Outlier Detection in Survival Analysis based on the Concordance C-index

João Diogo Pinto[1], Alexandra M. Carvalho[1,2] and Susana Vinga[3]
[1]*PIA, Instituto de Telecomunicações, Lisboa, Portugal*
[2]*DEEC, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal*
[3]*LAETA, IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal*

Keywords: Survival Analysis, Outlier Detection, Robust Regression, Cox Proportional Hazards, Concordance c-index.

Abstract: Outlier detection is an important task in many data-mining applications. In this paper, we present two parametric outlier detection methods for survival data. Both methods propose to perform outlier detection in a multivariate setting, using the Cox regression as the model and the concordance c-index as a measure of goodness of fit. The first method is a single-step procedure that presents a delete-1 statistic based on bootstrap hypothesis, testing for the increase in the concordance c-index. The second method is based on a sequential procedure that maximizes the c-index of the model using a a greedy one-step-ahead search. Finally, we use both methods to perform robust estimation for the Cox regression, removing from the regression a fraction of the data by their measure of outlyingness. Our preliminary results on three different datasets have shown to improve the estimation of the Cox Regression coefficients and also the model predictive ability.

1 INTRODUCTION

Survival analysis is the field that studies time-to-event data and has become a relevant topic in clinical and medical research. Usually there are three main goals when performing survival analysis (David G. Kleinbaum, Mitchel Klein, 2005): 1) to estimate survival/hazard functions from the data; 2) to compare survival/hazard functions between groups of patients; and 3) to assess the impact of explanatory variables on patients survival time. Goals 1) and 2) are dealt by recurring to non-parametric methods like Kaplan-Meier and Nelson-Aalen estimators in order to estimate survival curves. Log-rank tests are commonly used to compare survival curves. All these methods have good robustness to the presence of outlying observations. When modeling the data in relation to explanatory variables, the most popular method is the Cox proportional hazards (Cox, 1972). The robustness of the Cox regression has shown to be rather weak,with outlying observations severely affecting the Cox regression coefficients. Concerning robustness, one important concept is the breakdown point (Donoho and Huber, 1983; Hampel, 1971), which represents the fraction of corrupt observations needed to arbitrarily offset the estimation values. It has been pointed out that Cox partial likelihood estimator has a breakdown point of $\frac{1}{n}$ (Kalbfleisch and Prentice, 2011), this means that when fitting a Cox regression to n data points, one single outlier observation is enough to cause the estimator to take values arbitrarily far from their true value (Rousseeuw and Leroy, 1987).

Goal 3) will be the focus of this study, in particular, our goal is to improve the Cox regression estimation by identifying and removing outlying observations. This way the regression becomes more robust, thus providing more accurate relationships between explanatory variables and survival times, along with improving the global model predictive ability.

2 OUTLIERS IN SURVIVAL DATA

To fix notation, a dataset will be denoted by $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ with each X_i being a p-dimensional vector of covariates and Y_i the corresponding dependent variable value. In survival data, is very common the occurrence of censoring, i.e., the event of interest does not always occur for a given individual during the period of the study. To model censoring, it is common to add a binary variable which indicates if the event occurred or not.

There are many definitions of an outlier in the literature, both mathematical and more informal, as can be seen more thoroughly in (Ben-Gal, 2005). For example (Hawkins, 1980) defines an outlier *as an obser-*

Pinto J., Carvalho A. and Vinga S..
Outlier Detection in Survival Analysis based on the Concordance C-index.
DOI: 10.5220/0005225300750082
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 75-82
ISBN: 978-989-758-070-3

vation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism or (Johnson et al., 1992) that defines an outlier as *an observation in a dataset which appears to be inconsistent with the remainder of that set of data*. These definitions provide two different ways of detecting outliers: the first one considers only the values of X_i and Y_i, the second. assesses the relation between them by introducting the notion of the model's quality of fit. Of course the second notion of outlyingness needs a model to define this relationship between Y_i and X_i. The first perspective corresponds to a non-parametric approach to outlier detection, the second corresponds to a parametric or model-based perspective and will be the focus of the our proposal.

2.1 Swamping and Masking

Data sets with multiple outliers or clusters of outliers are subject to *masking* and *swamping* effects. Here we enunciate the following definitions (Acuna and Rodriguez, 2004):

Masking Effect. One outlier masks another outlier if the second outlier can be considered an outlier only by itself but not in the presence of the first outlier.

Swamping Effect. One outlier observation swamps a second observation if the latter can be considered as an outlier in presence of the first but not by itself.

As seen in (Fischler and Bolles, 1981), these effects are particularly harmful when developing sequential procedures for outlier detection, mainly because the subset of observations already deleted influences which observations will be deleted in the subsequent iterations.

2.2 Model-specific Outlier Detection: Cox Proportional Hazards

In this paper the model chosen to represent the data was the Cox proportional hazards due to its simplicity, good results and great power of interpretability.

Several works have been developed to increase the robustness of the estimation of the Cox Regression by performing outlier detection, for example through residual analysis, estimating the variation in regression parameters with the removal of a given observation (Therneau et al., 1990). The outliers can then be detected by selecting the observations that cause the largest variation in the parameters upon its removal. This approach is susceptible to masking and swamping and also needs the tuning of the outlier or non-outlier threshold.

In (Farcomeni and Viviani, 2011) outlying observations are defined as the individuals that have the smallest contributions to the Cox partial likelihood. In order to find these observations they first make a robust fitting of the Cox regression and then in the absence of masking, they employ residual analysis as in (Nardi and Schemper, 1999) to perform outlier detection. The robust fitting is done using an algorithm that maximizes the maximum partial likelihood. This maximization is made over all possible subsets of the trimmed set of observations.

2.3 Concordance C-index

To assess the predictive ability of a survival model, we will use Harrel's concordance c-index (Harrell et al., 1982). It measures the ability of the model to predict a higher relative risk to an individual whose event occurs first. The relative risk is estimated from the output of the model for each individual; in a Cox model for instance, the relative risk corresponds to the hazard ratio. The c-index is calculated using the following procedure:

1. Form all possible pairs of individuals.
2. Omit the pairs whose shorter survival time is censored and all pairs where both observations are censored. These are the permissible pairs, being $N_{permissible}$ its cardinality.
3. To calculate *Concordance*, for each permissible pair when $T_i \neq T_j$: count 1 if the shorter survival time has higher predicted risk, count 0.5 otherwise. For $T_i = T_j$ and both not censored: count 1 if the predicted risks are the same, 0.5 otherwise; if at least one is censored and it corresponds to a lower risk, count 1 (0.5, otherwise). *Concordance* is defined as the sum of all counts for each permissible pair.
4. The c-index is given by
   c-index = $Concordance/N_{permissible}$.

The c-index is a rank measure, thus it only measures how well predicted values are concordant with rank-ordered response variables. For example, the c-index for two patients with predicted hazard ratios of 0.4 and 0.6 is the same as if the patients had hazard ratios of 0.1 and 0.9 (Harrell, 2001), it only measures if the outcome is concordant with the response variables or not. Thus, unlike measures such as the sum of squared errors, one observation by itself has a limited contribution for the overall concordance. This robustness may allow for the maximization of the c-index

without worrying if it is being maximized at the cost of the majority of the data, only to fit better one or a cluster of outlying observations, as it can happen with the sum of squared errors (Fischler and Bolles, 1981).

3 METHODS FOR OUTLIER DETECTION

We propose two novel methods for outlier detection in survival data based on the concordance index, described in sections 3.1 and 3.2. Section 3.3 describes alternative proposals that will be further used for comparison purposes.

The proposed methods make use of an operational definition of outlier, defined as an observation that, when absent from the data, will likely decrease the prediction error of the fitted model. In a survival setting, this prediction error will be measured recurring to the concordance c-index, which has the particularity of using the predictive model as a black-box.

3.1 Bootstrap Hypothesis Testing (BHT)

Ideally we would know the underlying distribution of the observations X_i, Y_i and perform an hypothesis test about the difference in terms of concordance between the two distributions. Thus the idea is to perform n hypothesis tests about the concordance variation, one for each observation i, and sorting the resulting p-values.

The hypothesis tests will be made following the bootstrap approach (Efron, 1979). Each observation X_i, Y_i is considered a discrete random variable having a distribution equal to the empirical distribution given by the original dataset. We will consider n different empirical distributions, each distribution results from removing each observation i from the original data and adjust densities in order to sum one. Denoting by C the concordance c-index and $C_{original}$ the concordance in the original data, distributions $Data_i$ represent the adjusted empirical distributions having $P(X = X_i, Y = Y_i) = 0$. The hypothesis test for each observation is formulated as follows:

$$H_0 : C_{Model,(X,Y)\sim Data_i} \leq C_{original}$$
$$H_1 : C_{Model,(X,Y)\sim Data_i} > C_{original}$$

Writing $C_{Model,(X,Y)\sim Data_i}$ and $\delta C_i = C_i - C_{original}$ it is more useful to reformulate the hypothesis tests as:

$$H_0 : \delta C_i \leq 0$$
$$H_1 : \delta C_i > 0$$

The rejection of the null hypothesis given a significance level α corresponds to estimate a confidence interval for the values of δC for each distribution $Data_i$, if this interval does not contain values less or equal than zero we can reject the null hypothesis for the significance level α, alternatively we can calculate the test p-value.

These confidence intervals will be computed using Monte Carlo Bootstrap as explained in (Harrell, 2001), for each observation i the procedure is the following: 1) produce B bootstrap samples by sampling with replacement $n-1$ observations from the empirical distribution $Data_i$; 2) compute the concordance for each bootstrap sample; 3) the p-value corresponds to the proportion of bootstrap samples having $C_i - C_{original} \leq 0$.

The number of bootstrap samples B used has shown to be dependent on the number of individuals and number of covariates. In our tests the value for B was iteratively increased until p-values convergence.

Following the same reasoning provided in (Singh and Xie, 2003), given an outlying observation ξ the probability that a bootstrap sample does not contain ξ is approximately $(1-\frac{1}{n})^n \approx \frac{1}{e} (\approx 37\%)$ as $n \to \infty$. Thus, each observation will be absent in approximately 37% of the samples. A low p-value for the hypothesis test mentioned above, means that the given observation i improves the concordance c-index in a systematic way not depending on the cooperation of any other observation. On the other hand, if one outlier is masked by another, the masking outlier will not be present in approximately 37% of the bootstrap samples and thus we can expect a multimodal behavior for the expected δC. Thus an outlier subject to masking may not systematically improve concordance (present a high p-value for the hypothesis test) but if presents multimodality and one of the modes is relatively high, it is a candidate for an outlier.

To sum up, Bootstrap Hypothesis testing (BHT) on δC works as follows: for each observation, an hypothesis test by bootstrap is done. The resulting statistics for each observation will be a p-value and the expected value of δC. The p-value gives us the confidence level to reject the hypothesis that the removal of the observation causes no increase in the c-index. Experimentally we verified that these two values are correlated. When the p-value is low, the expected δC is usually very high, the opposite relation has shown to be weaker. So in order to obtain a 1-dimensional metric for outlyingness, we consider the observations with the lowest p-values the more outlying ones.

3.2 One-Step Deletion (OSD)

This method is a sequential procedure for outlier removal. We start with all data and at each iteration of the algorithm, the observation that, when ex-

cluded, causes the largest increase in concordance, is removed. The resulting subset is interpreted as containing the most outlying observations. This method is equivalent to do one-step-ahead greedy search for maximizing the c-index of the model in the data. The resulting subset of observations, will be considered the most outlying ones.

3.3 Alternative Methods

Here we present alternative methods for outlier detection in survival data that will be used to assess the performance of the proposed methods.

3.3.1 Martingale Residuals (MART)

These residuals are provenient from the counting process framework for censored survival, first a Martingale process is defined by the difference between observed and expected number of events (David W. Hosmer, Stanley Lemeshow, Susanne May, 2008). Let $N(t)$ be the number of events until t and $H(t)$ the cumulative hazard function, we have for each individual i the *Martingale* residual process:

$$M_i(t) = N_i(t) - H_i(t). \quad (1)$$

The martingale residual is defined as the value of process $M_i(t)$ at the time of failure/censoring, as $N(t)$ takes 1 if the event is observed and zero when censored (David Collett, 2003), their are given by:

$$r_{M_i} = \delta_i - H_i(t), \quad (2)$$

where δ_i is the censoring indicator for individual i. For the Cox model the residuals are given by:

$$r_{M_i} = \delta_i - \exp\{\beta X\} H_0(t). \quad (3)$$

3.3.2 Deviance Residuals (DEV)

The deviance residuals are an attempt (David Collett, 2003) to adjust the *Martingale* residuals to be more centered around zero, given by:

$$r_{D_i} = \mathrm{sgn}(r_{M_i})[-2\{r_{M_i} + \delta_i \log(\delta_i - r_{M_i})\}]^{\frac{1}{2}}. \quad (4)$$

3.3.3 Likelihood Displacement Statistic (LD)

Let $\hat{\beta}$ be the value of β that maximizes the partial Cox likelihood and $\hat{\beta}_{(-i)}$ the estimate when observation i is eliminated from the fitting. The likelihood displacement (Cook, 1977) statistic (LD) is given by:

$$LD_i = 2\log L(\hat{\beta}) - 2\log L(\hat{\beta}_{(-i)}). \quad (5)$$

Under the null hypothesis $\hat{\beta}_{(-i)} = \hat{\beta}$ the LD statistic follows a chi-square distribution with one degree of freedom. Therefore we calculate the p-value for this test for all observations, the ones having more significance are considered the most outlying ones.

4 DATASETS

4.1 Simulation Data (SIM)

Similarly to the simulation data in (Farcomeni and Viviani, 2011), we will generate datasets having as underlying probabilistic model, the Cox proportional hazards. Our goal is to recreate a realistic setting, with survival times and covariates as similar as real datasets. In order to approximate this conditions, each simulated dataset will have a *pure* model β that translates a a general trend of the observations, and two other Cox models with different parameter values. Each dataset consists in 200 observations having covariates X_1, X_2, X_3. These follow a 3-D normal distribution with zero mean and covariance matrix Σ, that will be equal to the identity matrix I for the pure model and $\sigma \cdot I$ for the outlier models.

For the survival times, the probabilistic model for the hazard of each individual follows one of three possible models: the pure model β and two outlier models β' and β''. Having $k < n$ outliers (k even), the hazard function for each observation i is generated by:

$$h_i(t) = \begin{cases} h_0(t)\exp\{\beta X\} & 1 \le i \le n-k \\ h_0(t)\exp\{\beta' X\} & n-k < i \le n-k/2 \\ h_0(t)\exp\{\beta'' X\} & n-k/2 < i \le n \end{cases}. \quad (6)$$

The baseline hazard $h_0(t)$ is given by a *Weibull* function with both shape and scale parameters equal to unity, defined in the interval from 0 to 1. The value for k will be set in order to have 10% of outliers.

The estimation of the cumulative hazard function $H_i(t)$ is then obtained:

$$H_i(t) = \int_0^t h_i(\tau) d\tau. \quad (7)$$

From each $H_i(t)$ we further calculate the corresponding survival curves by $S_i(t) = e^{-H_i(t)}$. Having this distribution, we generate 200 survival times according to the distribution given by $S_i(t)$ and generate a censoring vector $c_1, .., c_{200}$ following a Bernoulli with probability p, corresponding to the proportion of censored observations, typically a value around 0.2:

$$\begin{aligned} t_i &\sim 1 - S_i(t), \\ c_i &\sim Bernoulli(p). \end{aligned} \quad (8)$$

4.2 Clinical Data

In order to test the procedures in a more realistic setting, we have further applied the methods to real clinical data, focusing on two studies:

WHAS. Dataset from the Worcester Heart Attack Study, with 100 individuals each with 5 covariates. This data concerns the survival times of patients having their first heart attack. Data publicly available at https://www.umass.edu/statdata/statdata/data/.

BMT. Bone Marrow Transplant Data (Klein and Moeschberger, 1997): contains data about 137 leukemia patients each with 10 covariates. The data concerns the survival time after the bone marrow transplant. It is publicly available in the R (R Development Core Team, 2006) package *KMsurv*.

5 RESULTS AND DISCUSSION

In this section we assess the performance of the two proposed outlier detection methods BHT and OSD and we compare their results with MART, DEV and LD. We start by presenting the configuration of our simulation study for outlier detection. Then we apply all methods to two real datasets, performing outlier detection. We further use the detected outliers to perform a robust Cox regression by removing them from the data, the coefficients and p-values of the regression will be compared.

5.1 SIM Dataset

The outlier detection methods will be used on simulated datasets generated using the methodology described in Section 4.1. In order to test the outlier detection methods in a variety of conditions for the outlying models and for the general model, we will fix the general trend model $\beta = (1,1,1)$ and then we define a set of configurations for the two sources of outlying observations. Each parameter for the outlier sources is given by a three dimensional normal distribution with a diagonal covariance matrix, the values for the means and variances in each scenario are presented in Table 1.

Table 1: Tested scenarios for the outlier sources.

Scenario	β'	β''	σ
1	(-0.5,-0.5,-0.5)	(0.5,-0.5,-0.5)	0.25
2	(-2,-2,-2)	(-2,2,-2)	0.50
3	(-1,-1,-1)	(-1,1,-1)	0.50
4	(0.5,0.5,0.5)	(0.5,-0.5,0.5)	0.25
5	(2,2,2)	(2,-2,2)	0.50
6	(1,-1,1)	(1,1,-1)	0.50
7	(0.8,0.8,-1.6)	(-1.6,0.8,0.8)	0.50
8	(0.25,0.25,-0.50)	(-0.50,0.25,0.25)	0.10
9	(2,2,2)	(-2,-2,-2)	0.50
10	(2,2,2)	(-1,-1,-1)	0.50

Although the outlying values for the parameters may seem close to the general trend model it is worth noting that the Cox model defines the hazards as an exponential function of βX, thus the ratio between the hazard of an outlying and a general trend observation is given by $\exp\{\beta' X - \beta X\}$. The reasons behind the choice of this set of scenarios is to have a variety of combinations with different norms and contrasting parameters.

Table 2 reports the accuracy in terms of percentage of retrieved outliers or true positive rate. By inspecting Table 2 we see that the OSD algorithm is the one that has an overall better performance, overcoming the other methods in 6 out of 10 of the scenarios. For scenarios 5 and 10, MART achieves the best performance.

Table 2: Fraction of true positives averaged over 100 runs for each method in the 10 chosen scenarios.

Scenario	MART	DEV	LD	BHT	OSD
1	0.29	0.31	0.39	0.25	**0.42**
2	0.43	0.49	0.54	0.45	**0.62**
3	0.35	0.40	0.45	0.39	**0.52**
4	0.22	0.24	0.27	0.27	**0.28**
5	**0.26**	0.25	0.19	0.18	0.13
6	0.22	0.30	0.30	0.20	**0.32**
7	0.31	0.32	0.33	0.31	**0.39**
8	0.22	0.26	**0.34**	0.24	0.32
9	0.25	**0.26**	0.21	0.22	0.21
10	**0.22**	0.21	0.17	0.17	0.13

5.2 WHAS Dataset

The outliers detected by the methods in the WHAS dataset are presented in Table 3. The selection is based on the ten lowest p-values.

Table 3: Top 10% outliers detected by the methods in the WHAS dataset.

Nb.	MART	DEV	LD	BHT	OSD
1	93	1	97	67	1
2	51	31	67	1	67
3	90	56	1	78	97
4	33	85	52	56	51
5	11	97	23	69	23
6	27	93	7	8	31
7	40	30	57	45	93
8	1	78	78	93	52
9	31	51	56	30	56
10	56	90	17	32	57

It is noteworthy that all the methods identified observation 56.The estimates for the regression coefficients when fitting the Cox model to all observations are given in Table 4.

We observe that only two covariates are statistically significant corresponding to the age at the first

Table 4: Cox model fitted to the WHAS dataset.

	β	*p*-value
los	-0.022	0.3972
age	0.039	0.0025
gender	0.157	0.6066
bmi	-0.071	0.0497

hear attack (*age*) and the body mass index (*bmi*).

After removing 10% of the observations indicated in Table 3 for each of the methods, new models are obtained (Table 5 and Table 6). The goal is to unveil a trend model, unaffected by outlying observations.

Table 5: Cox estimates removing the top 10% outlier observations in the WHAS dataset for methods BHT and OSD.

	BHT		OSD	
	β	*p*-value	β	*p*-value
los	-0.166	0.006	-0.025	0.374
age	0.048	0.000	0.068	0.000
gender	0.003	0.992	0.042	0.899
bmi	-0.162	0.001	-0.137	0.002

Table 6: Cox estimates removing the top 10% outlier observations in the WHAS dataset for methods MART, DEV and LD.

	MART		DEV		LD	
	β	*p*-value	β	*p*-value	β	*p*-value
los	-0.016	0.498	-0.015	0.550	-0.016	0.506
age	0.045	0.001	0.032	0.012	0.069	0.000
gender	-0.082	0.800	0.155	0.653	-0.230	0.483
bmi	-0.082	0.029	-0.037	0.030	-0.146	0.001

The results show that in the proposed BHT method the length of stay (*los*) after the first heart attack appeared as significant, which did not occur for the other methods. These results show that BHT can potentially unveil covariates that were not considered useful.

The fact that *los* rose as a significant covariate in the Cox regression calls for a better analysis of this measure. There are several studies that relate the length of hospital stay with patient readmission. Also studied, is the association between *los* and the quality of hospital care, (Thomas et al., 1996) with data for 12 different conditions, that a longer *los* risk-adjusted for other covariates, is associated with poorer hospital care. In our case we have a negative coefficient, meaning that the hazard function decreases with a longer length of stay, thus this might be also a potential indicator that the hospital has a good quality of care.

5.3 BMT Dataset

The outliers detected by the methods in the BMT dataset are presented in Table 7. The selection is based, again, on the 10% lowest *p*-values. For BHT, a value of bootstrap samples $B = 2000$ has shown to be sufficient for the convergence.

Table 7: Top 10% outliers detected by the methods in the BMT dataset.

Nb.	MART	DEV	LD	BHT	OSD
1	65	129	129	129	129
2	103	35	132	103	132
3	99	108	89	99	30
4	97	65	90	65	130
5	13	132	26	30	26
6	42	87	30	132	28
7	63	84	28	13	65
8	40	103	130	130	13
9	92	30	17	16	103
10	14	99	105	136	14
11	43	97	136	15	72
12	39	28	116	26	89
13	49	109	72	97	50

The estimates for the regression coefficients when fitting the Cox model to all observation are given in Table 8.

Table 8: Cox model fitted to all BMT data.

	β	*p*-value
Age_Diagn	-0.0017	0.9357
Donor_Age	0.0316	0.1072
Sex	-0.2738	0.2651
Donor_Sex	0.0409	0.8662
CMV	-0.1701	0.4922
Donor_CMV	0.0038	0.9875
Wait_Time	-0.0001	0.8701
FAB	0.7917	0.0012
Hospital	-0.5570	0.0004
MTX	1.0062	0.0026

After removing 10% of the observations indicated in the Table 7 for each of the methods, new models are obtained (Table 9 and Table 10).

Table 9: Cox estimates removing the top 10% outlier observations in the BMT dataset for methods BHT and OSD.

	BHT		OSD	
	β	*p*-value	β	*p*-value
Age_Diagn	-0.017	0.418	0.027	0.222
Donor_Age	0.033	0.097	0.016	0.432
Sex	-0.412	0.115	-0.556	0.029
Donor_Sex	0.076	0.780	0.403	0.144
CMV	-0.541	0.047	-0.622	0.026
Donor_CMV	-0.024	0.926	0.116	0.651
Wait_Time	0.000	0.623	-0.001	0.472
FAB	1.260	0.000	1.157	0.000
Hospital	-0.991	0.000	-1.190	0.000
MTX	2.127	0.000	2.488	0.000

When using all the data, the statistically significant covariates are *FAB*, *Hospital* and *MTX* (Table 8). When the first top 10% outlier observations were removed, the results were very similar between the proposed methods BHT and OSD as both reduced the *p*-value of the variable *CMV* to 0.047 and 0.026, respectively. This possibly reveals that the variable *CMV* is

Table 10: Cox estimates removing the top 10% outlier observations in the BMT dataset for methods MART, DEV and LD.

	MART		DEV		LD	
	β	*p*-value	β	*p*-value	β	*p*-value
Age_Diagn	-0.009	0.640	0.029	0.181	0.006	0.777
Donor_Age	0.027	0.149	0.024	0.243	0.050	0.027
Sex	-0.443	0.078	-0.624	0.021	-0.325	0.235
Donor_Sex	0.053	0.833	0.257	0.345	0.361	0.195
CMV	-0.356	0.178	-0.460	0.094	-0.395	0.148
Donor_CMV	-0.432	0.867	0.075	0.771	0.032	0.910
Wait_Time	-0.000	0.866	0.000	0.321	-0.000	0.586
FAB	1.170	0.000	1.286	0.000	1.058	0.000
Hospital	-0.693	0.000	-0.794	0.000	-1.442	0.000
MTX	1.813	0.000	1.495	0.000	2.350	0.000

much more significant to the model than first expected using the complete dataset. The covariate *CMV* represents the cytomegalovirus immune status (positive or negative) and therefore might be a relevant feature to predict survival. It is noteworthy that the other methods did not retrieve this variable as significant.

In all these experiments, the choice of the outlier percentage threshold has obvious implications on the obtained Cox regression coefficients and a more detailed analysis is warranted to analyze the tradeoff between keeping and removing observations.

5.4 Leave-one-Out Cross-validation of the C-index

To assess the predictive ability of the model when facing new observations, we perform leave-one-out cross-validation of the c-index. The outliers also become part of the several test sets, but they are never present in the training used to estimate the models. Thus this measure takes into account the prediction performance of the model on outlying observations. The results are very positive, with the concordance showing a systematic increase while removing candidate outliers.

Table 11: Leave-one-out estimated c-indexes for the BHT method.

Dataset	All data	top-3	top-10	top-30
WHAS	0.6607	0.6813	0.6824	0.6900
BMT	0.6208	0.6314	0.6441	0.6668

Table 12: Leave-one-out estimated c-indexes for the OSD procedure.

Dataset	All data	top-3	top-10	top-30
WHAS	0.6607	0.6832	0.6853	0.6986
BMT	0.6208	0.6314	0.6441	0.6629

6 CONCLUSION

We proposed two methods for outlier detection in a survival setting. Both methods improve the performance of the Cox Regression using cross-validation. Overall, OSD has shown promising results in terms of *p*-value improvement of the regression coefficients. We think BHT can be improved in order to be a 2-D index possibly using multimodality measures (Singh and Xie, 2003) to identify the outliers that have a higher *p*-value (that do not systematically improve concordance when removed from the data, but still are outlying observations).

Finally, we use both methods to perform robust estimation for the Cox regression, removing from the regression a fraction of the data by their measure of outlyingness. Our preliminary results on three different datasets have shown to improve the estimation of the Cox Regression coefficients and also the model predictive ability.

ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e Tecnologia (FCT, Portugal) under contracts LAETA Pest-OE/EME/LA0022 and IT (PEst-OE/EEI/LA0008/2013), as well as project CancerSys (EXPL/EMS-SIS/1954/2013). SV acknowledges support by Programa Investigador FCT(IF/00653/2012) from FCT, co-funded by the European Social Fund (ESF) through the Operational Program Human Potential (POPH).

REFERENCES

Acuna, E. and Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*.

Ben-Gal, I. (2005). Outlier detection. In *Data Mining and Knowledge Discovery Handbook*, pages 131–146. Springer.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, pages 15–18.

Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistic Society*, B(34):187–202.

David Collett (2003). *Modelling survival data in medical research*. Boca Raton, Fla. : Chapman & Hall/CRC, c2003.

David G. Kleinbaum, Mitchel Klein (2005). *Survival analysis: a self-learning text*. New York, NY : Springer, c2005.

David W. Hosmer, Stanley Lemeshow, Susanne May (2008). *Applied survival analysis: regression modeling of time-to-event data*. Hoboken, N.J. : Wiley-Interscience, c2008.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, pages 157–184.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26.

Farcomeni, A. and Viviani, S. (2011). Robust estimation for the cox regression model based on trimming. *Biometrical Journal*, 53(6):956–973.

Fischler, M. and Bolles, R. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Johnson, R. A., Wichern, D. W., and Education, P. (1992). *Applied multivariate statistical analysis*, volume 4. Prentice hall Englewood Cliffs, NJ.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Klein, J. and Moeschberger, M. (1997). Survival analysis: techniques for censored and truncated regression.

Nardi, A. and Schemper, M. (1999). New residuals for cox regression and their application to outlier screening. *Biometrics*, 55(2):523–529.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reid, N. and Crépeau, H. (1985). Influence functions for proportional hazards regression. *Biometrika*, 72(1):1–9.

Rousseeuw, P. and Leroy, A. (1987). *Robust regression and outlier detection*. Wiley Series in probability and mathematical statistics. Wiley, New York [u.a.].

Singh, K. and Xie, M. (2003). Bootlier-Plot: Bootstrap Based Outlier Detection Plot. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, 65(3):532–559.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

Thomas, J. W., Guire, K. E., and Horvat, G. G. (1996). Is patient length of stay related to quality of care? *Hospital & health services administration*, 42(4):489–507.