$ is the set of labels related to the individuals at risk at time $t_i$.

In real situations, it is very common that many covariates are collected by the researcher which makes the analysis very complex. Therefore, variable selection becomes a crucial part of the statistical analysis. Parsimony and capacity of predicting are some desirable characteristics of statistical models. There are many techniques for variable selection in linear regression models (see, for instance, Draper & Smith [6]), and non-linear regression models (see, for instance, Goncalves & Macrino [7]). Some of them can naturally be applied to the context of censored survival data, such as the well known stepwise procedure. Other techniques have been extended for this type of data. More recently, Bayesian approaches for variable selection have been the subject of substantial research. An interesting approach can be found in George & McCulloch [8, 9]. Some of them have been extended to the Cox model, such as Faraggi & Simon [10], Ibrahim et al. [11] among others. In the Bayesian approach, specifying meaningful prior distributions for the parameters is a difficult task especially in the presence of many of them. Another drawback of some Bayesian approaches for variable selection, such as Bayes factor, is the intensive computational requirement.

An alternative procedure that has shown a nice variable selection performance is the Lasso (Least Absolute Shrinkage and Selection Operator) method proposed by Tibshirani [12]. This technique has attracted attention in the literature mainly due to its capability as providing automatic variable selection and optimized prediction performance. Some properties of it are presented by Knight & Fu [13]. An estimate of the shrinkage parameter was proposed by Foster et al. [14] by using a random linear model approach. In special for the Cox model, Tibshirani [15] applied the Lasso method for variable selection. Recently, adaptive Lasso methods have been proposed by Zhang & Lu [16] and Zou [17] for the proportional hazards model. These methods consider adaptive penalizations for the regression coefficients.

Originally, the Lasso method requires a Newton-Raphson step in its iteractive process [15]. This is a really difficult step especially in the presence of many covariates. Recently, Friedman et al. [18] proposed a fast algorithm named *glmnet* to estimate the Lasso solutions based on the Coordinate Descent algorithm [19]. Simon et al. [20] extended the Coordinate Descent algorithm to the Cox model.

This paper aims at proposing a new Lasso algorithm for variable selection in Cox model as a non-linear extension of the *Forward stagewise linear regression* algorithm Tibshirani et al. [21]. The algorithm presents good performance for this model and it can be easily extended for other sorts of models. Furthermore, the proposed algorithm can be extended to generate ALasso (Adaptive LASSO) estimates. The ultimate goal is to identify characteristics of the in debt clients of a Brazilian telephone company.

The structure of the present paper is as follows. The Lasso method is described in Section 2. Section 3 presents the original algorithm proposed by Tibshirani [15]. The Forward Stagewise Lasso algorithm is presented in Section 4 and also an extended version which generates ALasso solutions. Monte Carlo simulations are used in Section 5 in order to compare Lasso, ALasso and other variable selection techniques. In Section 6, variable selection techniques are applied to the data set reported in Fleming and Harrington [22]. The Brazilian telephone company data set is analyzed in Section 7. Discussion and Conclusion in Section 8 ends the paper.

## 2 THE LASSO METHOD

The Lasso method was proposed by Tibshirani [12]. It is originally aimed at minimizing the constrained residual sum of squares (error) where the constraint is represented by the sum of the absolute coefficients function as shown in the following:

$$\begin{aligned} \hat{\beta}^{lasso} &= \arg\min_{\beta} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \\ \text{subject to} &: \sum_{j=1}^{p} |\beta_j| \leq s. \end{aligned} \tag{3}$$

where $y_i$ is the response for the $i$-th individual in the sample, which is usually assumed to be normally distributed, and $s$ is the Lasso parameter whose maximum value is associated to the standard least squares estimates $s_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{ls}|$. The method is very similar to the norm constraint approach, named *ridge regression* [23] but with the important difference that the Lasso

algorithm is also able to provide coefficients that are exactly zero, therefore providing variable selection. In addition, it minimizes multi-collinearity and consequently leads to the improvement of the model prediction performance.

The Lasso method was previously used for variable selection in the Cox model [15] to maximize the partial likelihood subjected to the same constraint function presented in (3), that is,

$$\begin{aligned} \hat{\beta}^{lasso} &= \arg\max_{\beta} \; l(\beta), \\ \text{subject to} \quad &: \quad \sum_{j=1}^{p} |\beta_j| \leq s. \end{aligned} \tag{4}$$

It can be noticed from (3) and (4) that the solution depends on the constraint value $s$ or Lasso parameter, which must be chosen based on some quality measures. Originally, quality measures such as cross-validation, generalized cross-validation and analytical unbiased estimate of risk were used for linear models [12] and an approximated generalized cross validation statistic for the Cox model [15]. In all these cases an approximation for the number of parameters involved in the fitted $\hat{\beta}^{lasso}$ is calculated using matrix algebra, which can be a hard task.

In addition, Leng et al. [24] reported that prediction accuracy criterion is not suitable for variable selection using the Lasso method. Results were provided for linear regression models only. Alternatively, we propose and evaluate the use of BIC (*Bayesian Information Criteria*) as the optimization function for selecting the Lasso parameter.

### 2.1 The adaptive LASSO (ALASSO)

Zhang & Lu [16] proposed a weighted $L_1$ penalty on the regression coefficients such that the estimates can be found by solving the following maximization problem,

$$\begin{aligned} \hat{\beta}^{alasso} &= \arg\max_{\beta} \; l(\beta), \\ \text{subject to} \quad &: \quad \sum_{j=1}^{p} |\beta_j|/|\hat{\beta}_j^*| \leq s. \end{aligned} \tag{5}$$

where $\hat{\beta}_j^*$ is the maximizer of the log partial likelihood, $l(\beta)$. According to Zhang & Lu [16], the ALasso approach has the desired theoretical properties and computational convenience. In this work, we found that our algorithm can be extended to generate ALasso solutions. It is also worth noting that by including a weighted penalty, the ALasso estimates exclude parameters with low absolute values, and it therefore generates more compact models when compared to the original Lasso estimates.

## 3 THE ORIGINAL LASSO ALGORITHM

Tibshirani [15] proposes an algorithm to obtain Lasso estimates as an adaptation of the Newton-Raphson iterative least squares algorithm. Let $\eta = X\beta$ be the linear predictor, where $X$ is the regression matrix, $\beta$ is the coefficients vector. Let $u = \partial l/\partial \eta$ and $A = -\partial^2 l/\partial \eta \eta^T$. Consider $W$ as $W = diag\{a_{ii}\}$ the diagonal matrix whose elements are the diagonal components of $A$ and assume $z = \eta + A^{-1}u$. The algorithm has the following four steps:

**Step 1** Fix $s$ and start with $\hat{\beta} = 0$.

**Step 2** Compute $\eta$, $u$, $A$ and $z$ from $\hat{\beta}$.

**Step 3** Minimize $(z - X\beta)^T W(z - X\beta)$ subject to $\sum_{j=1}^{p} |\beta_j| \leq s$.

**Step 4** Repeat Steps 2 and 3 until $\hat{\beta}$ does not change.

To perform Step 3, consider such minimization problem as a weighted least squares problem subjected to a general linear inequality constraint, say,

$$\text{Minimize } (z - X\beta)^T W(z - X\beta), \text{ subject to } C\beta \leq D,$$

for some matrices $C$ and $D$. This approximation leads to the following solution:

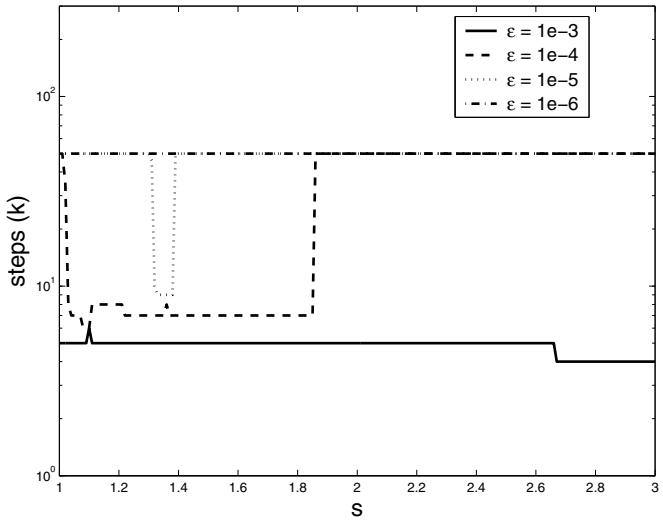
$$\widetilde{\beta} = \hat{\beta} + (X^T WX)^{-1} C^T \left[ C((X^T WX)^{-1}) C^T \right]^{-1} (D - C\hat{\beta}), \tag{6}$$

where $\hat{\beta} = (X^T WX)^{-1} X^T Wz$.

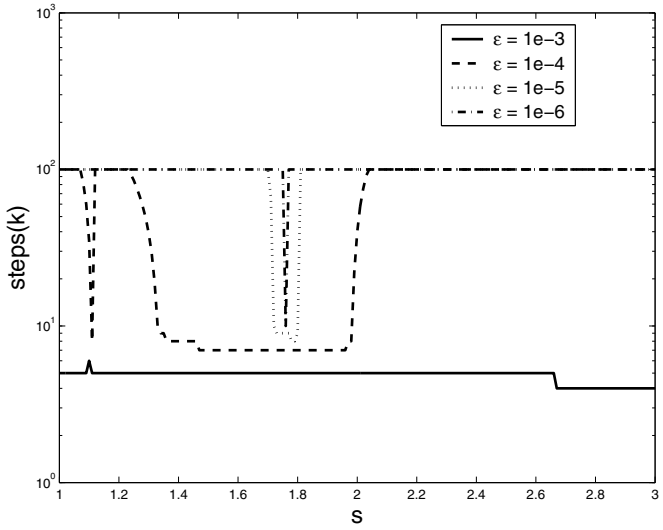
Instead of applying $2^p$ possible constraints representing all possible combinations of signals for the parameters $\beta$, Lawson & Hansen [25] suggest to grow matrices $C$ and $D$, sequentially, starting with $D = s$ and $C^T = sign(\hat{\beta}^0)$, where $sign()$ is the signal function, $\hat{\beta}^0$ is the non-constraint solution and from Equation 6 find a new estimate $\widetilde{\beta}$. If $\sum_{j=1} |\beta_j| > s$ include a new line $i$ in matrices $C$ and $D$ with $C_i^T = sign(\widetilde{\beta})$, $D = s\mathbf{1}$, $\mathbf{1} = (1, \ldots, 1)^T$, and obtain a new solution $\widetilde{\beta}_{k+1}$. The procedure stops at step $k$, $k = 1, 2, \ldots, m$, when $\sum_{j=1} |\beta_j| - s < \xi$, where $\xi$ is the numerical tolerance, and $m$ is the maximum number of steps. Both, $\xi$ and $m$, are specified by the user.

The tolerance $\xi$ and the maximum number of steps $m$ are used in both algorithms proposed by Tibshirani [15] and Lawson & Hansen [25]. These parameters are associated to the convergence speed and also to the estimates precision. It is worth noting that the overall algorithm first turns the Cox model estimation into a dynamic weighted linear problem and then it sequentially incorporates the constraint by means of an adaptive linear inequality. Therefore, the algorithm has two loops one for the Cox estimate and another to include the constraint. The tolerance and number of steps are selected in order to generate a single solution. The influence of the tolerance parameter is presented in Figure 1 for the *liver* data set used previously by Tibshirani [15]. This data was collected for the Mayo Clinic trial in PBC (Primary biliary cirrhosis) of the liver conducted between January 1974 and May 1984 comparing the drug D-penicillamine with a placebo. It has 17 variables and 276 observations. For the analysis four different values were chosen for the tolerance $\xi$, $\xi \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, and two different values for the maximum number of steps: $m = 50$ (1.a) and $m = 100$ (1.b), respectively. The maximum partial likelihood estimate is such that $\sum |\hat{\beta}_j^*| \approx 3.18$. We used the algorithm to generate 200 solutions homogenously distributed in the range $s \in [1, 3]$. According to Figure 1 when the tolerance is $10^{-3}$ the algorithm converges on average in 5 steps. If the tolerance is changed to $10^{-4}$, the algorithm stops mainly due to the maximum number of steps. The algorithm is also sensitive to the choice of $s$.

Results show that for a particular range of $s$ the algorithm reaches the tolerance before reaching the maximum number of steps. As the tolerance decreases, the algorithm stops when it reaches the maximum number of steps. Therefore, the algorithm is not capable of generating solutions if the tolerance is below a certain threshold, in this example this threshold is approximately $10^{-4}$. Figure 1(b) shows that even when selecting a higher number of steps ($m = 100$) the convergence problem remains.




(a) $m = 50$




(b) $m = 100$

**Figure 1** – Number of steps for different values of tolerance $\xi$ and constraint $s$ for maximum number of steps equal to $m = 50$ (a) and $m = 100$ (b).

Table 1 shows the average for both, the number of steps for different values of tolerance $\xi$ and the computational time for 200 solutions within the previous range of $s$ as shown in Figure 1(b). According to Table 1 a single solution with tolerance equal to $10^{-3}$ converges, on average, in 4.84 steps or 0.9 seconds. As expected, for small values of $\xi$ the mean number of steps tends to be close to the maximum number of steps $m$. This represents a lower bound for the tolerance not crossed by the algorithm.

**Table 1** – Mean computational time in seconds for original Lasso algorithm for different values of tolerance and maximum number of steps equals to 50.

| Tolerance ($\xi$) | Mean Steps ($m$) | Time (sec) |
|---|---|---|
| $10^{-3}$ | 4.84 | 0.90 |
| $10^{-4}$ | 32.2 | 5.24 |
| $10^{-5}$ | 48.56 | 6.38 |
| $10^{-6}$ | 50 | 6.62 |

Following Zhang & Lu [16], the ALasso optimization problem is strictly convex and therefore standard optimization packages available, for example, in Matlab and R can be used to generate the estimates. The authors also propose minimizing the penalized log partial likelihood using a variation of the Fu's shooting algorithm [26].

An alternative approach to solve both Lasso and ALasso optimization problems is to use Lagrange Multiplier and rewrite the constrained optimization problem as $\lambda \cdot l(\beta) + (1 - \lambda) \cdot \sum_{j=1}^{p} |\beta_j|$, for the Lasso estimate, and $\lambda \cdot l(\beta) + (1 - \lambda) \cdot \sum_{j=1}^{p} |\beta_j|/|\hat{\beta}_j^*|$ for the ALasso estimate. In this case, $0 \leq \lambda \leq 1$, and therefore estimates might be provided by standard optimization packages for different values of $\lambda$. Furthermore, in this case, the Lasso parameter, $s$, and the Lagrange Multiplier, $\lambda$, are related.

## 4 THE FORWARD STAGEWISE LASSO ALGORITHM

In order to avoid the problems mentioned in the previous section, we proposed a new algorithm that outperforms previous ones since it does not require computing first and second order derivatives, and relies on simple programming code based on a test-and-update approach. However, our approach increases the computational burden as compared to previous algorithms, which can be substantially minimized using a more efficient programming language such as C/C++. The algorithm is based on the *forward stagewise linear regression* [21]. The coefficients are initiated with zero and a small increment, $\epsilon$, is added to each one of them, separately. The log partial likelihood is then measured and the candidate that maximizes the log-function is updated. The process is continuously repeated until it reaches a maximum number of iterations, say $M$. The increment $\epsilon$ switches between positive and negative values in order to give flexibility to the coefficients' values. The proposed algorithm is presented in Figure 2.

**Algorithm**
**Initialize** $\hat{\beta}_j = 0$, $j = 1, ..., p$.
**Set** $\epsilon > 0$ **to some small constant and** $M$ **large**
**For** $m = 1$ **to** $M$
$lmax = l(\hat{\beta})$
**For** $j = 1$ **to** $p$
**For** $x = 1$ **to** 2
$\epsilon = -\epsilon$
$\beta^{aux} = \hat{\beta}$
$\beta_j^{aux} = \beta_j^{aux} + \epsilon$
**If** $l(\beta^{aux}) > lmax$ **then**
$k^* = j$
$\alpha^* = \beta_j^{aux}$
$lmax = l(\beta^{aux})$
**end If**
**end For**
**end For**
$\hat{\beta}_{k^*} = \alpha^*$
**end For**
**end Algorithm**

**Figure 2** – Forward Stagewise Lasso algorithm.

Despite increasing computational cost with such a test-and-update approach, it is remarkable the simplicity of the proposed algorithm and its capability of generating approximations of the Lasso solutions without the need of computing first or second order derivatives. Originally, the Lasso method was applied to the Cox model by means of an iterative Newton-Raphson update [15]. In that case, convergence is quite fast to a single solution, but multitudes of solutions have to be generated in order to select the final one with the proper constraint value for the sum of the absolute coefficients. Consequently, if a high resolution for the constraint is chosen, the computational effort might be greater than that for the proposed algorithm. Moreover, in our simulation study and for the real data fitting, the proposed algorithm did converge to the maximum log-likelihood estimate.

In a very similar way, our proposed Forward Stagewise Lasso algorithm has two parameters the step length $\epsilon$ and the maximum number of steps $M$. However, these parameters are related to the maximum amount of solutions $M$ generated by the algorithm and the difference between the sum of the absolute coefficients in two consecutive steps i.e. $\left|\sum|\hat{\beta}_{j(k)}| - \sum|\hat{\beta}_j(k-1)|\right| = \epsilon$. At each step, only one coefficient is updated adding to it a negative or positive value of $\epsilon$, as shown in Figure 3(b). The updated coefficient is the one that makes the resulting partial log likelihood $ł(\beta)$ at each step $k$ the highest one. The algorithm may stop before reaching $M$ steps if, in a particular step, the new solution does not generate a value for $ł(\beta)$ higher than the obtained in the previous step. This early stop event happens if the algorithm reaches a local maximum or

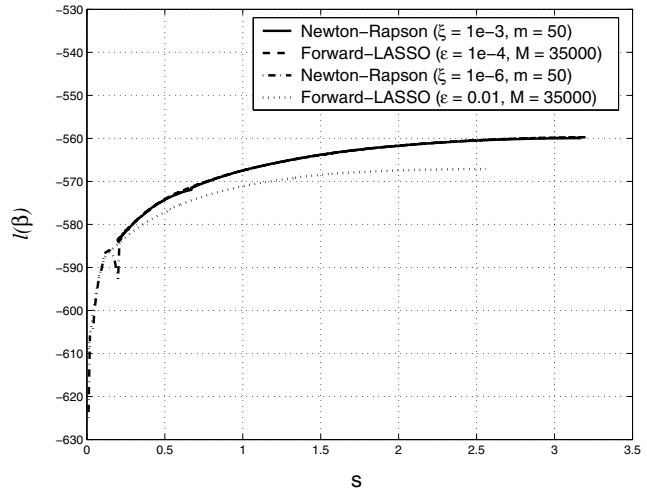
the global maximum. The Forward Stagewise Lasso parameters, $\epsilon$ and $m$, when compared to the original algorithm, do not control convergence. However, it is crucial to choose a suitable value for $\epsilon$ otherwise the set of solutions will be biased. Figure 3(a) compares the partial log-likelihood curve for original and the proposed algorithm. If $\epsilon = 0.01$ the proposed approach generates biased solutions and converges to a local maximum before reaching the maximum number of steps $M = 35,000$. If changed to $\epsilon = 10^{-4}$ the $l(\beta)$ curve achieves its maximum with $33,651$ steps. Indeed, Figure 3(b) shows that our proposal provides a fine resolution with slightly higher values for $l(\beta)$ when solutions are closer to the maximum partial likelihood estimate $\hat{\beta}_j^*$. In addition, our approach generates on average one single solution in 0.1675 seconds. The $33,651$ solutions took approximately 93.94 minutes using a code implemented in Matlab software. It is worth noting that computing speed can be substantially improved using C\C++ language. Although the smaller the $\epsilon$ the higher the computational cost required to generate $M$ solutions, empirical results have shown that after a certain small value $\epsilon^*$, the set of solutions becomes stable, not differing substantially if an even smaller $\epsilon$ value is applied. An empirical approach to select the parameter $\epsilon$ is to execute the algorithm several times, each one with a different $\epsilon$ value.

Comparing both algorithms, the original approach is very sensitive to the number of variables in the model. For a large number of variables such algorithm leads to numerical problems due to difficulties in the calculation of the inverse of matrices. Figure 4(a) shows the estimates for the *liver* data set. For small values of $s$ ($s < 0.3$), the algorithm is not able to generate solutions. Problems with the numerical convergence appear as spikes and non-smooth sequential estimates. However, this algorithm has the advantage of generating fast solutions for arbitrary values of $s$. The proposed approach does not present problems with the numerical convergence since no inverse matrices are needed. The estimates are very smooth as noticed in Figure 4(b). However, although it generates one solution much faster than the original approach, we need to estimate a large set of solutions with some relatively fine resolution starting from $\hat{\beta}_j = 0$ which increases computational cost considerably.

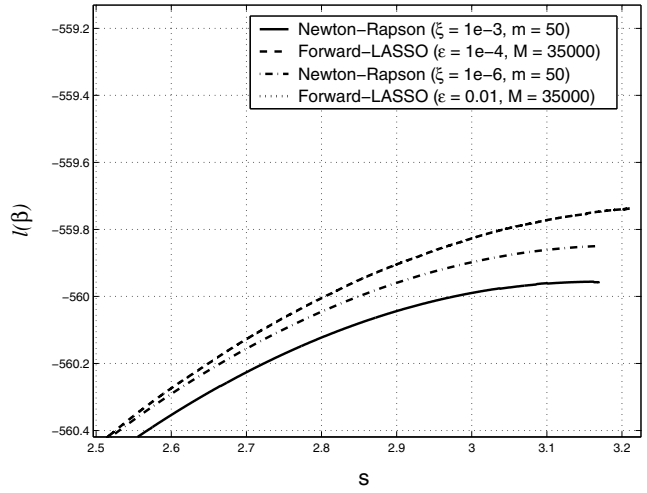
Furthermore, our approach can be extended to generate ALasso solutions. Briefly, the ALasso approach penalizes parameters with low absolute values and it prioritizes the parameters with large absolute values. In our algorithm this feature is implemented by replacing the parameters update step, defined as $\beta_j^{aux} = \beta_j^{aux} + \epsilon$, with $\beta_j^{aux} = \beta_j^{aux} + \epsilon \cdot |\hat{\beta}_j^*|$. Figure 5 shows ALasso solutions generated with an optimization package available in R software, as suggested by Zhang & Lu [16], and also the solutions provided by our algorithm. Results show that our approach minimizes instabilities due to numerical approximations.

## 5 COMPARING LASSO, GLMNET, ALASSO AND OTHER METHODS FOR VARIABLE SELECTION

In this section a Monte Carlo study is performed in order to compare the Lasso, ALasso methods and some standard procedures for variable selection, such as stepwise, AIC and BIC. All methods are applied for selecting variables in the Cox model. We also evaluate the *glmnet* package [20] which applies the Coordinate Descent algorithm to estimate Lasso solutions.



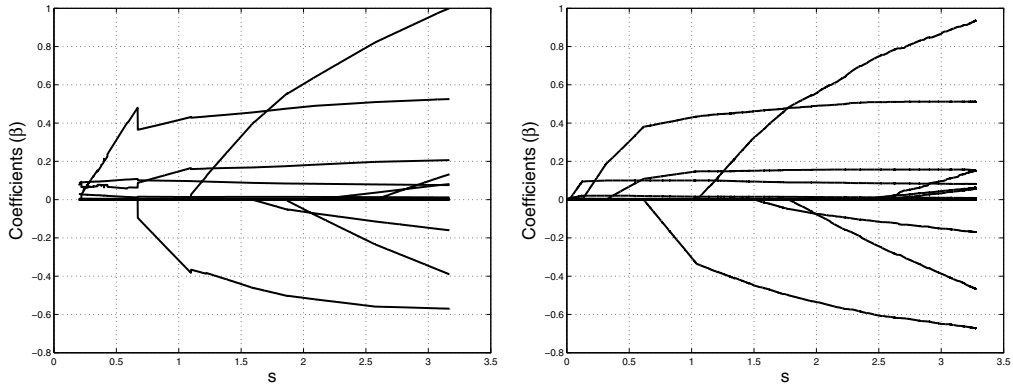

(a) original Lasso




(b) Forward-Lasso

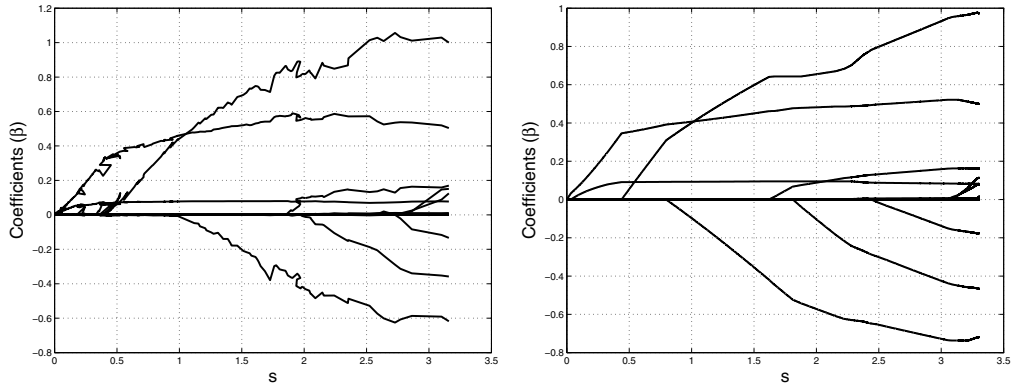
**Figure 3** – Partial log-likelihood as a function of $s$ for original (Newton-Raphson) and Forward-Lasso algorithms.

The BIC quality measure was used in this section as twofold: (1) in the Lasso-Cox approach, this measure is estimated for each solution generated by the proposed algorithm and the Lasso parameter is selected as the one that provides the highest value for BIC, and (2) as a variable selection technique. BIC is computed through the formula $BIC = 2l(\beta) - p \log n$. The expression for AIC is: $AIC = 2l(\beta) - 2p$. There are many stepwise procedures. We selected the one that is presented by Collett [27]. It consists in a procedure that removes covariates in one step and then tests again the ones that were excluded in the previous steps. This process goes on until no covariates enter or leave the model.




(a) original Lasso (b) Forward-Lasso

**Figure 4** – Estimates for original Lasso and Forward Lasso algorithms. Numerical approximations to the inverse of matrices leads to convergence problems in the original approach.




(a) estimated ALasso solutions (b) Forward-ALasso

**Figure 5** – Estimates for ALasso and Forward ALasso algorithms. Alasso coefficients were generated using a numerical optimization package available in the R software.

The AIC and BIC measures were applied as variable selection techniques as the following. Given $p$ candidate variables, all $2^p - 1$ possible combinations were tested. The combinations with the highest AIC and BIC values were compared to the true model.

For the *glmnet* package, we apply the *leave-one-out* cross-validation to select the Lasso parameter.

Four scenarios are considered in the study. They are built taking into account different percentages of censoring observations (0% and 30%), sample sizes ($n = 50$ and 100) and the numbers of covariates (4 and 5). The response variable is generated from the Weibull model [28] with the shape parameter $\alpha(x) = \exp\{\sum_{l=1}^{p} x_l \beta_l\}$ and scale parameter $\rho > 0$. Three different values

for $\rho$ are assumed: $\rho = 0.5$ (decreasing failure rate), $\rho = 1$ (constant failure rate) and $\rho = 2$ (increasing failure rate). Vector $\beta$ is set equal to $(1, 0, 0, 1)$ or $(1, 0, 0, 1, 1)$ for scenarios which consider 4 or 5 covariates, respectively. Covariates are generated from the following distributions: the standard normal (N), the Bernoulli (B) with parameter $\theta = 0.5$, and the exponential (E) with mean equal to 1. The scenarios are described in Table 2.

**Table 2** – Scenarios for the Monte Carlo study.

| Scenario | Sample size | Percentage of censoring | Response variables |
|---|---|---|---|
| 1 | 50 | 0.0 | (N, N, B, B) |
| 2 | 100 | 0.0 | (N, N, B, B) |
| 3 | 100 | 30.0 | (N, N, B, B) |
| 4 | 50 | 0.0 | (N, N, B, B, E) |

For each scenario, 5,000 replications generated using the Weibull model are considered and it is calculated the percentage of times that the model is correctly selected by each method and criteria. For the Lasso and ALasso methods, it is also assumed $\epsilon = 0.001$ and $M = 40,000$. MATLAB and R packages were used in order to get the results.

Tables 3 and 4 show the results for Scenarios 1 and 2. From these tables, there may be observed the effect of sample sizes as well as the failure rates in the variable selection, for all methods.

**Table 3** – Percentage of correct model selection, Scenario 1.

| Model | Method | Percentage |
|---|---|---|
| $Weibull(\alpha(x), 0.5)$ | Lasso | 40.72 |
| | ALasso | 72.24 |
| | glmnet | 35.74 |
| | stepwise | 64.62 |
| | AIC | 64.84 |
| | BIC | 79.48 |
| $Weibull(\alpha(x), 1.0)$ | Lasso | 40.94 |
| | ALasso | 73.22 |
| | glmnet | 34.28 |
| | stepwise | 66.14 |
| | AIC | 66.40 |
| | BIC | 79.48 |
| $Weibull(\alpha(x), 2.0)$ | Lasso | 41.80 |
| | ALasso | 73.04 |
| | glmnet | 35.62 |
| | stepwise | 66.42 |
| | AIC | 66.64 |
| | BIC | 79.26 |

**Table 4** – Percentage of correct model selection, Scenario 2.

| Model | Method | Percentage |
|---|---|---|
| *Weibull*($\alpha(x)$, 0.5) | Lasso | 52.92 |
| | ALasso | 87.68 |
| | glmnet | 36.68 |
| | stepwise | 69.24 |
| | AIC | 69.28 |
| | BIC | 92.16 |
| *Weibull*($\alpha(x)$, 1.0) | Lasso | 52.86 |
| | ALasso | 88.84 |
| | glmnet | 36.04 |
| | stepwise | 69.74 |
| | AIC | 69.84 |
| | BIC | 92.68 |
| *Weibull*($\alpha(x)$, 2.0) | Lasso | 52.66 |
| | ALasso | 88.72 |
| | glmnet | 36.26 |
| | stepwise | 69.72 |
| | AIC | 69.76 |
| | BIC | 93.08 |

Remarkably, the BIC approach presents the best performance, followed by the ALasso method. The *glmnet* approach provides the worst results for all scenarios, followed by the standard Lasso. It is worth noticing that *glmnet* applies *leave-one-out* cross validation to select the Lasso parameter whereas the Lasso approach applies the BIC statistic. As expected, for samples of size $n = 100$ and all methods and criteria, the frequency in which the model is selected correctly is higher than for smaller sample sizes. All methods tend to identify the correct model more frequently for decreasing failure rate scenarios when compared with increasing failure rate scenarios, for both sample sizes. However, it can also be observed that all methods present better performance when the data set is generated from the Weibull distribution with constant failure rate (which corresponds to the exponential distribution).

It is well known that the presence of censoring observations in the data set may affect the estimates of failure rates. The effect of different percentages of censoring observations in variable selection can be observed from Tables 4 and 5. Similar to what was observed for scenarios 1 and 2, the BIC method presents the best performance, followed by the ALasso. However, it is observed that the frequency by which the model is correctly identified is smaller for scenarios including censoring observations. The presence of censoring observations also affects the performance of BIC and ALasso – contrary to what was observed in scenarios with censoring observations (see Table 4). On the other hand, AIC did achieve some minor improvement in the presence of censoring observations.

**Table 5** – Percentage of correct model selection, Scenario 3.

| Model | Method | Percentage |
|---|---|---|
| $Weibull(\alpha(x), 1.0)$ | Lasso | 52.80 |
| | ALasso | 84.90 |
| | glmnet | 52.16 |
| | stepwise | 69.62 |
| | AIC | 75.40 |
| | BIC | 90.00 |

From Table 6 it can be observed that the percentage by which the true model is correctly identified is higher for scenarios that presented larger numbers of covariates (see also Table 3) for all methods. The BIC method is again the best procedure, followed again by ALasso. The AIC, in this scenario, does not present good results, and Lasso is even worse, followed by *glmnet*.

**Table 6** – Percentage of correct model selection, Scenario 4.

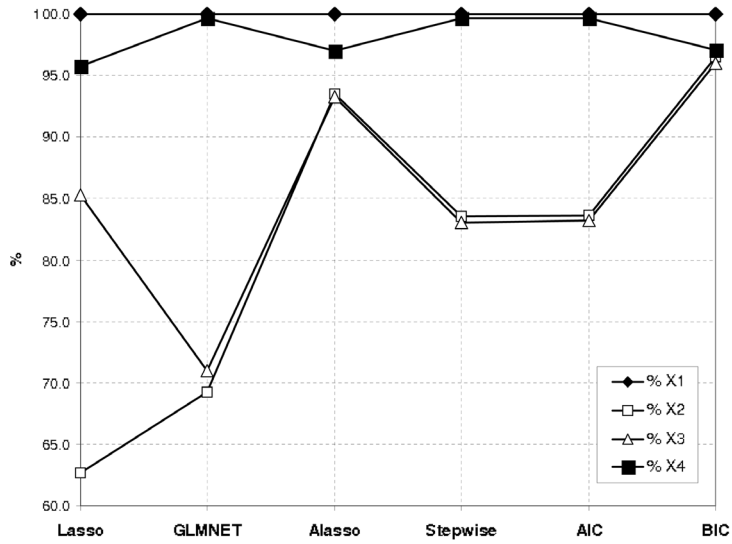
| Model | Method | Percentage |
|---|---|---|
| $Weibull(\alpha(x), 1.0)$ | Lasso | 45.04 |
| | ALasso | 86.94 |
| | glmnet | 41.80 |
| | stepwise | 69.86 |
| | AIC | 69.94 |
| | BIC | 92.68 |

Regarding the performance of Lasso method for selecting each covariate, Figure 6 shows that for scenarios 3 and 4 the Lasso and *glmnet* methods frequently select the true covariates. Nevertheless, it usually includes some additional covariates which compromises its capacity to detect only the true covariates. BIC and ALasso performances are very close but BIC provides the best results.

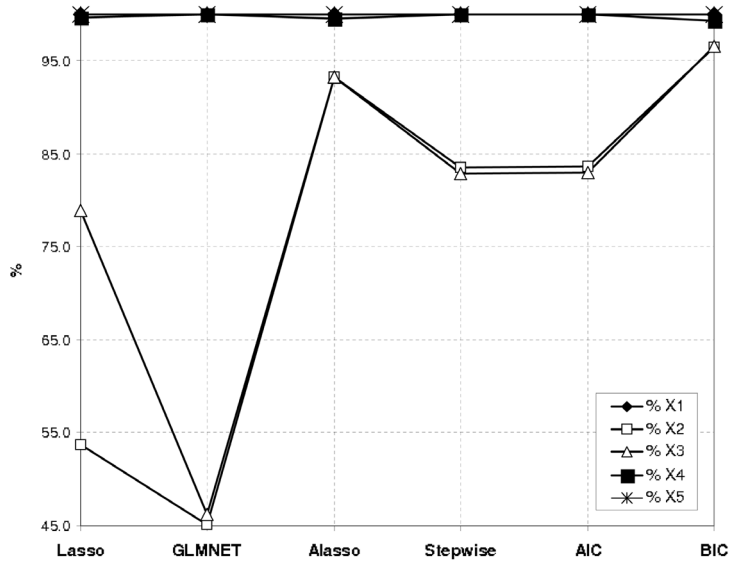
In order to illustrate the use of the proposed Lasso and ALasso algorithms, and the other variable selection methods, discussed previously, they are presented below to analyze the data set reported in Fleming and Harrington [22].

## 6 EXAMPLE: THE PBC DATA SET

The methods described in previous sections are applied to analyze the data set reported in Fleming and Harrington [22] which correspond to data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of liver. Biliary cirrhosis is a fatal and rare disease whose origin is unknown. The trial was conducted between 1974 and 1984. A total of 424 PBC patients met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. Missing data items were excluded from the data set providing a sample of total size 276. The response variable $T$ is




(a) Scenario 3




(b) Scenario 4

**Figure 6** – Percentage of correct covariate selection for scenarios 3 and 4.

the number of days between registration and the earliest death by cirrhosis of liver. Data related to patients submitted to liver transplantation or that died by other causes except cirrhosis were considered as censored observations. The covariates considered in the study are:

$X_1$: Treatment code, 1 = D-penicillamine, 2 = placebo;

$X_2$: Age in years;

$X_3$: Sex, 0 = male 1=female;
$X_4$: Presence of ascites;
$X_5$: Presence of hepatomegaly;
$X_6$: Presence of spiders;
$X_7$: Presence of edema;
$X_8$: Serum bilirubin, in mg/dl;
$X_9$: Serum Cholesterol, in mg/dl;
$X_{10}$:Albumin, in gm/dl;
$X_{11}$: Urine cupper, in $\mu$g/day;
$X_{12}$: Alkaline phosphatase, in U/liter;
$X_{13}$: SGOT, in U/ml;
$X_{14}$: Triglycerides, in mg/dl;
$X_{15}$: Platelet count;
$X_{16}$: Prothombin time, in seconds;
$X_{17}$: Histologic stage of disease, graded 1, 2, 3 or 4.

The most parsimonious models selected to describe the behavior of $T$ provided by each procedure are given in Table 7.

**Table 7** – Selected models for each method – PBC data set.

| Method | Model |
|---|---|
| Lasso | $X_2, X_7, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$ |
| ALasso | $X_7, X_8, X_{10}, X_{11}, X_{17}$ |
| glmnet | $X_4, X_7, X_8, X_{10}, X_{11}, X_{13}, X_{16}, X_{17}$ |
| Stepwise | $X_3, X_7, X_8, X_{10}, X_{11}, X_{13}, X_{16}, X_{17}$ |
| AIC | $X_3, X_7, X_8, X_{10}, X_{11}, X_{13}, X_{16}, X_{17}$ |
| BIC | $X_7, X_8, X_{10}, X_{11}, X_{17}$ |

Notice from Table 7 that all procedures indicate that variables $X_7$, $X_8$, $X_{10}$, $X_{11}$ and $X_{17}$ should be in the model. Comparing the models obtained using the ALasso and BIC procedures, they are exactly the same. However, the computing cost related to the BIC approach is 3.3 times higher than the ALasso approach. Since there are 17 variables, the BIC approach tested 131,071 possible combinations. The ALasso algorithm used a maximum number of steps equal to 40,000.

Table 8 shows estimates of the coefficients and standard deviations in parentheses. Standard deviations were estimated by fitting the regular Cox model with the selected covariates, as suggested by Tibshirani [12, 15].

In general, the BIC and ALasso methods selected covariates with larger absolute values, except for the $X_{11}$ covariate which has a smaller absolute value. However, results show that this variable is statistically significant as can be seen by its standard deviation. Therefore, for this data set, the BIC and ALasso methods did select statistically significant covariates regarding the absolute value of their estimates.

**Table 8** – Coefficients' estimates and standard deviations for PBC data set.

| Variables | Full Model | AIC, stepwise | LASSO | ALASSO, BIC | GLMNET |
|---|---|---|---|---|---|
| $X_1$ | –0.18 (0.20) | — | — | — | — |
| $X_2$ | 0.01 (0.01) | — | 0.02 (0.01) | — | — |
| $X_3$ | –0.46 (0.29) | –0.46 (0.26) | — | — | — |
| $X_4$ | 0.11 (0.37) | — | — | — | 0.08 (0.34) |
| $X_5$ | 0.08 (0.23) | — | — | — | — |
| $X_6$ | 0.01 (0.23) | — | — | — | — |
| $X_7$ | 0.97 (0.37) | 0.78 (0.34) | 0.89 (0.36) | 0.81 (0.32) | 0.67 (0.33) |
| $X_8$ | 0.08 (0.02) | 0.09 (0.02) | 0.07 (0.02) | 0.10 (0.02) | 0.08 (0.02) |
| $X_9$ | 4e-4 (4e-4) | — | 5e-4 (4e-4) | — | — |
| $X_{10}$ | –0.72 (0.29) | –0.74 (0.25) | –0.63 (0.26) | –0.70 (0.25) | –0.65 (0.27) |
| $X_{11}$ | 3e-3 (1e-3) | 3e-3 (1e-3) | 4e-3 (1e-3) | 4e-3 (1e-3) | 4e-3 (1e-3) |
| $X_{12}$ | -3e-5 (4e-5) | — | — | — | — |
| $X_{13}$ | 3e-3 (2e-3) | 3e-3 (2e-3) | 3e-3 (2e-3) | — | 3e-3 (2e-3) |
| $X_{14}$ | –1e-3 (1e-3) | — | –6e-4 (1e-3) | — | — |
| $X_{16}$ | 0.16 (0.11) | 0.17 (0.10) | 0.16 (0.10) | — | 0.17 (0.10) |
| $X_{17}$ | 0.50 (0.16) | 0.51 (0.13) | 0.51 (0.14) | 0.52 (0.13) | 0.50 (0.14) |

## 7 CASE STUDY: A BRAZILIAN TELEPHONE COMPANY DATA SET

A certain Brazilian telephone company used to have a small number of clients owing their bills. However, after a huge expansion of the company, the number of unpaid bills grew very fast leading it to change its policies for accepting a new client. In order to establish rules that are flexible enough to still allow the company to grow and also strict enough to avoid a large number of unpaid bills, a study was carried out to identify, among the variables related to the clients characteristics, the ones that could explain their tendency to be in debt with the company. A sample of 530 clients was randomly selected from the company data base. The response variable is the number of days between registration in the data base and its first unpaid bill. The percentage of failures is 60%. The covariates considered in the study are:

$X_1$: Client local calls restriction (1 yes; 0 no);
$X_2$: Client budget restriction (1 yes; 0 no);
$X_3$: Client general calls restriction (1 yes; 0 no);
$X_4$: Number of payment parcels made to the company;
$X_5$: Client has more than one unpaid bill in the past (1 yes; 0 no);
$X_6$: Client has automatic debt in his bank account (1 yes; 0 no);
$X_7$: Client has not paid the first bill (1 yes; 0 no);
$X_8$: Client has more products from the company (1 yes; 0 no).

We initially investigate using the Kaplan-Meier curve the characteristics of the data set with respect to potential cure fractions (clients that will never stop paying the bills in the due date), as shown in Figure 7. Looking at the Kaplan-Meier curve, this hypothesis cannot be confirmed, i.e.,

since the curve reaches zero which means that all clients in data base eventually stopped paying their bills. Thus, since the response variable is the number of days between registration in the data base and the first unpaid bill, eventually, most of the customers will pay at least one bill after due date. This is the most likely situation because only 12.3% of the clients in the database had automatic debit in banking account, known as an unpopular choice among Brazilians which do prefer mail bills.

In addition, since the date at which the customer enters the database might be different from the due date of the bill, and different customers can choose different due dates, it is natural to use the number of days as the measure of time for the Cox model. By using a daily basis unit as the measure of time and a follow-up around 10,000 days the standard Cox-model can be used in a continuous scale.

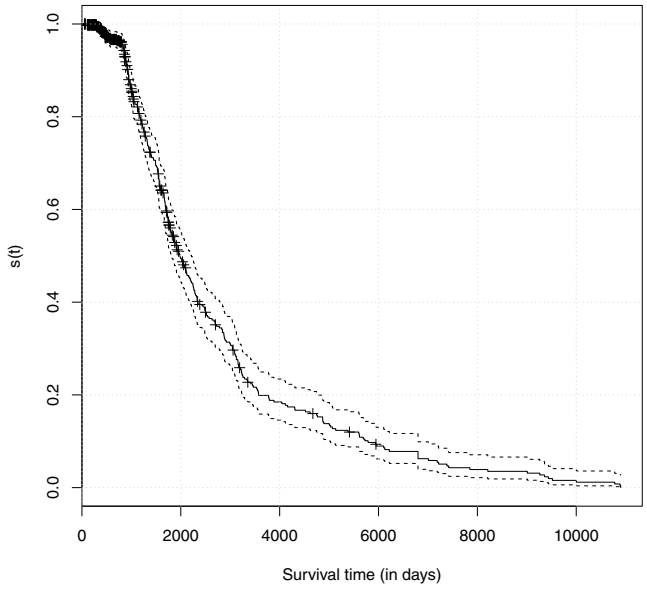



**Figure 7** – Kaplan-Meier curve for the analysis of the Brazilian Telephone Company data set.

Models selected by the proposed methods are presented in Table 9. Results show that ALasso, stepwise, AIC and BIC selected exactly the same model with three covariates: $X_1$, $X_5$ and $X_6$. The Lasso solution included the covariate $X_4$ which is not statistically significant, as shown in Table 10. The *glmnet* solution has the greatest number of variables.

The previous simulation study showed that the BIC achieved the best results but has a high computational cost as compared to our proposed methods. The ALasso approach achieved the second best results and presents a smaller computational cost. In our case study, the final model is the one estimated with BIC, ALasso and AIC. This result is consistent with the simulation study. It can be concluded based on the final model, that the hazard rates of unpaid bills for clients with local call restrictions is 90% higher ($e^{0.64} - 1$) as compared to clients with no local call restrictions. Clients with previous unpaid bills have hazard rates of future unpaid bills of 80%

higher ($e^{0.57}$) as compared to clients with no previous unpaid bills. Finally, clients with payments automatically debited from their bank accounts have hazard rates of 42% smaller ($1 - e^{-0.54}$) as compared to clients with no automatic debits.

**Table 9** – Selected models for each method – Telephone data set.

| Method | Model |
|---|---|
| Lasso | $X_1, X_4, X_5, X_6$ |
| ALasso | $X_1, X_5, X_6$ |
| glmnet | $X_1, X_3, X_5, X_6, X_7, X_8$ |
| Stepwise | $X_1, X_5, X_6$ |
| AIC | $X_1, X_5, X_6$ |
| BIC | $X_1, X_5, X_6$ |

**Table 10** – Coefficients' estimates and standard deviations for Telephone data set.

| Variables | Full Model | LASSO | GLMNET | others |
|---|---|---|---|---|
| $X_1$ | 0.62 (0.18) | 0.63 (0.18) | 0.61 (0.18) | 0.64 (0.17) |
| $X_2$ | 0.10 (0.28) | — | — | — |
| $X_3$ | 0.14 (0.35) | — | 0.09 (0.32) | — |
| $X_4$ | –0.03 (0.10) | 0.01 (0.08) | — | — |
| $X_5$ | 0.54 (0.23) | 0.56 (0.23) | 0.51 (0.22) | 0.57 (0.21) |
| $X_6$ | –0.52 (0.17) | –0.53 (0.17) | –0.52 (0.17) | –0.54 (0.17) |
| $X_7$ | 0.52 (0.42) | — | 0.49 (0.42) | — |
| $X_8$ | –0.10 (0.21) | — | -0.10 (0.21) | — |

## 8 DISCUSSION AND CONCLUSION

In this paper, new implementations of the Lasso and ALasso methods for variables selection in the Cox model were proposed to address the problem of selecting the profile of in debt clients of a Brazilian Telephone Company. Their performances were compared with some standard techniques for variable selection, such as stepwise, BIC and AIC, through a Monte Carlo study.

From the Monte Carlo study it could be concluded that the BIC method achieved the best performance followed by the ALasso. Lasso, in general, presented the worst performance. All methods had their performances affected by the sample sizes, the presence of censoring observations as well as by the characteristic of failure rate function.

For the Brazilian Telephone Company data set, the ALasso, stepwise, AIC and BIC selected the same model with three covariates. Since BIC and ALasso methods were pointed out as the best ones in the simulation study, the company should use the model as the reference to define future policies.

The proposed Forward-Lasso and Forward-ALasso algorithms have two parameters, the increment $\epsilon$ and maximum number of steps $M$. An empirical approach to select these parameters is to choose a maximum number of steps in such a way that the algorithm will stop before reaching this value and performing the algorithm for more than one value for $\epsilon$, usually $10^{-3}$ and $10^{-4}$.

Although the BIC method achieved the best results in the simulation study, it tests all possible combinations of the covariates. Consequenly, as the number of covariates increases the computing time gets very high. Therefore, the BIC method is unsuitable when the number of covariates is large. When compare to the BIC method, the ALasso approach is computationally affordable and provides similar results than BIC. The simulation study suggest that the difference between BIC and ALasso is that the latter usually includes some additional covariates. This analysis suggests that combing BIC and ALasso methods would provide better results than just applying BIC or ALasso alone. Therefore, we suggest the following approach: first apply ALasso method to select the initial set of covariates and then run BIC algorithm using the reduced number of covariates selected by the ALasso method.

Discussions highlight that the proposed and original Lasso and ALasso algorithms have distinct features as their main advantages. The original algorithms have the ability to generate arbitrary solutions and our approach is insensitive to numerical problems. Further work aims at combining both approaches by first reaching a reasonable estimator using the original approach and then providing a fine grid of solutions using the proposed approach, taking the best of each algorithm. This is an interesting topic for future research.

## ACKNOWLEDGEMENTS

The authors thank CNPq, CAPES and FAPEMIG for financial support.

## REFERENCES

[1] COX DR. 1972. Regression Models for Life Tables (with discussion). *Journal of the Royal Statistical Society, B*, **34**: 187–220.

[2] COX DR. 1975. Partial Likelihood. *Biometrika*, **62**: 269–276.

[3] COX DR & OAKES D. 1984. *Analysis of Survival Data*, London, Chapman and Hall.

[4] KALBFLEISH JD & PRENTICE RL. 2002. *The Statistical Analysis of Failure Time Data*, New York, Wiley.

[5] LAWLESS JF. 2002. *Statistical Models and Methods for Lifetime Data*, New York, Wiley.

[6] DRAPER NR & SMITH H. 1998. *Applied Regression Analysis*, New York, Wiley.

[7] GONCALVES LB & MACRINI LR. 2011. *Pesquisa Operacional*, **31**(3): 499–519.

[8] GEORGE EI & MCCULLOCH RE. 1995. Two Approaches to Bayesian Model Selection with Applications. *In Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner* (Edited by BERRY D, CHALONER K & GEWEKE J): 339–348, New York, Wiley.

[9] GEORGE EI & MCCULLOCH RE. 1997. Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**: 339–373.

[10] FARAGGI D & SIMON R. 1998. Bayesian Variable Selection Method for Censored Survival Data. *Biometrics*, **54**: 1475–1485.

[11] IBRAHIM JG, CHEN MH & MACEACHERN SN. 1999. Bayesian Variable Selection for Proportional Hazards Models. *The Canadian Journal of Statistics*, **27**: 701–717.

[12] TIBSHIRANI RJ. 1996. Regression Shrinkage via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**(1): 267–288.

[13] KNIGHT K & FU W. 2000. Asymptotics for lasso-type estimators. *Annals of Statistics*, **28**: 1356–1378.

[14] FOSTER SD, VERBYLA AP & PITCHFORD WS. 2008. A random model approach for the Lasso. *Computational Statistics*, **23**: 217–233.

[15] TIBSHIRANI RJ. 1996. The Lasso method for variable selection in the COX model. *Statistics in Medicine*, **16**: 385–395.

[16] ZHANG HH & LU W. 2007. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94**: 691–703.

[17] ZOU H. 2008. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, **95**: 241–247.

[18] FRIEDMAN J, HASTIE T & TIBSHIRANI R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1).

[19] FRIEDMAN J, HASTIE T, HOEFLING H & TIBSHIRANI R. 2007. Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, **2**(1): 302–332.

[20] SIMON N, FRIEDMAN J, HASTIE T & TIBSHIRANI R. 2011. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, **39**(5).

[21] TIBSHIRANI RJ, FRIEDMAN JH & HASTIE TJ. 2003. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

[22] FLEMING TR & HARRINGTON DP. 1991. *Counting Processes and Survival Analysis*, New York, Wiley.

[23] MARQUARDT DW & SNEE RD. 1975. Ridge Regression in Practice. *The American Statistician*, **29**(1): 3–20.

[24] LENG C, LIN Y & WAHBA G. 2004. A Note on the Lasso and Related Procedures in Model Selection. *Technical Report no. 1091r*.

[25] LAWSON C & HANSEN R. 1974. *Solving Least Squares Problems*, Englewood Cliffs: prentice Hall.

[26] FU W. 1998. Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**: 397–416.

[27] COLLETT D. 2003. *Modelling Survival Data in Medical Research*, London, Chapman and Hall.

[28] BENDER R, AUGUSTIN T & BLETTNER M. 2005. Generating survival times to simulate Cox proportional harzards models *Statistics in Medicine*, **24**: 1713–1723.