A Variable Selection Method for High-Dimensional Survival Data

Francesco Giordano, Sara Milito(✉), and Marialuisa Restaino

University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Salerno, Italy
{giordano,smilito,mlrestaino}@unisa.it

Abstract. Survival data with high-dimensional predictors are regularly collected in many studies. Models with a very large number of covariates are both infeasible to fit and likely to incur low predictability due to overfitting. The selection of significant variables plays a crucial role in estimating models. Even if several approaches that identify variables in presence of censored data are available in literature, there is not unanimous consensus on which method outperforms the others. Nonetheless, it is possible to exploit the advantages of methods to get the final set of covariates as good as possible. Therefore, we propose a method that combines different variable selection procedures by using the subsampling technique, for identifying as relevant those covariates that are selected most frequently by the different variable selectors on subsampled data. By a simulation study, we evaluate the performance of the proposed procedure and compare it with other techniques.

Keywords: Variable selection · High-dimension · Survival data

1 Introduction

In recent years the classical problem of variable selection has enjoyed increased attention thanks to a massive growth of high-dimensional data available in many scientific disciplines. In modern statistical applications, the number of variables often exceeds the number of observations. In such contexts, the true model is often assumed to be sparse, meaning that only a small fraction of the variables are actually related to the response. Therefore, the selection of the relevant variables is of fundamental importance in the analysis of high-dimensional data.

Survival analysis deals with the expected time until one or more events occur. It is frequently used in the field of economics, where the event of interest is the failure of companies (mainly due to bankruptcy) or the reasons for which customers choose to stop their relationship with company. In regression analysis of survival data, the Cox Proportional Hazard model, proposed by Cox in 1982 [2], is the most used to explore the relationship between subjects' survival and some explanatory variables.

Like linear regression models, traditional variable selection methods such as subset selection, forward selection, backward elimination, and a combination

M. Corazza et al. (Eds.): MAF 2022, *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pp. 303–308, 2022.
https://doi.org/10.1007/978-3-030-99638-3_49

of both are among the most common applied for choosing the set of relevant variables under survival framework. However, these methods have computational difficulties in presence of high-dimensional data. Therefore, other methods have been proposed to overcome this problem. Lasso, firstly proposed for linear regression models [5], is then extended to the Cox model [6]. Subsequently, some authors have developed some penalized shrinkage techniques such as SCAD introduced by [3] specifically for Cox models. On one hand, the above methods of variable selection have been shown to be successful in theoretical properties and numerous experiments. On the other hand, their performance is highly dependent on the correct choice of the tuning parameter and these approaches can be unstable, especially in the high-dimensional data setting.

Among the problems encountered in identifying relevant variables, the choice of the best selector from those available is the most relevant. Unfortunately, the set of covariates selected by one method may be different from that selected by another. Even if it might be seen as a disadvantage, analysing the differences and similarities among the various methods can provide useful information. For example, a covariate chosen from all methods can be considered as actually relevant, while ones selected only by one method cannot be related to the response. In order to take into account this insight, following the idea of [7] for linear model, we propose a method called Combined Variable Selector with Subsample (CVSS) that combines different variable selection procedures by using the subsampling technique. We record the percentage of times a covariate is selected among the procedures and we get the final set by identifying as relevant those covariates that are selected most frequently. The main difference between our procedure and [7] consists in the choice of the tuning parameter in the various methods used. In fact, while in [7] for each method the authors take into consideration some vectors of covariates selected by different penalty coefficients, we consider only one vector of betas referring to the best tuning parameter. Thus, we extract only one set of variables for each approach with the advantage that the procedure becomes very fast.

The paper is organized as follows. In Sect. 2, we introduce our proposed approach. In Sect. 3, we show the simulation results. We conclude this work with a discussion in Sect. 4.

2 The Proposed Procedure

Suppose there are n observations $\{(y_i, \mathbf{x}_i, \delta_i)\}_{i=1}^n$ of survival data. For an individual i, y_i denotes its survival time and $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ represents the observed data for the p covariates. At the same time, $\delta_i \in \{0, 1\}$ is a variable indicator of censorship, where $\delta_i = 0$ means that y_i is right-censored. We assume also that the censoring mechanism is non-informative and independent of the event process. Let $h(t)$ be the hazard rate at a time t; the generic form of the Cox proportional hazards model can be expressed as

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta})$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ denotes a p-dimensional vector of unknown regression coefficients and $h_0(t)$ is the baseline hazard function, that is the hazard function at time t when all the covariates take value zero. In general, $\boldsymbol{\beta}$ can be estimated by maximizing the partial likelihood function [2].

In order to identify the set of true relevant variables, it is possible to use a penalized variable selection method among those proposed in the last years. For example, the Lasso is able to select the non-zero components in setting with large p, it is computationally efficient and it uses an L_1 type penalty, while the SCAD is a regularized regression methods with non-convex penalties and it is designed to reduce estimation bias. Although in the literature there are several approaches for selecting variables in presence of censored data, there is not unanimous consensus on which method outperforms the others. Then, how to select a method remains an open question. Since choosing a method rather than another influences the selection of relevant variables, it is very important to identify the best variable selection method for the data under analysis.

In order to solve this open question, we propose to implement different variable selection methods on the sampled data and to check similarities between different variable selectors. Combining the models with subsampling is used to improve the variable selection performance of a single variable selection method. For example, RBVS proposed by [1] uses subsampling to identify the set of highly-ranked covariates, while Stability Selection proposed by [4] repeatedly samples observations and fits the sampling data using a variable selection method (e.g. the Lasso). It therefore keeps covariates with a selection frequency above a certain threshold.

Similarly to the methods above, our proposal fits variable selection methods to the subsampled data and it identifies as non-zero components those covariates appearing most frequently. Unlike these other approaches, however, our procedure uses various variable selection methods. In fact, we observe that no method outperforms all other methods in all settings, since different variable selection methods optimize different objective functions. In the case of regularized regression, the difference among methods is usually in terms of the penalty. If a covariate is selected by the majority of methods, it means that the covariate is chosen to minimize many various objective functions. We expect that a true covariate should frequently be chosen regardless of the objective function used. We repeat the fitting on subsampled data to incorporate the variability in selection due to the variability in the data.

The variable selection procedure proposed can be summarized as follows. First, we consider mutually exclusive subsets $I_{b1}, \ldots, I_{br}$ of size m, drawn uniformly from $\{1, \ldots, n\}$ without replacement, where $r = \lfloor n/m \rfloor$, $b = 1, \ldots, B$ and $B \in \mathbb{N}$ is the number of replicates. Assume that the sets of subsamples are independently drawn for each b. Second, we fit different variable selection methods on the sets $I_{b1}, \ldots, I_{br}$ and we collect the estimated model in $\mathcal{M}$, where $|\mathcal{M}| = r \times B \times k$ and k is the number of variable selector used. For each subset and for each procedure, we obtain a vector of $\hat{\beta}$. Third, we measure the relative frequency of times the jth covariate is selected given by

$$\hat{\tau}_j = \frac{1}{|\mathcal{M}|}\left(\sum_{M_i \in \mathcal{M}} I_{(\hat{\beta}_j^{M_i} \neq 0)}\right)$$

where $\hat{\beta}_j^{M_i}$ is the estimated coefficient of the jth covariate on the fitted model $M_i \in \mathcal{M}$, and I_x is the indicator function. Fourth, we identify as relevant those variable such that

$$\hat{S} = \{j : \hat{\tau}_j \geq q\}$$

where q is a fixed threshold. For the practical use, the number of replicates B should be large enough to stabilize the value of j and at the same time, it should be small enough to not increase the computational time. Following [1], we set $r = 2$ and $B = 50$, so we obtain 100 sets each with $n/2$ number of observations. In this paper, we set $q = 1/2$, which means that covariates with $\hat{\tau}_j \geq 1/2$ are selected.

The choice of the different methods to be used within our procedure is based on the following considerations. Each method must have good variable selection performance and it is required some variability among methods. In this article, we choose Lasso, MCP, SCAD, Elastic Net and Ridge since they optimize different objective functions, as they use various penalty terms. Furthermore, such methods are also computationally feasible in high-dimensional setting.

3 Simulation Study

We compare the variable selection performance among different methods by the number of false positive (FP), the number of false negative (FN), the total number of variable selection error (FN+FP) and the size of selected set. For comparison, we also consider other variable selector methods applied on the whole dataset: the Lasso, the Elastic net, the Ridge regression, the SCAD and the MCP.

In our simulation study we generate survival times $t_i, i = 1, 2, \ldots, n$, as exponential distributions with subject-specific parameters $h_i = h_0(t_i)\exp(\beta^T X_i)$, baseline $h_0(t_i) = 1$ and $\beta = (2_5, 0_{p-5})$. Thus the true size of model is $s = 5$. The variables $X_1, \ldots, X_p$ are sampled from a multivariate normal density $N(0, \Sigma)$ where the entries of Σ are fixed to $corr(X_j, X_k) = \rho^{|j-k|}$ with $\rho \in \{0, 0.3, 0.6\}$. The percentage of censorship c is setting to 20% or 40%. We set $n = 150$ and $p = \{100, 200\}$. The results are shown in Table 1.

In all scenarios our procedure has the best performance in terms of both total error FP+FN and FP. When $p = 100$ the highest value of FP for CVSS is 1.46, this means that at most 1.46 of the variables identified as relevant are not related to the response. MCP procedure is the only selector for which in Setting 3 the FN is not equal to zero: the final set contains in this case variables that are not relevant in the model. Looking at the size, our procedure selects a number of covariates that is very close to the real size 5. As we expected, the

Table 1. Simulation results for different combination of ρ, c and p. A dark grey cell represents the best results, while a grey one represents the worst. Standard errors are shown in the parentheses

Parameters	Methods	$p = 100$				$p = 200$			
		FP+FN	FP	FN	Size	FP+FN	FP	FN	Size
Setting 1	LASSO	12.91 (13.52)	12.91	0	16.92 (13.56)	13.26 (18.68)	13.26	0	17.27 (18.73)
$\rho = 0$	Elastic Net	10.22 (10.59)	10.22	0	14.32 (10.62)	11.57 (16.64)	11.57	0	15.50 (16.70)
$c = 20\%$	Ridge	10.13 (12.46)	10.13	0	14.14 (12.50)	13.28 (17.91)	13.28	0	17.29 (17.96)
	SCAD	2.54 (2.74)	2.54	0	6.55 (5.86)	3.01 (3.51)	3.01	0	7.02 (3.56)
	MCP	1.85 (1.67)	1.85	0	5.86 (1.71)	1.63 (1.40)	1.63	0	5.64 (1.45)
	CVSS	1.14 (0.38)	1.14	0	5.14 (0.38)	1.09 (0.29)	1.09	0	5.09 (0.29)
Setting 2	LASSO	11.32 (15.39)	11.32	0	15.33 (15.43)	12.89 (15.92)	12.89	0	16.90 (15.96)
$\rho = 0.3$	Elastic Net	9.9 (12.35)	9.90	0	13.91 (12.40)	9.38 (11.25)	9.38	0	13.39 (11.29)
$c = 20\%$	Ridge	12.54 (13.84)	12.54	0	16.55 (13.87)	11.26 (12.88)	11.26	0	15.27 (12.91)
	SCAD	2.16 (2.49)	2.16	0	6.17 (2.54)	2.53 (3.23)	2.53	0	6.54 (3.28)
	MCP	1.58 (1.40)	1.58	0	5.60 (1.50)	1.36 (0.80)	1.36	0	5.38 (0.87)
	CVSS	1.04 (0.20)	1.04	0	5.04 (0.20)	1.02 (0.14)	1.02	0	5.02 (0.14)
Setting 3	LASSO	9.84 (14.17)	9.84	0	13.85 (14.22)	11.61 (14.54)	11.61	0	15.62 (14.57)
$\rho = 0.6$	Elastic Net	8.17 (10.08)	8.17	0	12.18 (10.12)	12.11 (13.71)	12.11	0	16.12 (13.75)
$c = 20\%$	Ridge	12.97 (17.27)	12.97	0	16.98 (17.32)	13.49 (15.97)	13.49	0	17.50 (16.00)
	SCAD	2.57 (2.52)	2.57	0	6.58 (2.56)	2.50 (2.67)	2.50	0	6.51 (2.71)
	MCP	1.66 (1.27)	1.64	0.02	5.63 (1.32)	1.51 (1.01)	1.49	0.02	5.48 (1.07)
	CVSS	1.03 (0.17)	1.17	0	5.03 (0.17)	1.01 (0.10)	1.01	0	5.01 (0.10)
Setting 4	LASSO	13.05 (12.13)	13.05	0	17.06 (12.18)	18.22 (18.82)	18.22	0	22.23 (18.86)
$\rho = 0$	Elastic Net	13.66 (12.34)	13.66	0	17.67 (12.38)	16.19 (17.63)	16.19	0	20.20 (17.68)
$c = 40\%$	Ridge	12.67 (12.52)	12.67	0	16.68 (12.56)	14.81 (17.37)	14.81	0	18.83 (17.45)
	SCAD	2.61 (2.82)	2.61	0	6.62 (2.87)	3.34 (3.25)	3.34	0	7.35 (3.29)
	MCP	1.51 (1.27)	1.51	0	5.52 (1.32)	1.64 (1.48)	1.64	0	5.66 (1,56)
	CVSS	1.46 (0.70)	1.46	0	5.46 (0.70)	1.25 (0.48)	1.25	0	5.25 (0.48)
Setting 5	LASSO	10.76 (11.52)	10.76	0	14.77 (11.55)	13.46 (17.42)	13.46	0	17.47 (17.46))
$\rho = 0.3$	Elastic Net	11.59 (12.95)	11.59	0	15.60 (12.98)	13.80 (20.75)	13.80	0	17.81 (20.80)
$c = 40\%$	Ridge	9.36 (9.39)	9.36	0	13.37 (9.43)	13.19 (19.45)	13.19	0	17.20 (19.49)
	SCAD	2.17 (1.99)	2.17	0	6.18 (2.04)	2.38 (2.64)	2.38	0	6.40 (2.72)
	MCP	1.45 (0.99)	1.45	0	5.46 (1.03)	1.45 (0.99)	1.45	0	5.47 (1.08)
	CVSS	1.17 (0.45)	1.17	0	5.17 (0.45)	1.12 (0.38)	1.12	0	5.12 (0.38)
Setting 6	LASSO	14.02 (16.99)	14.02	0	18.03 (16.93)	15.98 (19.02)	15.98	0	19.99 (19.05
$\rho = 0.6$	Elastic Net	12.19 (13.33)	12.19	0	16.20 (13.32)	18.03 (21.05)	18.03	0	22.04 (21.09)
$c = 40\%$	Ridge	11.32 (13.43)	11.32	0	15.33 (13.51)	14.65 (16.69)	14.65	0	18.66 (16.73)
	SCAD	2.46 (2.03)	2.46	0	6.48 (2.10)	3.33 (2.86)	3.33	0	7.35 (2.92)
	MCP	1.54 (1.09)	1.54	0	5.55 (1.31)	1.68 (1.41)	1.68	0	5.69 (1.45)
	CVSS	1.07 (0.26)	1.07	0	5.07 (0.26)	1.02 (0.14)	1.02	0	5.02 (0.14)

procedures with highest FP (the Lasso, the Elastic Net and the Ridge) are also the procedures that select a higher number of covariates compared to s. In fact, as the total error increases, also the size increases. While the other approaches suffer when the correlation increases, CVSS, Lasso and Elastic Net give better results in terms of selection performances. On the other hand, the increase of censoring percentage worsens the selection for all the methods.

When $p = 200$, our procedure is still the best one. If we compare the total error for two values of p, it is possible to notice that FP+FN is lower when $p =$

200. This characteristic is not shared with the competitors. Other approaches, such as Lasso and Ridge, suffer the increase of the number of variables in the dataset. The size of CVSS is the closest to the true size $s = 5$ in all scenarios and the best performance is related at high correlation value.

4 Conclusion

In this work we proposed a new method to choose the relevant covariates with high-dimensional survival data. Although survival analysis was initially used to study death as a specific event in medical studies, these statistical techniques have increasingly been used in economics and social sciences. Given the relevance of the topic, it is important to be able to find a method that selects the relevant variables related to the response variable as good as possible. In particular, we proposed to combine several variable selectors available in literature with the subsample technique. Simulation study has shown that our approach works better than its competitors. For future work we will evaluate this approach from a theoretical point of view and apply it to real data.

References

1. Baranowski, R., Chen, Y., Fryzlewicz, P.: Ranking-based variable selection for high-dimensional data. Stat. Sin. **30**(3), 1485–1516 (2020)
2. Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**(2), 187–202 (1972)
3. Fan, J., Li, R.: Variable selection for Cox's proportional hazards model and frailty model. Ann. Stat. **30**(1), 74–99 (2002)
4. Meinshausen, N., Bühlmann, P.: Stability selection. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **72**(4), 417–473 (2010)
5. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
6. Tibshirani, R.: The lasso method for variable selection in the Cox model. Stat. Med. **16**(4), 385–395 (1997)
7. Yuen, C., Fryzlewicz, P.: Exploiting disagreement between high-dimensional variable selectors for uncertainty visualization. J. Comput. Graph. Stat. 1–9 (2021). https://doi.org/10.1080/10618600.2021.2000421