Robust estimation for the Cox regression model based on trimming

Alessio Farcomeni[1,2] and **Sara Viviani***[,3,4]

[1] Department of Public Health and Infectious Diseases, Sapienza – University of Rome, Italy
[2] Department of Public Health Sciences, University of Hawaii, Honolulu, USA
[3] Department of Statistics, Sapienza – University of Rome, Italy
[4] School of Nursing and Dental Hygiene, University of Hawaii, Honolulu, USA

Received 11 January 2011, revised 8 July 2011, accepted 10 August 2011

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 a Metropolis-type maximization routine, and show its convergence to a global optimum. We discuss global robustness properties of the approach, which is illustrated and compared through simulations. We finally fit the model on an original and on a benchmark data set.

Keywords: Cox model; Hepatic encephalopathy; Outliers; Robustness; Trimming.

1 Introduction

Robust estimation is a well-developed topic in different areas of statistics (e.g. Hawkins, 1980; Heritier et al., 2009; Hubert et al., 2008; Maronna et al., 2006). 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. Even a single malicious observation can be unduly influent. A single observation can lead to violation of the assumption of proportionality of hazard, and this departure may not be detected by common checking methods. The influence function of Cox regression model 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, i.e. lack of identification of true outliers (Becker and Gather, 1999), and swamping, i.e. identification of false positives. This is due to the fact that outliers influence parameter estimates. Robust estimates are a pre-requisite for distance-based outlier detection procedures.

*Corresponding author: e-mail: sara.viviani@uniroma1.it, Phone: +39-06-49910499, Fax: +39-06-4959241

Among the few attempts to robustify the Cox model 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, with different possible weighting schemes. Schemper et al. (2009) demonstrated the usefulness of weighted partial likelihood for computing average hazard ratios in the presence of non-proportionality of hazards.

The double weighting approach of Bednarski (1993) was refined by 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).

We focus on contamination through outliers, and do not consider robustness with respect to other violations of the model assumptions (e.g. the proportionality of hazards assumption).

To motivate, consider an example on simulated data. We generate $n = 60$ observations from the Cox model with constant baseline and a standard log-normal covariate, with log hazard ratio set equal to $\beta = 0.60$ and uniform censoring. We then add a censored observation with survival time equal to the maximum observed survival and a value of the covariate equal to the observed maximum plus a uniform variate. On this data, classical Cox model gives $\hat{\beta} = -0.06$, with $p = 0.519$. Even if not significant, the sign of the estimate is wrong. The Sasieni (1993b) approach assigning weights to the contributions of the likelihood proportional to the number of patients at risk yields a negative and not significant estimate ($\hat{\beta} = -0.01$, $p = 0.98$). Bednarki's estimator instead leads to $\hat{\beta} = 0.22$, with $p = 0.010$. When we set our trimming approach to remove one observation, we get $\hat{\beta} = 0.74$ with $p = 0.016$.

Outliers in survival studies can be interpreted as by 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 kinds 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. Furthermore, the covariates can be separately explored with methods for detecting multivariate outliers.

For a deeper discussion about robust methods in biometrical applications refer to the recent book (Heritier et al., 2009, Chapter 7) and to Farcomeni and Ventura (2011).

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 real data examples 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 Breslow partial likelihood (Breslow, 1974), which allows for tied survival and censoring times, where

$$l(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' X_i)}{\sum_{j \in R_i} \exp(\beta' S_j)} \right)^{\delta_i}, \tag{1}$$

and R_i is the risk set just before the i-th ordered event time t_i, $S_i = \sum_{l \in \mathcal{D}_i} X_l$ and $\mathcal{D}_i$ is the set of individuals that fail at time t_i. The resulting maximum partial likelihood estimator (MPLE) is consistent and asymptotically normal under regularity conditions (Andersen and Gill, 1982). 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

$$l_{\text{TRIM}}(\beta) = \max_{I \in H(\alpha)} \prod_{i \in I} \left(\frac{\exp(\beta' X_i)}{\sum_{j \in R_i} \exp(\beta' S_j)} \right)^{\delta_i}, \tag{3}$$

where R_i and S_j are restricted to the observations in I. In words, $\hat{\beta}$ gives 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.

The simple idea of trimming the smallest contributions to the likelihood is seen to lead to robust estimation in the Cox model. 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. It is not straightforward to show asymptotic properties of the robust estimator. We here give an intuition of why, under certain assumptions, the robust estimator should still be consistent

and asymptotically normal. In Section 4 we report Fig. 1 which illustrates the estimated distribution of the robust estimator, which is reasonable close to normal already for $n = 250$ in the simulation setting adopted. The reason why we expect the trimmed estimator to behave like the classical estimator (with an inflated variance) is that if a separability condition holds (i.e. the hazard rates of the outliers are far enough from the hazard rate which would be given by the model for the clean part for the data and from each other), and if the trimming level is correct, outliers *will* have a small individual contribution to the likelihood by definition. Consequently, they will be sorted out and β will be estimated based only on observations arising from the true model, i.e. just as if the sample size was $n(1–\alpha)$ rather than n.

In order to perform a formal outlier detection we rely on log-odds residuals. Log-odds residuals are defined as

$$w_i = \log\left(\frac{\hat{S}(t_i)}{1-\hat{S}(t_i)}\right) \tag{4}$$

when the i-th subject experiences the event, and

$$\tilde{w}_i = w_i - \log(1+e^{w_i})(1+e^{w_i})e^{-w_i} \tag{5}$$

for censored observations. 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. More details can be found in Nardi and Schemper (1999).

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 special purpose algorithms, e.g. using repeated concentration steps (Rousseeuw and van Driessen, 1999; see also Farcomeni, 2009). Here, individual contributions to the partial likelihood have cumbersome expressions (Verweij and Van Houwelingen, 1993), making it difficult to use these algorithms. A different method, suitable for our problem, has been recently described by Chakraborty and Chaudhury (2008), and we will now adapt it to the survival context.

Let

$$l(\beta, I) = \prod_{i\in I}\left(\frac{\exp(\beta' X_i)}{\sum_{j\in R_i}\exp(\beta' S_j)}\right)^{\delta_i}, \tag{6}$$

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, $\hat{\beta}$ 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. Let us indicate with D a tuning parameter to be explained below, and with $k_{\max}$ the maximum number of iterations. Furthermore, let $I^{(k)}$ the subset of observations corresponding to the k-th iteration, and with $C(I^{(k)})$ the set of observations which do not belong to $I^{(k)}$.

Algorithm 1 General iteration for maximization of trimmed partial likelihood

for $k = 1,\ldots,k_{\max}$ **do**

sample a candidate $i' \in C(I^{(k)})$ and a candidate $i \in I^{(k)}$ with uniform probabilities.

Let $I_{\text{cand}} = (I^{(k)} \backslash i) \cup i'$, i.e. obtained by replacing i with i' in $I^{(k)}$.

Let

$$\tau_k := \log(k+1)/D \tag{7}$$

$$\hat{\beta} := \arg\max_{\beta} l(\beta, I^{(k)})$$

$$\hat{\beta}_{\text{cand}} := \arg\max_{\beta} l(\beta, I_{\text{cand}})$$

$$p := \min\left(e^{\tau_k(\log(l(\hat{\beta}_{\text{cand}}, I_{\text{cand}})) - \log(l(\hat{\beta}, I^{(k)})))}, 1\right) \tag{8}$$

Let U be a random draw from a Bernoulli with parameter p.

if $U = 1$ **then**

$I^{(k+1)} = I_{\text{cand}}$.

else

$I^{(k+1)} = I^{(k)}$.

end if

end for

The algorithm follows an acceptance-rejection scheme similar to Metropolis–Hastings in MCMC. At each iteration, a new proposal for the optimal subset is chosen at random by changing a single entry of the current subset $I^{(k)}$. 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 escapes local optima.

This probability p, given in expression (8), 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 escape 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. This probability of acceptance is proportional to the likelihood ratio between the proposal I_{cand} and the current subset $I^{(k)}$ when the new proposal corresponds to a slightly lower likelihood than the current subset.

There are two tuning parameters: $k_{\max}$ and D. The first one controls the maximum number of iterations, and should be set large enough that in the last few iterations 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 $k_{\max} = 10\,000$ and $D = 0.1n(1–\alpha)$. We have also found that a more efficient algorithm is obtained with an additional stopping rule that terminates iterations when the current maximum is not updated for r iterations in a row. We set $r = 50$.

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 superset of all sets containing the indices of observations which lead to the global optimum of the trimmed partial likelihood.

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

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

The proposed algorithm satisfies conditions 1–3 in (Chakraborty and Chaudhury, 2008, p. 686). Define

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

where I and I' differ only by one coordinate, and $\tau_k = \log(k+1)/D$, as defined in (7). It only remains to prove that

$$\sum_k e^{-n(1-\alpha)\Delta_k} = \infty. \tag{9}$$

Now, (9) 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 by 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 (6), 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.

As suggested by one of the referees, the algorithm may converge quicker if uniform sampling is replaced by weighted sampling, with weights proportional e.g. to the leave-one-out partial likelihood or inversely proportional to the log-odds residuals. A lower bound is recommended so that no observation receive a numerically zero weight.

We now discuss standard errors for $\hat{\beta}$. 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. Note that we have considered the partial likelihood with ties, so that application of the bootstrap is straightforward. Algorithm 1 is replicated in each bootstrap sample, possibly with a starting solution based on the estimates obtained from the original data to speed up convergence. 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).

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 sample sizes are often moderate, many observations could be censored and the nonparametric baseline for the Cox model should capture mildly aberrant behaviours, we suggest setting α by default to 5 or 10%.

A data-driven choice of $0 \leq \alpha \leq \alpha_{\max}$ can be performed through the bootstrap. For each value of $\alpha = i/n$, $i = 0, \ldots, n\alpha_{\max}$, we record the the sum of squares of the differences between the parameter estimates and the estimates on the resampled observations, averaged over the $B = 999$ iterations. The final estimates are then those corresponding to the (smallest) value of α minimizing the average sum of squares. We pursue this approach in Section 5. Note that one this approach approximately corresponds to minimizing an estimate of the asymptotic variance of the estimates, as in Bednarski and Clarke (1993).

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 $\hat{\beta}$ as

$$\varepsilon_n^p(\hat{\beta}, (t, \delta, X)) = \frac{1}{n} \min \left\{ m : \sup_{(t', \delta', X')_m} \left\| \hat{\beta}((t, \delta, X)) - \hat{\beta}((t', \delta', X')_m) \right\| = \infty \right\} \tag{10}$$

where $\hat{\beta}((t, \delta, X))$ is the vector of regression parameters estimated on the original data, and $\hat{\beta}((t', \delta', X')_m)$ is instead estimated on the data in which m rows have been replaced by arbitrary values. Equation (10) directly measures the variation in the estimates which can result from arbitrary contamination by outliers.

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. This fact can be used to deduce that the partial breakdown point as defined in (10) for classical Cox regression is upper bounded as

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

and hence infinitesimal. Note that the breakdown point decreases with n: a single observation can arbitrarily ruin the estimates even in the face of a huge sample size. With a different definition of breakdown point, this is the same value obtained from estimating a population mean with the sample mean.

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 cannot be trimmed, and the procedure breaks down. Hence, for Algorithm 1,

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

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

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{1i} + \beta_2 X_{2i}), \tag{11}$$

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 $\mathrm{HR}_{\mathrm{low}}$ and $\mathrm{HR}_{\mathrm{high}}$, as simulated under model (11). 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

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

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`, see Bednarski and Borowicz (2006) and Ploner and Heinze (2009), respectively), other than classical Cox model, by fitting the three competitors on the same fabricated data. For the Bednarski method we use a quadratic weighting function, and option `trunc = 0.9` to allow trimming in the robust estimator. For the Sasieni method we use Breslow weights for all the covariates. All other tuning parameters are set to the default in the `R` functions.

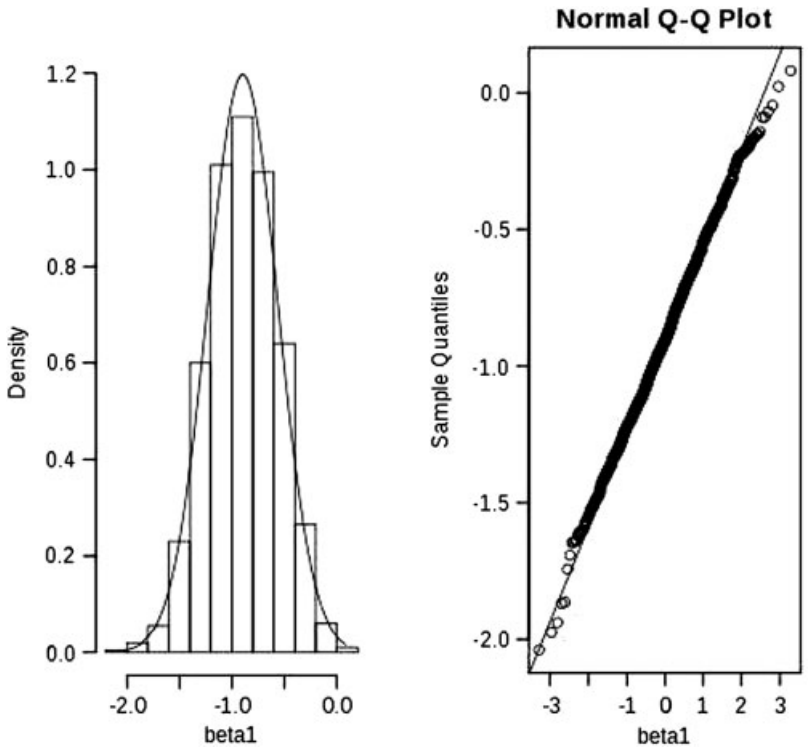

Figure 1 Histogram and normal qqplot for $\hat{\beta}_1$ based on 1000 simulated data sets.

We begin by evaluating the computational demands of our method for increasing sample sizes. Time needed for convergence with non-optimized R code on a laptop with a 2.4 GHz CPU (4 MB Cache) and 1 Gb RAM is given in Table 1.

It can be seen that computing time increases with the sample size at a slow rate, and is always reasonable. We underline that we report the sample times needed for convergence of the algorithm on a single data set. The computational demand is of course increased once using the bootstrap in order to estimate standard errors. Few hours are always sufficient anyway for any data set.

Second, we give a sample evaluation of the distribution of our estimators. We repeat data generation and model fit with $\alpha = \pi_{cont} = 10\%$, $n = 250$, $\beta = (1,-1)$ and $\pi_{cens} = 5\%$ and plot in Fig. 1 the histogram (with normal density superimposed) and a qqplot based on 1000 data sets for $\hat{\beta}_1$. It can be seen that the sampling distribution of the estimator is close to normality at least in this setting.

We now evaluate the performance of each method by recording the sum of squared errors (SSE), i.e.

$$\text{SSE} = (\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2.$$

Note that the SSE depends on the prior standardization of the variables, as the β's are equally weighted. In our simulation we do not standardize the generated covariates.

We replicate data generation and model fitting 5000 times for each simulation setting, and finally report the median SSE for each technique in Tables 2–4.

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 available information most efficiently.

Table 1 Time (in seconds) for convergence of the proposed algorithm on simulated data sets of increasing size, with $\beta = (1, 1)$, $\pi_{cont} = 5\%$, $\alpha = 10\%$ and $\pi_{cens} = 5\%$.

N	100	200	300	400	500	600	700	800	900	1000
Time (s)	25	27	29	30	35	39	39	41	48	49

Table 2 Median SSE for different proportions of contamination π_{cont}, censoring π_{cens}, sample size n and $\beta = (1, -1)$.

π_{cont}	π_{cens}	Cox	Bednarski	Sasieni	Trim	Cox	Bednarski	Sasieni	Trim
		$n = 250$				$n = 500$			
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
0.15	0.05	0.407	0.196	0.148	0.198	0.395	0.174	0.120	0.286
0.15	0.25	0.215	0.182	0.186	0.153	0.196	0.147	0.146	0.137

Cox stands for classical Cox regression, Trim for our proposal. The results are based on 5000 replications.

Table 3 Median SSE for different proportions of contamination π_{cont}, censoring π_{cens}, sample size n and $\beta = (1, -3)$.

π_{cont}	π_{cens}	Cox	Bednarski	Sasieni	Trim	Cox	Bednarski	Sasieni	Trim
		$n = 250$				$n = 500$			
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
0.15	0.05	3.792	2.174	1.144	0.116	3.880	2.206	1.149	0.051
0.15	0.25	2.634	1.800	1.485	0.275	2.697	1.814	1.476	0.659

Cox stands for classical Cox regression, Trim for our proposal. The results are based on 5000 replications.

Table 4 Median SSE for different proportions of contamination π_{cont}, censoring π_{cens}, sample size n and $\beta = (3, -3)$.

π_{cont}	π_{cens}	Cox	Bednarski	Sasieni	Trim	Cox	Bednarski	Sasieni	Trim
		$n = 250$				$n = 500$			
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
0.15	0.05	5.811	3.986	2.499	0.263	6.036	4.146	2.579	0.122
0.15	0.25	3.416	3.224	3.931	0.740	3.569	3.311	4.118	1.785

Cox stands for classical Cox regression, Trim for our proposal. The results are based on 5000 replications.

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 the 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, and the behaviour does not change in the case $\pi_{cont} > \alpha$. Reasonably, we obtain this as we give a zero weight to outliers. Consequently, when the method succeeds in selecting all outliers for the trimmed set, the results are basically those that would have been obtained by recording only a sample of clean observations of size $n(1–\alpha)$.

Table 5 Median SSE for misspecified model with censoring π_{cens}, sample size n and different β vectors.

β	π_{cens}	Cox	Bednarski	Sasieni	Trim	Cox	Bednarski	Sasieni	Trim
		$n = 250$				$n = 500$			
$(1, -1)$	0.05	0.034	0.034	0.043	0.027	0.021	0.021	0.021	0.015
$(1, -1)$	0.25	0.030	0.041	0.051	0.039	0.025	0.027	0.029	0.022
$(1, -3)$	0.05	0.109	0.161	0.182	0.048	0.110	0.165	0.191	0.057
$(1, -3)$	0.25	0.151	0.162	0.157	0.108	0.157	0.170	0.161	0.117
$(3, -3)$	0.05	1.434	1.637	1.613	0.972	1.462	1.621	1.630	1.161
$(3, -3)$	0.25	1.992	1.666	1.188	1.496	2.006	1.674	1.206	1.694

Data were generated using X_1 and X_2, but only X_1 was used for fitting. Cox for classical Cox regression, Trim for our proposal. The results are based on 5000 replications.

Table 6 Power and probability of type I error of the outlier test applied after robust estimation. The results are based on 1000 replications.

	$n = 250$		$n = 500$	
	Power	Type I errors	Power	Type I errors
$\beta = (1, -1)$	0.279	0.015	0.291	0.015
Cox				
Trim	0.290	0.017	0.297	0.016
Sasieni	0.103	0.101	0.101	0.103
Bednarski	0.411	0.112	0.426	0.098
$\beta = (3, -3)$				
Cox	0.489	0.001	0.491	0.001
Trim	0.500	0.044	0.502	0.033
Sasieni	0.257	0.259	0.264	0.259
Bednarski	0.510	0.353	0.520	0.344
$\beta = (1, -3)$				
Cox	0.505	0.000	0.510	0.001
Trimmed	0.513	0.027	0.512	0.033
Sasieni	0.251	0.109	0.252	0.252
Bednarski	0.515	0.003	0.515	0.004

A referee asked also how the trimmed approach compares under misspecification but no contamination. To illustrate this, we still simulate data from a Cox model with two covariates, but then only use one of the two for model fitting. The final estimates are obviously biased since an important covariate is omitted. The median SSE is reported in Table 5.

It can be seen that our trimmed approach leads to the lowest SSE except for the case $\beta = (3, -3)$, $\pi_{cens} = 0.25$, where the Sasieni method seems to be slightly better.

For the sake of completeness, we apply the outlier test based on log-odds residuals (as defined in (4) at the end of Section 2, with the bias correction (5) for censored observations) and compare the

type I error rate (i.e. proportion of false discoveries) and power (i.e. proportion of correctly identified outliers) after estimation with the four competitors in Table 6. The outlier tests are used at a nominal level of 5%.

The outlier test applied after maximizing the partial likelihood leads to a very specific test, as only outliers clearly standing out are identified while others are masked. Sasieni and Bednarski approaches do not seem to lead to satisfactory outlier tests, as the probability of type I error is well above the nominal level in all cases for Sasieni and in two out of three cases for the Bednarski approach. Note that the Bednarski approach leads to a uniformly better test in terms of power, but at the price of a very high likelihood of misclassification of good observations. For a detailed discussion on outlier identification in related contexts, among other see Cerioli and Farcomeni (2011). It can be noted that trimming leads to a valid test, as the probability of type I error is always below 5%. There is an improvement in power with respect to testing after estimation with the Cox model, but this seems to be very small, at least in this simulation setting.

5 Real data applications

5.1 Hepatic encephalopathy data

Transjugular intrahepatic portosystemic shunt (TIPS) is presently used in the treatment of complications of portal hypertension in cirrhotic patients. This procedure however is known to involve a high risk of hepatic encephalopathy, an alteration of neuropsychiatric functions. For a small fraction of patients, these alterations are refractary to cure. For a more detailed description of the the problem refer for instance to Riggio et al. (2008), where results are reported about a subset of the data we are about to present.

We analyze an original data set related to $n = 179$ patients who underwent TIPS in order to investigate risk factors associated with hepatic encephalopathy. We have recorded three binary predictors: age ($\geq 60, < 60$), serum sodium in mEq/dL (> 140, ≤ 140), and model for end-stage liver disease index (MELD, > 10, ≤ 10). MELD is a combined score of liver functioning, see e.g. Cholongitas et al. (2006). We record time to occurrence of first episode of hepatic encephalopathy, if any. We have observed an event for 57% of patients.

The results for the classical Cox model fit to the full data set are shown in Table 7.

We then use our partial likelihood trimming method, and choose α using the bootstrap approach described above. We end up setting $\alpha = 4/179 \cong 2.23\%$. The final estimates obtained with our robust Cox regression are reported in Table 8.

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 age and MELD may be underestimated by Cox model. Serum sodium could be controlled in order to decrease the risk of hepatic encephalopathy events.

As a consequence of trimming, standard errors and confidence intervals for the hazard ratios are slightly larger. We ran a sensitivity analysis fitting our trimmed Cox regression model with other values of α, obtaining essentially the same results.

Interestingly enough, both Cox and trimmed Cox regression lead to the identification of the same six outliers, which are reported in Table 9. The essential difference is that, even if residuals of the Cox model *can* in this case identify outliers, estimates give full weight to them, and are apparently biased.

In Table 9 we note the presence of three patients censored with very large and two with small follow-up times, which are clearly not predicted well by the models. The last patient has a configuration of covariates which puts him in the highest risk category (more than 60 years old, low sodium, large MELD), hence making the observed time to event of 90 days too large under the estimated model.

Table 7 Summary of Cox model for the hepatic encephalopathy data.

Variable	Parameter Estimate	Standard Error	p-Value	Hazard Ratio	95% Confidence Limits
Age	0.601	0.203	0.003	1.824	1.225–2.717
SerumSodium	−0.629	0.232	0.007	0.533	0.339–0.839
MELD	0.471	0.199	0.018	1.601	1.082–2.368

Table 8 Summary of trimmed Cox model for the hepatic encephalopathy data.

Variable	Parameter Estimate	Standard Error	p-Value	Hazard Ratio	95% Confidence Limits
Age	0.699	0.213	0.001	2.013	1.326–3.058
SerumSodium	−0.744	0.315	0.018	0.475	0.256–0.882
MELD	0.539	0.249	0.030	1.714	1.053–2.791

Table 9 Outliers for prostate cancer data according to classical Cox model and to trimmed Cox model with their log-odds residuals w_i^{Cox} and w_i^{trim}.

Patient	t_i	Status	w_i^{Cox}	w_i^{trim}
35	6	Event	4.53	4.61
37	35	Event	3.48	3.74
95	726	Censored	−3.48	−3.74
98	2035	Censored	−4.42	−4.44
110	1106	Censored	−4.42	−4.44
171	90	Event	−3.79	−3.98

As a final comparison, we compare the computational demand of the trimming algorithm and the exhaustive method that would evaluate all the subsets of cardinality $\lceil n(1-\alpha)\rceil$. The exhaustive method would need the evaluation of 41 356 876 subsets, while our method only evaluates 564 on average (computed over 20 runs from different randomly chosen starting solutions).

5.2 Prostate cancer data

Our second 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 (which was randomized), 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).

Data include patients with a survival time of 0 or 1 days, which may be generally questionable, as the therapy scheme may not have started to come into effect. Thus the early death cannot be

Table 10 Summary of Cox model for the prostate cancer data.

Variable	Parameter Estimate	Standard Error	p-Value	Hazard Ratio	95% Confidence Limits
History	0.509	0.146	<0.001	1.665	1.251–2.215
Size	0.783	0.209	<0.00	2.189	1.453–3.299
Grade	0.694	0.154	<0.001	2.002	1.479–2.708
Weight	−0.326	0.149	0.029	0.721	0.538–0.968
Hemoglobin	−0.246	0.184	0.180	0.782	0.545–1.121
PS	0.1405	0.249	0.573	1.151	0.706–1.187
Treatment	0.52	0.1676	0.757	1.053	0.758–1.463

Table 11 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}.

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 12 Summary of trimmed Cox model for the prostate cancer data, trimming level $\alpha = 3/n$.

Variable	Parameter Estimate	Standard Error	p-Value	Hazard Ratio	95% Confidence Limits
History	0.529	0.146	<0.001	1.697	1.273–2.261
Size	0.794	0.209	<0.001	2.212	1.465–3.336
Grade	0.718	0.155	<0.001	2.050	1.512–2.779
Weight	−0.333	0.151	0.027	0.717	0.533–0.963
Hemoglobin	−0.137	0.184	0.167	0.872	0.540–1.112
PS	0.137	0.249	0.582	1.790	0.703–1.871
Treatment	0.067	0.168	0.688	1.070	0.769–1.487

Estimates of standard errors, p-values and confidence limits are based on a non-parametric bootstrap with $B = 999$ replications.

attributed to this or that treatment. The results for the classical Cox model fit to the full data set are shown in Table 10. Note that our robust approach flags those patients as outliers (see Table 11). It may in fact be reasonable to expect that, according to model (2), those subjects follow a specific hazard which is not dependent on covariates, or at least not as the remaining bulk of data.

We now perform outlier detection. Through the computation of log-odds residuals, four patients are flagged as outliers at level 0.05 summary is given in the first four rows of Table 11.

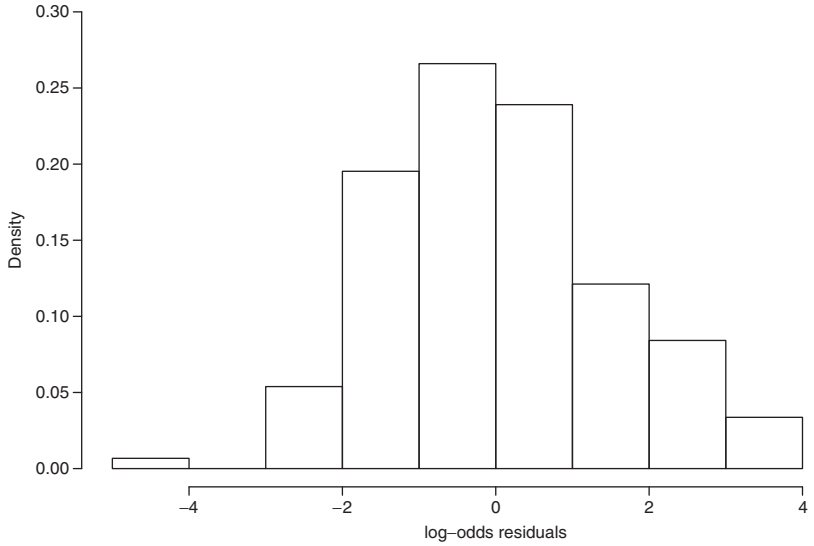

Figure 2 Histogram of log-odds residuals for Cox model for prostate cancer data.

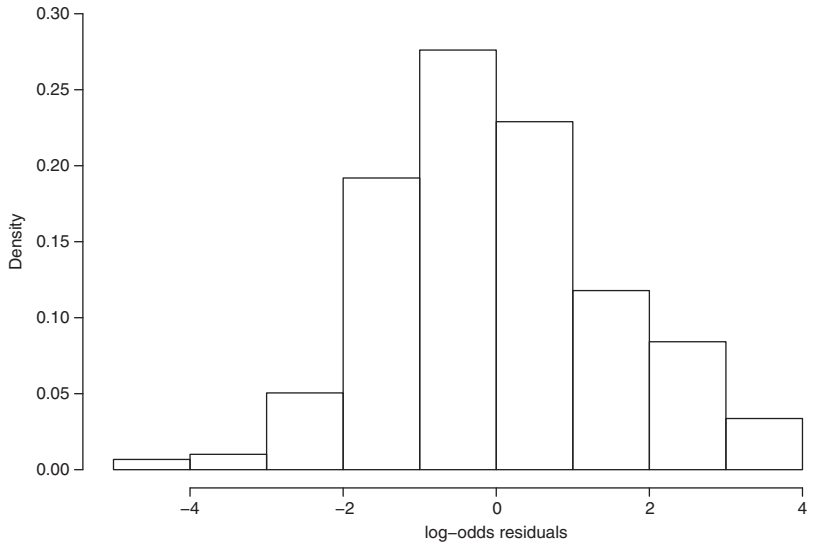

Figure 3 Histogram of log-odds residuals for trimmed Cox model for prostate cancer data.

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

We now proceed with an analysis based on our proposed method. We choose the trimming level α by minimizing the MSE as estimated with the bootstrap. Our procedure identifies $\alpha = 3/297 \cong 1\%$ as the optimal trimming level. The parameters estimates, the corresponding significance level and 95% confidence limits are shown in Table 12.

Interestingly enough, the robustly estimated parameters are very close to the parameters estimated using the Cox model. Nevertheless, these small departures allow us to identify the same four outliers as before, plus an additional two which were before masked by the fact that the outliers

influenced the MPLE. The six outliers identified are given in Table 11. The additional two outliers correspond to patients who *died too early*. In this case, both the very small time to event and the configuration of covariates (e.g. both patients have small tumors) indicate possible outlyingness of these two patients, whose history of treatment may deserve further investigation.

We have finally performed a sensitivity analysis, repeatedly fitting the model with different trimming levels, comparing the results shown in Table 12 also with the results obtained fixing $\alpha = 0.05$ and $\alpha = 0.01$. The results are fairly stable with respect to the trimming level. Interestingly enough, if we set the trimming level as $\alpha = 4/297$, 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 11.

As a final comparison, we plot in Fig. 2 the log-odds residuals for Cox model, and in Fig. 3 the log-odds residuals for the trimmed Cox model. 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.

For the sake of completeness, we point the attention of the reader on the fact that our algorithm evaluated 650 candidate subsets on average (computed over 20 runs from different randomly chosen starting solutions), while the exhaustive 'all subset' method would have evaluated 4 322 340.

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 the observations with smallest contributions to the likelihood. We have argued formal global robustness properties of our approach. In the simulated settings, the proposal has been seen 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.

When data may be contaminated our trimmed approach to survival analysis is expected to avoid bias in estimated effects, and to be more powerful in outlier identification. Robust survival analysis helps better understand relationships between covariates and survival times in the population, and to sort out outlying subjects for further study.

As illustrated in the real data applications, one perspective does not necessarily imply the other. In our first example, while identifying the same outliers, we obtained slightly different coefficient estimates. As a matter of fact, outliers should be at least downweighted in order to obtain reliable estimates. In the prostate cancer example, it was illustrated that the reverse is also true: even if robust and classical parameter estimates are very close, the use of robust estimates to build the outlier test lead to identification of two additional important outliers. This of course is a consequence of the multivariate and complex nature of survival data.

Acknowledgements The authors are grateful to two anonymous referees for stimulating comments, to Prof. Alessandra Nardi and Prof. Marco Alfó for suggestions on a first draft, and to Prof. Giovanna Jona Lasinio for support.

Conflict of Interest
The authors have declared no conflict of interest.

References

Andersen, P. K. and Gill, R. D. (1982). Cox regression model for counting processes: a large sample study. *The Annals of Statistics* **10**, 1100–1120.

Andrews, D. F. and Herzberg, A. M., (1985). *Data: A Collection of Problems From Many Fields for the Student and Research Worker*. Springer, New York.

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

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

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

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

Bednarski, T. and Borowicz, F. (2006). *coxrobust: Robust Estimation in Cox Model*. R package version 1.0.

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

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

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–99.

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

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

Cerioli, A. and Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Computational Statistics and Data Analysis* **55**, 544–553.

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

Cholongitas, E., Marelli, L., Shusang, V., Senzolo, M., Rolles, K., Patch, D. and Burroughs, A. (2006). A systematic review of the performance of the model for end-stage liver disease (MELD) in the setting of liver transplantation. *Liver Transplantation* **12**, 1049–1061.

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

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

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

Davison, C.A. and Hinkley, D. (2006). *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge.

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

Farcomeni, A. (2009). Robust double clustering: a method based on alternating concentration steps. *Journal of Classification* **26**, 77–101.

Farcomeni, A. and Ventura, L. (2011). An overview of robust methods in medical research. *Statistical Methods in Medical Research*, DOI:10.1177/0962280210385865.

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

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

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

Heritier, S., Cantoni,V., Copt, S. and Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. Wiley, Chichester, UK.

Hodges, J. L. Jr., (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **vol. 1**, University of California Press, pp. 163–186.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science* **23**, 92–119.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of the Failure Time Data*. Wiley, New York.

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

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

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

Ploner, M. and Heinze, G. (2009). *coxphw: Weighted Cox regression*. R package version 1.3.

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

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

Riggio, O., Angeloni, S., Salvatori, F.M., De Santis, A., Cerini, F., Farcomeni, A., Attili, A.F. and Merli, M. (2008). Encephalopathy after transjugular intrahepatic portosystemic shunt with polytetrafluoroethylene-covered stent grafts. *American Journal of Gastroenterology*, **103**, 2738–2746.

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

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

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

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

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

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

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

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

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

