Outlier detection in Cox proportional hazards models based on the concordance c-index

João Diogo Pinto[1], Alexandra M. Carvalho[1] and Susana Vinga[2]

[1]Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Portugal,
[2]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Portugal

Abstract. Outliers can have extreme influence on data analysis and so their presence must be taken into account. We propose a method to perform outlier detection on multivariate survival datasets, named *Dual Bootstrap Hypothesis Testing* (DBHT). Experimental results show that DBHT is a competitive alternative to state-of-the-art methods and can be applied to clinical data.

1 Introduction

Survival analysis, the field that studies time-to-event data, has become a relevant topic in clinical and medical research. In many medical studies time to death is the event of interest, hence, it is usually named *survival time*. However, other important measures may also be considered, such as the time between response to treatment or the time to the onset of a disease.

Survival analysis is specifically tailored to deal with unknown survival times for a subset of the study group, a phenomenon called *censoring*. The most common type is *right-censoring*, addressed in this work; it occurs when the event is beyond the end of the follow-up period. Survival data is typically denoted by $D = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, where each X_i is a p-dimensional vector of covariates and $Y_i = (t_i, \delta_i)$, where t_i is the event or censoring time and δ_i the censoring indicator.

There are several definitions of an outlier. Hawkins [6] defines it "as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism than the remaining data". In the survival field, Nardi and Schemper [8] define outlying observations as individuals whose survival time is too short, or too long, with respect to the values of its covariates.

In this work, we propose to perform outlier detection in survival analysis taking profit from Harrel's *concordance c-index* [5] and extending the work in [7]. The concordance c-index measures the model's ability of predicting a higher relative risks to individuals whose event occurs first. The relative risk is estimated from the output of the model for each individual; in a Cox Proportional hazards model, for instance, the relative risk corresponds to the hazard ratio.

2 DBHT

Bootstrapping [3] is a resampling technique to unveil the underlying distribution of the data. It is used when this distribution is unknown or simplifying assumptions are not reasonable. Given a dataset D with N observations, one *bootstrap sample* is obtained by sampling, with replacement, N observations from D.

We propose to improve the *bootstrap hypothesis test* (BHT) described in [7]. In the BHT method, the procedure removes one observation from the dataset and then assesses the impact of each removal on concordance. This has the undesired effect that, with less observations to fit, concordance tends to increase, which potentially increases the number of "false positives", signalling inliers as outliers.

The proposed method, called dual bootstrap hypothesis test (DBHT), overcomes this problem. In starts by generating two histograms from two antagonistic versions of the bootstrap procedure – the *poison* and the *antidote* bootstraps – and then compare them using a statistical test. The *antidote* bootstrap excludes the observation under test from every bootstrap sample. On the other hand, the *poison* bootstrap works by forcing the observation under test to be part of every bootstrap sample. Both the poison and antidote bootstraps have the same number of observations in each bootstrap sample.

The general strategy is as follows. For each observation i we make the hypothesis that the observation is "poison" (meaning that the observation is an outlier). To test it, we compare the histograms of concordance variation $\varDelta C$ between the *antidote* and *poison* bootstraps. If the observation is indeed an outlier, we expect the *antidote* bootstrap to push the histogram for higher values of $\varDelta C$. Conversely, we expect the *poison* bootstrap to generate lower values of $\varDelta C$. The more the *poison* histogram is to the left of the *antidote* histogram, the more outlying the observation is. We consider $\varDelta C_{antidote}$ and $\varDelta C_{poison}$ two real random variables with the following hypothesis:

$$H_0 : E\left[\varDelta C_{antidote}\right] > E\left[\varDelta C_{poison}\right];$$
$$H_1 : E\left[\varDelta C_{antidote}\right] \leq E\left[\varDelta C_{poison}\right].$$

To calculate the p-value of the test we use a independent two sample Welch's t-test.

DBHT is a *soft-classifier* and a *single-step* method with the output being an outlying measure for each observation. From this, it is possible to extract the k most outlying observations. Pseudo-code of the DBHT procedure can be found in Algorithm 1.

Algorithm 1 Dual Bootstrap Hypothesis Test

```
Input: input dataset $D$, the survival model and number of bootstrap samples $B$.
Output: a $p$-value for each observation
for all $d_i \in D$ do
    $D_{-i} = D \setminus d_i$  {remove observation $i$ from the original dataset}
    Generate $B$ poison bootstrap samples
    Generate $B$ antidote bootstrap samples from
    Compute the $B$ values of $\varDelta C_{poison}$ and store them in vector $psn$
    Compute the $B$ values of $\varDelta C_{antidote}$ and store them in vector $ant$
    From $psn$ and $ant$ compute the $p$-value using a $t$ test for equality of means
end for
return the vector of $p$-values
```

In Figure 1, *poison* and *antidote* histograms for an outlier (on the left) and inlier (on the right) can be found.

3 Results

Herein, we assess the performance of DBHT in 12 synthetic datasets. Its performance is compared with two concordance-based methods [7] – *one step deletion* (OSD) and

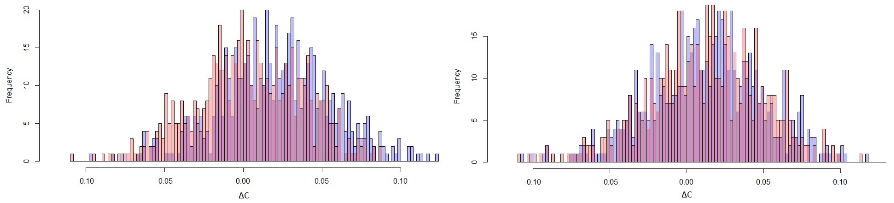

Fig. 1. (Colour online) On the left, contrast between antidote (blue) and poison (red) bootstrap histograms of concordance variation, for a typical outlier. On the right, antidote (blue) and poison (red) bootstrap histograms of concordance variation, for a typical inlier.

Bootstrap hypothesis test (BHT) – and with outlier detection methods commonly employed on survival data, namely, *martingale residuals* (MART), *deviance residuals* (DEV) , *likelihood displacement statistic* (LD) and *DFBETAS* (DFB).

The model chosen to recreate survival times was the Cox proportional hazards. The simulated observations were generated from two different Cox models, a general trend model $\beta = \beta^G$ and an outlier model $\beta = \beta'$. From the Cox hazard function, the distribution of T is given by $F(t|X) = 1 - \exp\left[-H_0(t) \cdot \exp(\beta X)\right]$. The vector of covariates X characterizing each individual was generated from a three-dimensional normal distribution with zero mean with identity covariance matrix. The survival times were generated using the methodology explained in [1], each observation time as function of the covariate vector X given by $T = H_0^{-1}\left[-\log(U) \cdot \exp(-\beta X)\right]$, where U is a uniform random variable distributed in the interval $[0, 1]$.

Several scenarios were simulated. For each one, the vector of covariates was given by $X_i \sim N(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Each simulated dataset contains 100 observations with hazard functions

$$h_i(t) = \begin{cases} h_0(t)\exp\{\boldsymbol{\beta}^{\boldsymbol{G}}\mathbf{X}\} & 1 \leq i \leq n-k \\ h_0(t)\exp\{\boldsymbol{\beta}'\mathbf{X}\} & n-k < i \leq n, \end{cases}$$

where the pure model $\beta^G = (1, 1, 1)$ and $\beta^{'}$ taking 12 different vectors; see Table 1.

When assessing the performance of outlier detection methods on the simulated data it has to be taken into account that the observations are randomly generated from distributions: the inliers from the general distribution β^G, and the outliers from an outlying distribution β'. It may happen that observations initially intended to be inliers may be drawn from the lower or upper tail of the distribution and may configure an outlier, and vice-versa. Our performance assessment assumes that for each scenario the observations generated from general distribution are inliers and the observations generated from the outlying distribution are outliers.

We used two metrics to analyse the results, the *true positive rate* (TPR), also known as *sensitivity*, and the *area under the ROC curve* (AUC). For datasets with k outliers the TPR will measure for each scenario the fraction of true outliers found in the top-k most outlying observations indicated by each method. The AUC provides us a threshold-independent outlier detection ability. The AUC is not applicable to the output of the OSD method, because it does not provide an outlying score for every observation. The

TPR and AUC are the mean of 50 runs per simulation configuration. Results are depicted in Table 1.

Table 1. Outlier configurations used in the simulated data (left). Average of TPR (middle) and average of AUC (right) grouped by outlier scenarios.

Scen.	Θ'	$\|\|\beta'\|\|/\|\|\beta^G\|\|$	β'	MART	DEV	LD	DFB	OSD	BHT	DBHT	MART	DEV	LD	DFB	BHT	DBHT
1	180°	1	(-1,-1,-1)	0.29	0.36	0.43	0.36	0.47	0.43	**0.47**	0.70	0.70	0.74	0.68	0.78	**0.82**
2	180°	0.2	(-0.2,-0.2,-0.2)	0.22	0.25	0.31	0.29	0.32	0.31	**0.34**	0.65	0.65	0.70	0.64	0.71	**0.75**
3	180°	5	(-5,-5,-5)	0.50	0.58	0.59	0.52	0.63	0.59	**0.65**	0.80	0.80	0.78	0.77	0.86	**0.90**
4	135°	0.2	(-0.143,0,-0.283)	0.22	0.23	0.30	0.28	0.30	0.29	**0.32**	0.64	0.64	0.69	0.63	0.71	**0.73**
5	135°	5	(-3,6,0,-7.07)	0.44	0.54	0.52	0.48	**0.58**	0.53	**0.58**	0.78	0.77	0.74	0.75	0.82	**0.84**
6	90°	0.2	(-0.245,0,-0.245)	0.21	0.22	**0.28**	0.26	0.27	0.26	**0.28**	0.63	0.63	0.67	0.63	0.68	**0.71**
7	90°	5	(6.12,0,-6.12)	0.40	**0.50**	0.40	0.41	0.44	0.37	0.42	**0.76**	**0.76**	0.66	0.73	0.70	0.72
8	0°	0.2	(0.2,0.2,0.2)	0.18	0.18	**0.23**	0.22	0.22	0.20	**0.23**	0.62	0.62	0.66	0.62	0.65	**0.68**
9	0°	5	(5,5,5)	0.32	**0.36**	0.18	0.25	0.09	0.06	0.07	**0.74**	0.72	0.61	0.69	0.60	0.60
10	180°	10	(-10,-10,-10)	0.53	0.63	0.64	0.57	0.68	0.60	**0.70**	0.83	0.83	0.80	0.81	0.87	**0.92**
11	0°	10	(10,10,10)	0.38	**0.46**	0.24	0.32	0.14	0.11	0.12	**0.78**	0.76	0.61	0.73	0.59	0.61
12	135°	10	(-7.15,0,-14.15)	0.49	**0.60**	0.54	0.51	**0.60**	0.52	**0.60**	0.80	0.80	0.74	0.78	0.81	**0.86**

4 Conclusion and future work

DBHT has shown promising results, being the best method in nine of the 12 simulated outlier scenarios. On the three scenarios where β' is collinear with β^G, the performance of DBHT, BHT and OSD is poor; in these scenarios outliers have the same hazard direction as inliers, and so concordance fails to capture them as it does note take into account the difference in predicted hazards. This kind of outliers are typically very well detected by residual-based methods, so DBHT may be useful when used jointly with these methods. Future applications include outliner detection for oncological patients.

Acknowledgments

Work supported by Fundação para a Ciência e a Tecnologia (FCT) under contracts LAETA (UID/EMS/50022/2013) and IT (UID/EEA/50008/2013), and by projects CancerSys (EXPL/EMS-SIS/1954/2013) and InteleGen (PTDC/DTP-FTO/1747/2012). SV acknowledges support by Program Investigador (IF/00653/2012) from FCT, co-funded by the European Social Fund through the Operational Program Human Potential.

References

1. R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.
2. B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.
3. F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, and R.A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.
4. D.M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.
5. J. Pinto, A.M. Carvalho and S. Vinga. Outlier detection in survival analysis based on the concordance c-index. In *Proceedings of BIOINFORMATICS, 2015*, pages 72–82, 2015.
6. A. Nardi and M. Schemper. New residuals for cox regression and their application to outlier screening. *Biometrics*, 55(2):523–529, 1999.