Outlier robust modeling of survival curves in the presence of potentially time-varying coefficients

Journal Title
XX(X):2–31

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Jorne Lionel Biccler[1, 2], Martin Bøgsted[1, 2], Stefan van Aelst[2], and Tim Verdonck[3, 4]

Abstract

In time to event studies censoring often occurs and models that take this into account are wide-spread. In the presence of outliers standard estimators of model parameters may be affected such that results and conclusions are not reliable anymore. This in turn also hampers the detection of these outliers due to masking effects. To cope with outliers when using proportional hazard models, we propose to use the Brier score as a loss function. Since the coefficients often vary over time, we focus on the piecewise constant hazard (PCH) model, which can flexibly model time-varying coefficients if a large number of cut-points is used. To prevent overfitting, we add a penalty term that potentially shrinks time-varying effects to constant effects. By fitting the coefficients of the PCH model using a penalized Brier score loss we obtain a robust model that can handle time-varying coefficients. Its good performance is illustrated in a simulation study and using two datasets from practice.

Keywords

Robust statistics, time-varying coefficients, piecewise constant hazard, Brier score, penalization, cut-point selection

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

1 Introduction

Modeling time to event data, also called survival data, is an important objective in many medical studies. Popular models for time to event data that take into account censoring include Cox proportional hazard models, Weibull models, and piecewise constant hazard (PCH) models.[1]

Real data sets often contain some observations that deviate from the pattern of the majority. Such outliers can heavily influence the estimators which may lead to incorrect conclusions in survival analysis. Robust statistical methods have been proposed to obtain estimates which remain reliable when a limited number of outliers are present in the data.[2] However, only few robust methods that can cope with censoring exist and these are rarely used in practice.[3,4]

In addition to obtaining correct estimates in the presence of outliers, also the detection of these atypical observations can be of interest. Outlying observations may be errors, may have been recorded under exceptional circumstances, or could belong to another population. Consequently, outlier detection can for example lead to the discovery of new disease subgroups or prognostic markers. One way to detect outliers is by inspection of the residuals. Since non-robust estimators can be attracted by the outliers, residuals obtained by using these estimates might not be detected as outliers. This is called the masking effect. Moreover, some regular observations might even appear to be outliers, which is known as swamping.[4] On the other hand, residuals based on robust estimates allow to reliably identify outlying observations in the data.

The first robust estimation procedures in survival analysis focused on the coefficients of the widely used Cox proportional hazard model.[5,6] Extensions of and variations on these initially proposed methods include

[1]Department of Heamatology, Aalborg University Hospital
[2]Department of Clinical Medicine, Aalborg University
[3]Department of Mathematics, KU Leuven
[4]Department of Mathematics, University of Antwerp

Corresponding author:
Jorne Lionel Biccler, Department of hematology, Aalborg University Hospital, Mølleparkvej 4, 9100 Aalborg, DK.
Email: jorne.bic@gmail.com

Prepared using sagej.cls

a trimmed estimator of the coefficients and a robust estimator of the cumulative baseline hazard.[7,8] Recently, robust estimation procedures have also been developed for alternatives to the Cox proportional hazard model, e.g. additive hazard and parametric accelerated failure time models.[9–12]

In this paper we propose robust estimators for proportional hazard models by adapting a minimum quadratic distance based robust estimation procedure, which was originally developed for logistic regression.[13] In particular, we use a loss function based on the quadratic error, also called the Brier score, between the patient status (dead or alive) and the predicted probability at specified time-points.[14–16] To take the censoring into account, an estimator based on inverse probability of censoring weighting (IPCW) is applied.[17] IPCW M-estimators have already been used to obtain robust estimators for accelerated failure time models.[11]

In medical studies the effect of covariates often varies over time. For example, the size of a tumor measured at diagnosis tends to have a less pronounced effect after surviving two years post-diagnosis. One model that naturally allows the incorporation of time-varying effects is the PCH model[18,19], which is often used in practice.[20,21] In a PCH model a unique hazard rate is applied in each (predefined) time interval. By utilizing a large number of intervals such models can be made flexible.[22] A drawback of increasing the number of cut-points is that it requires the estimation of a large number of parameters which can lead to instability issues. This can be solved by adding an additional penalty to the time-varying effects. Recent developments concerning the penalization of time-varying Cox proportional hazard models have focused on inducing sparsity in the sense that some covariate effects are set to be constant over time.[23,24] In this paper we induce sparsity in the same sense for PCH models by applying a penalty similar to the group lasso (GL).[25]

By estimating the coefficients of a PCH model using the Brier score with the group lasso-like penalty as loss function we obtain a flexible time-varying model which is more robust to the presence of outliers. To the best of our knowledge, this is the first robust model that explicitly allows time-varying coefficients.

The paper is structured as follows. In Section 2 the PCH model is introduced, the standard maximum likelihood estimator (MLE) is described, and the influence of leverage points is illustrated. In Section 3 we introduce the Brier score to obtain robust estimates and provide some intuition as to why this approach is more robust against leverage points. The penalization of the time-varying coefficients is described in Section 4 and an

optimization algorithm is described in Section 5. To test the performance of the new method a simulation study is performed in Section 6. Finally, in Section 7 we test the new method on data from an AIDS cohort and on a cohort of cutaneous malignant melanoma patients. R scripts for the estimators and analyses are available on https://github.com/JorneBiccler/RobustPCH.

2 Piecewise constant hazard models

The event and censoring times of the ith observation are denoted by T_i and C_i respectively, and the observed right-censored information is $U_i = \min(T_i, C_i)$ and $\Delta_i = \mathbb{1}(T_i \leq C_i)$. Throughout this paper we assume independent right-censoring (possibly conditional on covariates). The data is said to be generated from a PCH model if there exist p time intervals $[\alpha_0 = 0; \alpha_1), [\alpha_1; \alpha_2), \ldots, [\alpha_{p-1}; \alpha_p = \infty)$ and p constants $\tilde{\theta}_{1,0}, \ldots, \tilde{\theta}_{p,0}$ such that the hazard is

$$\alpha_i(t) = \lim_{h \to 0} \frac{\mathbb{P}(T_i \in [t; t+h) | T_i > t)}{h} = \sum_{j=1}^{p} \mathbb{1}(t \in [\alpha_{j-1}; \alpha_j)) \exp(\tilde{\theta}_{j,0}).$$

Inclusion of the effect of q covariates ($\tilde{\boldsymbol{X}}_i \in \mathbb{R}^q$) is typically obtained by assuming a proportional effect on the hazard, i.e.

$$\alpha_i(t) = \sum_{j=1}^{p} \mathbb{1}(t \in [\alpha_{j-1}; \alpha_j)) \exp\left(\tilde{\theta}_{j,0} + \tilde{\boldsymbol{X}}_i^t \tilde{\boldsymbol{\theta}}\right),$$

with $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^q$. Time-varying covariate effects can be included by allowing the coefficients to change in each interval

$$\alpha_i(t) = \sum_{j=1}^{p} \mathbb{1}(t \in [\alpha_{j-1}; \alpha_j)) \exp\left(\tilde{\theta}_{j,0} + \tilde{\boldsymbol{X}}_i^t \tilde{\boldsymbol{\theta}}_j\right), \tag{1}$$

with $\tilde{\boldsymbol{\theta}}_j \in \mathbb{R}^q$. The necessity of selecting cut-points is perhaps the biggest complication of this model. When there are no preferred cut-points, equally spaced quantiles of the event times are recommended.[26]

From now on we denote whether or not we observe that the i'th event happens in $[\alpha_{j-1}; \alpha_j)$ by $O_{ij} = \Delta_i \mathbb{1}(U_i \in [\alpha_{j-1}; \alpha_j))$ and the time at risk in the interval as $R_{ij} = \int_0^\tau \mathbb{1}(U_i > t)\mathbb{1}(t \in [\alpha_{j-1}; \alpha_j))dt$ in which τ is the

largest time of interest, often the longest follow-up time observed in the study. Furthermore, we write the covariate vector of observation i augmented with a 1 corresponding to the intercept as $\boldsymbol{X}_i = (1, \tilde{\boldsymbol{X}}_i^t)^t$. To ease the notation, we define $\boldsymbol{\theta}_j = (\tilde{\theta}_{j,0}, \tilde{\boldsymbol{\theta}}_j^t)^t$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^t$. It can be shown that the negative log-likelihood of the PCH model corresponding to (1) is [1]

$$\mathcal{L}_{MLE}(\boldsymbol{\theta}) = -l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p} O_{ij}\boldsymbol{X}_i^t\boldsymbol{\theta}_j - \exp\left(\boldsymbol{X}_i^t\boldsymbol{\theta}_j\right)R_{ij}.$$

The maximum likelihood estimator (MLE) is then defined as $\hat{\boldsymbol{\theta}}_{MLE} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{MLE}(\boldsymbol{\theta})$. The corresponding score equations,

$$\frac{\partial \mathcal{L}_{MLE}}{\partial \theta_{z,l}} = \sum_{i=1}^{n} O_{iz}X_{il} - R_{iz}\exp\left(\boldsymbol{X}_i^t\boldsymbol{\theta}_z\right)X_{il} = 0, \quad z \in \{1, \ldots, p\},\ l \in \{0, \ldots, q\}, \tag{2}$$

give some intuition into why the MLE is not robust against leverage points. When the score equations are evaluated at a certain $\boldsymbol{\theta}$, varying the covariate value of one observation can make the influence of that observation on the score equation arbitrarily large.

3 PCH model with the Brier score loss

We now introduce a method to obtain outlier robust estimators for proportional hazard models and work out the details for the PCH model. In logistic regression, estimates obtained by minimizing the squared distance, or Brier score, have been shown to be robust. [13] Prediction of the survival status at a single time t, $\mathbb{1}(T \geq t)$ can be seen as a classification problem with censoring. To measure the classification error at time t, the Brier score, $B(t) = E\left[\{\mathbb{1}(T \geq t) - S(t|\boldsymbol{X}, \boldsymbol{\theta})\}^2\right]$ can then be used. [16] Given an estimate of the conditional probability of an observation being uncensored, $G(\cdot|\boldsymbol{X})$, the Brier score can, under some regularity conditions, be consistently

estimated by an IPCW estimator[17]

$$\hat{B}(t) = \sum_{i=1}^{n} \frac{\mathbb{1}(\min(T_i, t) \leq C_i)}{\hat{G}(\min(T_i, t)|\boldsymbol{X}_i)} \left\{ \mathbb{1}(T_i \geq t) - S(t|\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}) \right\}^2.$$

Often it is assumed that the covariates do not influence the censoring distribution and the Kaplan-Meier (KM) estimator can be used as an estimator of $G(t)$. If the real survival function is $S_0(t|\boldsymbol{X})$, it can be shown that

$$\begin{aligned} & E\left[\frac{\mathbb{1}(\min(T, t) \leq C)}{G(\min(T, t)|\boldsymbol{X})} \left\{ \mathbb{1}(T \geq t) - S(t|\boldsymbol{X}, \boldsymbol{\theta}) \right\}^2 \right] \\ & = E\left[\left\{ \mathbb{1}(T \geq t) - S(t|\boldsymbol{X}, \boldsymbol{\theta}) \right\}^2 \right] \\ & = E\left[\left\{ \mathbb{1}(T \geq t) - S_0(t|\boldsymbol{X}) \right\}^2 \right] + E\left[\left\{ S_0(t|\boldsymbol{X}) - S(t|\boldsymbol{X}, \boldsymbol{\theta}) \right\}^2 \right]. \end{aligned} \tag{3}$$

Hence, the expected value of the Brier score decomposes into an irreducible variation term and a model-misspecification error.

Instead of minimizing the negative log-likelihood, we propose to select a number of time-points, $t_1, \ldots, t_s$ and use $\mathcal{L}_{BSL}(\boldsymbol{\theta}) = \sum_{k=1}^{s} \hat{B}(t_k)$, which we denote the Brier score loss (BSL), as loss function. The coefficients are then estimated as

$$\hat{\boldsymbol{\theta}}_{BSL} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{s} \hat{B}(t_k) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{BSL}(\boldsymbol{\theta}).$$

In the supplemental material it is shown that, under some regularity conditions, this leads to a consistent estimator. Denoting the IPCW weight of observation i at time-point t_k as $\text{IPW}_{ik} = \frac{\mathbb{1}(\min(T_i, t_k) \leq C)}{\hat{G}(\min(T_i, t_k)|\boldsymbol{X}_i)}$, we obtain that for a PCH model the first order conditions are

$$\begin{aligned} \frac{\partial \mathcal{L}_{BSL}}{\partial \theta_{zl}}(\boldsymbol{\theta}) &= 2 \sum_{k=1}^{s} \sum_{i=1}^{n} \text{IPW}_{ik} \left\{ \mathbb{1}(T_i \geq t_k) - S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \right\} \frac{\partial S}{\partial \theta_{zl}}(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \\ &= 2 \sum_{k=1}^{s} \sum_{i=1}^{n} \left[\begin{array}{c} \text{IPW}_{ik} \left\{ \mathbb{1}(T_i \geq t_k) - S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \right\} S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \\ \times \exp\left\{ \boldsymbol{X}_i^t \boldsymbol{\theta}_z \right\} \int_0^{t_k} \mathbb{1}(u \in [\alpha_z, \alpha_{z+1})) du X_{il} \end{array} \right] \\ &= 0. \end{aligned}$$

We get some insight into the effect of leverage points by inspecting the influence of a perturbed covariate value on this equation.[13] In the Supplementary material it is derived that when the score equations are evaluated at $\boldsymbol{\theta} \neq \mathbf{0}$ and we let $|\tilde{X}_{il}| \to \infty$ we get $S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \exp\{\boldsymbol{\theta}_z \boldsymbol{X}_i\} X_{il} \to 0$. Hence, observations with large covariate values have a minimal contribution to the score equations.

One drawback of using the Brier score as loss function is that, in the case of PCH models, the optimization problem is not convex, so advanced optimization methods to compute the estimator are needed (see Section 5).

3.1 Selection of evaluation time-points

There are multiple options for selecting the time-points, $t_1, \ldots, t_s$, at which the Brier score is evaluated. One option is to approximate the integrated Brier score (IBS), $\int_0^\tau \hat{B}(s)ds$ by using an equally spaced grid of time-points. A second option is to use the quantiles of the event times.[27]

To avoid multiple global minima one should aim to evaluate the Brier score in at least one time-point in each interval. When the cut-point selection is based on the quantiles of the event times this can be ensured by using a larger number of quantiles as the evaluation time-points. Therefore, in this paper we focus on using quantiles for both the cut- and evaluation time-point selection.

4 Penalizing time-varying effects

In order to deal with large amounts of parameters, penalization is a popular solution. This is done by adding a penalty term, $J(\lambda, \boldsymbol{\theta})$, to the minimization problem:

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + J(\lambda, \boldsymbol{\theta}). \tag{4}$$

Depending on the functional form of the penalty, certain desirable properties can be obtained. One example is the least absolute shrinkage and selection operator (lasso)[28],

$$J(\lambda, \boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} \sum_{l=0}^{q} |\theta_{j,l}|.$$

This penalty typically sets a number of the coefficients equal to zero, yielding a sparse solution. An extension of the lasso, the group lasso (GL), focuses on shrinking the coefficients of variables with an intrinsic group structure, e.g. categorical variables, towards zero as a group.[25] Although the GL penalty was initially proposed for categorical variables it has been adapted to solve other problems, e.g. change point detection in multidimensional signals.[29]

Often it is not a priori known whether a covariate effect should be modelled as time-varying. Therefore, it is natural to focus on penalties that can select between constant and time-varying coefficients.[23,24] In discrete survival models this has previously been accomplished by using a GL-like penalty.[30] The analogue of such a GL-like penalty for the PCH model is obtained by penalizing the differences between adjacent coefficients. A covariate-wise selection is then obtained by applying a GL-like penalty,

$$J_{GL}(\lambda, \boldsymbol{\theta}) = \lambda \sum_{l=0}^{q} \sqrt{\sum_{j=2}^{p} (\theta_{j,l} - \theta_{j-1,l})^2}.$$

The relation with the GL becomes clear when the model is re-parametrized using the original coefficients in the first interval and the differences between adjacent coefficients for the remaining intervals,

$$\begin{aligned} \zeta_{1,l} &= \theta_{1,l} && \forall l \in \{0, \dots, q\} \\ \zeta_{j,l} &= \theta_{j,l} - \theta_{j-1,l} && \forall j \in \{2, \dots, p\}, \forall l \in \{0, \dots, q\}. \end{aligned}$$

The corresponding transformation matrix is defined by $\boldsymbol{D\theta} = \boldsymbol{\zeta}$. For ease of notation we will use this re-parametrization from now on and write $\mathcal{L}(\boldsymbol{\zeta})$ instead of $\mathcal{L}(\boldsymbol{D}^{-1}\boldsymbol{\theta})$, the same is done for the penalty function which becomes $J_{GL}\left(\lambda, \boldsymbol{\zeta}\right) = \lambda \sum_{k=0}^{q} \sqrt{\sum_{l=2}^{p} \zeta_{j,l}^2}$.

5 Optimization algorithm

The optimization algorithm used to find a solution of the minimization problem in Equation 4 should be able to deal with non-convex loss functions, such as the BSL, and efficiently incorporate the proposed penalty. Proximal gradient algorithms are a class of algorithms that have shown their merit when used for penalized non-convex loss functions and lead to a simple update scheme when combined with the GL penalty.[31,32]

5.1 Proximal gradient method

Proximal gradient algorithms are generalizations of the gradient descent algorithm which is an iterative procedure that aims to minimize a differentiable function, e.g. the unpenalized loss $L(\boldsymbol{\zeta})$. At step $r+1$ the gradient descent method takes a step of size $\alpha^{(r)}$ in the direction of the negative gradient

$$\boldsymbol{\zeta}^{(r+1)} = \boldsymbol{\zeta}^{(r)} - \alpha^{(r)} \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right).$$

$\boldsymbol{\zeta}^{(r+1)}$ can be interpreted as the solution of a linear approximation to the minimization problem penalized by a quadratic term, i.e.[33]

$$\boldsymbol{\zeta}^{(r+1)} = \arg\min_{\boldsymbol{\eta}} \left[\mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right) + \left\langle \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right), \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\rangle + \frac{1}{2\alpha^{(r)}} \left\| \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\|_2^2 \right]. \tag{5}$$

When a non-differentiable penalty is added to the differentiable loss function the gradient method has to be adapted. One way of doing so is by the use of proximal gradient algorithms which focus on minimizing objective functions of the form $\mathcal{L}(\cdot) + J(\lambda, \cdot)$ in which $J(\lambda, \cdot)$ is not necesarily differentiable.[34] To achieve this, the

penalty is added to the approximate minimization problem (5),

$$\begin{aligned}\boldsymbol{\zeta}^{(r+1)} &= \arg\min_{\boldsymbol{\eta}} \left[\begin{array}{l} \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right) + \left\langle \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right), \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\rangle \\ + \frac{1}{2\alpha^{(r)}} \left\| \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\|_2^2 + J(\lambda, \boldsymbol{\eta}) \end{array} \right] \\ &= \arg\min_{\boldsymbol{\eta}} \left[\begin{array}{l} \left\langle \alpha^{(r)} \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right), \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\rangle + \alpha^{(r)} J(\lambda, \boldsymbol{\eta}) \\ + \frac{1}{2} \left\| \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\|_2^2 + \frac{1}{2} \left\| \alpha^{(r)} \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right) \right\|_2^2 \end{array} \right] \\ &= \arg\min_{\boldsymbol{\eta}} \left[\frac{1}{2} \left\| \alpha^{(r)} \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right) + \boldsymbol{\eta} - \boldsymbol{\zeta}^{(r)} \right\|_2^2 + \alpha^{(r)} J(\lambda, \boldsymbol{\eta}) \right]. \end{aligned} \tag{6}$$

By defining the so-called proximal operator of $\alpha^{(r)} J(\lambda, \cdot)$ as

$$\text{prox}_{\alpha^{(r)} J(\lambda, \cdot)}\left(\boldsymbol{\zeta}\right) = \arg\min_{\boldsymbol{\eta}} \left[\alpha^{(r)} J(\lambda, \boldsymbol{\eta}) + \frac{1}{2} \left\| \boldsymbol{\zeta} - \boldsymbol{\eta} \right\|_2^2 \right],$$

the updates from Equation 6 can be rewritten as

$$\boldsymbol{\zeta}^{(r+1)} = \text{prox}_{\alpha^{(r)} J(\lambda, \cdot)} \left[\boldsymbol{\zeta}^{(r)} - \alpha^{(r)} \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right) \right].$$

In the case of $J_{GL}\left(\lambda, \boldsymbol{\zeta}\right)$, the proximal operator can be derived from the proximal operator of the GL[32] and the updates can be explicitly computed as

$$\zeta_{1,l}^{(r+1)} = \zeta_{1,l}^{(r)} - \alpha^{(r)} \frac{\partial \mathcal{L}}{\partial \zeta_{1,l}}\left(\boldsymbol{\zeta}^{(r)}\right)$$

$$\boldsymbol{\zeta}_{-1,l}^{(r+1)} = \begin{cases} \boldsymbol{\zeta}_{-1,l}^{(r)} - \alpha^{(r)} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\zeta}_{-1,l}}\left(\boldsymbol{\zeta}^{(r)}\right) \left(1 - \frac{\lambda}{\left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{\zeta}_{-1,l}}\left(\boldsymbol{\zeta}^{(r)}\right) \right\|_2}\right) & \text{if } \left\| \boldsymbol{\zeta}_{-1,l}^{(r)} - \alpha^{(r)} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\zeta}_{-1,l}}\left(\boldsymbol{\zeta}^{(r)}\right) \right\|_2 > \lambda \alpha^{(r)} \\ 0 & \text{if } \left\| \boldsymbol{\zeta}_{-1,l}^{(r)} - \alpha^{(r)} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\zeta}_{-1,l}}\left(\boldsymbol{\zeta}^{(r)}\right) \right\|_2 \leq \lambda \alpha^{(r)} \end{cases}$$

where $\boldsymbol{\zeta}_{-1,l} = (\zeta_{2,l}, \ldots, \zeta_{p,l})^t$.

To speed up convergence we rely on the Barzilai-Borwein method, a Hessian-free approximation to Newton's method, to select the $\alpha^{(r)}$'s.[32,35] The complete minimization procedure is described in Algorithm 1.

Algorithm 1: Minimization algorithm of the penalized problem

input: initial values $\boldsymbol{\zeta}^{(0)}$; a fixed penalty parameter λ; an initial step-size $\alpha^{(0)}$, a maximum step-size α_{max}, and a minimum step-size α_{min} such that $\alpha_{max} \geq \alpha^{(0)} \geq \alpha_{min} > 0$; and a tolerance $\epsilon > 0$
initialization: $r \leftarrow 0$;
repeat
  $\boldsymbol{\zeta}^{(r+1)} = \text{prox}_{\alpha^{(r)} J(\lambda,\cdot)}\left(\boldsymbol{\zeta}^{(r)} - \alpha^{(r)} \nabla \mathcal{L}\left(\boldsymbol{\zeta}^{(r)}\right)\right)$;
  $\alpha^{(r+1)} = \frac{\langle \nabla L(\boldsymbol{\zeta}^{(r+1)}) - \nabla L(\boldsymbol{\zeta}^{r}), \boldsymbol{\zeta}^{(r+1)} - \boldsymbol{\zeta}^{(r)} \rangle}{\left\| \nabla L(\boldsymbol{\zeta}^{(r+1)}) - \nabla L(\boldsymbol{\zeta}^{(r)}) \right\|_2^2}$;
  if $\alpha^{(r+1)} > \alpha_{max}$ **then**
    $\alpha^{(r+1)} \leftarrow \alpha_{max}$;
  else if $\alpha^{(r+1)} < \alpha_{min}$ **then**
    $\alpha^{(r+1)} \leftarrow \alpha_{min}$;
  $r \leftarrow r + 1$;
until $\frac{\sum_{j=1}^{p} \sum_{l=0}^{q} |\zeta_{j,l}^{(r)} - \zeta_{j,l}^{(r+1)}|}{\sum_{j=1}^{p} \sum_{l=0}^{q} |\zeta_{j,l}^{(r)}|} < \epsilon$;
output: $\boldsymbol{\zeta}^{(r)}$

To select an optimal λ usually a regularization path is calculated and the λ corresponding to the minimum of a cross validated summary measure is selected. In the rest of this paper this summary measure is the log-likelihood when the MLE is used, when the BSL estimator (BSLE) is used, this is based on the BSL.

5.2 Initial values

Plugging $\mathcal{L}_{BSL}(\cdot)$ and $J_{GL}(\cdot, \cdot)$ into Equation 4 does not necessarily lead to a convex problem. Convergence problems can to some degree be avoided by using appropriate initial values. The approach taken here consists of two steps. First of all, a large penalty parameter is used which leads to a model without time-varying coefficients. Secondly, starting from the solution corresponding to such a large penalty parameter, a regularization path is calculated by relying on warm starts.

First of all, when there are no time-varying effects, i.e. $\boldsymbol{\zeta}_j = \mathbf{0},\ \forall j \in \{2, \ldots, p\}$, the PCH model corresponds to an exponential model. Indeed, the parameters of an exponential model estimated by the Brier score loss

$$\boldsymbol{\zeta}_{exp} = \arg\min_{\boldsymbol{\zeta}} \mathcal{L}_{BSL}(\boldsymbol{\zeta}) \qquad \text{subject to } \boldsymbol{\zeta}_j = 0,\ \forall j \in \{2, \ldots, p\},$$

are identical to the solution of

$$\arg\min_{\boldsymbol{\zeta}} \mathcal{L}_{BSL}(\boldsymbol{\zeta}) + J_{GL}(\lambda, \boldsymbol{\zeta}) \qquad \text{with } \lambda \geq \lambda_{max} = \max_{l \in \{0,\ldots,q\}} \left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{\zeta}_{-1,l}}(\boldsymbol{\zeta}_{exp}) \right\|_2 .$$

To estimate $\boldsymbol{\zeta}_{exp}$ we propose to use a gradient descent algorithm with initial values obtained by other robust methods. In particular, firstly the covariates are standardized using robust location and scale estimators, e.g. the median and median absolute deviation estimators. Secondly, estimates of the covariate coefficients are obtained from a robust estimator for the Cox proportional hazard model.[5] The estimator of the baseline parameter, $\zeta_{1,0}$, is based on the fact that if there are no covariate effects and then the p'th percentile of an exponentially distributed T is

$$\zeta_{1,0} = \log\left[\frac{-\log(1-p)}{F_T^{-1}(p)}\right]. \tag{7}$$

We suggest to estimate the marginal distribution function, $F_T(\cdot)$, using the KM estimator and pick $p = \frac{\hat{F}_T(\max_i(\Delta_i T_i))}{2}$. Note that when there is no censoring $p \approx 0.5$ and the denominator in Equation 7 will be the robust median estimator.

To find the estimates for arbitrary $\lambda < \lambda_{max}$ the regularization path over a grid of decreasing λ values can be calculated using the preceding solution to calculate the solution of a new λ.

6 Simulations

To inspect the estimation procedures and robustness properties of the proposed estimator, a simulation study is performed. Throughout this section the penalized MLE and BSLE are compared. The simulated data is generated from a PCH model which depends on two covariates X_1 and X_2 and has

$$\alpha(t) = \begin{cases} \exp(-3 + 2X_2) & \text{if } t < 2 \\ \exp(-2.5 + 0.75X_1 + 2X_2) & \text{if } 2 \leq t < 10 \\ \exp(-2 + 0.75X_1 + 2X_2) & \text{if } 10 \leq t \end{cases}$$

as hazard rate. Hence, the coefficient of X_1 is time-varying while the coefficient of X_2 is not. The covariates are independent and simulated using a standard normal distribution. To get a censoring percentage of approximately 40% the censoring time variable is simulated from an exponential distribution with mean $\exp(3.02)$. The robustness properties are inspected by contaminating a fraction ϵ of the data, in Section 6.1 this is done by simulating data from a PCH model with a different hazard rate while in Section 6.2 leverage points at pre-specified positions are added. We consider two different sample sizes, $n = 500$ and $n = 1000$, and every scenario is simulated 100 times. To measure the performance, we compute the integrated absolute error (IAE) for each coefficient

$$\text{IAE}_{j,i} = \frac{\int_0^\tau |\theta_{j,i} - \hat{\boldsymbol{\theta}}_i(u)| \mathbb{1}(u \in (\alpha_j, \alpha_{j+1}])du}{\alpha_j - \alpha_{j+1}},$$

where $\hat{\boldsymbol{\theta}}(u) = \sum_{j=1}^p \mathbb{1}(u \in [\alpha_{j-1}; \alpha_j))\hat{\boldsymbol{\theta}}_j$. Based on the decomposition of the prediction error in Equation (3), the quality of the overall fit is measured using the root of the integrated Brier score (RIBS)[16] between the estimated and data-generating survival functions

$$\text{RIBS} = \sqrt{\frac{E\left[\int_0^\tau \left\{S_0(t|\boldsymbol{X}) - \hat{S}(t|\boldsymbol{X}, \hat{\boldsymbol{\theta}})\right\}^2 dt\right]}{\tau}}.$$

The RIBS and IAEs are calculated by setting $\tau = 15$. The expected value is approximated by simulating 100000 values from a bivariate normal distribution, and the involved integral is approximated with the composite midpoint rule.

6.1 Outlying event times

In this section the contaminated data is generated from an exponential survival model with hazard function $h_{cont}(t) = \exp(-4 - 2X_1)$. To achieve a censoring rate of approximately 50% the censoring time is simulated using an exponential distribution with mean $\exp(4)$. Just as for the clean data the covariates are independently generated from independent standard normal distributions.

6.1.1 Known breaks The results of the simulations when the correct cut-points are provided can be found in Table 1. In the case without outliers, the MLE led to lower IAEs than the BSLE for nearly all coefficients. Furthermore, the RIBS was essentially equal for the MLE and the BSLE. When the contamination proportion is set to 5% or 10% the MLE for the estimates of the coefficients in the second and third intervals degrade noticeably while the performance of the BSLE remains acceptable. Finally, once the contamination rate is raised to 15% also the performance measures of the BSLE indicate a decrease in performance when compared to the non-contaminated scenario. Comparing the performance measures across sample sizes shows that both methods perform slightly better when larger sample sizes are used.

Table 1. Integrated absolute errors and RIBS of the MLE and Brier score loss estimator (BSLE) when the correct breaks are specified in the estimation process.

ϵ	Sample size	method	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$	RIBS
0	500	MLE	0.13	0.09	0.09	0.08	0.08	0.09	0.13	0.10	0.11	0.02
		Brier	0.15	0.15	0.11	0.09	0.10	0.14	0.18	0.17	0.23	0.03
	1000	MLE	0.08	0.06	0.06	0.05	0.06	0.06	0.08	0.07	0.10	0.01
		Brier	0.12	0.10	0.08	0.07	0.07	0.08	0.13	0.13	0.17	0.02
0.05	500	MLE	0.17	0.09	0.21	0.11	0.23	0.62	1.15	0.70	1.27	0.07
		Brier	0.20	0.13	0.21	0.10	0.16	0.24	0.35	0.24	0.44	0.04
	1000	MLE	0.13	0.06	0.17	0.07	0.26	0.71	1.27	0.77	1.40	0.07
		Brier	0.20	0.10	0.20	0.07	0.15	0.24	0.31	0.27	0.34	0.03
0.1	500	MLE	0.22	0.12	0.35	0.16	0.37	0.94	1.48	0.90	1.57	0.11
		Brier	0.30	0.14	0.35	0.14	0.30	0.47	0.64	0.52	0.66	0.07
	1000	MLE	0.19	0.10	0.33	0.14	0.36	0.96	1.53	0.94	1.60	0.11
		Brier	0.31	0.10	0.35	0.13	0.26	0.49	0.61	0.52	0.66	0.06
0.15	500	MLE	0.23	0.15	0.45	0.22	0.45	1.09	1.63	1.06	1.62	0.14
		Brier	0.38	0.15	0.52	0.22	0.44	0.75	0.93	0.75	0.93	0.10
	1000	MLE	0.24	0.14	0.47	0.20	0.47	1.11	1.66	1.06	1.68	0.14
		Brier	0.39	0.11	0.49	0.19	0.43	0.80	0.94	0.75	0.98	0.10

6.1.2 Unknown breaks In this section the cut-points are estimated. More specifically, 9 intervals are used. The results can be found in Table 2. The main difference between this and the previous scenario is that the IAEs indicate that both the BSLE and MLE lead to less accurate estimates when the cut-points are estimated. On the other hand, for both methods the RIBS is approximately on par with those reported in the previous section. The

effect of contamination on both methods is similar as before. While a small fraction of contamination already has a big impact on the MLE, the performance of the BSLE degrades much more slowly.

Table 2. Integrated absolute errors and RIBS of the MLE and Brier score loss estimator (BSLE) when the breaks are estimated in the estimation process.

ϵ	Sample size	method	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$	RIBS
0	500	MLE	0.49	0.58	0.21	0.49	0.25	0.27	0.37	0.18	0.21	0.03
		Brier	0.43	0.65	0.28	0.53	0.36	0.44	0.44	0.33	0.41	0.03
	1000	MLE	0.47	0.51	0.16	0.48	0.20	0.19	0.30	0.15	0.14	0.02
		Brier	0.42	0.55	0.22	0.48	0.29	0.35	0.37	0.26	0.34	0.02
0.05	5000	MLE	0.69	0.49	0.76	0.64	1.05	2.36	2.18	1.29	2.64	0.08
		Brier	0.56	0.56	0.46	0.30	0.58	0.69	0.82	0.53	0.74	0.04
	1000	MLE	0.69	0.49	0.78	0.60	1.08	2.54	2.23	1.36	2.80	0.08
		Brier	0.57	0.53	0.48	0.31	0.52	0.81	0.86	0.59	0.85	0.04
0.1	500	MLE	0.75	0.42	1.13	0.92	1.40	2.94	2.84	1.80	3.42	0.11
		Brier	0.73	0.46	0.85	0.37	0.98	1.41	1.49	1.14	1.57	0.07
	1000	MLE	0.84	0.46	1.26	0.91	1.38	2.97	2.83	1.87	3.64	0.11
		Brier	0.68	0.46	0.77	0.37	0.94	1.43	1.55	1.17	1.71	0.06
0.15	500	MLE	0.86	0.34	1.44	1.39	1.63	3.29	3.76	2.21	4.17	0.15
		Brier	0.79	0.35	1.22	0.71	1.39	2.10	2.34	1.72	2.51	0.11
	1000	MLE	0.88	0.45	1.60	1.07	1.56	3.18	3.38	2.34	4.25	0.14
		Brier	0.89	0.44	1.31	0.69	1.38	2.17	2.47	2.00	2.74	0.10

6.2 Leverage points

Instead of generating contamination data from a different survival distribution we now focus on a point contamination. More specifically, 5% of the data is replaced by observations with $t_i = 15,\ \delta_i = 0,\ X_1 = \nu,\ X_2 = \nu$ and the effect of varying ν over $[-5, 5]$ is inspected. In total 9 time-intervals are used and the cut-points are estimated. A box-plot of the RIBS values can be found in Figure 1. Note that for $\nu \leq 0$ the MLE is very stable and performs very well. This can be explained by inspecting the score equations of the MLE, i.e. Equation 2. Given that the contaminated points correspond to censored observations the first term is zero. Additionally, $\exp(\boldsymbol{X}s\boldsymbol{\theta})\boldsymbol{X} \to \mathbf{0}$ when $\boldsymbol{X} \to -\infty$ and, hence, the contaminated points have a near zero contribution to the score equations when $-\nu$ is sufficiently large. However, once $\nu > 0$ the performance of the

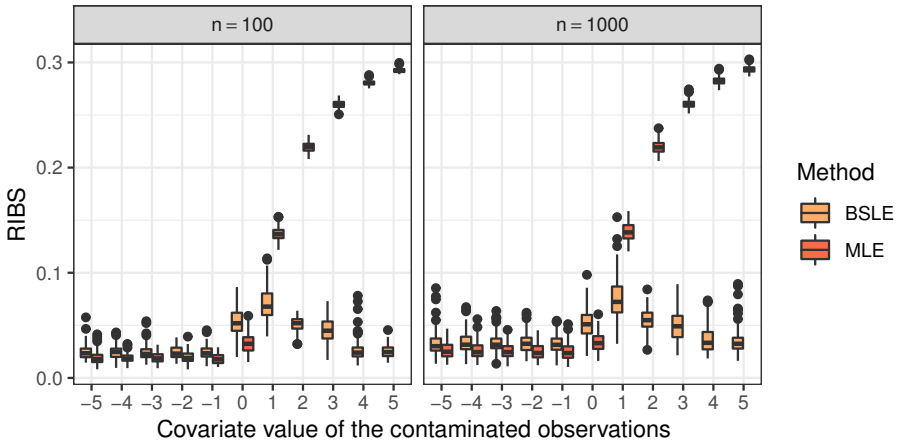

Figure 1. Root integrated squared error (RIBS) when contamination is added at a fixed position.

MLE quickly degrades while the performance of the BSLE is a lot less sensitive to the presence and position of the leverage point.

7 Applications

7.1 AIDS

The first example is based on a dataset of 2843 Australian AIDS patients diagnosed prior to 1991. The data can be found in the MASS R package and was originally extracted from the national AIDS register of the National Centre in HIV Epidemiology and Clinical Research.[36] The first registered patient was diagnosed in 1982 and hence the dataset describes the period during which the AIDS epidemic spread throughout Australia. The limited knowledge of transmission patterns of HIV in the eighties led to its largely unchecked spread throughout Australia by means of sexual activity, contaminated blood transfusions, transmission from mother to unborn, etc. Before 1987 antiretroviral treatment was not routinely administered in Australia and from 1987 until 1991 monotherapy with zidovudine was standard of care.

In this analysis we model the time to death and investigate the effect of age at diagnosis in years and gender. In $> 80\%$ of cases the cause of transmission was sexual contact between homo- or bisexual people, the median age at diagnosis was 36 years, and 3.3% of the included patients were female. Follow-up ended on the first of July 1991 and in total 613 events were observed. To model the coefficients and the baseline hazard 10 intervals were used. The BSLE was obtained by evaluating the Brier score in 31 time-points. The effect of the covariates on the censoring distribution was inspected with a robust Cox proportional hazard model and using the KM-estimator of the marginal censoring distribution was deemed reasonable.[5] The optimal shrinkage parameters were selected using a 10 times repeated 10-fold CV procedure.

A plot of the estimated coefficients can be found in Figure 2. Both the MLE and BSLE indicate that older patients initially had an increased hazard rate. Later on the effect of age at diagnosis decreased and became negligible two years post-diagnosis. The age effect is modeled differently by the MLE and BSLE. The BSLE leads to a monotonically decreasing age effect, while the MLE estimates an increasing hazard ratio during the first half year post-diagnosis. A comparison between the MLE and BSLE of the baseline indicates a similar discrepancy during the first six months post-diagnosis after which the estimates become more similar. For the gender effect, both the BSLE and MLE indicate that, for the majority of the follow-up period, males with AIDS had a larger hazard rate during the follow-up as compared to females with AIDS. Considering that in $> 80\%$ of the cases the infection occurred through sexual contact between homo- or bisexual people, this is in agreement with previous reports of the population of Australian AIDS patients infected through sexual contact between homo- or bisexual people having a worse prognosis than most patients infected in other ways.[37]

To find potential outliers, normal deviate residuals, defined as $\Phi^{-1}[S(U_i|\boldsymbol{X}_i)]\Delta_i + \Phi^{-1}\left[\frac{S(U_i|\boldsymbol{X}_i)}{2}\right](1-\Delta_i)$ were calculated.[38] When the MLE estimates were used, no observation with absolute value of the residual larger than three was detected. When the BSLE estimates were used, ten observations with an absolute value of their residual larger than three were detected. These observations are described in Table 3. Each of these potential outliers was a patient who died right after being diagnosed. This might indicate that a variable describing the state of the disease progression at diagnosis should be included. Furthermore, five of the described patients were remarkably young and did not get the disease through sexual contact. This could indicate that the disease

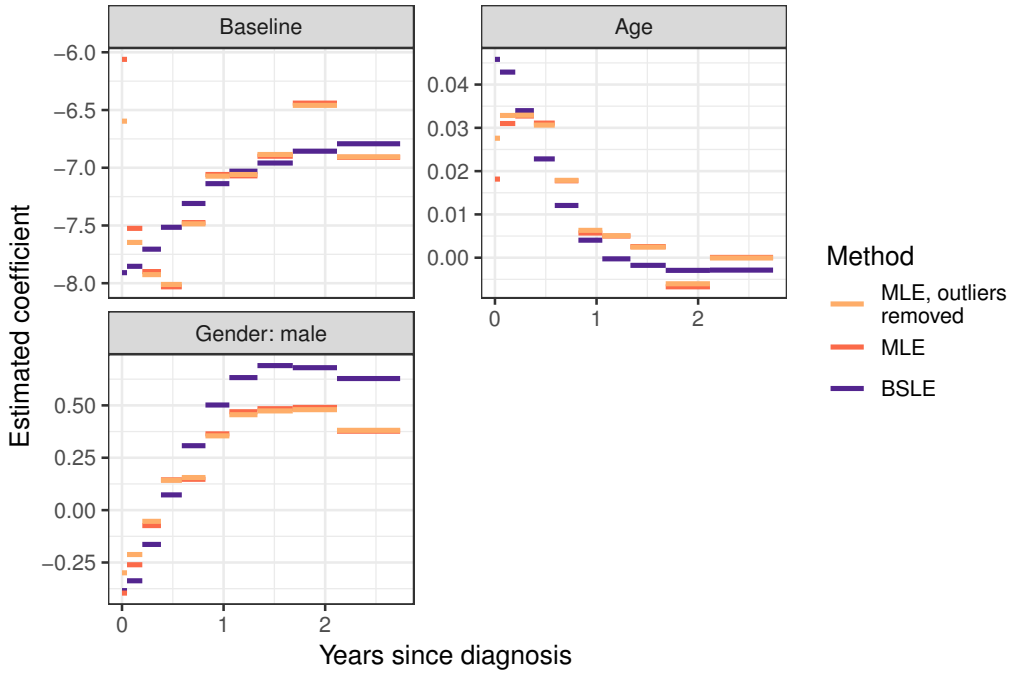

Figure 2. Estimated coefficients of a PCH model for the AIDS dataset.

progression of non-adolescent patients or patients who did not contract AIDS through sexual contact was significantly different from that of the majority of the AIDS patients. The coefficient paths obtained by fitting the MLE while excluding the nine potential outliers can be found in Figure 2. The MLE estimates in which the potential outliers were removed lie closer to the BSLE estimates than the original MLE estimates.

7.2 Cutaneous malignant melanoma

Cutaneous malignant melanoma (CMM) is a type of aggressive skin cancer. The application described here is based on a dataset of 205 CMM patients treated during the period 1964-1973 at the University Hospital of Odense. Patients were followed until the first of January 1978.[39] The dataset includes the covariates gender, presence of ulcerations, and tumor thickness (1/100 mm). The outcome variable of interest is the number of days from surgery till death from any cause or censoring. In total there were 71 events of which 57 were

Table 3. Observation from the AIDS dataset for which the absolute value of the corresponding normal deviate residuals was larger than three.

Status	Time (in days)	Gender	Age	Cause of transmission
Death	1	Female	0	Mother with HIV infection
Death	1	Male	11	Blood transfusion
Death	1	Male	0	Blood transfusion
Death	1	Male	12	Blood transfusion
Death	1	Male	31	Sexual contact between homo- or bisexual people
Death	1	Male	34	Sexual contact between homo- or bisexual people
Death	1	Male	25	Sexual contact between homo- or bisexual people
Death	1	Male	25	Sexual contact between homo- or bisexual people
Death	1	Male	36	Sexual contact between homo- or bisexual people
Death	2	Male	12	Haemophilia

melanoma related. The data is publicly available in the timereg R package[40] and has previously been analyzed by Martinussen and Scheike[41]. The analysis done by Martinussen and Scheike[41] focused on the cause-specific hazard of deaths from melanoma modeled by a non-parametric multiplicative model with time-varying effects. Tests and visual inspection of the time-varying coefficients in the model for melanoma related deaths indicated that the effect of tumor thickness decreased over time while the effects of the other covariates were relatively constant.

Instead of following the cause-specific hazard approach we modeled overall survival, i.e. time from surgery till death from any cause. The PCH model included 10 intervals and when using the BSLE 31 evaluation points were used. To inspect whether relying on the KM estimator for the IPCW weights used in the BSLE was reasonable, the influence of the covariates on the censoring was inspected with a robust Cox proportional hazard model and none of the effects were found to be significant.[5] Optimal values for the penalty parameters were selected using a 100 times repeated 10-fold cross-validation.

The estimated coefficients can be found in Figure 3. The gender effect is modeled as time-varying by the BSLE but the effect is constant when the MLE is used, indicating that the adjacent coefficients have been shrunk towards each other. All other coefficients were modeled as time-varying, demonstrating the GL-like shrinkage of the proposed penalty. The BSLE of the tumor thickness effect shows a large decrease over time while the

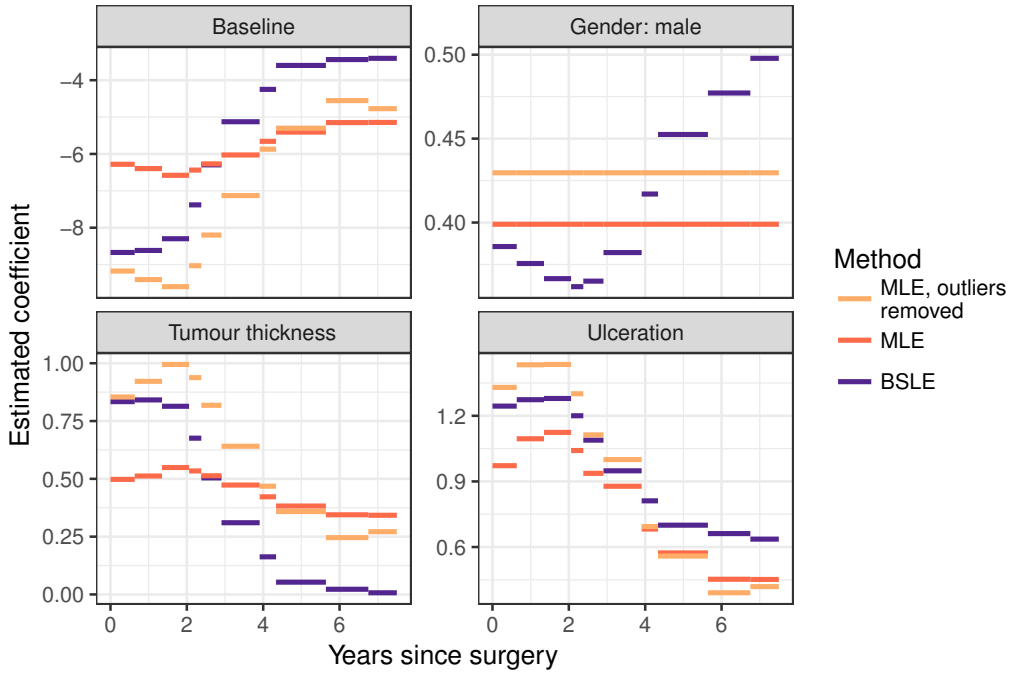

Figure 3. Estimated coefficients of a PCH model for the melanoma dataset.

MLE remains rather stable. A similar conclusion is obtained for the baseline effect. Finally, the hazard ratio of the presence of an ulceration is modeled as slowly decreasing by both the MLE and BSLE.

Based on the BSLE estimates, five observations with absolute residual values larger than two were detected. These observations are described in Table 4. Four out of the five patients died relatively early on from a non-melanoma related cause. This is rather interesting given that a minority of the deaths (20%) were non-melanoma related. Refitting the MLE without the five potential outliers leads to estimates closer to the BSLE (Figure 3). In particular, the baseline and tumour thickness effects of the MLE with the potential outliers removed are close to those obtained from the BSLE.

Table 4. Observations from the CMM dataset for which the absolute value of the corresponding normal deviate residual was larger than two.

Status	Time (in days)	Ulceration	Tumour thickness	Gender
Death: other cause	10	Present	676	Male
Death: other cause	30	Absent	65	Male
Death: other cause	99	Absent	290	Female
Death: other cause	355	Present	16	Female
Death: due to melanoma	817	Absent	32	Female

8 Discussion

In this paper we proposed to use the Brier score as loss function to obtain more robust estimators for proportional hazard models. The feasibility of the BSLE approach was illustrated for PCH models with penalized time-varying coefficients which was compared with the penalized MLE in a simulation study and on real data.

In the future this approach could be extended to different types of models. Based on the intuitive explanation of the robustness properties in Section 3 this approach is readily extendable to most parametric proportional hazard models. Furthermore, a different trade-off between efficiency and robustness could be obtained by replacing the quadratic error used in the Brier score by a different error measure. However, care will have to be taken when selecting the error measure to ensure it corresponds to a proper scoring rule. E.g. replacing the quadratic error with the absolute error would lead to an estimator which does not obtain its minimum at the true survival function.[42] Finally, only penalties that select between time-varying and constant effects were considered, this could be extended to include automatic variable selection.

One of the main weaknesses of this model is that standard errors are not available. Although e.g. the bootstrap could be used to obtain these, they are not particularly meaningful since the penalization leads to, potentially substantially, biased estimates. Another potential drawback is that the IPCW step might cause unstable estimates if the censoring rate is very high. Furthermore, as touched upon in the application, the IPCW step requires a reasonable model for the censoring distribution. When the estimator of the censoring distribution is heavily influenced by outliers, this might become problematic. However, in practice the censoring can often be assumed to be independent of the covariates and using the KM estimator seems to lead to relatively stable estimates.

Acknowledgements

The authors would like to thank Lasse Hjort Jakobsen for advice regarding the selection of cut-points. A significant amount of work was performed during a stay of Jorne Lionel Biccler at the ROBUST@Leuven research group of the Department of Mathematics at KU Leuven.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study was partly supported by a grant for the Danish Cancer Society and a grant from the Danish Lymphoma Group. This research was also funded by International Funds KU Leuven grant number C16/15/068.

References

1. Aalen OO, Borgan Ø and Gjessing HK. *Survival and Event History Analysis*. Statistics for Biology and Health, New York: Springer New York, 2008. ISBN 978-0-387-20287-7. DOI:10.1007/978-0-387-68560-1. URL http://link.springer.com/10.1007/978-0-387-68560-1.
2. Maronna RA, Martin RD and Yohai VJ. *Robust Statistics*. Wiley Series in Probability and Statistics, Chichester, UK: John Wiley & Sons, Ltd, 2006. ISBN 9780470010945. DOI:10.1002/0470010940. URL http://doi.wiley.com/10.1002/0470010940.
3. Farcomeni A and Ventura L. An overview of robust methods in medical research. *Statistical Methods in Medical Research* 2012; 21(2): 111–133. DOI:10.1177/0962280210385865. URL http://journals.sagepub.com/doi/10.1177/0962280210385865.
4. Heritier S, Cantoni E, Copt S et al. *Robust Methods in Biostatistics*. Chichester, UK: John Wiley & Sons, Ltd, 2009. ISBN 0470027266.

5. Bednarski T. Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics* 1993; : 213–225URL http://www.jstor.org/stable/4616277.

6. Sasieni P. Some New Estimators for Cox Regression. *The Annals of Statistics* 1993; 21(4): 1721–1759. DOI: 10.1214/aos/1176349395. URL http://projecteuclid.org/euclid.aos/1176349395.

7. Farcomeni A and Viviani S. Robust estimation for the Cox regression model based on trimming. *Biometrical Journal* 2011; 53(6): 956–973. DOI:10.1002/bimj.201100008. URL http://doi.wiley.com/10.1002/bimj.201100008.

8. Bednarski T. On a robust modification of Breslow's cumulated hazard estimator. *Computational Statistics & Data Analysis* 2007; 52(1): 234–238. URL http://www.sciencedirect.com/science/article/pii/S0167947306004798.

9. Locatelli I, Marazzi A and Yohai VJ. Robust accelerated failure time regression. *Computational Statistics & Data Analysis* 2011; 55(1): 874–887. DOI:10.1016/j.csda.2010.07.017.

10. Álvarez EE and Ferrario J. Robust estimation in the additive hazards model. *Communications in Statistics - Theory and Methods* 2016; 45(4): 906–921. DOI:10.1080/03610926.2013.853790. URL http://www.tandfonline.com/doi/full/10.1080/03610926.2013.853790.

11. Wang S, Hu T, Xiang L et al. Generalized M-estimation for the accelerated failure time model. *Statistics* 2016; 50(1): 114–138. DOI:10.1080/02331888.2015.1032970. URL http://www.tandfonline.com/doi/full/10.1080/02331888.2015.1032970.

12. Agostinelli C, Locatelli I, Marazzi A et al. Robust estimators of accelerated failure time regression with generalized log-gamma errors. *Computational Statistics & Data Analysis* 2017; 107: 92–106. DOI:10.1016/j.csda.2016.10.012.

13. Chi EC and Scott DW. Robust Parametric Classification and Variable Selection by a Minimum Distance Criterion. *Journal of Computational and Graphical Statistics* 2014; 23(1): 111–128. DOI:10.1080/10618600.2012.737296. URL http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.737296.

14. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather review* 1950; 78(1): 1–3. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(1950)078{%}3C0001:VOFEIT{%}3E2.0.CO;2.

15. Graf E, Schmoor C, Sauerbrei W et al. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; 18(17-18): 2529–2545. DOI:10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2529::AID-SIM274⟩3.0.CO;2-5. URL http://doi.wiley.com/10.1002/{%}28SICI{%}291097-0258{%}2819990915/30{%}2918{%}3A17/18{%}3C2529{%}3A{%}3AAID-SIM274{%}3E3.0.CO{%}3B2-5.

16. Mogensen UB, Ishwaran H and Gerds TA. Evaluating Random Forests for Survival Analysis using Prediction Error Curves. *Journal of statistical software* 2012; 50(11): 1–23. URL http://www.ncbi.nlm.nih.gov/pubmed/25317082http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4194196.

17. Gerds TA and Schumacher M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal* 2006; 48(6): 1029–1040. DOI:10.1002/bimj.200610301. URL http://doi.wiley.com/10.1002/bimj.200610301.

18. Holford TR. Life Tables with Concomitant Information. *Biometrics* 1976; 32(3): 587. DOI:10.2307/2529747.

19. Holford TR. The Analysis of Rates and of Survivorship Using Log-Linear Models. *Biometrics* 1980; 36(2): 299. DOI:10.2307/2529982. URL http://www.jstor.org/stable/2529982?origin=crossref.

20. Kessing LV, Thomsen AF, Mogensen UB et al. Treatment with antipsychotics and the risk of diabetes in clinical practice. *The British Journal of Psychiatry* 2010; 197(4): 266–271. URL http://bjp.rcpsych.org/content/197/4/266.long.

21. Nagot N, Kankasa C, Tumwine JK et al. Extended pre-exposure prophylaxis with lopinavir–ritonavir versus lamivudine to prevent HIV-1 transmission through breastfeeding up to 50 weeks in infants in Africa (ANRS 12174): a randomised controlled trial. *The Lancet* 2016; 387(10018): 566–573. DOI:10.1016/S0140-6736(15)00984-8. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673615009848.

22. Geman S and Hwang CR. Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics* 1982; 10(2): 401–414. DOI:10.1214/aos/1176345782. URL http://projecteuclid.org/euclid.aos/1176345782.

23. Xiao W, Lu W and Zhang HH. Joint structure selection and estimation in the time-varying coefficient Cox model. *Statistica Sinica* 2016; 26(2): 547–567. DOI:10.5705/ss.2013.076. URL http://www.ncbi.nlm.nih.

gov/pubmed/27540275http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4987133http://www3.stat.sinica.edu.tw/statistica/J26N2/J26N26/J26N26.html.

24. Yan J and Huang J. Model Selection for Cox Models with Time-Varying Coefficients. *Biometrics* 2012; 68(2): 419–428. DOI:10.1111/j.1541-0420.2011.01692.x. URL http://doi.wiley.com/10.1111/j.1541-0420.2011.01692.xhttp://www.ncbi.nlm.nih.gov/pubmed/22506825http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3384767.
25. Yuan M and Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006; 68(1): 49–67. DOI:10.1111/j.1467-9868.2005.00532.x. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00532.x.
26. Royston P and Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002; 21(15): 2175–2197. DOI:10.1002/sim.1203. URL http://doi.wiley.com/10.1002/sim.1203.
27. Wey A, Connett J and Rudser K. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics* 2015; 16(3): 537–549. DOI:10.1093/biostatistics/kxv001. URL http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxv001.
28. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1996; : 267–288URL http://www.jstor.org/stable/2346178.
29. Bleakley K and Vert JP. The group fused Lasso for multiple change-point detection. *arXiv preprint arXiv:11064199* 2011; URL https://arxiv.org/pdf/1106.4199.pdf. arXiv:1106.4199v1.
30. Tutz G and Gertheiss J. Regularized regression for categorical data. *Statistical Modelling* 2016; 16(3): 161–200. DOI:10.1177/1471082X16642560. URL http://journals.sagepub.com/doi/10.1177/1471082X16642560.
31. Yang Z, Wang Z, Liu H et al. Sparse nonlinear regression: Parameter estimation under nonconvexity. In *International Conference on Machine Learning*, volume 48. pp. 2472–2481. URL http://www.jmlr.org/proceedings/papers/v48/yangc16.pdfhttp://jmlr.org/proceedings/papers/v48/yangc16.pdf.

32. Mosci S, Rosasco L, Santoro M et al. Solving Structured Sparsity Regularization with Proximal Methods. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, pp. 418–433. DOI:10.1007/978-3-642-15883-4_27. URL http://link.springer.com/10.1007/978-3-642-15883-4{_}27.

33. Hastie T, Tibshirani R and Wainwright M. *Statistical learning with Sparsity : the lasso and generalizations*. Chapman and Hall/CRC, 2015. ISBN 9781498712163.

34. Parikh N and Boyd S. Proximal algorithms. *Foundations and Trends® in Optimization* 2014; 1(3): 127–239. URL https://www.nowpublishers.com/article/Details/OPT-003.

35. Barzilai J and Borwein JM. Two-point step size gradient methods. *IMA journal of numerical analysis* 1988; 8(1): 141–148. URL https://academic.oup.com/imajna/article-abstract/8/1/141/802460.

36. Venables WN and Ripley BD. *Modern Applied Statistics with S*. Statistics and Computing, New York, NY: Springer New York, 2002. ISBN 978-1-4419-3008-8. DOI:10.1007/978-0-387-21706-2. URL http://link.springer.com/10.1007/978-0-387-21706-2.

37. Ripley BD and Solomon P. *A note on Australian AIDS survival*. University of Adelaide, Department of Statistics, 1994.

38. Nardi A and Schemper M. New Residuals for Cox Regression and Their Application to Outlier Screening. *Biometrics* 1999; 55(2): 523–529. DOI:10.1111/j.0006-341X.1999.00523.x. URL http://doi.wiley.com/10.1111/j.0006-341X.1999.00523.x.

39. Drzewiecki KT, Ladefoged C and Christensen HE. Biopsy and prognosis for cutaneous malignant melanomas in stage I. *Scandinavian journal of plastic and reconstructive surgery* 1980; 14(2): 141–4. URL http://www.ncbi.nlm.nih.gov/pubmed/7221482.

40. Scheike TH and Zhang MJ. Analyzing Competing Risk Data Using the R timereg Package. *Journal of Statistical Software* 2011; 38(2): 1–15. DOI:10.18637/jss.v038.i02. URL http://www.jstatsoft.org/v38/i02/.

41. Martinussen T and Scheike TH. *Dynamic Regression Models for Survival Data*. Statistics for Biology and Health, New York: Springer New York, 2006. ISBN 978-0-387-20274-7. DOI:10.1007/0-387-33960-4. URL http://link.springer.com/10.1007/0-387-33960-4.

42. Buja A, Stuetzle W and Shen Y. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November* 2005; URL https://pdfs.semanticscholar.org/d670/6b6e626c15680688b0774419662f2341caee.pdf.

43. Newey WK and McFadden D. Large sample estimation and hypothesis testing. *Handbook of Econometrics* 1994; URL http://www.sciencedirect.com/science/article/pii/S1573441205800054.

Supplemental material

Consistency of the Brier score loss estimator

Theorem 1. *If the data are i.i.d. and come from a PCH model with coefficients* $\dot{\boldsymbol{\theta}}$*, then the BSLE is consistent given that the following conditions are met:*

1. *The estimator of the censoring distribution is uniformly consistent on* $[0, t_s]$*, i.e.* $\sup_{t\in[0,t_s]} \sup_{\boldsymbol{x}} |\hat{G}(t|\boldsymbol{x}) - G(t|\boldsymbol{x})| \to 0$.
2. $G(t_s|\boldsymbol{x}) > \delta > 0$ *for all* $\boldsymbol{x}$.
3. *The parameter space* Θ *is compact.*
4. $\forall j \in \{1, \ldots, p\}, \exists k \in \{1, \ldots, s\}$ *such that* $t_k \in [\alpha_{j-1} - \alpha_j)$.
5. *There is no hyperplane in* $\mathbb{R}^q$ *which contains the covariates with probability one, i.e.* $\forall \boldsymbol{\gamma} \in \mathbb{R}^{q+1}$ *we have that* $P(\boldsymbol{X}^t\boldsymbol{\gamma} \neq 0) > 0$.

Proof. Following Wey et al[27] we start by showing that the difference between the empirical and expected versions of the BSL converges uniformly to zero,

$$
\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{k=1}^{s}B(t_k|\boldsymbol{\theta})-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\frac{\mathbb{1}(\min(T_i,t_k)\leq C_i)}{\hat{G}(\min(T_i,t_k)|\boldsymbol{X}_i)}\left\{\mathbb{1}(T_i\geq t_k)-S(t_k|\boldsymbol{X}_i,\boldsymbol{\theta})\right\}^2\right| \\
=&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{k=1}^{s}B(t_k|\boldsymbol{\theta})-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\left[\begin{array}{c}\mathbb{1}(\min(T_i,t_k)\leq C_i)\left\{\mathbb{1}(T_i\geq t_k)-S(t_k|\boldsymbol{X}_i,\boldsymbol{\theta})\right\}^2\\ \times\left(\frac{1}{\hat{G}(\min(T_i,t_k)|\boldsymbol{X}_i)}-\frac{1}{G(\min(T_i,t_k)|\boldsymbol{X}_i)}+\frac{1}{G(\min(T_i,t_k)|\boldsymbol{X}_i)}\right)^2\end{array}\right]\right| \\
\leq&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{k=1}^{s}B(t_k|\boldsymbol{\theta})-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\frac{\mathbb{1}(\min(T_i,t_k)\leq C_i)}{G(\min(T_i,t_k)|\boldsymbol{X}_i)}\left\{\mathbb{1}(T_i\geq t_k)-S(t_k|\boldsymbol{X}_i,\boldsymbol{\theta})\right\}^2\right|+ \\
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\left[\begin{array}{c}\mathbb{1}(\min(T_i,t_k)\leq C_i)\left\{\mathbb{1}(T_i\geq t_k)-S(t_k|\boldsymbol{X}_i,\boldsymbol{\theta})\right\}^2\\ \times\left(\frac{1}{\hat{G}(\min(T_i,t_k)|\boldsymbol{X}_i)}-\frac{1}{G(\min(T_i,t_k)|\boldsymbol{X}_i)}\right)^2\end{array}\right]\right|.
\end{aligned}
$$

Condition 1 gives that the second supremum converges to zero. If we can show that

$$
E\left[\sum_{k=1}^{s}\frac{\mathbb{1}(\min(T,t_k)\leq C)}{G(\min(T,t_k)|\boldsymbol{X}_i)}\left\{\mathbb{1}(T\geq t_k)-S(t_k|\boldsymbol{X},\boldsymbol{\theta})\right\}^2\right]
$$

is bounded, then Lemma 2.4 from Newey and McFadden[43] gives that the first supremum converges to zero. This follows easily from Condition 2,

$$
E\left[\sum_{k=1}^{s}\frac{\mathbb{1}(\min(T,t_k)\leq C)}{G(\min(T,t_k)|\boldsymbol{X})}\left\{\mathbb{1}(T\geq t_k)-S(t_k|\boldsymbol{X},\boldsymbol{\theta})\right\}^2\right]<E\left[\sum_{k=1}^{s}\frac{1}{\delta}\left\{\mathbb{1}(T\geq t_k)-S(t_k|\boldsymbol{X},\boldsymbol{\theta})\right\}^2\right]
$$

which is clearly bounded. We therefore get

$$
\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{s}\frac{\mathbb{1}(\min(T_i,t_k)\leq C_i)}{\hat{G}(\min(T_i,t_k)|\boldsymbol{X}_i)}\left\{\mathbb{1}(T_i\geq t_k)-S(t_k|\boldsymbol{X}_i,\boldsymbol{\theta})\right\}-\sum_{k=1}^{s}B(t_k|\boldsymbol{\theta})\right|\to 0.
$$

Hence, to inspect the asymptotic behaviour we can focus on the minimizer of the expected BSL. We now show that the minimizer of the expected BSL is $\dot{\boldsymbol{\theta}}$ and that this is the only solution. That $\dot{\boldsymbol{\theta}}$ is a solution of the minimization problem

follows directly from the decomposition of the Brier score loss into

$$\sum_{k=1}^{s} B(t_k|\boldsymbol{\theta}) = \sum_{k=1}^{s} \left(E\left[\{\mathbb{1}(T \geq t_k) - S(t_k|\boldsymbol{X}, \dot{\boldsymbol{\theta}})\}^2\right] + E\left[\{S(t_k|\boldsymbol{X}, \dot{\boldsymbol{\theta}}) - S(t_k|\boldsymbol{X}, \boldsymbol{\theta})\}^2\right] \right).$$

The first term within the sum is independent of $\boldsymbol{\theta}$ and the second term reaches a minimum when $\boldsymbol{\theta} = \dot{\boldsymbol{\theta}}$. To show the uniqueness of this minimizer assume that there is a second solution $\ddot{\boldsymbol{\theta}} \neq \dot{\boldsymbol{\theta}}$ for which

$$\sum_{k=1}^{s} B(t_k|\ddot{\boldsymbol{\theta}}) = \sum_{k=1}^{s} B(t_k|\dot{\boldsymbol{\theta}}).$$

Let j be the smallest integer for which the coefficients within the corresponding interval $[\alpha_{j-1}, \alpha_j)$ are different, i.e. $\ddot{\boldsymbol{\theta}}_j \neq \dot{\boldsymbol{\theta}}_j$ Furthermore, from condition 3 we know that there is a $w \in \{1, \ldots, s\}$ such that the evaluation point $t_w \in [\alpha_{j-1}, \alpha_j)$. Given condition 4 we have that

$$P[(\boldsymbol{X}^t \dot{\boldsymbol{\theta}}_j - \ddot{\boldsymbol{\theta}}_j) \neq 0] > 0,$$

which implies that

$$P[S(t_w|\boldsymbol{X}, \dot{\boldsymbol{\theta}}) \neq S(t_w|\boldsymbol{X}, \ddot{\boldsymbol{\theta}})] > 0.$$

And hence $E\left[\left\{S(t_w|\boldsymbol{X}, \dot{\boldsymbol{\theta}}) - S(t_w|\boldsymbol{X}, \ddot{\boldsymbol{\theta}})\right\}^2\right] > 0$ contradicting that $\ddot{\boldsymbol{\theta}}$ was a minimizer of the expected BSL.

Robustness of the Brier score loss against leverage points

For each observation i, the contribution to the BSL estimating equation consists of a finite sum of terms with the following form:

$$\text{IPW}_{ik} \left\{\mathbb{1}(T_i \geq t_k) - S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta})\right\} S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \exp\left(\boldsymbol{X}_i^t \boldsymbol{\theta}_z\right) \int_0^{t_k} \mathbb{1}(u \in [\alpha_z, \alpha_{z+1})) du X_{il}.$$

Note that if $t_k \leq \alpha_z$ this equals zero and we will focus on the case in which $t_k > \alpha_z$ for which $\int_0^{t_k} \mathbb{1}(u \in [\alpha_z, \alpha_{z+1}))du = \zeta > 0$. By defining $c_{ik} = \text{IPW}_{ik} \int_0^{t_k} \mathbb{1}(u \in [\alpha_z, \alpha_{z+1}))du$ we get that

$$
\begin{aligned}
&\left|\text{IPW}_{ik} \left\{\mathbb{1}(T_i \geq t_k) - S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta})\right\} S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta}) \exp\left(\boldsymbol{X}_i^t \boldsymbol{\theta}_z\right) \int_0^{t_k} \mathbb{1}(u \in [\alpha_z, \alpha_{z+1}))du X_{il}\right| \\
&\leq |c_{ik}| \times \left|\mathbb{1}(T_i \geq t_k) - S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta})\right| \times \left|S(t_k|\boldsymbol{X}_i, \boldsymbol{\theta})\right| \times \left|\exp\left(\boldsymbol{X}_i^t \boldsymbol{\theta}_z\right) X_{il}\right| \\
&\leq |c_{ik}| \times \left|\exp\left(-\zeta \exp\left[\boldsymbol{X}_i^t \boldsymbol{\theta}_z\right]\right) \exp\left(\boldsymbol{X}_i^t \boldsymbol{\theta}_z\right) X_{il}\right|.
\end{aligned}
$$

Since c_{ik} is independent of $\boldsymbol{X}$ we can focus on the behavior of $\left|\exp\left(-\zeta \exp\left[\boldsymbol{\theta}_z \boldsymbol{X}_i\right]\right) \exp\left(\boldsymbol{X}_i^t \boldsymbol{\theta}_z\right) X_{il}\right|$. To simplify the notation, we restrict ourselves to the univariate case and drop the i, z, and l subscripts. We now inspect the behaviour of $\frac{\exp(\theta X) X}{\exp(\zeta \exp[\theta X])}$ when $X \to \pm\infty$.

$\theta > 0,\ X \to +\infty$ or $\theta < 0,\ X \to -\infty$

$$
\begin{aligned}
\lim_{X \to \pm\infty} \frac{\exp(\theta X) X}{\exp(\zeta \exp[\theta X])} &= \lim_{X \to \pm\infty} \frac{\exp(\theta X)(1 + \theta X)}{\zeta \theta \exp(\theta X) \exp(\zeta \exp[\theta X])} \\
&= \lim_{X \to \pm\infty} \frac{(1 + \theta X)}{\zeta \theta \exp(\zeta \exp[\theta X])} \\
&= \lim_{X \to \pm\infty} \frac{\theta}{\zeta^2 \theta^2 \exp(\theta X) \exp(\zeta \exp[\theta X])} \\
&= 0.
\end{aligned}
$$

$\theta > 0,\ X \to +\infty$ or $\theta < 0,\ X \to +\infty$

$$\begin{aligned}\lim_{X\to\pm\infty}\frac{\exp(\theta X)\,X}{\exp(\zeta\exp[\theta X])} &= \lim_{X\to\pm\infty}\frac{X}{\exp(-\theta X+\zeta\exp[\theta X])}\\ &= \lim_{X\to\pm\infty}\frac{1}{(-\theta+\zeta\exp[\theta X]\,\theta)\exp(-\theta X+\zeta\exp[\theta X])}\\ &= 0.\end{aligned}$$

All the equations follow from applying L'Hôpital's rule. When $\theta = 0$ we obtain $\lim_{X\to\pm\infty}\frac{\exp(\theta X)X}{\exp(\zeta\exp[\theta X]))} \to \pm\infty$. Hence, leverage points can only have a large influence on the estimating equations of the BSL when $\theta = 0$.