Trimmed Cox regression for Robust Estimation in Survival Studies

Sara Viviani
Department of Statistics, Probability and Applied Statistics
Sapienza - University of Rome
Alessio Farcomeni
Department of Hygene and Public Health
Sapienza - University of Rome

Abstract

We propose a robust Cox regression model with outliers. The model is fit by trimming the smallest contributions to the partial likelihood. To do so, we implement an ad-hoc algorithm, and show its convergence to a global optimum. We discuss global robustness properties of the approach, which is illustrated and compared through simulations. Finally, we develop an application to survival in a randomized study about prostate cancer.

Keywords: Cox model, Partial Likelihood, Prostate cancer, Outliers, Robustness, Survival Analysis, Trimming

1 Introduction

Robust estimation is a well developed topic in different areas of statistics (e.g., Maronna *et al.* (2006), Hubert *et al.* (2008), Hawkins (1980)). It is well known that many statistical procedures can be sensitive to violation of underlying assumptions, and even non-parametric non-robust procedures may break down due to outliers or departures from model assumptions. Visual inspection may not reveal masked outliers, and it is practically infeasible in large dimensions, thereby making it very hard to detect covariate outliers even when the number of predictors is small.

Survival analysis makes no exception. Despite the unspecified baseline in the Cox model may be able to capture some aberrant behaviours, it can still happen that even a single malicious observation is unduly influent, with dramatic effects on the parameter estimates. A single observation sufficies for violation of the assumption of proportionality of hazard, and this departure may not be detected by common checking methods. The Cox model is sensitive even to slight departures from the assumptions: its influence function is not bounded (Reid and Crépeau, 1985). Lack of robustness of the Cox model is clearly pointed out in the literature, see for instance Samuels (1978), Bednarski (1989) and Minder and Bednarski (1996). Many studies are devoted to diagnostics and assessing of robustness (for instance, to influential outliers) of the Cox model, e.g., Cain and Lange (1984), Reid and Crépeau (1985). Many of these proposals rely on residual analysis (Schoenfeld (1982), Grambsch and Therneau (1994), Therneau *et al.* (1990), Nardi and Schemper (1999)). It is well known anyway that residual analysis cannot be directly used for outlier identification, since there is a very high likelihood of masking (Becker and Gather, 1999) and swamping due to the fact that outliers, when present, are used to obtain the parameter estimates. Robust estimates are a pre-requisite for distance based outlier detection procedures.

Among the few attempts to robustify the Cox model estimators we point the reader to Bednarski (1993), who proposes an approach based on a smooth modification of the partial likelihood; and to Sasieni (1993a,b), who uses a weighted partial likelihood method, e.g., a Wilcoxon-type weighting scheme. Schemper *et al.* (2009) have demonstrated the usefulness of weighted partial likelihood for computing average hazard ratios in presence of non-proportionality of hazards.

The double weighting approach of Bednarski (1993) was refined in Bednarski (2007) to make it adaptive and invariant to time-transformation.

In this paper we describe a different approach for robust estimation in the Cox model, which is based on trimming. Our approach is simple, but very effective in terms of robustness. The idea of adaptively trimming observations which are least likely to occur as indicated by the likelihood has also been investigated in other contexts (e.g., Bednarski and Clarke (1993), Clarke (2000)).

Outliers in survival studies can be interpreted as in Nardi and Schemper (1999), who define outliers as individuals whose failure time is too short, or too long; with respect to the median survival as predicted by the model. This definition is very useful since it serves as a unified approach to the problem of treating covariate outliers, patients responding differently to a covariate combination, misclassified covariates, misclassified events, and plain gross outlying (too short or too long) survival times.

Our definition is slightly more general than that of Nardi and Schemper (1999). We define outliers as individuals whose contribution to the (partial) likelihood is small when compared to the other subjects. Hence, they can be "too long living", "too early dying", or belong to any other configuration of covariates and survival times which is unusual with respect to the fitted model. Valsecchi *et al.* (1996) provide a detailed illustration on how long surviving outliers may affect the estimates. Too long living individuals are only one of the possible kind of outliers, but they probably are the most harmful (i.e., influent) to the parameter estimates since they are present in almost all risk sets.

We stress that we focus on robust estimation. In many survival studies, outlier detection may follow (robust) estimation. Outliers may unveil important clinical information. In our approach, outliers are confined to the trimmed set of observations, together with possibly few clean observations. After robust estimation, we suggest formally assessing which observations are outliers through residual analysis as in Nardi and Schemper (1999). After robust estimation, residual analysis is not expected to suffer from problems related to masking. Further, the covariates can be separately explored with methods for detecting multivariate outliers.

The rest of the paper is as follows: in Section 2 we illustrate our methodology for robust survival analysis. In Section 3 we discuss robustness properties. We illustrate the method with a brief simulation study in Section 4 and on a real data example on prostate cancer in Section 5.

Non-optimized R (R Development Core Team, 2009) code for fitting the proposed model is available from the authors upon request.

2 A Proportional Hazards model with outliers

Suppose we observe time to an event of interest for n independent subjects, and let (t_i, δ_i) denote the observed time and the event indicator for the i-th subject. Denote also by X_i a vector of subject specific covariates. In Cox proportional hazard regression (Cox, 1972) the effects of covariates on the hazard rate $\lambda(t|x_i)$ for the i-th individual is of the form:

$$\lambda(t|x_i) = \lambda_0(t)\exp(\beta' X_i),$$

where $\lambda_0(t)$ denotes a non-parametric baseline hazard.

Regression parameters β are estimated by maximizing the partial likelihood $l(\beta)$, where

$$l(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum\limits_{t_j > t_i} \exp(\beta' x_j)} \right)^{\delta_i}. \tag{1}$$

The resulting maximum partial likelihood estimator (MPLE) is consistent and asymptotically normal under regularity conditions.

Here we consider a Cox model with possible contamination. We denote with I^* the set of clean observations. Then, we have that

$$\begin{cases} \lambda(t|x_i) = \lambda_0(t)\exp(\beta' x_i) & \text{If } i \in I^* \\ \lambda(t|x_i) = \lambda_i(t) & \text{If } i \notin I^*. \end{cases} \tag{2}$$

Contaminated observations arise from an unknown and observation-specific unspecified hazard rate $\lambda_i(t)$. This leads contaminated observations not to give useful information for estimating the effects of covariates on the survival times.

Suppose the set of clean observations is of cardinality $\lceil n(1-\alpha)\rceil$, for a fixed $0 < \alpha < 1$. Denote with $H(\alpha)$ the set of all subsets of the vector of integers $(1, \ldots, n)$, where each of these subsets is of cardinality $\lceil n(1-\alpha)\rceil$.

The MPLE for model (2) is the maximizer of

$$l_{TRIM}(\beta) = \max_{I \in H(\alpha)} \prod_{i \in I} \left(\frac{\exp(\beta' x_i)}{\sum\limits_{t_j > t_i, j \in I} \exp(\beta' x_j)} \right)^{\delta_i} \tag{3}$$

That is, $\widehat{\beta}$ is the largest maximum over all possible maxima of the partial likelihoods computed only on subsets of $\lceil n(1-\alpha)\rceil$ observations.

In practice, the proportion of contaminated observations α is not known, and the user will set α slightly larger than the expected proportion of contaminated observations. See Section 2.2 for further discussion on this point.

The simple idea of trimming the smallest contributions to the likelihood is seen to lead to robust estimation in the Cox model. It is straightforward to check that the resulting estimator is still consistent and asymptotically normal under the assumption of no contamination. Of course, the asymptotic variance will be inflated, resulting in a small loss of efficiency with respect to the classical MPLE. This is the price that is (always) paid for robustness.

2.1 Model fit

Maximization of (3) is a much harder optimization problem than maximization of (1). In order to obtain the maximum likelihood estimates for the trimmed model we should solve a formidable combinatorial problem and compare the maxima obtained under all the possible subsets of the data of size $\lceil n(1-\alpha)\rceil$. This is obviously infeasible, as the number of such subsets is $\binom{n}{\lceil n(1-\alpha)\rceil}$, and grows very rapidly with n.

This kind of optimization problem is common in robust statistics, and it is usually tackled via the use of ad-hoc algorithms. These algorithms use repeated concentration steps (Rousseeuw and van Driessen, 1999), see also Farcomeni (2009). Here it is not straightfoward to use these algorithms, since individual contributions to the partial likelihood have cumbersome expressions (Verweij and Van Houwelingen, 1993). A different general method, suitable for our problem, has been recently described in Chakraborty and Chaudhury (2008), and we will now adapt it to the survival context.

Let

$$l(\beta, I) = \prod_{i \in I} \left(\frac{\exp(\beta' x_i)}{\sum\limits_{t_j > t_i, j \in I} \exp(\beta' x_j)} \right)^{\delta_i} \tag{4}$$

denote the trimmed partial likelihood for the regression parameters β computed in a given subset I.

Our algorithm (whose general iteration is summarized in Algorithm 1) is initialized from a set I of observations of cardinality $\lceil n(1-\alpha)\rceil$. This initialization set can be chosen at random, or a set of likely clean observations can be used if available. Given I, $\widehat{\beta}$ is estimated via the usual score equations for the Cox model, restricted to the observations in I, and the corresponding maximum partial likelihood is recorded.

The algorithm then follows an acceptance-rejection scheme similar to Metropolis-Hastings in MCMC. At each iteration, a new proposal for the

Algorithm 1 General iteration for maximization of trimmed partial likelihood

```
for $k = 1, \dots, k_{\max}$ do
  for $i \in I$ do
    randomly sample a candidate $i' \in I^C$.
    let $I_{cand}$ be the set of positions equal to $I$, except that $i$ is replaced with $i'$.
    Let
        $\tau_k := \log(k+1)/D$    (5)
    $\widehat{\beta} := \arg\max_{\beta} l(\beta, I)$
    $\widehat{\beta}_{cand} := \arg\max_{\beta} l(\beta, I_{cand})$
        $p := \min(e^{\tau_k(\log(l(\widehat{\beta}_{cand}, I_{cand})) - \log(l(\widehat{\beta}, I)))}, 1)$    (6)
    Let $U$ be a random draw from a Bernoulli with parameter $p$.
    if $U = 1$ then
      $I(i) := i'$.
    end if
  end for
end for
```

optimal subset is chosen at random by changing a single entry of the current subset I. The maximum likelihood corresponding to the randomly sampled candidate subset I_{cand} is then recorded. Whenever this maximum is larger than the maximum partial likelihood corresponding to the current subset, it is accepted with a probability $p = 1$. If the likelihood is not increased, the candidate is accepted with a probability $p < 1$, so that the algorithm is not trapped in local optima.

This probability p, given in expression (6), is a function of the iteration number k and of the difference between the current and candidate log-likelihoods. More precisely, it decreases with the iteration number k so that in the first few iterations of the algorithm it is possible to excape local optima while in the last iterations, when the global mode is more likely to have been found, it is unlikely to explore other regions of the parameter space. Of course, this probability of acceptance is also always proportional to the likelihood ratio between the proposal and the current subset; and when the new proposal corresponds to a slightly lower likelihood than the current subset, p is very small regardless of the iteration number k.

There are two tuning parameters: k_{max} and D. The first tuning parameter controls the maximum number of iterations, and should be set large enough that in the last few iterations cycle p is always either equal to 1 or approximately zero. The second is instead related to the maximal expected change in the log partial likelihood when a single observation is changed in the subset I (refer to Chakraborty and Chaudhury (2008) for further details). The choice of D has consequences only on the speed of convergence and acceptance ratio for the candidate subsets. Unless stated otherwise, we will set $k_{\text{max}} = 10000$ and $D = 0.1n(1-\alpha)$.

Formally, we can prove the following theorem:

Theorem 1. *Fix $0 < D < n(1-\alpha)$. For any initial subset I_0, for Algorithm 1 we have that*

$$P(I_k \in H) \to 1$$

as $k \to \infty$, where I_k is the subset obtained at the k-th iteration and H denotes the set of all global optima of the trimmed partial likelihood.

Proof. Note that if $\max\{|(\log(l(\widehat{\beta}_{I'}, I')) - \log(l(\widehat{\beta}, I)))|\} = 0$ then the thesis trivially holds.

Now suppose $\max\{|(\log(l(\widehat{\beta}_{I'}, I')) - \log(l(\widehat{\beta}, I)))|\} \neq 0$.

The proposed algorithm trivially satisfies conditions 1-3 in (Chakraborty and Chaudhury, 2008, Pag. 686). Define

$$\Delta_k = \max\{|\tau_k(\log(l(\widehat{\beta}_{I'}, I')) - \log(l(\widehat{\beta}, I)))|\},$$

where I and I' differ only by one coordinate, and $\tau_k = \log(k+1)/D$, as defined in (5). It only remains to prove that

$$\sum_k e^{-n(1-\alpha)\Delta_k} = \infty. \tag{7}$$

Now, (7) holds since

$$\sum_k e^{-n(1-\alpha)\Delta_k} \geq \sum_k \frac{1}{k+1} = \infty,$$

and the trimmed partial likelihood $l(\beta, I)$ is identifiable. Consequently, all conditions of Theorem 1 in Chakraborty and Chaudhury (2008) are satisfied, and the thesis follows. □

An implication of Theorem 1 is that if the number of iterations $k_{\max}$ is large enough, the algorithm converges to the global maximum for (4), provided the maximum partial likelihood estimators computed on each subset are the true global maxima for each subset. The initial set I is consequently irrelevant if $k_{\max}$ is large enough. This is confirmed by our experience: Algorithm 1 is not heavily dependent on the starting values, and even when the initial set I is contaminated, outliers are dropped quite soon. Nevertheless, we adopt a multistart strategy in order to increase the odds of finding soon (i.e., for smaller values of $k_{\max}$) the global maximum: the algorithm is replicated (say 10 times) from different randomly chosen starting solutions and the solution corresponding to the largest trimmed partial likelihood retained.

We now discuss standard errors for $\widehat{\beta}$. Since there is additional uncertainty related to the composition of the set of clean observations, the standard errors obtained from the score equations of the selected set I may grossly underestimate the true standard errors. A formal derivation of the standard errors for the trimmed estimators is cumbersome. On the other hand, we can propose a simple strategy based on the bootstrap. There are different approaches to bootstrap for censored data (refer for instance to Davison and Hinkley (2006)). In our implementation we used a simple case resampling. The n vectors (t_i, δ_i, X_i) are resampled with replacement and the robust Cox model is estimated on the resampled data. The operation is repeated a large number of times (say $B = 999$ times). Standard errors and confidence intervals can be directly estimated from the vector of estimates obtained after resampling. See also Burr (1994).

2.2 Choosing the trimming level

A general rule for choosing the trimming level is still an open issue in robust statistics. It is anyway acknowledged that if the trimming level is set too low malicious outliers can break down the estimates; while on the other hand, a high level may lead to a mild loss of efficiency, which is not substantial in presence of a moderate to large number of observations. A rule of thumb in related contexts is to set $\alpha = 0.2$ or even $\alpha = 0.25$. In the survival context, where many observations could be censored and the nonparametric baseline for the Cox model should capture mildly aberrant behaviours, we suggest setting α even lower. In the simulations and real data application we will always set $\alpha = 0.1$.

A more formal approach would be given by minimizing an estimate of the asymptotic variance of the estimates, as in Bednarski and Clarke (1993). Unfortunately, an explicit expression is not available for our trimmed Cox model. A formal choice could hence be performed through a bootstrap for each value on a grid of possible trimming levels, but we only expect a mild improvement in performance and do not pursue this approach further in this paper.

3 Robustness properties

In this section we study the global robustness properties of the proposed procedure. An important concept in global robustness is the one of *breakdown point.* Hodges (1967) and Donoho and Huber (1983) define a finite-sample breakdown value as the smallest fraction of outliers that can break down the estimate in a sample. The asymptotic breakdown value (Hampel, 1971) is the breakdown value of a procedure for an infinite number of observations. An infinitesimal asymptotic breakdown point denotes a non-robust procedure.

Applying this concept to proportional hazards regression, we define the finite sample *partial breakdown point* of $\widehat{\beta}$ as

$$\varepsilon_n^p(\widehat{\beta},(t,\delta,X))=\frac{1}{n}\min\left\{m:\sup_{(t',\delta',X')_m}\left\|\widehat{\beta}((t,\delta,X))-\widehat{\beta}((t',\delta',X')_m)\right\|=\infty\right\} \tag{8}$$

where $\widehat{\beta}((t,\delta,X))$ is the vector of regression parameters estimated on the original data, and $\widehat{\beta}((t',\delta',X')_m)$ is instead estimated on the data in which m rows have been replaced by arbitrary values.

It has been pointed out (e.g., Kalbfleisch and Prentice (2002), p. 144) that the addition of a single divergent observation may breakdown the classical MPLE. Hence the partial breakdown point for classical Cox regression is upper bounded as

$$\varepsilon_m^p(\widehat{\beta},(t,\delta,X)))\leq\frac{1}{n},$$

and hence infinitesimal.

Let us now focus on our approach. Let us suppose that $(n\alpha)$ observations are contaminated. If Algorithm 1 is applied with a trimming level exactly equal to α, the outliers can be discarded and do not contribute to the computation of the estimate. If there is just one additional contaminated observation, this can not be trimmed, and the procedure breaks down.

Hence, for Algorithm 1,

$$\varepsilon_m^p(\widehat{\beta}, (t, \delta, X)) \leq \frac{n\alpha + 1}{n} = \alpha + \frac{1}{n}$$

Consequently, the maximal asymptotic partial breakdown point for our procedure is positive and coincides with the trimming level.

4 Simulations

We now illustrate the properties of the trimmed procedure with a brief simulation study.

We simulate clean data from the Cox model

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_1 x_{1i} + \beta_2 x_{2i}), \tag{9}$$

with $\lambda_0(t) = 1$, x_{1i} generated randomly from a uniform random variable and x_{2i} from a Bernoulli with parameter $p = 0.4$, and sample size n.

We then record the largest and smallest values of $\exp(\beta_1 x_{1i} + \beta_2 x_{2i})$, call them HR_{low} and HR_{high}, as simulated under model (9). Then, we select at random a proportion π_{cont} of observations, and regardless of their true covariate configuration we generate their survival times according to

$$\lambda(t|X_i) = \lambda_0(t)(u_i HR_{low} + (1 - u_i)HR_{high}),$$

where u_i is a random draw from a Bernoulli with parameter 0.5. These observations can be considered as outlying (and possibly influent).

In order to simulate censoring, we generate a vector $C_1, \ldots, C_n$ of i.i.d. random variables, uniformly distributed in $[0, T_{\max}]$. The parameter $T_{\max}$ is set as a function of β so to have a censoring proportion π_{cens} of approximately either 0.05 or 0.25.

We then fit our Algorithm 1 on the fabricated data, initialized from a randomly chosen starting solution and trimming level $\alpha = 0.1$.

We also compare our approach with Bednarski and Sasieni methods (as in R packages `coxrobust` and `coxphw`), other than the classical Cox model, by fitting the three competitors on the same fabricated data.

We evaluate the performance of each method by recording the Sum of Squared Errors (SSE), i.e.,

$$SSE = \left(\widehat{\beta}_1 - \beta_1\right)^2 + \left(\widehat{\beta}_2 - \beta_2\right)^2$$

We replicate data generation and model fitting for 5000 times for each simulation setting, and finally report the median SSE in Table 1 for each technique.

		Cox	Bednarski	Sasieni	Trim	Cox	Bednarski	Sasieni	Trim
		$n = 250$				$n = 500$			
π_{cont}	π_{cens}	$\beta_1 = 1, \beta_2 = -1$							
0	0.05	0.050	0.061	0.076	0.086	0.025	0.030	0.038	0.039
0	0.25	0.060	0.075	0.097	0.091	0.030	0.036	0.046	0.043
0.05	0.05	0.130	0.079	0.083	0.079	0.122	0.049	0.046	0.045
0.05	0.25	0.085	0.083	0.098	0.082	0.055	0.048	0.054	0.044
0.075	0.05	0.192	0.098	0.095	0.089	0.189	0.068	0.056	0.048
0.075	0.25	0.105	0.100	0.111	0.093	0.080	0.063	0.069	0.058
0.1	0.05	0.267	0.199	0.153	0.090	0.260	0.098	0.072	0.062
0.1	0.25	0.137	0.175	0.181	0.095	0.113	0.089	0.090	0.080
π_{cont}	π_{cens}	$\beta_1 = 1, \beta_2 = -3$							
0	0.05	0.072	0.087	0.131	0.150	0.037	0.043	0.064	0.064
0	0.25	0.089	0.100	0.164	0.160	0.041	0.049	0.081	0.074
0.05	0.05	1.711	0.660	0.232	0.109	1.785	0.661	0.207	0.048
0.05	0.25	0.762	0.435	0.324	0.118	0.826	0.440	0.282	0.103
0.075	0.05	2.273	1.011	0.402	0.098	2.413	1.036	0.385	0.048
0.075	0.25	1.163	0.708	0.553	0.103	1.214	0.702	0.537	0.157
0.1	0.05	2.904	1.445	0.648	0.100	2.970	1.449	0.618	0.048
0.1	0.25	1.612	1.058	0.897	0.149	1.634	1.040	0.881	0.366
π_{cont}	π_{cens}	$\beta_1 = 3, \beta_2 = -3$							
0	0.05	0.085	0.099	0.154	0.178	0.042	0.048	0.072	0.076
0	0.25	0.111	0.127	0.208	0.204	0.053	0.061	0.100	0.095
0.05	0.05	2.205	1.061	0.380	0.140	2.649	1.244	0.404	0.108
0.05	0.25	0.686	0.639	0.665	0.203	0.861	0.701	0.702	0.211
0.075	0.05	3.328	1.795	0.783	0.159	3.609	1.960	0.801	0.101
0.075	0.25	1.306	1.182	1.368	0.251	1.425	1.226	1.404	0.460
0.1	0.05	4.322	2.637	1.357	0.180	4.464	2.699	1.353	0.093
0.1	0.25	2.013	1.887	1.980	0.306	2.095	1.904	2.061	0.509

Table 1: Median SSE for different proportions of contamination π_{cont}, censoring π_{cens}, sample size n and $\beta = (\beta_1, \beta_2)$. Cox stands for classical Cox regression, Trim for our proposal. The results are based on 5000 replications.

It is interesting to note that under no contamination all methods more or less yield the same SSE, with robust methodologies slightly outperformed by classical Cox regression. This could be expected since classical Cox regression uses all of the available information.

Under contamination, Cox regression breaks down. Robust methods, instead, lead to SSE values much smaller than the one obtained with Cox regression. Sasieni method seems to be slightly more robust than Bednarski method, at least in this setting. This is particularly evident when the proportion of censoring is small. A deeper comparison between the two methods can be found in Bednarski and Nowak (2003).

Finally, at least in this settings, our proposed method always returns the smallest SSE values in contaminated settings. The differences are more and more evident as π_{cont} grows. We discard possible outliers (i.e., give a zero weight to them). Consequently, when the method succeds in putting all outliers in the trimmed set, the results are basically those that would have been obtained by recording only a sample of clean observations of size $n(1-\alpha)$.

5 Prostate cancer data

Our example comes from Andrews and Herzberg (1985) and was used by Nardi and Schemper (1999) to illustrate outlier detection in the Cox model by means of residuals analysis.

Survival times were recorded for $n = 297$ patients with prostate cancer, together with seven binary prognostic factors: treatment, performance status (PS), serum Hemoglobin level in g/100 ml ($>$ 12-$\leq$ 12), weight index, history of cardiovascular disease, tumor size (Small-Large), and a combined index of tumor stage and grade. A detailed description of the data can be found in Andrews and Herzberg (1985).

The results for the classical Cox model fit to the full data set are shown in Table 2.

We now perform outlier detection. Log-odds residuals are defined as $w_i = \log\left(\frac{\widehat{S}(t_i)}{1-\widehat{S}(t_i)}\right)$. Under the null hypothesis of no contamination for the i-th subject, w_i asymptotically follows a standard logistic distribution for subjects experiencing the event. There a different options to accomodate censored subjects, more details can be found in Nardi and Schemper (1999).

Through the computation of log-odds residuals, four patients are flagged

Variable	*Parameter Estimate*	*Standard Error*	p-value	*Hazard Ratio*	95%*Confidence Limits*
History	0.5096	0.1457	0.0005	1.665	1.251 – 2.215
Size	0.7835	0.2093	0.0002	2.189	1.453 – 3.299
Grade	0.6940	0.1542	< 0.0001	2.002	1.479 – 2.708
Weight	-0.3265	0.1498	0.0293	0.721	0.538 – 0.968
Hemoglobin	-0.2462	0.1838	0.1805	0.782	0.545 – 1.121
PS	0.1405	0.2495	0.5731	1.151	0.706 – 1.187
Treatment	0.0518	0.1676	0.7572	1.053	0.758 – 1.463

Table 2: Summary of Cox model for the prostate cancer data.

as outliers at level 0.05. A summary is given in the first four rows of Table 3.

Patient	t_i	*Status*	w_i^{Cox}	w_i^{trim}
50	72	*Censored*	-4.51	-4.41
293	76	*Censored*	-4.24	-4.93
437	0	*Dead*	3.94	3.96
451	4	*Dead*	3.70	3.87
243	1	*Dead*	3.56	3.81
362	1	*Dead*	3.46	3.90

Table 3: Outliers for prostate cancer data according to classical Cox model (first four) and to trimmed Cox model (all six), with their log-odds residuals w_i^{Cox} and w_i^{trim}.

The two censored patients also have high values for the DFBetas (Collett, 2003), so they can be deemed as influent outliers.

We can now proceed with an analysis based on our proposed method. The estimates and standard errors for a trimming level $\alpha = 0.1$ are reported in Table 4. We can observe that robustly estimated hazard ratios for the significant covariates are slightly more extreme than the hazard ratios estimated by the Cox model. Consequently, the effect of risk factors like size and grade may be underestimated by Cox model, resulting in overly optimistic survival prognosis and risk assessment for prostatic cancer patients. As a consequence of trimming, standard errors and confidence intervals for the hazard ratios are slightly larger.

$Variable$	$Parameter$ $Estimate$	$Standard$ $Error$	p-value	$Hazard$ $Ratio$	95%$Confidence$ $Limits$
$History$	0.5456	0.1825	0.0028	1.726	1.301 – 2.638
$Size$	1.0031	0.2248	< 0.0001	2.727	1.633 – 3.864
$Grade$	0.8753	0.1971	< 0.0001	2.400	1.496 – 3.182
$Weight$	-0.3872	0.1834	0.0347	0.679	0.486 – 0.992
$Hemoglobin$	-0.3090	0.2593	0.2334	0.734	0.451 – 1.257
PS	-0.0608	0.4066	0.8811	0.941	0.557 – 2.646
$Treatment$	0.0916	0.2016	0.6495	1.096	0.735 – 1.590

Table 4: Summary of trimmed Cox model for the prostate cancer data, trimming level $\alpha = 0.1$. Estimates of standard errors and confidence limits are based on a non-parametric bootstrap with $B = 999$ replications.

We have repeatedly fit the model with different trimming levels, comparing the results shown in Table 4 also with the results obtained fixing $\alpha = 0.05$, $\alpha = 0.15$ and $\alpha = 0.20$. The estimates are fairly stable with respect to the trimming level.

For what concers outlier detection, the robust model identifies the same four outliers as before, plus an additional two which were before masked by the fact that the other outliers influenced the MPLE. The six outliers identified are given in Table 3.

Interestingly enough, if we set the trimming level as $\alpha = 4/n$, our method leads to trim the very same observations that were identified by Cox model as outliers. Nevertheless, residual analysis applied to the resulting model leads to the identification of the other 2 outliers in Table 3.

As a final comparison, we plot in the right panel of Figure 1 the log-odds residuals for Cox model, and in the left panel the log-odds residuals for the trimmed Cox model (with $\alpha = 0.01$). It can be appreciated that, after trimming, generally extreme residuals are more extreme while the central part of the histogram becomes more concentrated and more peaked than before.

6 Discussion

We have proposed a semiparametric model that allows for a fraction of outlying observations. Inference on the hazard ratios β is performed by trimming

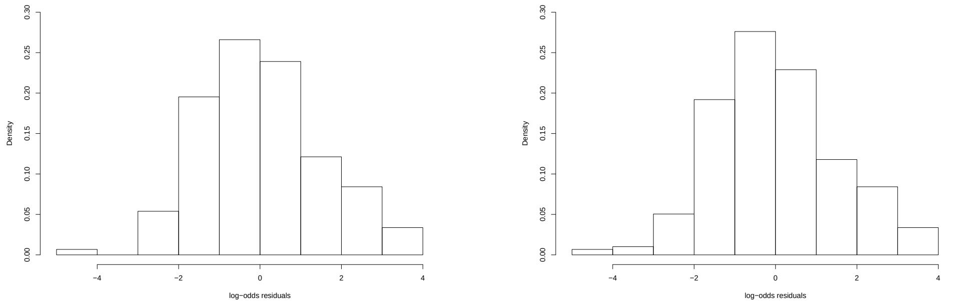

Figure 1: Histograms of log-odds residuals for Cox model (right panel) and trimmed Cox model (left panel) for prostate cancer data

the observations with smallest contributions to the likelihood. We have argued formal global robustness properties of our approach. The proposal has been seen in simulations to perform approximately like the classical MPLE under no contamination, and to compare very well with other robust techniques, with respect to SSE, when data are contaminated.

As illustrated in the real data application, outlier identification is expected to be more effective after robust estimation. Robust survival analysis can help to better understand relationships between covariates and survival times in the population, and to sort out outlying subjects for further study.

Acknowledgements

This work is part of a project of the Ph.D. program in Methodological Statistics of the Department of Statistics, Sapienza - University of Rome. Authors are grateful to Prof. Giovanna Jona Lasinio for organization and support. The authors are also grateful to Prof. Alessandra Nardi for suggestions on a first draft.

References

D.F. Andrews and A.M. Herzberg (1985). *Data : a collection of problems from many fields for the student and research worker*. Springer-Verlag, New York.

C. Becker and U. Gather (1999). The masking breakdown point of multivariate outliers. *Journal of the American Statistical Association*, **94**, 945–955.

T. Bednarski (1989). On sensitivity of Cox's estimator. *Statistics and Decisions*, **7**, 215–228.

T. Bednarski (1993). Robust estimation in Cox regression model. *Scandinavian Journal of Statistics*, **20**, 213–225.

T. Bednarski (2007). On a robust modification of Breslow's cumulated hazard estimator. *Computational Statistics and Data Analysis*, **52**, 234–238.

T. Bednarski and B.R. Clarke (1993). Trimmed likelihood estimation of location and scale of the normal distribution. *Australian Journal of Statistics*, **35**, 141–153.

T. Bednarski and M. Nowak (2003). Robustness and efficiency of Sasieni-type estimators in the Cox model. *Journal of Statistical Planning and Inference*, **115**, 261–272.

D. Burr (1994). A comparison of certain bootstrap confidence intervals in the Cox model. *Journal of the American Statistical Association*, **89**, 1290–1302.

K.C. Cain and N.T. Lange (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, **40**, 493–499.

B. Chakraborty and P. Chaudhury (2008). On an optimization problem in robust statistics. *Journal of Computational and Graphical Statistics*, **17**, 683–702.

B.R. Clarke (2000). An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes. *Probab. Statist.*, **20**, 25–50.

D. Collett (2003). *Modelling Survival Data in Medical Research.* Chapman & Hall/CRC, New York.

D. R. Cox (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

A. C. Davison and D. Hinkley (2006). *Bootstrap Methods and their Applications*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.

D.L. Donoho and P.J. Huber (1983). The notion of breakdown point. In: P. Bickel, K. Doksum, and J.L.Jr. Hodges, eds., *A Festschirift for Erich L. Lehmann*, 157–184. Wadsworth.

A. Farcomeni (2009). Robust Double Clustering: A Method Based On Alternating Concentration Steps. *Journal of Classification*, **26**, 77–101.

P. M. Grambsch and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.

F.R. Hampel (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, **42**, 1887–1896.

D. M. Hawkins (1980). *Identification of Outliers*. Chapman and Hall, London.

J.L.Jr. Hodges (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, vol. 1, 163–186. Univ. California Press.

M Hubert, Rousseeuw P.J., and S. Van Aelst (2008). High-Breakdown Robust Multivariate Methods. *Statistical Science*, **23, 1**, 92–119.

J.D Kalbfleisch and R.L Prentice (2002). *The Statistical Analysis of the failure time data.* Wiley-interscience, New York.

R. A. Maronna, R. D. Martin, and V. J. Yohai (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

C.E. Minder and T. Bednarski (1996). A robust method for proportional hazards regression. *Statistics in Medicine*, **15**, 1033–1047.

A. Nardi and M. Schemper (1999). New Residuals for Cox Regression and Their Application to Outlier Screening. *Biometrics*, **55**, 523–529.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

P.J. Rousseeuw and K. van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

S. Samuels (1978). *Robustness for survival estimators*. Ph.D. thesis, Department of Biostatistics, University of Washington.

P.D. Sasieni (1993a). Maximum weighted partial likelihood estimates for the Cox model. *Journal of the American Statistical Association*, **88**, 144–152.

P.D. Sasieni (1993b). Some new estimators for Cox regression. *Annals of Statistics*, **21**, 1721–1759.

M. Schemper, S. Wakounig, and G. Heinze (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, **28**, 2473–2489.

D. Schoenfeld (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**, 239–241.

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

M.G. Valsecchi, D. Silvestri, and P. Sasieni (1996). Evaluation of long-term survival: use of diagnostics and robust estimators with Cox's proportional hazards models. *Statistics in Medicine*, **15**, 2763–2780.

P.J.M. Verweij and H.C. Van Houwelingen (1993). Cross-validation in survival analysis. *Statistics in Medicine*, **12**, 2305–2314.