Fast approximations of pseudo-observations in the context of right-censoring and interval-censoring

Olivier Bouaziz

Fast approximations of pseudo-observations in the context of right-censoring and interval-censoring

Olivier Bouaziz[1]

[1]MAP5 (UMR CNRS 8145), Université de Paris

Abstract

In the context of right-censored and interval-censored data we develop asymptotic formulas to compute pseudo-observations for the survival function and the Restricted Mean Survival Time (RMST). Those formulas are based on the original estimators and do not involve computation of the jackknife estimators. For right-censored data, Von Mises expansions of the Kaplan-Meier estimator are used to derive the pseudo-observations. For interval-censored data, a general class of parametric models for the survival function is studied. An asymptotic representation of the pseudo-observations is derived involving the Hessian matrix and the score vector of the density. The formula is illustrated on the piecewise-constant-hazard model for the RMST. The proposed approximations are extremely accurate, even for small sample sizes, as illustrated on Monte-Carlo simulations and real data. We also study the gain in terms of computation time, as compared to the original jackknife method, which can be substantial for large dataset.

Keywords: Pseudo-observations; Restricted Mean Survival Time; Von Mises expansions; Jackknife; interval-censoring.

1 Introduction

In order to study censored data in time to event analysis it is common to model the hazard rate. This allows to correctly take into account censoring in the estimation procedure and provides hazard ratio estimates in the framework of proportional hazard models. However, in some contexts other quantities, that have a more direct interpretation related to the studied problem, might be of interest. One example is the Restricted Mean Survival Time (RMST) which is defined as the average survival time up to a fixed point. In that case, it is common (see [1], [2], [3]) to first model the hazard rate using for instance a Cox model, to derive a survival estimator from the estimated hazard rate and to obtain an estimator of the RMST by integrating out this function. This procedure results in a cumbersome computation where it might be difficult to disentangle the effect of each covariate on the RMST. This is a serious drawback for medical applications and there is a need for more direct approaches. There are several other contexts that are concerned by the difficulty of direct modelling of the quantity of interest. This is typically the case for cumulative incidence functions in a competing risk setting or transition probabilities in a multi-state framework.

Pseudo-observations have been developed in the seminal work of [4] to answer this problem. Those pseudo-observations are constructed using the jackknife method from an estimator of the survival function. Theoretical results in [5] show that the pseudo-observations can then be used as response variables in a regression model for the quantity of interest, such as the conditional RMST, the cumulative incidence functions in a competing risk setting or the transition probabilities in a multi-state framework. This offers the possibility to directly model the quantity of interest and it is often performed by use of a generalised linear model.

Another more recent areas of development involving pseudo-observations concerns the study of machine learning methods for time to event analysis. In this context, the problematic is similar: one aims at deriving a complex model, based for instance on neural networks, for quantities of interest such as the survival function (see [6], [7], [8]), or the RMST (see for instance [9]). The use of pseudo-observations is then appealing since, once the pseudo-observations are obtained, it is possible to directly use any standard machine learning algorithm by considering those pseudo-observations as (non-censored) response variables.

Methods based on pseudo-observations are also attractive for interval-censored data. With those data, it is challenging to build a regression model based on semi-parametric methods, for quantities of interest such as the RMST. This is due to the lack of informations induced by interval-censoring. As a matter of fact, even in a nonparametric setting it may be problematic to perform estimation of the survival function. In this context, one usually relies on the Turnbull estimator or the convex minorant method which were introduced in [10] and [11], respectively. In [11] it has been proved that these estimators achieve the slow rate of convergence of order $n^{1/3}$ and their distribution is not Gaussian and cannot be explicitly computed. In a regression context, for the estimation of the hazard rate, the Cox model with nonparametric baseline was studied in [12] but again, the baseline survival function has the $n^{1/3}$ slow rate of convergence and the asymptotic distribution of this estimator could not be derived. As a result, it is common to rely on fully parametric models for modelling quantities such as the survival function or the hazard rate in a regression context. In [13] and [14] a Cox model was studied using parametric baselines such as Weibull or piecewise constant. The methods used to perform estimation are based on maximum likelihood theory where the parametric estimators are derived by maximising the likelihood of the observed data. This allows to recover the classical $\sqrt{n}$ rate of convergence of the parametric estimators. However, the derivation of the estimators is not explicit, even in the absence of covariates, and rely on a maximisation algorithm such as the Newton-Raphson procedure. In [15] a different approach was proposed based on the EM algorithm by considering the true event times as unobserved variables. This method has the advantage that direct estimators can be computed in the E-step of the algorithm when no covariates are present, which results in a stable and robust estimation procedure. All the aforementioned methods consider estimation of the survival function or of the hazard rate through proportional hazard assumptions, but they are not suited for direct modelling of the RMST, in a regression context. However, this can be achieved by using the pseudo-observations approach. In [16], an illness-death model was considered, and conditional transition probabilities or RMST were computed based on this approach. In order to compute the pseudo-observations, the cumulative transition intensities were estimated using either a penalised spline approach or assuming a Weibull distribution.

The key concept about pseudo-observations is that they are built based on the unconditional jackknife estimator of the quantity of interest. While applying the jackknife is straightforward in practice, a limitation of this method comes from the computation burden of calculating the initial estimator n times, where n is the sample size. This is especially true for interval-censored data where there is no direct calculation of the estimators, even in the absence of covariates. In this paper, we develop approximated formulas for pseudo-observations where the jackknife technique does not need to be implemented. In our formulas, the pseudo-observations can be directly computed based on the initial estimator. In the case of right-censored data, we provide formulas based on Von Mises expansion of the Kaplan-Meier estimator. In the case of interval-censored data, we derive general formulas for parametric models that only involve the original estimator, the score function and the Hessian of the density. Those formulas are approximations of the original jackknife procedure in the sense that they are equal to the original pseudo-observations up to a remainder term that tends towards 0 as n tends to infinity.

However, they turn out to have a very high precision even for moderate sample sizes. Since they only involve the original estimator, the score vector and the Hessian matrix in a parametric

context, they are extremely fast to compute thus resulting in a drastic reduction of time.

In the next section, we present a brief summary on the pseudo-values approach. In Section 3 we develop asymptotic formulas for computing pseudo-observations of the survival function and the RMST in the context of right-censored data. The case of interval censored data is studied in Section 4. We first discuss the context of nonparametric estimation of the survival function in Section 4.1. Then the asymptotic pseudo-observations formulas are developed for general parametric models in Section 4.2. Simulations studies for modelling the conditional RMST in the context of right-censored or interval-censored data are conducted in Section 5 where precision and computation time of the approximate formulas are evaluated. Finally, two real data are analysed using the proposed methodology in Section 6.

2 Backgrounds on pseudo-regression estimation methods

Let $X_1, \ldots, X_n$ be n independent and identically distributed (i.i.d.) variables, let θ be a parameter of the form $\theta = \mathbb{E}[f(X_i)]$, where f is a known function. Then introduce $Z_1, \ldots, Z_n$ n i.i.d. covariates and define the conditional expectation $\theta_{(l)} = \mathbb{E}[f(X_l) \mid Z_l]$. We further assume there exists an invertible link function g such that $g(\theta_{(l)}) = Z_l^T\beta$, where β is a vector of regression parameters of interest. The l^{th} pseudo-observation is given by:

$$\hat{\theta}_{(l)} = n\hat{\theta} - (n-1)\hat{\theta}^{(-l)}, \tag{1}$$

where $\hat{\theta}^{(-l)}$ is the jackknife estimator of $\hat{\theta}$, that is the estimator $\hat{\theta}$ computed on the sample where the l^{th} observation has been removed.

It has been suggested (see [4]) to estimate β based on the estimating equation

$$U(\beta) = \sum_{l=1}^{n} \left(\dot{\theta}_{(l)}\right)^T V_l^{-1}(\hat{\theta}_{(l)} - \theta_{(l)}) = 0,$$

where $\dot{\theta}_{(l)}$ denotes the derivative with respect to β of $\theta_{(l)} = g^{-1}(Z_l^T\beta)$ and V_l is a weight matrix. As a result, the estimator $\hat{\beta}$ verifies the equation $U(\hat{\beta}) = 0$ and it has been suggested to use a sandwich estimator to estimate the variance of $\hat{\beta}$ (see for instance [4] for more details).

In the context of right-censored data, based on the Kaplan-Meier estimator for estimating θ, it has been proved in [5] that the resulting estimating function has a mean asymptotically equal to zero. More specifically, the authors have proved that:

$$\hat{\theta}_{(l)} = \theta + \dot{\psi}(X_l) + o_{\mathbb{P}}(1), \tag{2}$$

where $\dot{\psi}$ is a first order influence function that verifies $\mathbb{E}[\dot{\psi}(X_l) \mid Z_l] = \theta_{(l)} - \theta$. On the other hand, the same authors have shown that the sandwich estimator used to estimate the variance of $\hat{\beta}$ is asymptotically biased. However, they have concluded that the difference between their corrected variance estimator and the usual sandwich estimator is of minor importance and as a consequence it is customary to use the sandwich estimator for pseudo-regression.

Once the pseudo-observations have been computed, implementation of the estimating equation along with the sandwich variance estimator can be easily performed from the `geese` function in the `geepack` R package.

In this article, we will present approximate formulas for computing pseudo-observations in the context of right-censoring in Section 3 and in the context of interval-censoring in Section 4. Instead of directly using Equation (1) to obtain the pseudo-observations, we will present approximated formulas that only involve the estimator $\hat{\theta}$ computed on the whole sample. In both Sections 3 and 4, we will first focus our attention on the problem of modelling $\theta_{(l)} = S(t \mid Z_l)$, the conditional survival function evaluated at time t given the covariate Z_l. The pseudo-observations can be computed using an estimator of $\theta = S(t)$ the unconditional survival function. A standard

link function is $g(\cdot) = \log(-\log(\cdot))$ which gives rise to the Cox model. More complex functions can be chosen for g, such as neural-networks (see for instance [6], [7], [8]), which will provide performant prediction methods for the conditional survival function. Based on those results we will also consider the problem of modelling

$$\theta_{(l)} = \mathbb{E}[T^* \wedge \tau \mid Z_l] = \int_0^\tau S(t \mid Z_l)dt, \tag{3}$$

for some $\tau > 0$. This allows to estimate the Restricted Mean Survival Time in a regression context by considering for instance the identity function for g or again more complex link functions such as neural-networks (see for instance [9]). In the context of right-censored data only, θ will be estimated based on the Kaplan-Meier estimator and in the context of interval-censored data, it will based on parametric models.

3 Approximate pseudo-observations for right-censored data

Let T^* be a time to event of interest and suppose that instead of observing T^*, one observes $T = \min(T^*, C)$ and $\Delta = I(T^* \leq C)$ where C is a right-censoring variable. Using the notations of Section 2 we set $X_i = (T_i, \Delta_i)$, $i = 1, \ldots, n$ be n i.i.d replications of (T, Δ). Introduce $H_1(t) = \mathbb{P}(T \leq t, \Delta = 1)$, $H_0(t) = \mathbb{P}(T \leq t, \Delta = 0)$, $H(t) = \mathbb{P}(T \geq t)$ and their empirical counterparts, $\hat{H}_1(t) = \sum_i I(T_i \leq t, \Delta_i = 1)/n$, $\hat{H}_0(t) = \sum_i I(T_i \leq t, \Delta_i = 0)/n$, $\hat{H}(t) = \sum_i I(T_i \geq t)/n$, where $I(\cdot)$ is the indicator function. Introduce also the survival function $S(t) = \mathbb{P}(T^* > t)$ and the Kaplan-Meier estimator $\hat{S}(t)$ [see 17].

In order for Equation (2) to hold true one would usually assume (see [18] or [5]) C to be independent of (T^*, Z) and that estimation of $S(t)$ will be carried out for t in $[0, \tau]$ where τ is such that $H(\tau) > \nu$, with ν a positive constant. The independence assumption can be equivalently decomposed into the two conditions that C is conditionally independent of T^* given Z and that C is independent of Z. The first condition along with the boundness assumption on H are the two standard hypothesis for the Kaplan-Meier estimator to be consistent. On the other hand, the extra independent assumption between C and Z comes from the use of the pseudo-regression. This last assumption can be alleviated by using the methods presented in [19] or in [20]. Nevertheless it should be noted that none of those assumptions are needed for the approximations presented in the next proposition and corollary to hold true.

Now, we define for $l = 1, \ldots, n$, the l^{th} pseudo-estimates $\hat{S}^{(-l)}$, $\hat{H}_1^{(-l)}$ and $\hat{H}^{(-l)}$of $\hat{S}$, $\hat{H}_1$ and $\hat{H}$ as the estimators constructed when omitting the l^{th} observation $X_l = (T_l, \Delta_l)$. We have the following result.

Proposition 1. *The l^{th} pseudo-observation of the Kaplan-Meier estimator $\hat{S}$ satisfies the relation:*

$$\begin{aligned}\hat{S}_{(l)}(t) &= n\hat{S}(t) - (n-1)\hat{S}^{(-l)}(t) \\ &= \hat{S}(t) + \hat{S}(t)\left(\int_0^{T_l \wedge t} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} - \frac{I(T_l \leq t, \Delta_l = 1)}{\hat{H}(T_l)}\right) + o_{\mathbb{P}}(1)\cdot\end{aligned}$$

From this formula it is clear that the pseudo-observations can be approximated from quantities computed on the original sample. In other words pseudo-observations can be computed without performing the jackknife procedure. We will see in the simulation section that this approximation is very accurate even for moderate sample sizes. Besides, the main interest for using this formula is for large sample sizes, in particular in a machine learning context where computing pseudo-observations is the first step of the procedure before applying algorithms such as neural networks. In those contexts, the order of the sample size is often in millions.

We can then directly derive pseudo-observations for the RMST based on the previous proposition.

Corollary 1. *Let $\tau > 0$. The l^{th} pseudo-observation of the estimated RMST $\int_0^\tau \hat{S}(t)dt$ based on the Kaplan-Meier estimator $\hat{S}$ satisfies the relation*

$$\begin{aligned}\int_0^\tau \hat{S}_{(l)}(t)dt &= n\int_0^\tau \hat{S}(t)dt - (n-1)\int_0^\tau \hat{S}_{(l)}(t)dt \\ &= \int_0^\tau \hat{S}(t)dt + \int_{T_l}^\tau \hat{S}(t)dt I(T_l \leq \tau)\int_0^{T_l} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} \\ &\quad + \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i \int_{T_i}^{T_l\wedge\tau} \hat{S}(t)dt}{(\hat{H}(T_i))^2} I(T_i \leq T_l \wedge \tau) - \int_{T_l}^\tau \hat{S}(t)dt \frac{\Delta_l I(T_l \leq \tau)}{\hat{H}(T_l)} + o_{\mathbb{P}}(1)\cdot\end{aligned}$$

Again this formula shows that the pseudo-observations for the RMST can be directly computed from quantities based on the original sample. This results in a drastic reduction of the computation time of those pseudo-observations as illustrated in the simulation section. Proofs of Proposition 1 and Corollary 1 are deferred to the Appendix section.

4 Approximate pseudo-observations for interval-censored data

In this section, we suppose that instead of directly observing T^* we observe a random interval $[L, R]$, $L \geq 0$ and $L \leq R$, which almost surely contains the event time: $\mathbb{P}(T^* \in [L, R]) = 1$. The right end interval is allowed to take the infinite value such that:

- if $0 < L < R < \infty$ the data are strictly interval-censored,
- if $0 = L < R < \infty$ the data are left-censored,
- if $0 < L < R = \infty$ the data are right-censored,
- if $0 < L = R < \infty$ the data are exactly observed.

Using the notations of Section 2 the data then consist of i.i.d. replications $X_i = (L_i, R_i)$, $i = 1, \ldots, n$. This situation is often called interval-censoring case 2 (see [21]) when exact observations are not allowed and mixed interval censoring (see [22]) or partly interval censoring (see [23]) otherwise. In order to derive consistent estimators of the survival function under interval censoring one will usually assume independent censoring in the following way (see for instance [24]): $\mathbb{P}(T^* \leq t \mid L = l, R = r) = \mathbb{P}(T^* \leq t \mid l \leq T^* \leq r)$. This supposes that the variables (L, R) do not convey additional information on the law of T^* apart from assuming T^* to be bracketed by L and R.

4.1 Comments on the nonparametric case

It seems appealing to use the same methodology as in Section 3 for interval-censored data. We would first need to compute a nonparametric estimator of the survival function and we would then use Von-Mises expansion in order to derive approximated formulas for the pseudo-observations. A natural nonparametric estimator in the context of interval-censored data is the Turnbull estimator which can be seen as an EM estimator and is consequently rather slow to compute. The gain for avoiding computing n times the Turnbull estimator would therefore be highly significant.

However, in the general context of interval-censored data, the use of pseudo-observations as defined by Equation (1) is not theoretically valid due to the slow convergence of the nonparametric estimator. In [21] and [25] it has been showed that the nonparametric maximum

likelihood estimator converges at the $n^{1/3}$ or $(n\log(n))^{1/3}$ rates. Therefore it will not be possible to derive a relation of the following type:

$$\hat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \dot{\psi}(X_i) + o_{\mathbb{P}}(1),$$

with ψ verifying $\mathbb{E}[\dot{\psi}(X_i)] = 0$ such as derived in [5] (see also Equation (8) in [18]). Because if that would be the case, the convergence rate of $\hat{\theta}$ would be of the order $n^{1/2}$ due to the central limit theorem. However, this result is crucial to derive Equation (2) and assess the validity of the procedure. Finally, more problems would also arise in the estimation of the asymptotic variance of $\hat{\beta}$ due to the terms contained in the $o_{\mathbb{P}}(1)$ term above (which were omitted for sake of simplicity) and the sandwich variance estimator will certainly not be valid here.

An alternative could be to use results from [23] where it is further assumed that n_1/n tends to a positive constant as n tends to infinity, where n_1 is the number of exact observations. Under this assumption, the authors retrieved a $n^{1/2}$ rate of convergence for the nonparametric maximum likelihood estimator which converges toward a centred gaussian process. However, the covariance function of this process is not explicit and can only be determined as the solution of two integrals. This is caused by the construction of the nonparametric estimator that has no closed form but verifies a self-consistency equation. The asymptotic distribution of the nonparametric estimator was derived using results from infinite dimensional M-estimators in [26]. The same properties in M-estimators could be used here to derive approximated formulas for the nonparametric survival estimator. However, a careful examination of the proofs in [23] shows that such formulas would lead again to implicit expressions of the pseudo-observations in the same form as the asymptotic limit of the nonparametric survival estimator. Since it does not seem possible to approach those expressions in a straightforward manner we will not pursue this idea. We will focus instead in the next section in modelling the survival function using parametric models.

4.2 Parametric modelling of the survival function

We now assume that $X_1, \ldots, X_n$ have a common density function $f(t, \alpha_0)$ where α_0 is the true model parameter of dimension d. We will respectively denote $\Lambda(t;\alpha_0)$ and $S(t;\alpha_0) = \exp(-\Lambda(t;\alpha_0))$ the true cumulative hazard and survival functions. We will use the notations $\nabla f(t;\alpha_0)$ and $\nabla^2 f(t;\alpha_0)$ to represent the score vector and the Hessian matrix where the derivatives are taken with respect to the model parameter α and are evaluated at $\alpha = \alpha_0$. Subject to regularity conditions, the maximum likelihood estimator $\hat{\alpha}$ of α_0 verifies the following equality:

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = \frac{1}{\sqrt{n}}\left(-\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(X_i;\tilde{\alpha})\right)^{-1} \sum_{i=1}^{n} \nabla f(X_i;\alpha_0),$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0. Let $I = -\mathbb{E}\left(\nabla^2 f(X;\alpha_0)\right)$ be the Fisher information and consider the jackknife version $\hat{\alpha}^{(-l)}$ of the maximum likelihood estimator. It is then straightforward to write:

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = \frac{1}{\sqrt{n}}\, I^{-1} \sum_{i=1}^{n} \nabla f(X_i;\alpha_0) + \varepsilon_n,$$

$$\sqrt{n-1}(\hat{\alpha}^{(-l)} - \alpha_0) = \frac{1}{\sqrt{n-1}}\, I^{-1} \sum_{i\neq l}^{n} \nabla f(X_i;\alpha_0) + \varepsilon_n^{(-l)},$$

where

$$\varepsilon_n = \frac{1}{\sqrt{n}}\left(\left(-\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f(X_i;\tilde{\alpha})\right)^{-1} - I^{-1}\right) \sum_{i=1}^{n} \nabla f(X_i;\alpha_0),$$

and $\varepsilon_n^{(-l)}$ is the jackknife version of ε_n. As a result, the l^{th} pseudo-observation of $\hat{\alpha}$ verifies the relation:

$$n\hat{\alpha} - (n-1)\hat{\alpha}^{(-l)} = \alpha_0 + I^{-1}\nabla f(X_l;\alpha_0) + \sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)}. \tag{4}$$

In the Appendix section, it is proved that the term $\sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)}$ tends towards 0 in probability as n tends to infinity. This entails that asymptotically, the pseudo-observation of $\hat{\alpha}$ only depends on the true parameter, the Fisher information and the score vector, this latter quantity being only evaluated at the observation l. Since $\hat{\alpha}$ is a consistent estimator of α_0 and $\hat{I} = -\sum_{i=1}^n \nabla^2 f(X_i;\hat{\alpha})/n$ is a consistent estimator of I, a natural asymptotic approximation for the pseudo-observation of $\hat{\alpha}$ is simply:

$$\hat{\alpha} + \hat{I}^{-1}\nabla f(X_l;\hat{\alpha}).$$

While this result is interesting on its own, more work needs to be done in order to derive the pseudo-observations of $S(t;\hat{\alpha})$. The following proposition is derived based on this latter expression of the approximate pseudo-observation for $\hat{\alpha}$. The notation $\cdot^\top$ is used to denote the transpose of a vector or a matrix.

Proposition 2. *Under standard regularity conditions for maximum likelihood theory, the l^{th} pseudo-observation of the parametric estimator $S(t;\hat{\alpha})$ verifies the relation*

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\hat{\alpha}) - S(t;\hat{\alpha})\nabla\Lambda(t;\hat{\alpha})^\top \hat{I}^{-1}\nabla f(X_l;\hat{\alpha}) + o_{\mathbb{P}}(1).$$

The proof of this proposition can be found in the Appendix section. As in Section 3, the main interest in this result lies in the fact that the approximated version of the pseudo-observation only depends on the parameter estimator $\hat{\alpha}$ and not on its jackknife version. This means that pseudo-observations in parametric models can be obtained without actually computing the n^{th} jackknife estimators. Only the estimator of α_0, along the Hessian matrix, the gradient of Λ and of the density are needed. This is particularly interesting in the context of interval-censored data since parametric estimators cannot be derived explicitly and numeric methods must be implemented. Two different strategies exist for those types of data: either a direct maximisation of the likelihood can be performed using the Newton-Raphson algorithm (see [13] and [14] for instance) or the complete likelihood (based on the unobserved true times) can be used through the EM algorithm in order to maximise the likelihood (see [15]). But in either case the method is iterative. Also, it should be noted that the Newton-Raphson algorithm requires to compute the score vector and Hessian matrix. Therefore the computational cost for implementing the approximated pseudo-observations is similar to the cost of simply computing the pointwise estimate from the Newton-Raphson algorithm.

Pseudo-observations for the RMST are directly derived from Proposition 2 in the next corollary.

Corollary 2. *Let $\tau > 0$. The l^{th} pseudo-observation of the estimated RMST $\int_0^\tau S(t;\hat{\alpha})dt$ based on the parametric estimator $\hat{\alpha}$ of α_0 satisfies the relation*

$$n\int_0^\tau S(t;\hat{\alpha})dt - (n-1)\int_0^\tau S(t;\hat{\alpha}^{(-l)})dt = \int_0^\tau S(t;\hat{\alpha})dt - \int_0^\tau S(t;\hat{\alpha})\nabla\Lambda(t;\hat{\alpha})dt\,\hat{I}^{-1}\nabla f(X_l;\hat{\alpha}) + o_{\mathbb{P}}(1).$$

Those approximated formulas are general and work for any parametric model. As an illustration, the piecewise-constant hazard (pch) model will be used in the simulation section. This models assumes that the hazard function verifies $\lambda(t;\alpha) = \sum_{k=1}^K \alpha_k I_k(t)$ where $I_k(t) = I(c_{k-1} < t \leq c_k)$, $c_0 = 0 < c_1 < \cdots < c_K = +\infty$ represent $K+1$ cuts and $I(\cdot)$ denotes the indicator function. We do not specify precisely the regularity conditions for maximum

likelihood theory to hold. However, two important assumptions are first to assume the model identifiable and second to impose the Fisher information to be positive definite. For the pch model in the context of interval-censored data, two necessary conditions for those regularity assumptions to hold are:

$$\begin{aligned} \mathbb{P}(R < +\infty, [L, R] \cap (c_{k-1}, c_k] \neq \emptyset) > 0, \forall k = 1, \ldots, K, \\ \mathbb{P}(L > c_{k-1}) > 0, \forall k = 1, \ldots, K. \end{aligned} \tag{5}$$

The first assumption is quite natural: in order to estimate α_k, the probability that an interval intersects $[c_{k-1}, c_k]$ should be positive. The second assumption is necessary for the existence of a maximum of the likelihood function. It should be noted that those conditions are also valid when exact observations $L = R$ are allowed. Exact expressions of the score vector and Hessian matrix for the pch model along with the derivation of condition (5) are detailed in Section 9.5 of the Appendix. Details on the implementation of Proposition 2 and Corollary 2 for the pch model are given in Section 9.4 of the Appendix.

Precision and computational cost of the approximation for the RMST are evaluated and compared to the actual jackknife version of the pseudo-observations in the simulation section. In particular, it is seen that the approximation is much faster than the jackknife method and is very accurate even for small sample sizes.

5 Simulation studies for the Restricted Mean Survival Time

We study two different simulation scenarios for the RMST: one with right-censored data and another one with interval-censored data. In the first scenario, the approximate pseudo observations are based on the Kaplan-Meier estimator (using Corollary 1) while in the second scenario they are based on the pch model (using Corollary 2). In both settings, the performance of the estimators derived from the approximated formulas and the ones obtained from the standard jackknife method is compared based on 500 replications. Implementation of the generalised estimation equation is performed through the `geese` function in the `geepack` R package.

5.1 Right-censored data

The simulation setting is based on the one in [27]. We assume that

$$T_i^* = \tilde{\beta}_0^\top X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

with $\tilde{\beta}_0 = (5.5, 0.25, 0.25)^\top$, $X_i = (1, X_i^1, X_i^2)^\top$, X_i^1 and X_i^2 are Bernoulli variables with parameter 0.5 and $\varepsilon_i \sim \mathcal{U}[-\sigma, \sigma]$, with $\sigma = 3$. Under this model it can easily be seen that

$$\mathbb{E}[T_i^* \wedge \tau \mid X_i] = \beta_{00} + \beta_{01} X_i^1 + \beta_{10} X_i^2 + \beta_{11} X_i^1 X_i^2, \tag{6}$$

where $\beta_0 = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})^\top$ can be determined computationally using Monte-Carlo samples with size 10 million. We further set $\tau = 6$ which corresponds to the 54.2% quantile of T^* and to the value $\beta_0 = (4.98, 0.14, 0.14, 0.27)^\top$. Right-censored data were simulated from an exponential distribution with parameter 0.07 yielding 33% of censoring on average. The results are presented in Table 1.

It is seen that the approximated formula gives similar results as compared to the standard jackknife method for $n = 100$. For larger sample sizes, the results are almost identical. In terms of computation times, there is a clear advantage for the approximated formula which goes 10.6, 17.1, 18.3 and 77.2 times faster for $n = 100$, $n = 500$, $n = 1,000$ and $n = 10,000$ respectively. Clearly the computation time for the original jackknife method is not a linear function of the

sample size and the gain for using the approximated method is considerable for large sample sizes. It should be noted that the computation time was evaluated for the pseudo-regression procedure, but it does not include the computation of the initial survival estimator, it only takes into account the computation of the pseudo-observations along with the implementation of the generalised estimating equations.

	Jackknife				Approximated formula			
n	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time	Bias($\hat{\beta}$)	SE($\hat{\beta}$)	MSE($\hat{\beta}$)	Time
100	-0.002	0.282	0.079	0.223 s	0.001	0.278	0.077	0.021 s
	0.001	0.382	0.146		-0.001	0.375	0.140	
	-0.010	0.368	0.136		-0.012	0.361	0.131	
	0.007	0.356	0.127		0.002	0.349	0.122	
500	-0.008	0.120	0.015	1.536 s	-0.007	0.120	0.015	0.090 s
	0.010	0.158	0.025		0.010	0.158	0.025	
	0.005	0.161	0.026		0.005	0.161	0.026	
	0.009	0.158	0.025		0.008	0.158	0.025	
1,000	-0.003	0.081	0.007	3.629 s	-0.003	0.081	0.007	0.198 s
	-0.003	0.111	0.012		-0.004	0.110	0.012	
	0.003	0.112	0.013		0.003	0.112	0.012	
	0.001	0.109	0.012		0.001	0.109	0.012	
$10,000$	-0.001	0.026	0.001	8.659 min	-0.001	0.026	0.001	6.726 s
	0.002	0.034	0.001		0.002	0.034	0.001	
	0.001	0.036	0.001		0.001	0.036	0.001	
	0.003	0.034	0.001		0.003	0.034	0.001	

Table 1: Simulation results for the estimation of β in the RMST model (6) based on pseudo-regression with the Kaplan-Meier estimator on 33% of right-censored data. In the pseudo-regression, the true jackknife is compared to the approximated pseudo-estimates.

5.2 Interval-censored data

For interval censored data the survival function is estimated from the pch model, as detailed in Section 4.2. Using this model, estimation of the model parameter α_0 is performed using the EM algorithm, as presented in [15]. An alternative method could be to directly maximise the observed likelihood but this would result in implementing the Newton-Raphson algorithm for each jackknife sample with inversion of a Hessian matrix of full rank which, in turn, would result in unstable results. In the EM algorithm, the M-step is explicit and as a result the computation of the jackknife methods is always stable. We refer the reader to [15] for more details on the two methods. The approximated method is implemented from the result in Corollary 2 and details on the computation of the score vector and Hessian matrix are detailed in Section 9.4 of the appendix.

5.2.1 First model with a fixed value of τ

We first assume Model (6) with the same values of σ and τ. Then, in order to simulate interval-censored data, a total of $K = 5$ visits were simulated such that $V_1 \sim \mathcal{U}[0, 6]$ and $V_k = V_{k-1} + U[0, 2]$, for $k = 2, \ldots, K$. The observations for which $T_i^* < V_1$ correspond to left-censored observations with $L_i = 0$ and $R_i = V_1$, the observations for which $T_i^* > V_K$ correspond to right-censored observations with $L_i = V_K$ and $R_i = \infty$, and the observations for which $V_{k-1} < T_i^* < V_k$ $(k = 2, \ldots, K)$ correspond to strictly interval-censored observations with $L_i = V_{k-1}$ and $R_i = V_k$. This resulted in 14.6% of left-censored data, 52.07% of interval-censored data and 33.33% of right-censored data. For interval-censored data, the average length of the

intervals was approximately equal to 1.34. The pch model with cuts equal to $4, 5, 6, 7$ was used for the computation of the survival estimator. The pseudo-observations were generated based on the standard jackknife and on the approximated formulas and the results for the RMST model are presented in Table 2.

Again the results between the jackknife and the approximate formula are almost identical while there is a huge gain in terms of computational time for the approximated formula. The approximated formula is 1 688, 2 131, and 3 901 times faster than the jackknife method for $n = 200$, $n = 500$ and $n = 1,000$ respectively. It should be noted that the cuts must be carefully chosen in the pch model. In particular, the regularity conditions of Equation (5) must be satisfied. If there are only few values of L_i and R_i that intersect a cut $[c_{k-1}, c_k]$ or if the proportion of L_i's such that $L_i > c_{k-1}$ is too low then the pseudo-values can be incorrect (both for the jackknife method or using our approximated formula) which will in turn result in a poor performance of the parameters estimation. On the other hand, if the regularity conditions hold, the choice of the cuts will only have a minor impact on the performance of the estimator of β_0 and will lead to similar results.

	Jackknife				Approximated formula			
n	Bias	SE	MSE	Time	Bias	SE	MSE	Time
200	-0.187	0.222	0.085	6.219 min	-0.186	0.223	0.084	0.221 s
	0.045	0.298	0.091		0.046	0.297	0.090	
	0.049	0.307	0.096		0.049	0.305	0.095	
	0.067	0.299	0.094		0.065	0.298	0.093	
500	-0.187	0.150	0.057	23.589 min	-0.187	0.150	0.057	0.664 s
	0.040	0.198	0.041		0.040	0.198	0.041	
	0.037	0.209	0.045		0.037	0.209	0.045	
	0.077	0.186	0.041		0.077	0.186	0.041	
1,000	-0.177	0.104	0.042	87.717 min	-0.177	0.104	0.042	1.349 s
	0.029	0.138	0.020		0.029	0.138	0.020	
	0.041	0.142	0.022		0.041	0.142	0.022	
	0.072	0.130	0.022		0.072	0.130	0.022	

Table 2: Simulation results for the estimation of β in the RMST model (6) based on pseudo-regression with 14.6% of left-censored data, 52.07% of interval-censored data and 33.33% of right-censored data. The piecewise constant hasard model with cuts equal to $4, 5, 6, 7$ was used for the estimation of the survival function in the computation of the pseudo-observations. In the pseudo-regression, the true jackknife is compared to the approximated pseudo-estimates.

5.2.2 Second model with τ equal to infinity

In this scenario we assume a standard linear model for the time of interest:

$$T_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{7}$$

where $\beta_0 = 6$, $\beta_1 = 4$, $X_i \sim \mathcal{U}[0, 2]$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. Here $\tau = \infty$, and for the interval-censored data the values of L_i and R_i were determined through a visit process with a total of $K = 5$ simulated visits such that $V_1 \sim \mathcal{U}[0, 10]$ and $V_k = V_{k-1} + U[0, 4]$, for $k = 2, \ldots, K$. The left, interval and right-censored observations where obtained as in Section 5.2.1. This simulation setting corresponds to 10% of left-censoring, 26% of right-censoring and 64% of interval-censoring. For interval-censored data, the average length of the intervals was approximately equal to 3.5. The results are presented in Table 3 where the pch estimator was used with cuts equal to $6, 8, 10, 12, 14$.

This scenario is challenging both due to the fact that τ equals infinity (and thus causing estimation problems in the tails) and to the width of the intervals that are larger on average than in Section 5.2.1. As a result, the algorithm seems to fail in some rare cases for $n = 500$ and generates a drastic overestimation of the parameter value. This seemed to be caused by the generation of samples for which too few values of L_i and R_i satisfy the regularity conditions of Equation 5. In Table 3, one sample was removed for both the jackknife method and for the approximated formula. Also, 6 other samples were removed for the jackknife method. If all those samples would have been kept, then the results would have been in favour of the approximated formula: while its MSE would have been nearly unchanged, the MSE for the jackknife would have been equal to 3.863 instead of 0.058 (result not shown in the table). Table 3 only displays the results with those 7 samples generating drastic overestimations removed, but even in that case the MSE of the approximated formula is still slightly better than the jackknife method. All samples were kept for $n = 1,000$ and the results are identical for both methods. In terms of computation time, the approximated formula is $2,659$ and $3,791$ times faster than the jackknife method for $n = 500$ and $n = 1,000$ respectively.

	Jackknife				Approximated formula			
n	Bias($\hat{\theta}$)	SE($\hat{\theta}$)	MSE($\hat{\theta}$)	Time	Bias($\hat{\theta}$)	SE($\hat{\theta}$)	MSE($\hat{\theta}$)	Time
500	-0.130	0.202	0.058	24.461 min	-0.114	0.186	0.047	0.552 s
	0.094	0.174	0.039		0.083	0.153	0.030	
1,000	-0.113	0.113	0.025	68.998 min	-0.110	0.113	0.025	1.092 s
	0.080	0.102	0.017		0.078	0.102	0.016	

Table 3: Simulation results for the estimation of β in the RMST model (7) based on pseudo-regression with 10% of left-censored data, 26% of interval-censored data and 64% of right-censored data. The piecewise constant hasard model with cuts equal to $6, 8, 10, 12, 14$ was used for the estimation of the survival function in the computation of the pseudo-observations. In the pseudo-regression, the true jackknife is compared to the approximated pseudo-estimates. In the simulation setting with $n = 500$, 7 samples were removed.

6 Illustrative real data examples

6.1 The Cardiovascular Health Study (CHS)

In this data example, we mimic the analysis of the Cardiovascular Health Study (CHS) as it was performed in [9]. This study was initiated in 1987 to determine the risk factors for development and progression of cardiovascular disease (CVD) in older adults. The event of interest was time to CVD. In [9], the author considers a subsample of $5,380$ individuals of whom 65.2% had CVD during the study period and the others were right-censored. The aim of the study was to estimate the conditional RMST with 29 covariates and $\tau = 5$ years.

The methodology proposed in [9] uses pseudo-observations and implements a deep neural network directly on the pseudo-observations of the RMST, that is the g link function presented in Section 2 is a neural network. Moreover, a training dataset including 75% of the observations and a test set based on the remaining 25% of the data are built in order to evaluate the prediction performance of the method. This split of the data between training and test sets is repeated 10 times. At each repetition, the pseudo-observations must be entirely computed but only on the training datasets. This results in computing the pseudo-observations for the RMST for 10 samples of size $4,035$. We computed those pseudo observations from the jackknife method and the approximated formula. The former was computed in approximately 8.9 minutes while the latter took 38.9 seconds. Therefore, our approximated formula is more than 13.7 times faster than the original jackknife method. Since building a neural network is computationally expen-

sive and needs to be implemented for all the training samples, this reduction in computation time is a major advantage for our approximated formula. Of note, the results of the analysis implemented with the approximated formula are identical to the original analysis (based on the jackknife method) and are therefore omitted.

6.2 The Signal Tandmobiel® data

In this section, we analyse the Signal Tandmobiel® data using the conditional RMST model in Equation (3). This dataset is part of the `bayesSurv` R package. Those data were collected from a longitudinal dental survey of $4,468$ school children born in 1989, who were annually examined by a dentist. The time scale is age in years. The dataset is composed of 0.68% of left-censored data, 61.69% of strictly interval-censored data and 37.63% of right-censored data. Our aim is to study the emergence of the tooth number 14 which is a permanent first premolar. The covariates used for the analysis are: gender (binary variable equal to 1 for boys, 0 for girls) and the number of decayed or missing deciduous first molars due to caries among teeth 54, 64, 74, 84 of the dataset. This covariate is thus discrete taking values between 0 and 4. The survival function is estimated from the pch model using the whole dataset and the pseudo-observations are then computed from the approximated formula in Corollary 2. There are 126 individuals with missing covariates and the generalised estimating equation used to implement the RMST is therefore applied to this reduced dataset composed of $4,342$ pupils.

In the pch model, the number of cuts and locations were chosen using the adaptive-ridge algorithm developed in [15]. This led to the selection of the four cuts 7.6, 8.4, 9 and 10. Since the maximum likelihood estimator has converged this entails that the regularity conditions of Equation (5) are satisfied. We can also easily check them empirically: in particular there are 3% of strictly interval censored observations whose left intervals fell before 7.6, 40% of left intervals that fell after 10 and the percentage of strictly interval censored observations that intersect each other is high (values not shown). The corresponding estimated hazard and survival functions are displayed in Figure 1. We observe a low estimated hazard value (equal to $6{\cdot}10^{-4}$) from age 0 until age 7.6 due the low percentage of left intervals that fell before 7.6. This yields a very flat decay of the survival function on this time period, then the decay increases drastically for the four other time periods $[7.6, 8.4]$, $[8.4, 9]$, $[9, 10]$ and $[10, \infty)$. For illustration, we estimate from the survival function that approximately 83.39% of the teeth will emerge between age 7.6 and 12.

Based on the survival estimate, we took $\tau = 9$ and $\tau = 12$ in the RMST analysis which respectively correspond to the 11% and 84% estimated quantiles of T^* thus corresponding to early and late emergence of the tooth. The estimated regression parameters in the RMST model along with their Wald test are presented in Table 4. For $\tau = 9$ we observe a weak effect of the covariates with an intercept that is almost equal to τ, highlighting that most emergences of the tooth will occur after 9 years of age. As a matter of fact, gender is not significant and the number of decayed deciduous first molars is highly significant but with a weak effect. The number of decayed deciduous first molars will accelerate the emergence of the tooth with 1 decayed molar (respectively 4 decayed molars) yielding a reduction of 0.0097 years (respectively 0.0390 years) for the emergence of the tooth. For $\tau = 12$ the effect of gender is now highly significant, meaning that gender only plays a role for late emergence of the tooth. The emergence of the tooth for boys arrives on average 0.3336 years earlier than for girls. The number of decayed deciduous first molars is also highly significant with 1 decayed molar (respectively 4 decayed molars) yielding a reduction of 0.1303 years (respectively 0.5211 years) for the emergence of the tooth. We also tried to repeat the procedure using different cut values in the pch model and as already observed in the simulation study, this lead to very similar results.

Finally, based on the approximated formulas developed in this paper, the whole procedure (computation of the pseudo-observations and implementation of the generalised estimating equa-

tions) took about 1.78 seconds. The method was not implemented using the classical jackknife method but according to the simulation study it would have taken more than 4 hours to obtain the pseudo-observations, since in the simulation study the time for the jackknife procedure was evaluated at more than 1 hour for $n = 1,000$ (see Sections 5.2.1 and 5.2.2). Also, the results would have been identical, thus highlighting the relevance of the proposed approach in practical situations.

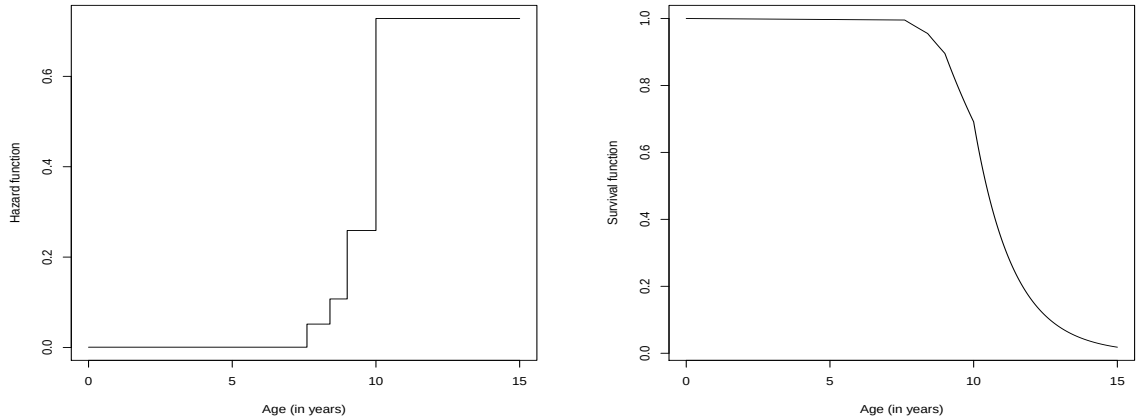

Figure 1: Distribution of time to emergence of the tooth number 14. On the left: estimated hazard function. On the right: estimated survival function. Those estimates were obtained from the pch model with cuts equal to $7.6, 8.4, 9$ and 10.

	$\tau = 9$			$\tau = 12$		
Covariates	**effect**	**se**	**p-value**	**effect**	**se**	**p-value**
Intercept	8.9851	0.0047	$< 10^{-15}$	10.8755	0.0306	$< 10^{-15}$
Gender (1 = boy)	-0.0097	0.0066	0.1422	-0.3336	0.0361	$< 10^{-15}$
Nb of decayed molars	-0.0180	0.0024	8.1379×10^{-14}	-0.1303	0.0120	$< 10^{-15}$

Table 4: Restricted Mean Survival Time Model for the time to emergence of the tooth 14 with the covariates gender and number of decayed or missing deciduous first molars due to caries among teeth 54, 64, 74, 84. Two values of τ are analysed in Equation (3). **se** represents the standard estimate of the regression parameter.

7 Conclusion

In this paper, we presented asymptotic formulas for computing pseudo-observations for time to event data. In the context of right-censored data, those formulas are based on the Kaplan-Meier estimator of the survival function. When dealing with interval-censoring our formulas were developed for a general class of parametric models. Pseudo-regression is an appealing tool when the goal is to directly model a complex quantity of interest, such as the RMST, cumulative incidence functions in a competing risk setting or transition probabilities for multi-state models. Our formulas were precisely developed for the RMST but they could be easily extended for those other quantities of interest. While the pseudo-values approach is originally based on the jackknife procedure, our formulas only involve quantities computed on the initial sample. This results in a drastic reduction of the computational time, which is an interesting feature when dealing with large dataset or when the data are interval-censored, since in that case, the estimators are computationally intensive.

There has been an increasing interest of the pseudo-values approach in the machine learning community. After having computed the pseudo-observations, standard machine learning models can be applied to those new observations, by simply ignoring the censoring. In particular, this methodology has been applied for estimating the survival function in [6], [7], [8] or the RMST in [9] based on neural networks that were directly applied on the pseudo-observations. Therefore, our formulas are particularly interesting in those settings where the dataset can be extremely large and the algorithm usually relies on a cross-validation procedure. Using our approximated formulas results in a significant gain in terms of computation time as illustrated on the real data analysis. Also, the approximations made by our formulas are extremely precise, even for moderate sample sizes, as shown in the simulation study. Surprisingly, we also saw that our formulas are more robust than the original jackknife method which sometimes fails due to some rare extreme values. For all these reasons, we advocate the use of our asymptotic formulas in practical situations.

8 Software

The asymptotic formulas developed in this paper for the pseudo-values of the survival function and the RMST can be implemented using the GitHub package `FastPseudo` available at https://github.com/obouaziz/FastPseudo. The package can deal with both right-censored or interval-censored data. In the latter case, the formulas are implemented for the pch model.

9 Appendix

9.1 Proof of Proposition 1

Let $\psi(A)(s,t] = \prod_{s<u\leq t}(1+dA(u))$, such that $\psi(\Lambda)(0,t] = S(t)$, where $\Lambda(t)$ is the cumulative hazard function and $\psi(\hat{\Lambda})(0,t] = \hat{S}(t)$. We have the following Von-Mises expansion [see 28, 29]:

$$\hat{S}^{(-l)}(t) = \hat{S}(t) - \hat{S}(t)(\hat{\Lambda}^{(-l)}(t) - \hat{\Lambda}(t)) + o_{\mathbb{P}}(\hat{\Lambda}^{(-l)}(t) - \hat{\Lambda}(t)).$$

We now derive a Von-Mises expansion for $\hat{\Lambda}^{(-l)}(t) - \hat{\Lambda}(t)$. The cumulative hazard function and its estimator can be defined as functions of H, H_1 and of $\hat{H}_1$, $\hat{H}$ respectively where $\Lambda(t) = g(H_1, H) := \int_0^t dH_1(u)/H(u)$ and $\hat{\Lambda}(t) = g(\hat{H}_1, \hat{H})$. We have the following Von-Mises expansion:

$$\hat{\Lambda}^{(-l)}(t) = \hat{\Lambda}(t) + g'_{(\hat{H}_1,\hat{H})}(\hat{H}_1^{(-l)} - \hat{H}_1, \hat{H}^{(-l)} - \hat{H}) + o_{\mathbb{P}}(n^{-1}),$$

where g' is the Hadamard derivative of g, which is equal to [see 28, 29]:

$$g'_{(H_1,H)}(h_1,h) = \int_0^t \frac{dh_1}{H} - \int_0^t \frac{h_2 dH_1}{H^2}.$$

The $o_{\mathbb{P}}(n^{-1})$ term above comes from the expressions:

$$\hat{H}_1^{(-l)}(t) - \hat{H}_1(t) = \frac{1}{n(n-1)}\sum_{i=1}^{n} I(T_i \leq t, \Delta_i = 1) - \frac{I(T_l \leq t, \Delta_l = 1)}{n-1}$$

and

$$\hat{H}^{(-l)}(t) - \hat{H}(t) = \frac{1}{n(n-1)}\sum_{i=1}^{n} I(T_i \geq t) - \frac{I(T_l \geq t)}{n-1},$$

which entail as a consequence that $\hat{H}_1^{(-l)}(t) - \hat{H}_1(t)$ and $\hat{H}^{(-l)}(t) - \hat{H}(t)$ are $O_{\mathbb{P}}(n^{-1})$. Moreover, using those expressions we have

$$\begin{aligned} g'_{(\hat{H}_1,\hat{H})}(\hat{H}_1^{(-l)} - \hat{H}_1, \hat{H}^{(-l)} - \hat{H}) &= \frac{1}{n-1}\int_0^t \frac{d\hat{H}_1(u)}{\hat{H}(u)} - \frac{1}{n-1}\frac{I(T_l \le t, \Delta_l = 1)}{\hat{H}(T_l)} \\ &\quad - \frac{1}{n-1}\int_0^t \frac{d\hat{H}_1(u)}{\hat{H}(u)} + \frac{1}{n-1}\int_0^t \frac{I(T_l \ge u)d\hat{H}_1(u)}{(\hat{H}(u))^2}, \\ &= -\frac{1}{n-1}\frac{I(T_l \le t, \Delta_l = 1)}{\hat{H}(T_l)} + \frac{1}{n-1}\int_0^t \frac{I(T_l \ge u)d\hat{H}_1(u)}{(\hat{H}(u))^2}. \end{aligned}$$

The result follows gathering all the different parts.

9.2 Proof of Corollary 1

From Proposition 1 we directly have

$$\int_0^\tau \hat{S}_{(l)}(t)dt = \int_0^\tau \hat{S}(t)dt + \int_0^\tau \hat{S}(t)\left(\int_0^{T_l \wedge t} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} - \frac{I(T_l \le t, \Delta_l = 1)}{\hat{H}(T_l)}\right)dt + o_{\mathbb{P}}(1).$$

We first use the following decomposition:

$$\int_0^{T_l \wedge t} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} = \int_0^{T_l} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} I(T_l \le t) + \int_0^t \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} I(T_l > t),$$

such that

$$\int_0^\tau \hat{S}(t)\int_0^{T_l \wedge t} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2}dt = \int_{T_l}^\tau \hat{S}(t)dt \int_0^{T_l} \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} I(T_l \le \tau) + \int \hat{S}(t) I(t < \tau \wedge T_l)\int_0^t \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2}dt.$$

In the second part of the equation, write:

$$\int_0^t \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2} = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i I(T_i \le t)}{(\hat{H}(T_i))^2}.$$

Therefore,

$$\int \hat{S}(t) I(t < \tau \wedge T_l)\int_0^t \frac{d\hat{H}_1(u)}{(\hat{H}(u))^2}dt = \frac{1}{n}\sum_{i=1}^n \frac{\Delta_i \int_{T_i}^{T_l \wedge \tau} \hat{S}(t)dt}{(\hat{H}(T_i))^2} I(T_i \le T_l \wedge \tau).$$

Finally,

$$\int_0^\tau \hat{S}(t)\frac{I(T_l \le t, \Delta_l = 1)}{\hat{H}(T_l)}dt = \int_{T_l}^\tau \hat{S}(t)\frac{I(\Delta_l = 1)}{\hat{H}(T_l)}dt I(T_l \le \tau),$$

which concludes the proof.

9.3 Proof of Proposition 2 and Corollary 2

Starting with Equation (4) we will first prove that $\sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)}$ tends to 0 in probability as n tends to infinity. Set

$$\tilde{I}_n = -\frac{1}{n}\sum_{i=1}^n \nabla^2 f(X_i; \tilde{\alpha}), \quad \tilde{I}_n^{(-l)} = -\frac{1}{n-1}\sum_{i \ne l} \nabla^2 f(X_i; \tilde{\alpha}),$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0. We have:

$$\begin{aligned}\sqrt{n}\,\varepsilon_n - \sqrt{n-1}\,\varepsilon_n^{(-l)} &= \left(\tilde{I}_n^{-1} - I^{-1}\right)\sum_{i=1}^n \nabla f(X_i;\alpha_0) - \left(\left(\tilde{I}_n^{(-l)}\right)^{-1} - I^{-1}\right)\sum_{i\neq l}^n \nabla f(X_i;\alpha_0)\\ &= \left(\left(\tilde{I}_n^{(-l)}\right)^{-1} - I^{-1}\right)\nabla f(X_l;\alpha_0) + \left(\tilde{I}_n^{-1} - \left(\tilde{I}_n^{(-l)}\right)^{-1}\right)\sum_{i=1}^n \nabla f(X_i;\alpha_0).\end{aligned}$$

Clearly for I positive definite, $\left(\left(\tilde{I}_n^{(-l)}\right)^{-1} - I^{-1}\right)\nabla f(X_l;\alpha_0)$ tends to 0 in probability. Since $\sum_{i=1}^n \nabla f(X_i;\alpha_0)/n$ tends to $\mathbb{E}[\nabla f(X;\alpha_0)] = 0$ in probability, we just need to prove that $\tilde{I}_n^{-1} - \left(\tilde{I}_n^{(-l)}\right)^{-1} = O_{\mathbb{P}}(1/n)$ to conclude the proof. Write:

$$\tilde{I}_n^{-1} - \left(\tilde{I}_n^{(-l)}\right)^{-1} = \tilde{I}_n^{-1}(\tilde{I}_n^{(-l)} - \tilde{I}_n)\left(\tilde{I}_n^{(-l)}\right)^{-1}.$$

From the law of large numbers, $\tilde{I}_n^{-1}$ and $\left(\tilde{I}_n^{(-l)}\right)^{-1}$ tend towards I^{-1} in probability and

$$\tilde{I}_n^{(-l)} - \tilde{I}_n = -\frac{1}{n(n-1)}\sum_{i\neq l}\nabla^2 f(X_i;\tilde{\alpha}) + \frac{1}{n}\nabla^2 f(X_l;\tilde{\alpha}) = O_{\mathbb{P}}\left(1/n\right).$$

This proves that

$$n\hat{\alpha} - (n-1)\hat{\alpha}^{(-l)} = \alpha_0 + I^{-1}\nabla f(X_l;\alpha_0) + o_{\mathbb{P}}(1). \tag{8}$$

Using the consistency of $\Lambda(t;\hat{\alpha})$ towards $\Lambda(t;\alpha_0)$ from standard maximum likelihood theory, we now write a Taylor expansion for the cumulative hazard function around α_0:

$$\Lambda(t;\hat{\alpha}) = \Lambda(t;\alpha_0) + (\hat{\alpha}-\alpha_0)^\top\nabla\Lambda(t;\alpha_0) + \frac{1}{2}(\hat{\alpha}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0), \tag{9}$$

where $\tilde{\alpha}$ lies between $\hat{\alpha}$ and α_0. We also write a Taylor expansion for the function $x\mapsto\exp(-x)$ around 0:

$$\begin{aligned}\exp(-(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0))) = 1 - \big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) + \frac{1}{2}\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2\\ - \frac{1}{6}e^{\xi_n}\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^3,\end{aligned}$$

with ξ_n tends to 0 in probability as n tends to infinity. This can be rewritten as:

$$S(t;\hat{\alpha}) = S(t;\alpha_0) + S(t;\alpha_0)\left(-\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) + \frac{1}{2}\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2\right) + o_{\mathbb{P}}(1/n),$$

using the fact that $\sqrt{n}(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0))$ converges in distribution towards a centred gaussian variable with finite variance from standard results on maximum likelihood theory and the delta-method. As a result,

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\alpha_0) + A_{n,1} + A_{n,2} + o_{\mathbb{P}}(1), \tag{10}$$

where

$$\begin{aligned}A_{n,1} &= -S(t;\alpha_0)\left(n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big)\right),\\ A_{n,2} &= S(t;\alpha_0)\left(n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big)^2\right)\frac{1}{2}.\end{aligned}$$

We start with the $A_{n,2}$ term. From Equation (9) we have:

$$\begin{aligned} n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 &= n(\hat{\alpha}-\alpha_0)^\top \nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(\hat{\alpha}-\alpha_0) \\ &\quad + n(\hat{\alpha}-\alpha_0)^\top \nabla\Lambda(t;\alpha_0)(\hat{\alpha}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0) \\ &\quad + \frac{n}{4}\Big((\hat{\alpha}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0)\Big)^2. \end{aligned}$$

Using the consistency of $\hat{\alpha}-\alpha_0$ and the asymptotic normality of $\sqrt{n}(\hat{\alpha}-\alpha_0)$ from standard maximum likelihood theory, each of the last two terms in the above equation tends to 0 in probability as n tends to infinity. Therefore

$$\begin{aligned} &n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big)^2 - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big)^2 \\ &\quad = (\hat{\alpha}^{(-l)}-\alpha_0)^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(n(\hat{\alpha}-\alpha_0)-(n-1)(\hat{\alpha}^{(-l)}-\alpha_0)) \\ &\qquad + n(\hat{\alpha}-\hat{\alpha}^{(-l)})^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(\hat{\alpha}-\alpha_0) + o_{\mathbb{P}}(1) \\ &\quad = (\hat{\alpha}^{(-l)}-\alpha_0)^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(I^{-1}\nabla f(X_l;\alpha_0)+R_n) \\ &\qquad + (\alpha_0-\hat{\alpha}^{(-l)}+I^{-1}\nabla f(X_l;\alpha_0)+R_n')^\top\nabla\Lambda(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top(\hat{\alpha}-\alpha_0)+o_{\mathbb{P}}(1), \end{aligned}$$

where the last two lines were derived from Equation (8) and R_n, R_n' both tend to 0 in probability. The consistency of $\hat{\alpha}$ and $\hat{\alpha}^{(-l)}$ shows that $A_{n,2} = o_{\mathbb{P}}(1)$. We now study the term $A_{n,1}$. From Equation (9),

$$\begin{aligned} &n\big(\Lambda(t;\hat{\alpha}) - \Lambda(t;\alpha_0)\big) - (n-1)\big(\Lambda(t;\hat{\alpha}^{(-l)}) - \Lambda(t;\alpha_0)\big) \\ &\quad = (n(\hat{\alpha}-\alpha_0)-(n-1)(\hat{\alpha}^{(-l)}-\alpha_0))^\top\nabla\Lambda(t;\alpha_0) \\ &\qquad + \frac{1}{2}(\hat{\alpha}^{(-l)}-\alpha_0)^\top\nabla^2\Lambda(t;\tilde{\alpha})(n(\hat{\alpha}-\alpha_0)-(n-1)(\hat{\alpha}^{(-l)}-\alpha_0)) \\ &\qquad + \frac{1}{2}n(\hat{\alpha}-\hat{\alpha}^{(-l)})^\top\nabla^2\Lambda(t;\tilde{\alpha})(\hat{\alpha}-\alpha_0). \end{aligned}$$

Using similar arguments as before, the last two lines of this Equation tend to 0 in probability from Equation (8) and from the consistency of $\hat{\alpha}$ and $\hat{\alpha}^{(-l)}$. Finally, using again Equation (8)

$$A_{n,1} = -S(t;\alpha_0)\nabla f(X_l;\alpha_0)^\top I^{-1}\nabla\Lambda(t;\alpha_0) + o_{\mathbb{P}}(1).$$

This equality combined with Equation (10) give

$$nS(t;\hat{\alpha}) - (n-1)S(t;\hat{\alpha}^{(-l)}) = S(t;\alpha_0) - S(t;\alpha_0)\nabla\Lambda(t;\alpha_0)^\top I^{-1}\nabla f(X_l;\alpha_0) + o_{\mathbb{P}}(1).$$

The final result of Proposition 2 is obtained by simply replacing each quantity by its consistent estimator. Integrating the equation in Proposition 2 directly yields Corollary 2.

9.4 Log-likelihood, score vector and Hessian matrix in the piecewise constant hazard model

In this section, we study the parametric piecewise constant hazard model defined as follows: $\lambda(t;\alpha) = \sum_{k=1}^K \alpha_k I_k(t)$ where $I_k(t) = I(c_{k-1} < t \leq c_k)$, $c_0 = 0 < c_1 < \cdots < c_K = +\infty$. The cumulative hazard function is then equal to

$$\Lambda(t;\alpha) = \sum_{k=1}^K \alpha_k(c_k \wedge t - c_{k-1})I(c_{k-1} \leq t).$$

Under the mixed-case of interval-censored and exact data, we can directly write the log-likelihood as the sum between the log-likelihood of strictly interval-censored observations and

the log-likelihood of exact observations. For the latter part see [30]. Recall that $X_i = (L_i, R_i)$ and $f(X_i;\alpha)$ denotes the density of the observations with parameter α evaluated at X_i. For strictly interval-censored data ($L_i \neq R_i$), the log-likelihood $\ell(\alpha)$ can be written as (see [14] or [15])

$$\ell(\alpha) = \sum_{i=1}^{n} f(X_i;\alpha) = \sum_{i=1}^{n} \Big\{ -(1-\Delta_i)\Lambda(L_i;\alpha) + \Delta_i \Big(\log\Big(1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)\Big) - \Lambda(L_i;\alpha)\Big) \Big\},$$

where we used the notation $\Delta_i = I(R_i < +\infty)$ to denote uncensored observations. The k^{th} component of the score vector is equal to:

$$\frac{\partial \ell(\alpha)}{\partial \alpha_k} = \sum_{i=1}^{n} \frac{\partial f(X_i;\alpha)}{\partial \alpha_k} = \sum_{i=1}^{n} \Big\{ -(c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i) + \Delta_i \frac{(c_k \wedge R_i - c_{k-1}) I(c_{k-1} \leq R_i) - (c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i)}{1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)} \times \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big) \Big\}. \tag{11}$$

The $k \times k'$ component of the Hessian matrix is equal to:

$$\begin{aligned} \frac{\partial^2 \ell(\alpha)}{\partial \alpha_{k'} \partial \alpha_k} = -\sum_{i=1}^{n} \Delta_i \Bigg\{ & \frac{(c_k \wedge R_i - c_{k-1}) I(c_{k-1} \leq R_i) - (c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i)}{1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)} \\ & \times \big((c'_k \wedge R_i - c_{k'-1}) I(c_{k'-1} \leq R_i) - (c'_k \wedge L_i - c_{k'-1}) I(c_{k'-1} \leq L_i) \big) \\ & \times \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big) \\ & + \frac{(c_k \wedge R_i - c_{k-1}) I(c_{k-1} \leq R_i) - (c_k \wedge L_i - c_{k-1}) I(c_{k-1} \leq L_i)}{\Big(1 - \exp\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)\Big)^2} \\ & \times \big((c'_k \wedge R_i - c_{k'-1}) I(c_{k'-1} \leq R_i) - (c'_k \wedge L_i - c_{k'-1}) I(c_{k'-1} \leq L_i) \big) \\ & \times \exp\Big(2\big(\Lambda(L_i;\alpha) - \Lambda(R_i;\alpha)\big)\Big) \Bigg\}. \end{aligned} \tag{12}$$

The Fisher information is equal to the expectation of minus the Hessian matrix divided by n. Looking at its expression, we directly see that a necessary condition for the Fisher information to be positive definite is to assume that

$$\mathbb{P}(\Delta = 1, [L,R] \cap (c_{k-1}, c_k] \neq \emptyset) > 0, \forall k = 1, \ldots, K.$$

Another important condition for the model to be identifiable is to assume that $\mathbb{E}_{\alpha_0}[f(X;\alpha)]$ has a unique maximum with respect to α, equal to α_0, where the notation $\mathbb{E}_{\alpha_0}$ means the expectation is taken with respect to the true parameter α_0. However, it is clear from Equation (11) that $\mathbb{E}_{\alpha_0}[\partial f(X;\alpha)/\partial\alpha]$ cannot vanish if $\mathbb{P}(L > c_{k-1}) = 0$. Therefore, a second necessary condition for the model to be identifiable is to assume

$$\mathbb{P}(L > c_{k-1}) > 0, \forall k = 1, \ldots, K.$$

Those two conditions have opposite effects on the estimation method if they are violated. In case the first one is not valid for a given k then it will not be possible to compute the corresponding estimator $\hat{\alpha}_k$ from the Newton-Raphson algorithm (since the Hessian will not be invertible) while using the EM algorithm (which does not involve the Score vector nor the Hessian matrix), the

estimator $\hat{\alpha}_k$ will become smaller at each iteration step until eventually reaching the value 0. This situation can be numerically resolved in the latter case, by simply setting the iterated estimate $\hat{\alpha}_k$ to 0 when it reaches a value below a fixed threshold. However this situation is problematic for the computation of the pseudo-values. This can be easily seen by recalling that pseudo-values should average to the initial estimator. In Proposition 2 and Corollary 2 this simply follows from the fact that $\sum_{l=1}^{n} \nabla f(X_l; \hat{\alpha}) = 0$ from regularity conditions for maximum likelihood estimation. However, the k^{th} component of the score vector will never vanish if the first condition is not valid, leading to incorrect pseudo-values.

On the other hand, if the second condition is not valid for a given k, the algorithm will attempt to minimise the term $\exp(-\Lambda(R_i; \alpha))$ from Equation (11) and as a consequence the corresponding estimator $\hat{\alpha}_k$ will become larger at each iteration step of the EM algorithm, diverging to infinity.

Finally, note that if the log-likelihood only include exact observations $L = R$, the conditions then translate to $\mathbb{P}(c_{k-1} < L < c_k) > 0, \forall k = 1, \dots, K$.

9.5 Implementation of the pseudo-observations for the survival function and the RMST in the pch model

In this section we provide the precise expression of the terms involved in Proposition 2 and Corollary 2 for the pch model. We have

$$S(t; \alpha) = \exp\Big(- \sum_{k=1}^{K} \alpha_k (t \wedge c_k - c_{k-1}) I(c_{k-1} \leq t) \Big)$$
$$\frac{\partial \Lambda(t; \alpha)}{\partial \alpha_k} = (c_k \wedge t - c_{k-1}) I(c_{k-1} \leq t),$$

while the expression of the gradient of the density $\nabla f(X_l; \alpha)$ is given by the term between brackets in Equation (11) and $\hat{I}$ is equal to minus the Hessian matrix (see Equation (12)) divided by n.

For Corollary 2 we need to precise how to compute the integral between 0 and τ of $S(t; \alpha)$ and the integral between 0 and τ of $S(t; \alpha) \nabla \Lambda(t; \alpha)$. We first notice that

$$S(t; \alpha) = \exp\Big(- \sum_{k=1}^{K} \alpha_k (c_k - c_{k-1}) I(c_k \leq t) \Big) \exp\Big(- \sum_{k=1}^{K} \alpha_k (t - c_{k-1}) I(c_{k-1} \leq t \leq c_k) \Big),$$

and

$$\begin{aligned}
\int_0^\tau S(t; \alpha) dt &= \sum_{l=1}^{K} \int_{c_{l-1}}^{c_l \wedge \tau} S(t; \alpha) dt \; I(\tau > c_{l-1}) \\
&= \sum_{l=1}^{K} \int_{c_{l-1}}^{c_l \wedge \tau} \exp\Big(- \sum_{k=1}^{K} \alpha_k (c_k - c_{k-1}) I(c_k \leq t) \Big) \exp\Big(- \alpha_l (t - c_{l-1}) \Big) dt \, I(\tau > c_{l-1}).
\end{aligned}$$

Set $A_1 = 0$ and for $l \geq 2$, define

$$A_l = - \sum_{k=1}^{l-1} \alpha_k (c_k - c_{k-1}) + c_{l-1} \alpha_l.$$

For the first term we now have:

$$\begin{aligned}
\int_0^\tau S(t; \alpha) dt &= \sum_{l=1}^{K} \exp(A_l) \int_{c_{l-1}}^{c_l \wedge \tau} \exp(-\alpha_l t) dt \, I(\tau > c_{l-1}) \\
&= \sum_{l=1}^{K} \exp(A_l) \alpha_l^{-1} \Big(\exp(-\alpha_l c_{l-1}) - \exp\big(- \alpha_l (c_l \wedge \tau) \big) \Big) I(\tau > c_{l-1}).
\end{aligned}$$

For the second term we have:

$$\begin{aligned}\int_0^\tau S(t;\alpha)\frac{\partial\Lambda(t;\alpha)}{\partial\alpha_k}dt &= \sum_{l=1}^{K}\int_{c_{l-1}}^{c_l\wedge\tau} S(t;\alpha)(c_k\wedge t - c_{k-1})I(c_{k-1}\le t)dt\, I(\tau > c_{l-1}) \\ &= \sum_{l=k+1}^{K}\int_{c_{l-1}}^{c_l\wedge\tau} S(t;\alpha)dt\,(c_k - c_{k-1})I(\tau > c_{l-1}) \\ &\quad + \int_{c_{k-1}}^{c_k\wedge\tau} tS(t;\alpha)dt\, I(\tau > c_{k-1}) - c_{k-1}\int_{c_{k-1}}^{c_k\wedge\tau} S(t;\alpha)dt\, I(\tau > c_{k-1}).\end{aligned}$$

From the previous calculation on the first term, we easily see that on the one hand

$$\int_{c_{l-1}}^{c_l\wedge\tau} S(t;\alpha)dtI(\tau > c_{l-1}) = \exp(A_l)\alpha_l^{-1}\Big(\exp(-\alpha_l c_{l-1}) - \exp\big(-\alpha_l(c_l\wedge\tau)\big)\Big)\, I(\tau > c_{l-1}).$$

On the other hand, we have:

$$\begin{aligned}&\int_{c_{l-1}}^{c_l\wedge\tau} tS(t;\alpha)dtI(\tau > c_{l-1}) \\ &= \exp(A_l)\int_{c_{l-1}}^{c_l\wedge\tau} t\exp(-t\alpha_l)dtI(\tau > c_{l-1}) \\ &= \exp(A_l)\Bigg(\alpha_l^{-2}\Big(\exp\big(-c_{l-1}\alpha_l\big) - \exp\big(-(c_l\wedge\tau)\alpha_l\big)\Big) \\ &\quad + \alpha_l^{-1}\Big(c_{l-1}\exp\big(-c_{l-1}\alpha_l\big) - (c_l\wedge\tau)\exp\big(-(c_l\wedge\tau)\alpha_l\big)\Big)\Bigg)I(\tau > c_{l-1}),\end{aligned}$$

where the last equation was obtained using integration by parts. Gathering all elements allows to implement the equation in Corollary 2.

References

[1] David M Zucker. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442):702–709, 1998.

[2] Pei-Yun Chen and Anastasios A Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001.

[3] Min Zhang and Douglas E Schaubel. Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*, 67(3):740–749, 2011.

[4] Per Kragh Andersen, John P Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

[5] Martin Jacobsen and Torben Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862, 2016.

[6] Lili Zhao and Dai Feng. Dnnsurv: Deep neural networks for survival analysis using pseudo values. *arXiv preprint arXiv:1908.02337*, 2019.

[7] Lili Zhao and Dai Feng. Deep neural networks for survival analysis using pseudo values. *IEEE journal of biomedical and health informatics*, 24(11):3308–3314, 2020.

[8] Dai Feng and Lili Zhao. Bdnnsurv: Bayesian deep neural networks for survival analysis using pseudo values. *arXiv preprint arXiv:2101.03170*, 2021.

[9] Lili Zhao. Deep neural networks for predicting restricted mean survival times. *Bioinformatics*, 2021.

[10] Bruce W Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

[11] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science and Business Media, 1992.

[12] Jian Huang and Jon A Wellner. Efficient estimation for the proportional hazards model with" case 2" interval censoring. Technical report, 1995.

[13] JK Lindsey. A study of interval censoring in parametric regression models. *Lifetime data analysis*, 4(4):329–354, 1998.

[14] Jianguo Sun. *The statistical analysis of interval-censored failure time data*. Springer Science and Business Media, 2007.

[15] Olivier Bouaziz, Eva Lauridsen, and Grégory Nuel. Regression modelling of interval-censored data based on the adaptive-ridge procedure. *Journal of Applied Statistics*, To appear.

[16] Camille Sabathe, Per K Andersen, Catherine Helmer, Thomas A Gerds, Hélène Jacqmin-Gadda, and Pierre Joly. Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical methods in medical research*, 29(3):752–764, 2020.

[17] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[18] Frederik Graw, Thomas A Gerds, and Martin Schumacher. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255, 2009.

[19] Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis. *Statistical methods in medical research*, 19(1):71–99, 2010.

[20] Nadine Binder, Thomas A Gerds, and Per Kragh Andersen. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis*, 20(2):303–315, 2014.

[21] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media, 1992.

[22] Qiqing Yu, Linxiong Li, and George YC Wong. On consistency of the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, 27(1):35–44, 2000.

[23] Jian Huang. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, pages 501–519, 1999.

[24] Zhigang Zhang, Liuquan Sun, Xingqiu Zhao, and Jianguo Sun. Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, 33(1):61–70, 2005.

[25] Jon A Wellner. Interval censoring, case 2: alternative hypotheses. *Lecture Notes-Monograph Series*, pages 271–291, 1995.

[26] AW Van der Vaart. Efficiency. of infinite dimensional m-estimators. *Statistica Neerlandica*, 49(1):9–30, 1995.

[27] Xin Wang and Douglas E Schaubel. Modeling restricted mean survival time under general censoring mechanisms. *Lifetime data analysis*, 24(1):176–199, 2018.

[28] Richard D Gill. Lectures on survival analysis. In *Lectures on Probability Theory*, pages 115–241. Springer, 1994.

[29] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[30] O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer, 2008.