# Regression modelling of interval censored data based on the adaptive ridge procedure

Olivier Bouaziz, Eva Lauridsen, Grégory Nuel

# Regression modelling of interval censored data based on the adaptive ridge procedure

Olivier Bouaziz[1], Eva Lauridsen[2] and Grégory Nuel[3]

[1]Laboratory MAP5, University Paris Descartes and CNRS, Sorbonne Paris Cité, Paris, France
[2]Ressource Center for Rare Oral Diseases, Copenhagen University Hospital, Rigshospitalet, Denmark
[3]LPSM, CNRS 7599, 4 place Jussieu, Paris, France

**Abstract**

A new method for the analysis of time to ankylosis complication on a dataset of replanted teeth is proposed. In this context of left-censored, interval-censored and right-censored data, a Cox model with piecewise constant baseline hazard is introduced. Estimation is carried out with the EM algorithm by treating the true event times as unobserved variables. This estimation procedure is shown to produce a block diagonal Hessian matrix of the baseline parameters. Taking advantage of this interesting feature of the estimation method a $L_0$ penalised likelihood method is implemented in order to automatically determine the number and locations of the cuts of the baseline hazard. This procedure allows to detect specific areas of time where patients are at greater risks for ankylosis. The method can be directly extended to the inclusion of exact observations and to a cure fraction. Theoretical results are obtained which allow to derive statistical inference of the model parameters from asymptotic likelihood theory. Through simulation studies, the penalisation technique is shown to provide a good fit of the baseline hazard and precise estimations of the resulting regression parameters.

**Keywords**: Adaptive Ridge procedure; Cure model; EM algorithm; Interval censoring; Penalised likelihood; Piecewise constant hazard.

## 1 Introduction

Interval censored data arise in situations where the event of interest is only known to have occurred between two observation times. These types of data are commonly encountered when the patients are intermittently followed up at medical examinations. This is the case for instance in AIDS studies, when HIV infection onset is determined by periodic testing, or in oncology where the time-to-tumour progression is assessed by measuring the tumour size at periodic testing. Dental data are another examples which are usually interval-censored because the teeth status of the patients are only examined at visits to the dentist. While interval-censored data are ubiquitous in medical applications it is still a common practice to replace the observation times with their midpoints or endpoints and to consider these data as exact. This allows to analyse the data using standard survival approach but may result in a large bias of the estimators. In the present paper we develop a new method for the analysis of time to ankylosis complication on a dataset of replanted teeth. The three main goals for our method is to adequately take into account interval-censoring, to be able to identify time ranges where patients are particularly at high risk of developing the complication and to investigate if a sub-population of non susceptible patients exists.

In the context of interval-censored data, [24] introduced an iterative algorithm for the non-parametric estimation of the survival function. As a different estimation method, the iterative

convex minorant was proposed by [11] and [14]. In [11], the authors derived the slow rate of convergence of order $n^{1/3}$ for the non-parametric survival estimator. Moreover, the obtained law is not Gaussian and cannot be explicitly computed. Many methods were also developed in a regression setting. In particular, the Cox model with non-parametric baseline was studied in [13]. The authors derived a $n^{1/2}$ convergence rate for the regression parameter with a Gaussian limit but the problem of estimation and inference of the baseline survival function pertains in this regression context: the baseline survival function has the $n^{1/3}$ slow rate of convergence and even more problematic, the asymptotic distribution of this function could not be derived. The same conclusions were observed in [5] where the authors use the more general Cox-Aalen model with non-parametric baseline. As a consequence, alternatives to the non-parametric baseline have been introduced. In [16] and [21] parametric baselines such as Weibull or piecewise constant are introduced. In that case, the convergence rate of the global parameters is of order $n^{1/2}$ and the asymptotic distribution is Gaussian (see [21]). In [4] a local likelihood is implemented which results in a smooth estimation of the baseline hazard using a kernel function. However, asymptotic properties of the estimators were not derived in their work and the performance of the estimators depends on the choice of the kernel bandwidth. In [27], monotone B-splines are implemented in order to estimate the cumulative baseline hazard. The authors introduce a two stage data augmentation which allows them to use the Expectation Maximisation algorithm [EM, see 9] in order to perform estimation. Asymptotics with $n^{1/2}$ rate of convergence of the estimators are derived. However, the number and location of the splines knots are pre-determined by the user and the estimators performance depend on the choice of these tuning parameters. A similar two stage data augmentation approach was developed in [28] where the authors study the more general class of semi-parametric transformation models, using a non-parametric baseline and allowing for time dependent covariates. The $n^{1/2}$ rate of convergence of the regression parameter is derived but the asymptotic distribution of the non-parametric baseline was not obtained.

In this work, we study the Cox model with piecewise constant baseline hazard. Treating the unobserved true event times as missing variables we use the EM algorithm to perform estimation. As a result, the Hessian of the log-likelihood to be maximised is seen to be diagonal. This is a remarkable feature of the method that easily allows to perform estimation with the piecewise constant baseline using arbitrarily large set of cuts. In contrast, this model had been already introduced in [8] and [16] but maximisation of the model parameters was achieved using the observed likelihood which resulted in a full rank Hessian matrix. In [8] for example, the authors warn against computational issues which may force the user to reduce the number of cuts by combining adjacent intervals. Using the EM algorithm to perform estimation in the piecewise constant hazard model is new to our knowledge and easy to implement. Also, all the quantities involved in the E-step can be explicitly computed in our method, contrary to previous works (see [4] for example) which require to approximate integrals. In comparison with [27] the E-step is more natural and directly applicable using the complete likelihood. Moreover, taking advantage of the sparse structure of the Hessian matrix, our method can be combined with a $L_0$ penalty designed to detect the location and number of cuts. This is performed through the adaptive ridge procedure, a regularisation method that was introduced in [20], [10] and then applied in a survival context (without covariates) in [6]. This penalisation technique results in a flexible method where the cuts and locations of the piecewise constant baseline are automatically chosen from the data, thus providing a good compromise between purely non-parametric and parametric baseline functions. This is in contrast with existing techniques such as in [27] where the location and number of knots of splines basis are fixed by the user. Finally we also emphasise the advantage of the $L_0$ method in terms of interpretability: by detecting the relevant set of cuts of the baseline the method highlights the different regions of time where the risk of failure varies. This is of great interest for the dental application in order for the dentists to precisely detect time intervals where patients are at a higher risk of ankylosis.

Another advantage of using the EM algorithm is to provide direct extensions of the Cox model. In this work we also consider the inclusion of exact data in the estimation method. This mixed case of exact and interval-censored data is usually not easy to analyse as standard methods for interval-censoring do not directly extend to exact data. However, using our method, inclusion of exact data is straightforward through the E-step and the likelihood can be decomposed into the contribution of exact and interval-censored observations. Another extension that is developed in this work is the inclusion of a fraction of non-susceptible patients. This situation is modelled using the cure model of [22] and [19], with a logit link for the probability of being cured. Little attention has been paid to this model in the case of interval-censored data. In [12] the authors consider a partially linear transformation model where the baseline is modelled using spline basis but the number and location of knots are chosen in an ad-hoc manner. In [17] a different cure model was introduced where the marginal survival function (without conditioning on the susceptible group) is modelled. However, the asymptotic distribution of the estimated parameters were not derived under this model. With our method, estimation in the cure Cox model is straightforward. The E-step results in a weighted log-likelihood with the weights corresponding to the probability of being cured such that our estimation method readily extends to the cure model. This model is especially useful on the dental dataset to assess if there exists a subpopulation of patients who are not at risk of developing the ankylosis complication.

In Section 2 the piecewise constant hazard model is introduced. The estimation method based on the EM algorithm is presented in Section 3 for interval censored data and fixed cuts of the hazard. Estimation in the non-parametric case, in the regression model and extensions for exact data and the cure model are also developed in this section. Then, the $L_0$ penalised likelihood that allows to select the location and number of cuts from the data is presented in Section 4. Asymptotic properties of the penalised estimator are discussed in Section 5. In particular, these results show that confidence intervals and tests can be constructed by considering the selected cuts as fixed. In Section 6, an extensive simulation study is presented where our adaptive ridge estimator is compared with the midpoint estimator and the ICsurv estimator from [27]. Finally, the dental dataset on ankylosis complications for replanted teeth is analysed in Section 7 using the proposed methodology.

## 2 A piecewise constant hazard model for interval censored data

Let $T$ denote the time to occurrence of the event of interest. We consider a situation where all individuals are subject to interval censoring defined by the random variables $(L, R)$ such that $L$ and $R$ are observed and $\mathbb{P}(T \in [L, R]) = 1$. The situation $L = 0$ and $R < \infty$ corresponds to left-censoring, $0 < L < R < \infty$ corresponds to strictly interval censoring and $L < R = \infty$ to right censoring. The special case $L = R$ is also allowed which corresponds to exact observations of the time of interest. We introduce a column covariate vector $Z$ of dimension $d_Z$ and for convenience we also introduce $\delta$ which equals 0 if an individual is right censored and 1 if he/she is left, interval censored or exactly observed. The variable $T$ is considered continuous and we assume independent censoring in the following way (see for instance [29]): $\mathbb{P}(T \leq t \mid L = l, R = r, Z) = \mathbb{P}(T \leq t \mid l \leq T \leq r, Z)$. 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 $R$. Finally, we assume non-informative censoring in the sense that the distribution of $L$ and $R$ does not depend on the model parameters involved in the distribution of $T$.

We consider the following Cox proportional hazard model for the time variable $T$:

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta Z), \tag{1}$$

where $\beta$ is an unknown row parameter vector of dimension $d_Z$. We model the baseline function

$\lambda_0$ through a piecewise constant hazard. Let $c_0, c_1, \ldots, c_K$ represent $K+1$ cuts, with the convention that $c_0 = 0$ and $c_K = +\infty$. Let $I_k(t) = I(c_{k-1} < t \leq c_k)$, with $I(\cdot)$ denoting the indicator function. We suppose that $\lambda_0(t) = \sum_{k=1}^{K} I_k(t) \exp(a_k)$. Under this model, note that the survival and density functions are respectively equal to:

$$S(t \mid Z) = \exp\Big( - \sum_{k=1}^{K} e^{a_k + \beta Z} (t \wedge c_k - c_{k-1}) I(c_{k-1} \leq t) \Big),$$

$$f(t \mid Z) = \sum_{k=1}^{K} I_k(t) \exp\Big( a_k + \beta Z - \sum_{j=1}^{k} e^{a_j + \beta Z} (t \wedge c_j - c_{j-1}) \Big).$$

We set $\boldsymbol{\theta} = (a_1, \ldots, a_K, \beta)$ the model parameter we aim to estimate. In the following, we will also study the so-called nonparametric situation, when no covariates are available, which is encompassed in our modelling approach as the special case where $Z = 0$. In this context the hazard function is simply equal to $\lambda_0$ which is assumed to be piecewise constant and the model parameter is $\boldsymbol{\theta} = (a_1, \ldots, a_K)$. The observed data consist of data $= \{\text{data}_i, i = 1, \ldots, n\}$ with $\text{data}_i = (L_i, R_i, \delta_i)$ in the nonparametric context and $\text{data}_i = (L_i, R_i, \delta_i, Z_i)$ in the regression context, while $T_i$ is considered as incompletely observed. In the latter context, we introduce the notation $a_{i,k} = a_k + \beta Z_i$.

# 3 Estimation procedure with fixed cuts

For the sake of simplicity, we first consider the scenario when no exact data are observed (which means there only are left, interval and right censored data). The estimation method is based on the EM algorithm and is presented in Section 3.1 in the general regression context since the nonparametric context can be easily derived by setting $Z = 0$. The nonparametric context is discussed in Section 3.2, the implementation of the M step for the regression context is presented in Section 3.3 and the method when exact observations are also available is developed in Section 3.4. Finally, the inclusion of a fraction of non-susceptible individuals is studied in Section 3.5.

## 3.1 The EM algorithm for left, right and interval censored observations

The observed likelihood is defined with respect to the observed data by:

$$\begin{aligned} \mathrm{L}_n^{\text{obs}}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} (S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})) \\ &= \prod_{i=1}^{n} \left\{ \exp\Big( - \int_0^{L_i} \lambda_0(t) dt \, e^{\beta Z_i} \Big) \left( 1 - \exp\Big( - \int_{L_i}^{R_i} \lambda_0(t) dt \, e^{\beta Z_i} \Big) \right) \right\}^{\delta_i} \\ &\quad \times \left\{ \exp\Big( - \int_0^{L_i} \lambda_0(t) dt \, e^{\beta Z_i} \Big) \right\}^{1-\delta_i}, \end{aligned}$$

with the slight abuse of notation $S(R_i \mid Z_i, \boldsymbol{\theta}) = 0$ if $R_i = \infty$ (for a right-censored observation). The Maximum Likelihood Estimator (MLE) can be derived from maximisation of this observed log-likelihood with respect to the model parameters, as in [8] for instance. The obtained parameter estimates are not explicit but a Newton-Raphson algorithm can be easily implemented. However, in this optimisation problem, the block of the Hessian matrix corresponding of the baseline coefficients $a_1, \ldots, a_K$ will be of full rank and can lead to intractable solutions if the number of cuts $K$ is large. An alternative method to compute the MLE is therefore to use the EM algorithm based on the complete likelihood of the unobserved true event times. This algorithm will result into a diagonal block matrix of the baseline coefficients.

The EM algorithm is based on the complete likelihood, defined by: $\mathrm{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^n f(T_i \mid Z_i, \boldsymbol{\theta})$. Denote by $\boldsymbol{\theta}_{\text{old}}$ the current parameter value. The E-step takes the expectation of the complete log-likelihood with respect to the $T_i$'s, given the $L_i$'s, $R_i$'s, $\delta_i$'s, $Z_i$'s and $\boldsymbol{\theta}_{\text{old}}$. Write

$$Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) := \mathbb{E}[\log(f(T_i \mid Z_i, \boldsymbol{\theta})) \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}}] = \int f(t \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}}) \log f(t \mid Z_i, \boldsymbol{\theta}) dt,$$

where $f(t \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}})$ represents the conditional density of $T_i$ given $\text{data}_i$ and $\boldsymbol{\theta}_{\text{old}}$, evaluated at $t$. Under the independent censoring assumption,

$$f(t \mid \text{data}_i, \boldsymbol{\theta}_{\text{old}}) = \frac{f(t \mid Z_i, \boldsymbol{\theta}_{\text{old}}) I(L_i < t < R_i)}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})}.$$

The E-step consists of computing the quantity $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_i Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})$. We have:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^n \frac{\int_{L_i}^{R_i} f(t \mid Z_i, \boldsymbol{\theta}_{\text{old}}) \log f(t \mid Z_i; \boldsymbol{\theta}) dt}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})}$$

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^n \left\{ \frac{1}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})} \times \sum_{k=1}^K J_{k,i} \int_{c_{k-1} \vee L_i}^{c_k \wedge R_i} \exp\left(a_{i,k}^{\text{old}} - \sum_{j=1}^k e^{a_{i,j}^{\text{old}}} (t \wedge c_j - c_{j-1})\right) \left(a_{i,k} - \sum_{j=1}^k e^{a_{j,k}} (t \wedge c_j - c_{j-1})\right) dt \right\},$$

where $J_{k,i}$ is the indicator $I\{(L_i, R_i) \cap (c_{k-1}, c_k) \neq \emptyset\}$ and $b_1 \wedge b_2$, $b_1 \vee b_2$ respectively denote $\min(b_1, b_2)$, $\max(b_1, b_2)$. Finally, the M-step corresponds of maximising, with respect to $\boldsymbol{\theta}$, the quantity

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^n \sum_{k=1}^K \left\{ \left(a_{i,k} - \sum_{j=1}^{k-1} (c_j - c_{j-1}) e^{a_{i,j}}\right) A_{k,i}^{\text{old}} - e^{a_{i,k}} B_{k,i}^{\text{old}} \right\},$$

where exact expressions of the statistics $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$ can be found in the Supplementary Material.

### 3.2 Estimation in the absence of covariates

In the absence of covariates, the previous results hold with $Z_i = 0$, $a_{i,k} = a_k$ and the model parameters we aim to estimate are just $\boldsymbol{\theta} = (a_1, \ldots, a_K)$. The objective function in the M-step can be defined with respect to the sufficient statistics $\bar{A}_k^{\text{old}} = \sum_i A_{k,i}^{\text{old}}$ and $\bar{B}_k^{\text{old}} = \sum_i B_{k,i}^{\text{old}}$:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_{k=1}^K \left\{ \left(a_k - \sum_{j=1}^{k-1} (c_j - c_{j-1}) e^{a_j}\right) \bar{A}_k^{\text{old}} - e^{a_k} \bar{B}_k^{\text{old}} \right\}.$$

The derivatives of $Q$ with respect to $a_k$, $k = 1, \ldots, K$, equal

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k} = \bar{A}_k^{\text{old}} - (c_k - c_{k-1}) e^{a_k} I(k \neq K) \sum_{l=k+1}^K \bar{A}_l^{\text{old}} - e^{a_k} \bar{B}_k^{\text{old}}.$$

As a consequence, in the absence of covariates, one gets the explicit parameters estimators:

$$\exp(\hat{a}_k) = \frac{\bar{A}_k^{\text{old}}}{I(k \neq K)\sum_{l=k+1}^{K} \bar{A}_l^{\text{old}}(c_k - c_{k-1}) + \bar{B}_k^{\text{old}}}, k = 1, \ldots, K,$$

at each step of the EM algorithm. At convergence, this provides an estimator of the hazard function from which quantities of interest, such as the survival function, can be easily derived.

### 3.3 Estimation in the general regression framework

In the regression framework, each step of the EM algorithm is solved through a Newton-Raphson procedure. The first and second order derivatives of $Q$ with respect to $a_k$ and $\beta$ are equal to

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k} = \sum_{i=1}^{n} \left\{ A_{k,i}^{\text{old}} - (c_k - c_{k-1}) e^{a_k} I(k \neq K) \sum_{l=k+1}^{K} A_{l,i}^{\text{old}} e^{\beta Z_i} - e^{a_k} B_{k,i}^{\text{old}} e^{\beta Z_i} \right\},$$

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial \beta} = \sum_{i=1}^{n} Z_i \sum_{l=1}^{K} \left( A_{l,i}^{\text{old}} - \left\{ \sum_{j=1}^{l-1} (c_j - c_{j-1}) e^{a_j} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_l} B_{l,i}^{\text{old}} e^{\beta Z_i} \right\} \right),$$

and

$$\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k^2} = -\sum_{i=1}^{n} \left\{ (c_k - c_{k-1}) e^{a_k} I(k \neq K) \sum_{l=k+1}^{K} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_k} B_{k,i}^{\text{old}} e^{\beta Z_i} \right\},$$

$$\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial \beta^2} = -\sum_{i=1}^{n} Z_i Z_i^t \sum_{l=1}^{K} \left( \sum_{j=1}^{l-1} (c_j - c_{j-1}) e^{a_j} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_l} B_{l,i}^{\text{old}} e^{\beta Z_i} \right),$$

$$\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k \partial \beta} = -\sum_{i=1}^{n} Z_i \left( (c_k - c_{k-1}) e^{a_k} I(k \neq K) \sum_{l=k+1}^{K} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_k} B_{k,i}^{\text{old}} e^{\beta Z_i} \right).$$

The block matrix of the Hessian corresponding to the second order derivatives with respect to the $a_k$'s is diagonal while the three other blocks are of full rank. Inversion of the Hessian matrix is then achieved using the Schurr complement which takes advantage of this sparse structure of the Hessian. When considering a large number of cuts, that is $K >> d_Z$, the total complexity of the inversion of the Hessian is of order $\mathcal{O}(K)$. The exact formula of the Schurr complement is given in the Supplementary Material.

### 3.4 Inclusion of exact observations

It is straightforward to deal with exact observations since they can be directly included in the EM algorithm. For an exact observation $i$, $\mathbb{E}[\log(f(T_i \mid Z_i; \boldsymbol{\theta})) \mid \text{data}, \boldsymbol{\theta}_{\text{old}}] = \log(f(T_i \mid Z_i; \boldsymbol{\theta})) = \sum_{k=1}^{K} \left\{ O_{i,k} a_{i,k} - \exp(a_{i,k}) R_{i,k} \right\}$, with $O_{i,k} = I(c_{k-1} < T_i < c_k)$ and $R_{i,k} = T_i \wedge c_k - c_{k-1}$. Note that this corresponds to the classical contribution of an exact observation to the log-likelihood in the standard Poisson regression for right censored observations (see for instance [1]). As a result, $Q$ can be decomposed as

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \sum_{i \text{ not exact}} \sum_{k=1}^{K} \left\{ \left( a_{i,k} - \sum_{j=1}^{k-1} (c_j - c_{j-1}) e^{a_{i,j}} \right) A_{k,i}^{\text{old}} - e^{a_{i,k}} B_{k,i}^{\text{old}} \right\}$$
$$+ \sum_{i \text{ exact}} \sum_{k=1}^{K} \left\{ O_{i,k} a_{i,k} - \exp(a_{i,k}) R_{i,k} \right\}.$$

The estimation method follows as previously. In particular, in the absence of covariates, the explicit parameters estimator of $(a_1, \ldots, a_K)$ are equal to:

$$\exp(\hat{a}_k) = \frac{\bar{A}_k^{\text{old}} + \bar{O}_k}{I(k \neq K) \sum_{l=k+1}^{K} \bar{A}_l^{\text{old}}(c_k - c_{k-1}) + \bar{B}_k^{\text{old}} + \bar{R}_k}, k = 1, \ldots, K,$$

where $\bar{O}_k = \sum_{i \text{ exact}} \bar{O}_{i,k}$ and $\bar{R}_k = \sum_{i \text{ exact}} \bar{R}_{i,k}$.

In the regression setting, maximisation over the $\beta$ and $a_1, \ldots, a_K$ parameters is performed through the Newton-Raphson algorithm as before. Full expressions of the score vector and Hessian matrix are given in the Supplementary Material. The Schurr complement is used again to invert the Hessian matrix (see the Supplementary Material).

### 3.5 Inclusion of a fraction of non-susceptibles (cure fraction)

Taking into account non-susceptible individuals is possible using the cure model from [22]. This is achieved by modelling the latent status (susceptible/non-susceptible) of the individuals through a variable $Y$ which equals 1 for patients that will eventually experience the event and 0 for patients that will never experience the event. Since the estimation method uses the EM algorithm, this latent variable can be easily dealt with through the E-step.

We assume that $Y$ is independent of $T$ conditionally on $(L, R)$. The proportional hazard Cox model for the susceptibles is defined as

$$\lambda(t \mid Y = 1, Z) = \lambda_0(t) \exp(\beta Z). \tag{2}$$

The cure model specifies the hazard, conditional on $Y$ and $Z$, to be equal to $\lambda(t \mid Y, Z) = Y\lambda(t \mid Y = 1, Z)$. The baseline function $\lambda_0$ is assumed to be piecewise constant as in Section 2 and the conditional density and survival functions of the susceptibles are respectively noted $f(t \mid Y = 1, Z)$ and $S(t \mid Y = 1, Z)$. If one wants to model the effect of covariates on the probability of being cured, a logistic link can be used:

$$p(X) = \mathbb{P}[Y = 1 \mid X] = \frac{\exp(\gamma X)}{1 + \exp(\gamma X)}, \tag{3}$$

where $X$ is a covariate vector including the intercept and $\gamma$ is a row parameter vector, both of dimension $d_X$. The observed data then consist of data $= (L_i, R_i, \delta_i, Z_i, X_i)_{1 \leq i \leq n}$ while $T_i$ and $Y_i$ are respectively incompletely observed and non observed data. The model parameter is $\boldsymbol{\theta} = (a_1, \ldots, a_L, p)$ in the completely nonparametric context (no covariates $X$ nor $Z$), $\boldsymbol{\theta} = (a_1, \ldots, a_L, \beta, p)$ if only the covariate $Z$ is used or $\boldsymbol{\theta} = (a_1, \ldots, a_L, \beta, \gamma)$ in the full regression context (with covariates $X$ and $Z$). In the later case, we introduce the notation $p_i = \mathbb{P}[Y_i = 1 \mid X_i]$. The other situations are encompassed in our modelling approach by setting $X = 0$ and/or $Z = 0$. Note that our cure model is identifiable and does not require additional constraints such as in [22] where the authors had to impose $S(t \mid Y = 1, Z)$ to be null for $t$ greater than the last event time in the context of exact and right-censored data.

Under the cure model with interval-censored and exact observations, the observed likelihood is now defined as

$$\begin{aligned} \mathrm{L}_n^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i \text{ not exact}} & \left\{ p_i \exp\Big( - \int_0^{L_i} \lambda_0(t) dt e^{\beta_0 Z_i} \Big) \left( 1 - \exp\Big( - \int_{L_i}^{R_i} \lambda_0(t) dt e^{\beta_0 Z_i} \Big) \right) \right\}^{\delta_i} \\ & \times \left\{ (1 - p_i) + p_i \exp\Big( - \int_0^{L_i} \lambda_0(t) dt e^{\beta_0 Z_i} \Big) \right\}^{1 - \delta_i} \prod_{i \text{ exact}} p_i f(T_i \mid Y_i = 1, Z_i; \boldsymbol{\theta}) \end{aligned}$$

and the complete likelihood is defined as: $\mathrm{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{1 - Y_i} \{ f(T_i \mid Y_i = 1, Z_i; \boldsymbol{\theta}) \}^{Y_i}$. The E-step consists of computing the function $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \mathbb{E}[\log(\mathrm{L}_n(\boldsymbol{\theta})) \mid \text{data}, \boldsymbol{\theta}_{\text{old}}]$. Let

$\pi_i^{\text{old}} = \mathbb{E}[Y_i \mid \text{data}, \boldsymbol{\theta}_{\text{old}}]$, we have:

$$\pi_i^{\text{old}} = \delta_i + \frac{(1-\delta_i)p_{\text{old}}S(L_i \mid Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}})}{1 - p_{\text{old}} + p_{\text{old}}S(L_i \mid Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}})}.$$

In the case of interval-censored and exact observations,

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = &\sum_{i=1}^{n} \left\{ \pi_i^{\text{old}} \log(p_i) + (1 - \pi_i^{\text{old}}) \log(1 - p_i) \right\} \\ &+ \sum_{i \text{ not exact}} \pi_i^{\text{old}} \sum_{k=1}^{K} \left\{ \left( a_{i,k} - \sum_{j=1}^{k-1} (c_j - c_{j-1}) e^{a_{i,j}} \right) A_{k,i}^{\text{old}} - e^{a_{i,k}} B_{k,i}^{\text{old}} \right\} \\ &+ \sum_{i \text{ exact}} \sum_{k=1}^{K} \left\{ O_{i,k} a_{i,k} - \exp(a_{i,k}) R_{i,k} \right\}, \end{aligned}$$

where $A_{k,i}^{\text{old}}$, $B_{k,i}^{\text{old}}$ are defined as in the Supplementary Material with the quantity $S(\cdot \mid Z_i, \boldsymbol{\theta}_{\text{old}})$ replaced by $S(\cdot \mid Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}})$. The terms $O_{i,k}$ and $R_{i,k}$ were defined in Section 3.4.

The $Q$ function separates the terms with $\gamma$ and the terms involving $(a_1, \ldots, a_K, \beta)$ such that maximisation of these terms can be performed separately. Let $\bar{A}_k^{\pi,\text{old}} = \sum_i \pi_i^{\text{old}} A_{k,i}^{\text{old}}$, $\bar{B}_k^{\pi,\text{old}} = \sum_i \pi_i^{\text{old}} B_{k,i}^{\text{old}}$ and $\bar{\pi}^{\text{old}} = \sum_i \pi_i^{\text{old}}$. In the nonparametric setting, explicit estimators of the parameters can be computed at each step of the EM algorithm through the formulas:

$$\begin{aligned} \hat{p} &= \frac{\bar{\pi}^{\text{old}}}{n}, \\ \exp(\hat{a}_k) &= \frac{\bar{A}_k^{\pi,\text{old}} + \bar{O}_k}{I(k \neq K) \sum_{l=k+1}^{K} \bar{A}_l^{\pi,\text{old}} (c_k - c_{k-1}) + \bar{B}_k^{\pi,\text{old}} + \bar{R}_k}, k = 1, \ldots, K. \end{aligned}$$

In the general regression context, a Newton-Raphson procedure is implemented separately to maximise both parts of $Q$. The first and second order derivatives of $Q$ with respect to $\gamma$ are equal to:

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial \gamma} &= \sum_{i=1}^{n} X_i \left( \pi_i^{\text{old}} - \frac{\exp(\gamma X_i)}{1 + \exp(\gamma X_i)} \right), \\ \frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial \gamma^2} &= -\sum_{i=1}^{n} X_i X_i^t \frac{\exp(\gamma X_i)}{(1 + \exp(\gamma X_i))^2}. \end{aligned}$$

Exact expressions of the first and second order derivatives of $Q$ with respect to $a_k$ and $\beta$ are given in the Supplementary Material. They are expressed as weighted versions with respect to $\pi_i^{\text{old}}$ of the derivatives obtained in the context where all individuals are susceptibles. As previously, the block matrix corresponding to the second order derivatives with respect to the $a_k$s of the Hessian is diagonal and inversion of the Hessian matrix is achieved using the Schurr complement.

# 4 Estimation procedure using the adaptive ridge method

In this section we present a penalised estimation method to detect the number and location of the cuts of the baseline hazard, when those are not known in advance. The proposed methodology is based on the work of [20], [10] and [6] and can be applied to any of the previous scenarios (with exact observations, with a cure fraction, in a nonparametric setting, in a regression setting) where the function $Q$ represents the objective function associated with the context under study.

### 4.1 A penalised EM algorithm

If the number of cuts is not known in advance, we choose a large grid of cuts (i.e $K$ large) and we penalise the log-likelihood in the manner of [10], [20] and [6]. This penalisation is designed to enforce consecutive values of the $a_k$s that are close to each other to be equal. It is defined in the following way:

$$\ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k (a_{k+1} - a_k)^2, \tag{4}$$

where $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_{K-1})$ are non-negative weights that will be iteratively updated in order for the weighted ridge penalty term to approximate the $L_0$ penalty. The pen term is a tuning parameter that describes the degree of penalisation. Note that the two extreme situations pen$= 0$ and pen$= \infty$ respectively correspond to the unpenalised log-likelihood model of Section 3 and to the Cox model with exponential baseline.

Only the maximisation over $(a_1, \ldots, a_K)$ is affected by the penalty. The first and second order derivatives of $\ell^{\text{pen}}$ with respect to $a_1, \ldots, a_K$ are equal to:

$$\begin{aligned}
\frac{\partial \ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k} &= \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k} + (\hat{w}_{k-1} a_{k-1} - (\hat{w}_{k-1} + \hat{w}_k) a_k + \hat{w}_k a_{k+1}) \text{pen}, \\
\frac{\partial^2 \ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k^2} &= \frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k^2} - (\hat{w}_{k-1} + \hat{w}_k) \text{pen},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k a_{k+1}} &= \frac{\partial^2 \ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_{k+1} a_k} = \hat{w}_k \text{pen}, \\
\frac{\partial^2 \ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k a_{k'}} &= 0 \text{ for } k, k' \text{ such that } \mid k - k' \mid \geq 2.
\end{aligned}$$

The block matrix corresponding to the second order derivatives with respect to the $a_k$s is therefore tridiagonal. For a given value of pen and of the weight vector $\hat{\boldsymbol{w}}$, inversion of the Hessian matrix is performed using the Schurr complement as previously (see the Supplementary Material) and the Newton-Raphson algorithm is implemented to derive $\hat{\boldsymbol{\theta}}$. Once the Newton-Raphson algorithm has reached convergence, the weights are updated at the $l$th step from the equation

$$\hat{w}_k^{(l)} = \left( (\hat{a}_{k+1}^{(l)} - \hat{a}_k^{(l)})^2 + \varepsilon^2 \right)^{-1}, \tag{5}$$

for $k = 1, \ldots, K-1$ with $\varepsilon = 10^{-5}$ (recommended value from [10]) and where the $\hat{a}_k^{(l)}$'s represent the estimates of the $a_k$'s obtained through the Newton-Raphson algorithm. This form of weights is motivated by the fact that $w_k(a_{k+1} - a_k)^2$ is close to 0 when $\mid a_{k+1} - a_k \mid < \varepsilon$ and close to 1 when $\mid a_{k+1} - a_k \mid > \varepsilon$. Hence the penalty term tends to approximate the $L_0$ norm. The weights are initialized by $\hat{w}_k^{(0)} = 1$, which gives the standard ridge estimate of $\boldsymbol{a}$.

Finally, for a given value of pen, once the adaptive ridge algorithm has reached convergence, a set of cuts is found for the $\hat{a}_k$'s verifying $\hat{w}_k(\hat{a}_{k+1} - \hat{a}_k)^2 > 0.99$. This hard thresholding allows to provide a sparse collection of cuts. The non-penalised log-likelihood $Q$ is then maximised using this set of cuts and the final maximum likelihood estimate is derived using the results of Section 3. It is important to stress that the penalised likelihood is used only to select a set of cuts. Reimplementing the non-penalised log-likelihood $Q$ in the final step enables to reduce the bias classically induced by penalised maximisation techniques.

### 4.2 Choice of the penalty term

A Bayesian Information Criterion (BIC) is introduced in order to choose the penalty term. As explained in the previous section, for each penalty value the penalised EM likelihood (4) selects a set of cuts. For a selected set of cuts we denote by $m$ the total number of parameters to be estimated and by $\hat{\boldsymbol{\theta}}_m$ the corresponding non-penalised estimated model parameter obtained by maximisation of the $Q$ function. The BIC is then defined as: $\text{BIC}(m) = -2\log(\text{L}_n^{\text{obs}}(\hat{\boldsymbol{\theta}}_m)) + m\log(n)$.

Note that the BIC is expressed here in terms of selected models. Since different penalty values can yield the same selection of cuts, the BIC needs only to be computed for all different selected models (and not for all different penalties). As an illustration of the model selection procedure, a full regularisation path is displayed in Section A.4 of the Supplementary Material on a simulated data sample, where for each penalty value correspond a selection of cuts and parameter estimates. The final set of cuts along with its estimator $\hat{\boldsymbol{\theta}}_{\hat{m}}$ is chosen such that $\text{BIC}(\hat{m})$ is minimal.

## 5 Asymptotic results

Theoretical properties of the derived estimator are presented in this section for interval-censored observations which can also include exact data. Theoretical results for the cure model are omitted for the sake of presentation. Two main results are established: it is first shown that the penalised estimator asymptotically detects the true support of the baseline, in the case where the true baseline is piecewise constant and the grid used to implement the estimator contains the true cuts of the baseline hasard. In the second step of the algorithm, using the cuts obtained from the penalised estimator, the non-penalised estimator from Section 3 is implemented. It is then shown that the resulting estimator is asymptotically normal and unbiased. The limiting variance is optimal in the sense that it is equal to the variance one would obtain from implementing the non-penalised estimator with the true cuts.

In the presence of interval-censored and exact data, the observed likelihood is equal to:

$$\text{L}_n^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i \text{ not exact}} (S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})) \prod_{i \text{ exact}} f(T_i \mid Z_i, \boldsymbol{\theta}),$$

with the slight abuse of notation $S(R_i \mid Z_i, \boldsymbol{\theta}) = 0$ if $R_i = \infty$ (for a right-censored observation). We assume that the EM procedure converges which entails that the penalised estimator that maximises Equation (4) also verifies

$$\hat{\boldsymbol{\theta}} = (\hat{a}_1, \ldots, \hat{a}_K, \hat{\beta}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^{K+d_Z}}{\arg\max} \left\{ \log(\text{L}_n^{\text{obs}}(\boldsymbol{\theta})) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k^{(1)} (a_{k+1} - a_k)^2 \right\}. \tag{6}$$

In the above formula, we consider only one iteration of the adaptive ridge procedure (5) where $\hat{\boldsymbol{a}}^{(1)}$ is supposed to be a consistent estimator (for example the unpenalised estimator or the ridge estimator). We now define a true parameter $\boldsymbol{\theta^*} = (a_1^*, \ldots, a_{K^*}^*, \beta^*)$ which is assumed to be in a compact set and a true baseline hazard function $\lambda_0^*(t) = \sum_{k=1}^{K^*} I(c_{k-1}^* < t \leq c_k^*) \exp(a_k^*)$ with true cuts $\mathcal{A}^* = \{c_1^*, \ldots, c_{K^*}^*\}$. Solving (6) provides, after detecting the consecutive values of $\hat{a}_k$ that are equal, an estimated set of cuts denoted $\mathcal{A}_n = \{\hat{c}_1, \ldots, \hat{c}_{\hat{K}}\}$. Note that the size of $\mathcal{A}_n$ and $\mathcal{A}^*$ might be different and typically smaller than $K$. The unpenalised estimator obtained when using $\mathcal{A}_n$ is noted $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}_n} = (\hat{\hat{a}}_{1,\mathcal{A}_n}, \ldots, \hat{\hat{a}}_{\hat{K},\mathcal{A}_n}, \hat{\hat{\beta}}_{\mathcal{A}_n})$. We also define $\hat{\hat{\lambda}}_{0,\mathcal{A}_n}(t) = \sum_{k=1}^{\hat{K}} I(\hat{c}_{k-1} < t \leq \hat{c}_k) \exp(\hat{\hat{a}}_{k,\mathcal{A}_n})$. In order to state our theorem we first introduce

$$h_{\boldsymbol{\theta}}^*(L_i, R_i, Z_i) = I(L_i \neq R_i) \log(S^*(L_i \mid Z_i, \boldsymbol{\theta}) - S^*(R_i \mid Z_i, \boldsymbol{\theta})) + I(L_i = R_i) \log(f^*(T_i \mid Z_i, \boldsymbol{\theta}))$$

and the matrices $\Sigma = -\mathbb{E}[\nabla^2_{\boldsymbol{\theta}} h^*_{\boldsymbol{\theta}}(L_i, R_i, Z_i))|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}]$ of dimension $(K^* + d_Z) \times (K^* + d_Z)$ and $\Sigma_{\beta^*} = \{\Sigma_{i,j} : K^* + 1 \leq i \leq K^* + d_Z, K^* + 1 \leq j \leq K^* + d_Z\}$ . In the formulas, $S^*$ and $f^*$ represent the survival and density functions computed using the true set of cuts for a $\boldsymbol{\theta}$ of dimension $K^* + d_Z$. Finally we let $\tau$ represents the endpoint of the study.

**Theorem 5.1** *Assume that $\mathcal{A}^* \subset \{c_1, \ldots, c_K\}$, $\mathbb{P}[\{R > \tau, R < \infty\} \cup \{L > \tau\}] > 0$, $Z$ is almost surely bounded and $\Sigma$ is a non-singular matrix. Then, if $pen/\sqrt{n} \to 0$ as $n \to \infty$ we have:*

1. $\lim_{n\to\infty} \mathbb{P}[\mathcal{A}_n = \mathcal{A}^*] = 1$.
2. *for all $t \in [0, \tau]$, $\sqrt{n}(\hat{\hat{\lambda}}_{0,\mathcal{A}_n}(t) - \lambda^*_0(t))$ converges in distribution toward a centered Gaussian variable with variance equal to $\sum_{k=1}^{K^*} I(c^*_{k-1} < t \leq c^*_k) \exp(a^*_k)(\Sigma_{k,k})^{-1}$.*
3. *$\sqrt{n}(\hat{\hat{\beta}}_{\mathcal{A}_n} - \beta^*)$ converges in distribution toward a centered Gaussian variable with variance equal to $(\Sigma_{\beta^*})^{-1}$.*

Two important remarks can be made from this theorem. Firstly, the asymptotic variances in 2. and 3. are identical to the variances obtained in the parametric piecewise constant hazard model using the true cuts. Secondly, these two variances can be consistently estimated by

$$-n \times \sum_{k=1}^{\hat{K}} I(\hat{c}_{k-1} < t \leq \hat{c}_k) \exp(\hat{\hat{a}}_{k,\mathcal{A}_n})(\partial^2 \log(\mathrm{L}^{\mathrm{obs}}_{\mathcal{A}_n}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}_n}))/\partial a_k^2)^{-1},$$

and

$$-n(\nabla^2_{\beta} \log(\mathrm{L}^{\mathrm{obs}}_{\mathcal{A}_n}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}_n})))^{-1},$$

where $\mathrm{L}^{\mathrm{obs}}_{\mathcal{A}_n}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}_n})$ represents the observed likelihood evaluated at the estimated parameter $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}_n}$ with the estimated cuts. In other words, this theorem states that inference on the model parameters can be achieved after selection of the cuts of the baseline function by considering these cuts as fixed parameters. The proof of the theorem is inspired from [31] and is provided in the Supplementary Materials.

A direct method for deriving confidence intervals or statistical tests can therefore be based on the normal approximation of the model parameter after computing the Hessian matrix of the observed log-likelihood. However since the calculation of the Hessian matrix is tedious under the piecewise constant hazard model, we prefer to use a likelihood ratio test approach. This approach and the explicit expression of the Hessian are detailed in the Supplementary Material. See also [30] for more details about the likelihood ratio test approach for constructing confidence intervals. Finally, note that bootstrap methods can also be implemented to derive confidence intervals. This technique is particularly relevant when the interest lies in the estimation of the survival function in a non-parametric or regression context. In order to derive the asymptotic distribution of such functional one would need to use the delta-method which may result in complicated formula for the variance estimator. The bootstrap alternative avoids these technicalities.

# 6 Simulation study

In this section we study the performance of the proposed estimators on simulated data. In what follows, two models including two scenarios with exact, left, interval-censored and right-censored data are presented. More scenarios considering the inclusion of a cure fraction can be found in the Supplementary Material.

We consider the Cox regression setting of Equation (1) where the aim is to correctly estimate the regression coefficient $\beta$ and the baseline function $\lambda_0$. We set the baseline as a piecewise

constant function with three cuts in Model M1 and as a Weibull function in Model M2 in the following way:

$$\text{M1: } \lambda_0(t) = \Big(0.5\, I(0 < t \leq 20) + I(20 < t \leq 40) + 2\, I(40 < t \leq 50) + 4\, I(50 < t)\Big) \cdot 10^{-2}$$

$$\text{M2: } \lambda_0(t) = \frac{\mu}{\kappa}\left(\frac{\mu}{\kappa}\right)^{(\mu-1)},\ \mu = 8,\ \kappa = 50.$$

In both models, the covariate vector $Z$ is of dimension $d_Z = 2$ with the first component simulated as a Bernoulli variable with parameter 0.6 and the second component is independently simulated as a uniform variable with parameters $[0, 2]$. The regression parameter is equal to $\beta = (\log(2), \log(0.8))$. The values of $L_i$ and $R_i$ were determined through a visit process defined in the following way. Let $\mathcal{U}$ denote the uniform distribution. Two visits were simulated such that the first one $V_1 \sim \mathcal{U}[0, 60]$ and the other one $V_2 = V_1 + \mathcal{U}[0, 120]$. Then 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 $T_i > V_2$ correspond to right-censored observations with $L_i = V_2$ and $R_i = \infty$, and the observations for which $V_1 < T_i < V_2$ correspond to strictly interval-censored observations with $L_i = V_1$ and $R_i = V_2$. This simulation setting corresponds to Scenario S1 and gave a proportion of 25% of left-censored observations, 52% of interval-censored observations and 23% of right-censored observations in Model M1 and a proportion of 2% of left-censored observations, 76% of interval-censored observations and 22% of right-censored observations in Model M2. In Scenario S2, 18% of exact observations were first sampled and then the same simulation scheme for the visit process was used. The percentage of right-censored observations remains identical under this scenario for both models.

Our adaptive ridge estimator was constructed from a grid of cuts ranging from $c_0 = 10$ to $c_{17} = 90$, with all cuts equally spaced of size 5. The set of penalty terms was taken, on the log scale, as the set of 200 equally spaced values ranging from $\log(0.1)$ to $\log(10\,000)$. For the EM algorithm, the $a_k$ and $\beta$ parameters were initialised to 0. As described in Section 4, the BIC was used to find an estimated set of cuts and the non penalised estimator was reimplemented with this set of cuts in order to derive our final estimator. This estimator was compared with the midpoint estimator and the ICsurv estimator from [27]. The midpoint estimator consists of replacing the interval-censored observations by their midpoint $(L_i + R_i)/2$. The data then consist of exact and right-censored observations and can be dealt with by implementing the standard Cox regression estimators. The ICsurv estimator models the cumulative baseline function using monotone splines and uses a two-stage data augmentation method to perform estimation through the EM algorithm. This estimator is implemented using a more recent version of the `fast.PH.ICsurv.EM` function provided from the maintainer of the `ICsurv` package. Following the guidelines from the maintainer of the `ICsurv` package this estimator was computed using basis splines having degree 3 with 5 interior knots placed evenly across the range of endpoints of the observed intervals. The $\beta$ parameters and the spline coefficients were respectively initialised to 0 and 1. A very fine grid of time was used for the calculation of the cumulative baseline hazard from time 0 to time 200 with a step equal to 0.1. This estimator cannot include exact observations and is computed only for the Scenario S1 in Models M1 and M2.

A total of $M = 500$ replications were implemented and the bias and the empirical standard error (SE) of $\hat{\beta}$ were computed for each estimator. Confidence intervals at the 95% level were constructed for $\hat{\beta}$ using the likelihood ratio test approach, as described in the Supplementary Material (see also Section 5), and the coverage probability (CP) was reported. In order to assess the quality of estimation of $\lambda_0$, the baseline survival function $S_0(t) = \exp(-\int_0^t \lambda_0(u)du)$ was also estimated with each estimator. Then, as a measure of precision, the Integrated Mean

Squared Error (MISE) was decomposed as $\text{MISE}(\hat{S}_0) = \text{IBias}^2(\hat{S}_0) + \text{IVar}(\hat{S}_0)$, where

$$\text{IBias}^2(\hat{S}_0) = \int_0^{60} \left( \frac{1}{M} \sum_{m=1}^{M} \hat{S}_0^{(m)}(u) - S_0(u) \right)^2 du,$$

$$\text{IVar}(\hat{S}_0) = \frac{1}{M} \sum_{m=1}^{M} \int_0^{60} \left( \hat{S}_0^{(m)}(u) - \frac{1}{M} \sum_{m'=1}^{M} \hat{S}_0^{(m')}(u) \right)^2 du.$$

The $\hat{S}_0^{(m)}$, $m = 1, \ldots, M$, represent the estimates for each replication. Finally, the total variation between $\hat{\lambda}_0$ and $\lambda_0$ was also computed for our adaptive ridge estimator. For a given estimate $\hat{\lambda}_0^{(m)}$, the quantity $\text{TV}^{(m)}(\hat{\lambda}_0^{(m)}) = \sum_{k=1}^{K} (c_k - c_{k-1}) \mid \exp(\hat{a}_k) - \exp(a_k) \mid$ was calculated in Model M1 and the average over all estimates $\text{TV}(\hat{\lambda}_0) = \sum_m \text{TV}^{(m)}(\hat{\lambda}_0^{(m)})/M$ was reported. The results are presented in Tables 1, 2 for Model M1 and Tables 3, 4 for Model M2. Results on the performance of cuts detection are displayed in Tables 5 and 6. Three different sample sizes ($n = 200, 400, 1\,000$) were considered in all models and scenarios, for the midpoint, the ICsurv and the adaptive ridge estimators.

From the simulation results, it is seen that the midpoint estimate has a lower variance than our adaptive ridge estimator both for $\hat{\beta}$ and $\hat{S}_0$. However, the midpoint estimator is systematically biased and this bias does not get smaller as the sample size increases. On the other hand, our estimator always has a smaller bias for all scenarios and models and both the bias and the variance decrease as the sample size increases. For example, in Scenario S1, Model M1, for $n = 400$, which corresponds to the sample size of the real data analysis of Section 7 and to similar proportions of left, interval and right censoring, our estimator exhibits a bias for $\beta = (\log(2), \log(0.8))$ that is 15 and 4 times smaller than the bias from the midpoint estimator. For the estimation of $S_0$ the bias of our estimator is more than 40 times smaller than the midpoint estimator. The ICsurv estimator shows similar performance as our adaptive ridge estimator in Model M1. However in Model M2, our estimator has a lower bias than ICsurv but a bigger variance, and a slightly bigger MSE. In Scenarios S2 the effect of adding exact observations is seen to decrease the bias and variance of our estimator. For $n = 400$ in Model M1, Scenario S2 the bias for our estimator of $\beta$ is divided by 4 and 23 and the bias for our estimator of $S_0$ is divided by 3.

Finally, the likelihood ratio test approach seems to provide adequate coverage probabilities for $\beta$ especially for $n = 400$ and $n = 1\,000$, in all scenarios and models. Tables 5 and 6 show that, in the piecewise constant baseline scenario (Model M1), a majority of one cut is found for $n = 200$ and $n = 400$, most of the time in the set $[35, 55]$ and a majority of two cuts are found for $n = 1\,000$, with 44% of chances to detect at least one cut in the set $[10, 30]$ and 96% of chances to detect at least one cut in the set $[10, 30]$. Due to the wide range of the two visits variables $V_1$ and $V_2$, the algorithm is able at best to detect two cuts under this scenario, and miss most of the time one cut in the set $[35, 55]$. More simulations were conducted: scenarios including a cure fraction can be found in the Supplementary Material along with a discussion on computational complexity.

## 7 Ankylosis complications for replanted teeth

The method is illustrated on a dental dataset. 322 patients with 400 avulsed and replanted permanent teeth were followed-up prospectively in the period from 1965 to 1988 at the university hospital in Copenhagen, Denmark. The following replantation procedure was used: the avulsed tooth was placed in saline as soon as the patient was received at the emergency ward. If the tooth was obviously contaminated, it was cleansed with gauze soaked in saline or rinsed with a flow of saline from a syringe. The tooth was replanted in its socket by digital pressure. The

patients were then examined at intermittent visits to the dentist. In this study, we focused on a complication called ankylosis characterized by the fusion of the tooth to the bone such that the variable of interest $T$ is the time from replantation of the tooth to ankylosis. This complication may occur if the cells on the root surface is damaged in which case, healing of the periodontal ligament surrounding the tooth will be impaired, leading to local ingrowth of bone. Ankylosis cannot be arrested and gradually the root of the tooth will be replaced by bone which will eventually lead to tooth loss. The data are described in great details in [2] and were analysed using our adaptive ridge method in [15].

A total of 28% of the data were left censored, 35.75% were interval censored and 36.25% were right censored. Four covariates were included in the study: the stage of root formation (72.5% of mature teeth, 27.5% of immature teeth), the length of extra-alveolar storage (mean time is 30.9 minutes), the type of storage media (85.25% physiologic, 14.75% non physiologic) and the age of the patient (the mean age for mature teeth is 16.81 years). There is no need for a cure fraction in this analysis since all different models (non-parametric or regression models) estimated the cure fraction to 0%. The adaptive ridge method found four cuts for the baseline hazard at time points 100, 500, 800 and 900 where the initial grid search was composed of 10 spaced time points from 0 to 200 and then of 100 spaced time points from 200 to 2 000 ($K_{\max} = 40$). The initial grid search was motivated by the data: for 71% of the left and interval-censored data, the right endpoint is lower than 200.

Non-parametric survival estimates were first computed, one for the whole population and two for each subgroup defined by the stage of root formation (see Figure 1). Confidence intervals were also computed using the boostrap method with 500 replications. These plots illustrate an interesting feature of the adaptive ridge procedure: by selecting a parsimonious set of cuts, the method highlights the different regions of time where the risk of failure varies. There is in particular a very high risk of ankylosis before 100 days as shown by the very steep survival curve on this time interval. On the global survival curve, the risk of developing ankylosis (one minus the survival function) before 100 days is estimated to 48.35% [43.39%; 53.67%]. Then the slope of the survival curve decreases from 100 days to 500 days, with a risk to develop ankylosis before 500 days estimated to 59.94% [54.96%; 64.57%]. The risk of ankylosis after 900 days is almost null (as shown by the plateau of the survival curve) suggesting that if a patient has not yet developed ankylosis after 900 days he/she is almost no longer at risk for this complication.

When looking at the two subgroups defined by stage of root formation we can see that the risk of ankylosis is much higher in the mature group than in the immature group. This is a very interesting result as it confirms the finding from [3] where periodontal ligament healing was seen to be less frequent with advanced stages of root development. From our analysis, it is seen that the risk is in particular higher in the interval $[100, 500]$ for the mature group than for the immature group, with ankylosis coming mostly from the mature group in this time range. For the immature group, the risk of developing ankylosis before 100 days is estimated to 35.54% [26.85%; 45.13%] and to 52.84% [46.26%; 59.03%] for the mature teeth. Then the slope of the survival curve decreases from 100 days to 500 days, with a risk to develop ankylosis before 500 days estimated to 38.74% [28.97%; 47.62%] for the immature teeth and to 67.92% [62.36%; 73.31%] for the mature teeth. The risk gets very low after 500 days for all groups.

Finally a Cox model was implemented with all the covariates included. Since age shows little variation for immature teeth, this last variable was only included in interaction with the stage of root formation such that the baseline value corresponds to immature teeth and the covariate is defined as age greater than 20 years for mature teeth only. The results for the effects of the covariates are shown in Table 7. Statistical tests and confidence intervals for each variable were implemented using the log-ratio statistic test as explained in the Supplementary Material (see also Section 5). It can be seen that the stage of root formation is highly significant with a two-fold increased risk for mature teeth to develop ankylosis. The storage time is also highly significant with a 1.23 increase of risk per hour. The type of storage media seems to

have no effect on ankylosis and age is not significant even at the 10% level. The baseline hazard values along with their 95% confidence intervals are also displayed in Table 8. This hazard corresponds to the risk of immature teeth with non-physiologic type of storage and a storage time of 20 minutes. We can see how the risk is much higher before 100 days than at any other time period. Prediction curves for any specific individual can be plotted using these values.

# 8 Conclusion

The estimation method proposed in this paper is very general and allows to deal with a wide range of situations. We first introduced the method for the mixed case of left-censored, interval-censored and right-censored data and we then directly extended it to consider the inclusion of exact observations and a cure fraction. We showed that treating the true event times as unobserved and using the EM algorithm to perform estimation resulted in a diagonal block matrix of the baseline hazard in the piecewise constant Cox model. This is a very interesting feature of our approach since the standard estimation method for this model (see for instance [21]) results in a full rank Hessian matrix, which can pose some serious computational problems for a moderate number of baseline cuts. Moreover, this allowed us to use the $L_0$ penalisation technique developed in [10] and [20] which was also implemented for exact and right censored data in [6]. Starting from a large grid of baseline cuts this penalisation technique forces two similar adjacent values to be equal. This results in a very flexible model since the location and number of cuts of the baseline are directly determined from the data. As compared to the ICsurv method from [27], the EM algorithm is readily applicable without need of a data augmentation step. Even though our cumulative baseline hazard does not result in a smooth function as compared to their spline approach, our method was shown to perform greatly on simulated data and even to outperform the method from [27] especially in terms of bias of the estimated parameters. It should be mentioned that their method could probably be improved by using an automatic procedure to choose the location and number of knots from the data. However, this is a complicated problem and there is currently no available method that could be directly applied on this estimator (see [26] for a review on selection methods of knots for spline estimators). On the dental dataset we also showed the interesting feature of the adaptive ridge procedure: by detecting the different time regions where the hazard for ankylosis changes, it revealed a very high risk of failure from replantation of the tooth until 100 days after replantation and a risk near to zero after 900 days. Finally, theoretical results were also provided for the adaptive ridge estimator. They show that the asymptotic distribution of the parameters can be determined by considering the estimated set of cuts as fixed and by using standard asymptotic likelihood theory for the piecewise constant hazard model.

By use of a logit link we developed the general cure model introduced by [22] and [19], for interval-censored data. From this model the effect of covariates on the odds of being cured and on the hazard risk of the susceptibles can be assessed. Interestingly, the combination of the piecewise constant baseline hazard and the adaptive ridge procedure produce a very flexible model in this context and avoids the use of arbitrary constraints such as in [22] where the authors had to require that the conditional survival function is set to zero beyond the last event time.

Another type of heterogeneity could be modelled with the use of frailty models (see [23] for instance). The EM approach for frailty models could then be used as a direct extension of our estimation method. However, it would require to compute the conditional value of the frailty variable given the observed data, a work that is left to future research. Similarly the standard mixture problem where one assumes the population to be composed of two (or more) subgroups with different hazards could be considered (see for instance [7] for this model in a high dimensional setting). The use of the piecewise constant baseline hazard would be crucial for this problem as the model is only identifiable for parametric baselines. The implementation

of the adaptive ridge procedure would then result in a very flexible model for this problem.

## References


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

[2] J. Andreasen, M. Borum, H. Jacobsen, and F. Andreasen. Replantation of 400 avulsed permanent incisors. 1. diagnosis of healing complications. *Dental Traumatology*, 11(2):51–58, 1995.

[3] J. Andreasen, M. K. Borum, H. Jacobsen, and F. Andreasen. Replantation of 400 avulsed permanent incisors. 4. factors related to periodontal ligament healing. *Dental Traumatology*, 11(2):76–89, 1995.

[4] R. A. Betensky, J. C. Lindsey, L. M. Ryan, and M. Wand. A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, 21(2):263–275, 2002.

[5] A. Boruvka and R. J. Cook. A cox-aalen model for interval-censored data. *Scandinavian Journal of Statistics*, 42(2):414–426, 2015.

[6] O. Bouaziz and G. Nuel. L0 regularization for the estimation of piecewise constant hazard rates in survival analysis. *Applied Mathematics*, 8(3), 2017.

[7] S. Bussy, A. Guilloux, S. Gaïffas, and A.-S. Jannot. C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical methods in medical research*, 2017.

[8] B. Carstensen. Regression models for interval censored survival data: application to hiv infection in danish homosexual men. *Statistics in Medicine*, 15(20):2177–2189, 1996.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[10] F. Frommlet and G. Nuel. An adaptive ridge procedure for $l_0$ regularization. *PLoS ONE*, 11(2), 2016.

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

[12] T. Hu and L. Xiang. Partially linear transformation cure models for interval-censored data. *Computational Statistics & Data Analysis*, 93:257–269, 2016.

[13] J. Huang and J. A. Wellner. Efficient estimation for the proportional hazards model with "case 2" interval censoring. Technical Report 290, Department of Statistics, University of Washington, Seattle, 1995.

[14] G. Jongbloed. The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7(3):310–321, 1998.

[15] E. Lauridsen, J. O. Andreasen, O. Bouaziz, and L. Andersson. Risk of ankylosis of 400 avulsed and replanted human teeth in relation to length of dry storage. a re-evaluation of a long-term clinical study. *Dental Traumatology*, 2019.

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

[17] H. Liu and Y. Shen. A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association*, 104(487):1168–1178, 2009.

[18] T. A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982.

[19] Y. Peng and K. B. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

[20] R. C. Rippe, J. J. Meulman, and P. H. Eilers. Visualization of genomic changes by segmented smoothing using an l0 penalty. *PloS one*, 7(6), 2012.

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

[22] J. P. Sy and J. M. Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[23] T. M. Therneau and P. M. Grambsch. *Modeling survival data: extending the Cox model*. Statistics for Biology and Health. Springer-Verlag, New York, 2000.

[24] B. 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.

[25] R. Varadhan and C. Roland. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353, 2008.

[26] M. P. Wand. A comparison of regression spline smoothing procedures. *Computational Statistics*, 15(4):443–462, 2000.

[27] L. Wang, C. S. McMahan, M. G. Hudgens, and Z. P. Qureshi. A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, 72(1):222–231, 2016.

[28] D. Zeng, L. Mao, and D. Lin. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103(2):253–271, 2016.

[29] Z. Zhang, L. Sun, X. Zhao, and J. Sun. Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, 33(1):61–70, 2005.

[30] M. Zhou. *Empirical likelihood method in survival analysis*. Chapman and Hall/CRC, 2015.

[31] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Table 1: Simulation results for the estimation of $\beta$ in Model M1 (piecewise constant baseline hazard), for Scenarios S1 and S2 with 100% of susceptible individuals. S1: no exact data, 25% of left-censoring, 52% of interval-censoring, 23% of right-censoring. S2: 18% of exact data, 19% of left-censoring, 40% of interval-censoring, 23% of right-censoring.

| | | Adaptive Ridge estimate | | | | Midpoint estimate | | | ICsurv estimate | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) | CP($\hat{\beta}$) | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) |
| S1 | 200 | 0.032 | 0.235 | 0.056 | 0.942 | −0.174 | 0.184 | 0.064 | 0.038 | 0.229 | 0.054 |
| | | −0.010 | 0.181 | 0.033 | 0.924 | 0.057 | 0.141 | 0.023 | −0.017 | 0.184 | 0.034 |
| | 400 | 0.012 | 0.166 | 0.028 | 0.946 | −0.177 | 0.127 | 0.047 | 0.016 | 0.160 | 0.026 |
| | | −0.014 | 0.120 | 0.015 | 0.938 | 0.050 | 0.096 | 0.012 | −0.013 | 0.121 | 0.015 |
| | 1 000 | 0.007 | 0.099 | 0.010 | 0.948 | −0.171 | 0.075 | 0.035 | 0.007 | 0.096 | 0.009 |
| | | −0.003 | 0.075 | 0.006 | 0.946 | 0.056 | 0.062 | 0.007 | −0.003 | 0.075 | 0.006 |
| S2 | 200 | 0.033 | 0.213 | 0.047 | 0.945 | −0.128 | 0.181 | 0.049 | | | |
| | | −0.006 | 0.169 | 0.029 | 0.954 | 0.045 | 0.147 | 0.024 | | | |
| | 400 | 0.003 | 0.153 | 0.023 | 0.947 | −0.138 | 0.128 | 0.035 | | | |
| | | −0.001 | 0.119 | 0.014 | 0.952 | 0.046 | 0.104 | 0.013 | | | |
| | 1 000 | 0.006 | 0.092 | 0.009 | 0.948 | −0.136 | 0.078 | 0.025 | | | |
| | | 0.002 | 0.071 | 0.005 | 0.949 | 0.051 | 0.062 | 0.006 | | | |

Table 2: Simulation results for the estimation of $S_0$ in Scenarios S1 and S2 in Model M1 (piecewise constant baseline hazard), with 100% of susceptible individuals. S1: no exact data, 25% of left-censoring, 52% of interval-censoring, 23% of right-censoring. S2: 18% of exact data, 19% of left-censoring, 40% of interval-censoring, 23% of right-censoring.

| | | Adaptive Ridge estimate | | | Midpoint estimate | | ICsurv estimate | |
|---|---|---|---|---|---|---|---|---|
| | $n$ | IBias$^2$($\hat{S}_0$) | IVar($\hat{S}_0$) | TV($\hat{\lambda}_0$) | IBias$^2$($\hat{S}_0$) | IVar($\hat{S}_0$) | IBias$^2$($\hat{S}_0$) | IVar($\hat{S}_0$) |
| S1 | 200 | 0.002 | 0.266 | 0.784 | 0.124 | 0.122 | 0.003 | 0.438 |
| | 400 | 0.003 | 0.138 | 0.600 | 0.124 | 0.061 | 0.002 | 0.213 |
| | 1 000 | 0.002 | 0.059 | 0.416 | 0.126 | 0.023 | 0.001 | 0.077 |
| S2 | 200 | 0.001 | 0.196 | 0.646 | 0.074 | 0.114 | | |
| | 400 | 0.001 | 0.103 | 0.484 | 0.074 | 0.060 | | |
| | 1 000 | 0.000 | 0.038 | 0.277 | 0.075 | 0.022 | | |

Table 3: Simulation results for the estimation of $\beta$ in Model M2 (Weibull baseline hazard), for Scenarios S1 and S2 with 100% of susceptible individuals. S1: no exact data, 25% of left-censoring, 52% of interval-censoring, 23% of right-censoring. S2: 18% of exact data, 19% of left-censoring, 40% of interval-censoring, 23% of right-censoring.

| | | Adaptive Ridge estimate | | | | Midpoint estimate | | | ICsurv estimate | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) | CP($\hat{\beta}$) | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) |
| S1 | 200 | 0.027 | 0.572 | 0.328 | 0.916 | −0.596 | 0.168 | 0.383 | −0.267 | 0.307 | 0.166 |
| | | −0.032 | 0.516 | 0.267 | 0.922 | 0.184 | 0.146 | 0.055 | 0.091 | 0.258 | 0.075 |
| | 400 | 0.022 | 0.412 | 0.171 | 0.930 | −0.609 | 0.116 | 0.384 | −0.263 | 0.234 | 0.124 |
| | | −0.021 | 0.298 | 0.089 | 0.934 | 0.193 | 0.104 | 0.048 | 0.087 | 0.174 | 0.038 |
| | 1 000 | 0.021 | 0.206 | 0.043 | 0.948 | −0.611 | 0.075 | 0.379 | −0.251 | 0.158 | 0.088 |
| | | 0.009 | 0.170 | 0.029 | 0.954 | 0.198 | 0.062 | 0.043 | 0.078 | 0.112 | 0.018 |
| S2 | 200 | −0.085 | 0.295 | 0.094 | 0.936 | −0.581 | 0.157 | 0.362 | | | |
| | | 0.012 | 0.239 | 0.057 | 0.941 | 0.192 | 0.149 | 0.059 | | | |
| | 400 | −0.066 | 0.217 | 0.052 | 0.942 | −0.582 | 0.115 | 0.352 | | | |
| | | 0.015 | 0.159 | 0.025 | 0.950 | 0.181 | 0.096 | 0.042 | | | |
| | 1 000 | −0.048 | 0.134 | 0.020 | 0.949 | −0.587 | 0.072 | 0.349 | | | |
| | | −0.004 | 0.103 | 0.011 | 0.950 | 0.190 | 0.061 | 0.040 | | | |

Table 4: Simulation results for the estimation of $S_0$ in Scenarios S1 and S2 in Model M2 (Weibull baseline hazard), with 100% of susceptible individuals. S1: no exact data, 25% of left-censoring, 52% of interval-censoring, 23% of right-censoring. S2: 18% of exact data, 19% of left-censoring, 40% of interval-censoring, 23% of right-censoring.

| | | Adaptive Ridge estimate | | Midpoint estimate | | ICsurv estimate | |
|---|---|---|---|---|---|---|---|
| | $n$ | $\text{IBias}^2(\hat{S}_0)$ | $\text{IVar}(\hat{S}_0)$ | $\text{IBias}^2(\hat{S}_0)$ | $\text{IVar}(\hat{S}_0)$ | $\text{IBias}^2(\hat{S}_0)$ | $\text{IVar}(\hat{S}_0)$ |
| S1 | 200 | 0.026 | 0.647 | 1.857 | 0.077 | 0.082 | 0.229 |
| | 400 | 0.005 | 0.391 | 1.856 | 0.043 | 0.069 | 0.148 |
| | 1 000 | 0.005 | 0.169 | 1.931 | 0.015 | 0.050 | 0.060 |
| S2 | 200 | 0.016 | 0.196 | 1.033 | 0.087 | | |
| | 400 | 0.010 | 0.104 | 1.046 | 0.040 | | |
| | 1 000 | 0.003 | 0.044 | 1.056 | 0.017 | | |

Table 5: Proportions of the number of cuts found by the adaptive ridge algorithm in Scenario S1 Model M1 (piecewise constant baseline hazard). The true number of cuts is 3.

| Number of cuts | Proportions found for: $n = 200$ | $n = 400$ | $n = 1\,000$ |
|---|---|---|---|
| 1 | 0.690 | 0.598 | 0.400 |
| 2 | 0.288 | 0.358 | 0.560 |
| 3 | 0.020 | 0.036 | 0.038 |
| 4 | 0.002 | 0.006 | 0.002 |

Table 6: Probabilities that a cut value has been selected by the adaptive ridge algorithm in the sets $[10, 30]$ and $[35, 55]$ in Scenario S1 Model M1 (piecewise constant baseline hazard). The true cuts are located at positions 20, 40 and 50.

| | | $n = 200$ | $n = 400$ | $n = 1\,000$ |
|---|---|---|---|---|
| Number of cuts in $[10, 30]$ | 0 | 0.718 | 0.710 | 0.560 |
| | 1 | 0.280 | 0.286 | 0.434 |
| | 2 | 0.020 | 0.004 | 0.006 |
| Number of cuts in $[35, 55]$ | 0 | 0.198 | 0.094 | 0.040 |
| | 1 | 0.782 | 0.844 | 0.860 |
| | 2 | 0.020 | 0.062 | 0.100 |

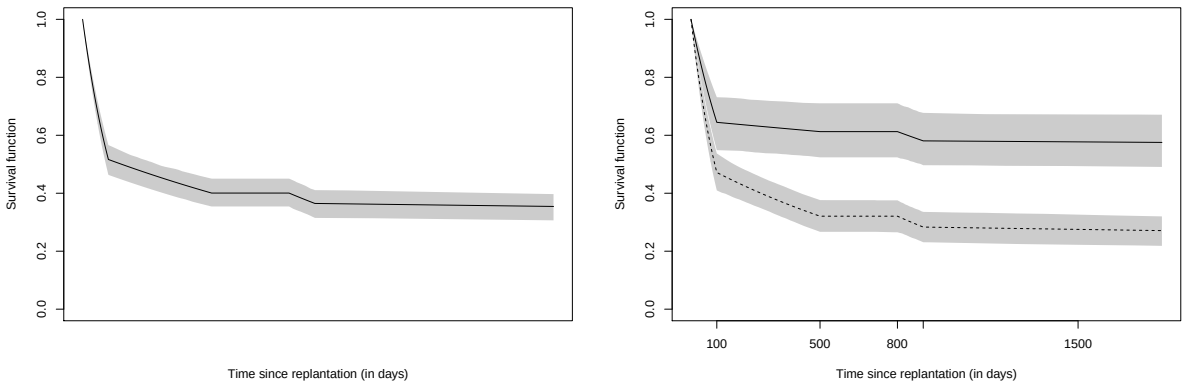


Figure 1: On the left panel, estimate of the survival function of time to ankylosis for the whole population. On the right panel, estimates of the survival function for the immature teeth (solid line) and for the mature teeth (dotted lines). Confidence intervals are plotted along the curves in shaded areas using the bootstrap approach.

| Covariates | HR | 95% CI | p-value |
|---|---|---|---|
| Mature | 2.00 | $[1.74; 2.29]$ | $1.89 \times 10^{-5}$ |
| Storage time (hours) | 1.23 | $[1.11; 1.34]$ | 0.0017 |
| Physiologic storage | 0.93 | $[0.81; 1.06]$ | 0.6980 |
| Age>20 (mature teeth) | 1.27 | $[0.99; 1.61]$ | 0.1272 |

Table 7: Regression modelling of time to ankylosis on the dental dataset (HR: Hazard Ratio, CI: Confidence Interval). The adaptive ridge found four cuts for the baseline hazard at times 100, 500, 800 and 900.

| Cuts | $\exp(\hat{a}_k) \times 10^3$ | 95% CI $\times 10^3$ |
|---|---|---|
| $(0, 100]$ | 3.71 | $[3.19; 4.28]$ |
| $(100, 500]$ | 0.39 | $[0.28; 0.52]$ |
| $(500, 800]$ | 0.00 | $[0.00; 0.00]$ |
| $(800, 900]$ | 0.62 | $[0.31; 1.07]$ |
| $(900, +\infty)$ | 0.02 | $[0.01; 0.04]$ |

Table 8: Baseline hazard from the regression modelling of time to ankylosis on the dental dataset (CI: Confidence Interval). This hazard corresponds to the risk of immature teeth with non-physiologic type of storage and a storage time of 20 minutes.

# Supplementary Material

## A.1 Expressions of the statistics $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$

For $k = 1, \dots, K$, $i = 1, \dots, n$, define

$$\begin{aligned} A_{k,i}^{\text{old}} &= \frac{\exp\Big(e^{a_{i,k}^{\text{old}}} c_{k-1} + a_{i,k}^{\text{old}} - \sum_{j=1}^{k-1} e^{a_{i,j}^{\text{old}}} (c_j - c_{j-1})\Big) J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})} \int_{c_{k-1} \vee L_i}^{c_k \wedge R_i} \exp\big(-e^{a_{i,k}^{\text{old}}} t\big) dt \\ &= \exp\Big(-e^{a_{i,k}^{\text{old}}} c_{k-1} \vee L_i\Big)\Big(1 - \exp\big(-e^{a_{i,k}^{\text{old}}} (c_k \wedge R_i - c_{k-1} \vee L_i)\big)\Big) \\ &\quad \times \frac{\exp\big(e^{a_{i,k}^{\text{old}}} c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}^{\text{old}}} (c_j - c_{j-1})\big) J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})} \end{aligned}$$

and

$$\begin{aligned} B_{k,i}^{\text{old}} &= \frac{\exp\Big(e^{a_{i,k}^{\text{old}}} c_{k-1} + a_{i,k}^{\text{old}} - \sum_{j=1}^{k-1} e^{a_{i,j}^{\text{old}}} (c_j - c_{j-1})\Big) J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})} \int_{c_{k-1} \vee L_i}^{c_k \wedge R_i} (t - c_{k-1}) \exp(-e^{a_{i,k}^{\text{old}}} t) dt \\ B_{k,i}^{\text{old}} &= \Big\{\big(\exp(-a_{i,k}^{\text{old}}) + c_{k-1} \vee L_i - c_{k-1}\big) \exp(-e^{a_{i,k}^{\text{old}}} c_{k-1} \vee L_i) \\ &\quad - \big(\exp(-a_{i,k}^{\text{old}}) + c_k \wedge R_i - c_{k-1}\big) \exp(-e^{a_{i,k}^{\text{old}}} c_k \wedge R_i)\Big\} \\ &\quad \times \frac{\exp\big(e^{a_{i,k}^{\text{old}}} c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}^{\text{old}}} (c_j - c_{j-1})\big) J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i \mid Z_i, \boldsymbol{\theta}_{\text{old}})}. \end{aligned}$$

The function $Q$ is then expressed as a function of these two statistics (see Section 3 of the main paper).

## A.2 The Schurr complement

The Schurr complement is used to compute the inverse of the Hessian matrix of $Q$, in the case of fixed cuts (Section 3 of the main paper) and of $\ell^{\text{pen}}$, for the adaptive ridge estimator (Section 4 of the main paper). It makes use of the special structure of the block matrix corresponding to the second order derivatives with respect to the $a_k$s which is either diagonal (for $Q$) or tri-diagonal (for $\ell^{\text{pen}}$).

Let $\mathcal{I}(a, \beta)$ be minus the Hessian matrix of $Q$ or $\ell^{\text{pen}}$ for the maximisation problem with respect to $a_1, \dots, a_L$ and $\beta_1, \dots, \beta_{d_Z}$. Let $A$ be of dimension $K \times K$, $B$ of dimension $K \times d_Z$ and $C$ be of dimension $d_Z \times d_Z$ such that

$$\mathcal{I}(a, \beta) = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}$$

Let $U(a, \beta)$ be the score vector of $Q$ or $\ell^{\text{pen}}$ and $b_1$ be the column vector of dimension $K$, $b_2$ be the column vector of dimension $d_Z$ such that $U(a, \beta) = (b_1, b_2)^t$. Using the Schurr complement,

we have

$$\mathcal{I}(a,\beta)^{(-1)}U(a,\beta) = \begin{pmatrix} A^{-1}b_1 - A^{-1}B(C - B^tA^{-1}B)^{-1}(b_2 - B^tA^{-1}b_1) \\ (C - B^tA^{-1}B)^{-1}(b_2 - B^tA^{-1}b_1) \end{pmatrix}.$$

For the inversion of the Hessian matrix of $Q$ and $\ell^{\text{pen}}$, the $K \times K$ matrix $A$ is either diagonal (for $Q$) or a band matrix of bandwidth equal to 1 (for $\ell^{\text{pen}}$). Its inverse can be efficiently computed using a fast C++ implementation of the LDL algorithm. This is achieved in linear complexity using the R `bandsolve` package. As a result, the total complexity for the computation of $\mathcal{I}(a,\beta)^{(-1)}U(a,\beta)$ is of order $\mathcal{O}(K)$ in the case $K >> d_Z$.

## A.3 Score vector and Hessian matrix for the function $Q$ when including exact observations and a cure fraction

In the presence of exact observations and a cure fraction, the score vector and the Hessian matrix are given from the following formulas:

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k} &= \sum_{i \text{ not exact}} \pi_i^{\text{old}} \left\{ A_{k,i}^{\text{old}} - (c_k - c_{k-1})e^{a_k} I(k \neq K) \sum_{l=k+1}^{K} A_{l,i}^{\text{old}} e^{\beta Z_i} - e^{a_k} B_{k,i}^{\text{old}} e^{\beta Z_i} \right\} \\
&\quad + \sum_{i \text{ exact}} \Big\{ O_{i,k} - \exp(a_k + \beta Z_i) R_{i,k} \Big\}, \\
\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial \beta} &= \sum_{i \text{ not exact}} \pi_i^{\text{old}} Z_i \sum_{l=1}^{K} \left( A_{l,i}^{\text{old}} - \left\{ \sum_{j=1}^{l-1} (c_j - c_{j-1}) e^{a_j} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_l} B_{l,i}^{\text{old}} e^{\beta Z_i} \right\} \right) \\
&\quad + \sum_{i \text{ exact}} Z_i \sum_{l=1}^{K} \Big\{ O_{i,l} - \exp(a_l + \beta Z_i) R_{i,l} \Big\},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k^2} &= - \sum_{i \text{ not exact}} \pi_i^{\text{old}} \left\{ (c_k - c_{k-1}) e^{a_k} I(k \neq K) \sum_{l=k+1}^{K} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_k} B_{k,i}^{\text{old}} e^{\beta Z_i} \right\} \\
&\quad - \sum_{i \text{ exact}} \exp(a_k + \beta Z_i) R_{i,k}, \\
\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial \beta^2} &= - \sum_{i \text{ not exact}} \pi_i^{\text{old}} Z_i Z_i^t \sum_{l=1}^{K} \left( \sum_{j=1}^{l-1} (c_j - c_{j-1}) e^{a_j} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_l} B_{l,i}^{\text{old}} e^{\beta Z_i} \right) \\
&\quad - \sum_{i \text{ exact}} Z_i Z_i^t \sum_{l=1}^{K} \exp(a_l + \beta Z_i) R_{i,l}, \\
\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})}{\partial a_k \partial \beta} &= - \sum_{i \text{ not exact}} \pi_i^{\text{old}} Z_i \left( (c_k - c_{k-1}) e^{a_k} I(k \neq K) \sum_{l=k+1}^{K} A_{l,i}^{\text{old}} e^{\beta Z_i} + e^{a_k} B_{k,i}^{\text{old}} e^{\beta Z_i} \right), \\
&\quad - \sum_{i \text{ exact}} Z_i \exp(a_k + \beta Z_i) R_{i,k}.
\end{aligned}$$

## A.4 Full regularisation path on a simulated dataset

We illustrate in this section the full regularisation path of the algorithm. As explained in Section 4 of the main paper the algorithm consists of the detection of the set of cuts from the penalised estimator combined with the non-penalised estimator using this estimated set of cuts.

We consider one sample generated from Model M1, Scenario S1 of Section 6 of the main paper in the absence of covariates and we estimate the hasard function using both the ridge and the adaptive ridge algorithm. More precisely, the first algorithm uses the weights $\hat{w}_k$ equal to 1 while the second algorithm iteratively updates the $\hat{w}_k$ using Equation (5) of the main paper. A set of penalty is chosen, on the log scale, as the set of 200 equally spaced values ranging from $\log(0.1)$ to $\log(10\,000)$. Figure 2 displays the regularisation path for the ridge on the left and for the adaptive ridge on the right where the $y$-axis represents the values of the estimated $a_k$'s for each penalty value of the $x$-axis. We clearly see that the ridge procedure produces a smooth estimation and the adaptive ridge procedure provides a selection of the cuts along with an estimated piecewise constant hazard. Both estimators converge toward the same constant model as pen tends to infinity. Figure 3 shows the resulting estimated hazard from the adaptive ridge procedure after selection of the cuts using the BIC. On the left panel it is seen that the BIC chooses a model with three cuts and four values of $a_k$'s. On the right panel we see that, on this sample, the adaptive ridge estimator follows closely the true value of the hazard.

## A.5 Proof of Theorem 5.1 of the main document

Proof of 1.

For this proof, we only consider the initial fixed set of cuts $\{c_1, \ldots, c_K\}$. In order to avoid confusion, we denote by $\boldsymbol{\theta}^\dagger = (a_1^\dagger, \ldots, a_K^\dagger, \beta^*)$ the true parameter using this set of cuts. This means that there might exist several $k$'s for which $a_k^\dagger = a_{k+1}^\dagger$. Note that removing the equal consecutive values of $a_k^\dagger$ will yield $\boldsymbol{\theta}^*$. In the following, we will prove that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^\dagger$ in probability.

For interval-censored, left or right-censored data, the full likelihood function can be written as

$$\tilde{\mathrm{L}}_n^{\mathrm{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} (f_{L,R,\delta}(L_i, R_i, 1))^{\delta_i} (f_{L,R,\delta}(L_i, R_i, 0))^{1-\delta_i},$$

where $f_{L,R,\delta}(L_i, R_i, 1), f_{L,R,\delta}(L_i, R_i, 0)$ represent the joint density of the mixed distribution $(L, R, \delta)$ respectively evaluated at $(L_i, R_i, 1)$ and $(L_i, R_i, 0)$. It is then seen that $f_{L,R,\delta}(L_i, R_i, 1) = \mathbb{P}[\delta = 1 \mid L = L_i, R = R_i, Z_i, \boldsymbol{\theta}] f_{L,R,Z}(L_i, R_i, Z_i)$ where $f_{L,R,Z}$ represents the joint density of $(L, R, Z)$ and $\mathbb{P}[\delta = 1 \mid L = L_i, R = R_i, Z_i, \boldsymbol{\theta}] = (S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta}))^{\delta_i}$ under the independent censoring assumption. The same kind of reasoning holds for $f_{L,R,\delta}(L_i, R_i, 0)$ such that

$$\begin{aligned}
\tilde{\mathrm{L}}_n^{\mathrm{obs}}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} (S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta}))^{\delta_i} (S(L_i \mid Z_i, \boldsymbol{\theta}))^{1-\delta_i} f_{L,R,Z}(L_i, R_i, Z_i), \\
&= \prod_{i=1}^{n} g_{\boldsymbol{\theta}}(L_i, R_i, Z_i),
\end{aligned}$$

where $g_{\boldsymbol{\theta}}(L_i, R_i, Z_i) := (S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})) f_{L,R,Z}(L_i, R_i, Z_i)$ with the slight abuse of notation $S(R_i \mid Z_i, \boldsymbol{\theta}) = 0$ if $R_i = \infty$ (for a right-censored observation). The above equation shows that the full likelihood is simply the observed likelihood $\mathrm{L}_n^{\mathrm{obs}}(\boldsymbol{\theta})$ of Section 3.1 of the main document multiplied by the quantity $f_{L,R,Z}(L_i, R_i, Z_i)$ which does not depend on $\boldsymbol{\theta}$. In case of exact observations, the full likelihood can be rewritten as:

$$\tilde{\mathrm{L}}_n^{\mathrm{obs}}(\boldsymbol{\theta}) = \prod_{i \text{ not exact}} g_{\boldsymbol{\theta}}(L_i, R_i, Z_i) \prod_{i \text{ exact}} f(L_i \mid Z_i, \boldsymbol{\theta}).$$

It should be noted that $g_{\boldsymbol{\theta}}(L_i, R_i, Z_i)$ and $f(L_i \mid Z_i, \boldsymbol{\theta})$ are densities. For $g_{\boldsymbol{\theta}}$, write

$$\begin{aligned}\iiint_{l\neq r} g_{\boldsymbol{\theta}}(l,r,z)dldrdz &= \mathbb{E}_{\boldsymbol{\theta}}\Big[I(L_i \neq R_i)\mathbb{E}_{\boldsymbol{\theta}}[S(L_i \mid Z_i,\boldsymbol{\theta}) - S(R_i \mid Z_i,\boldsymbol{\theta}) \mid L,R,Z]\Big] \\ &= \iiint \mathbb{P}[T \in (l,r) \mid L=l, R=r, Z=z, \boldsymbol{\theta}) f_{L,R,Z}(l,r,z)dldrdz.\end{aligned}$$

From the independent censoring assumption, $\mathbb{P}[T \in (l,r) \mid L=l, R=r, Z=z, \boldsymbol{\theta})] = 1$ and consequently $g_{\boldsymbol{\theta}}$ is a density.

Now the penalised estimator defined in (6) of the main document verifies $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_n^{\text{pen}}(\boldsymbol{\theta})$, where

$$\ell_n^{\text{pen}}(\boldsymbol{\theta}) = \left\{ \ell_n(\boldsymbol{\theta}) - \frac{\text{pen}}{2n} \sum_{k=1}^{K-1} \hat{w}_k^{(1)} (a_{k+1} - a_k)^2 \right\},$$

with $\ell_n(\boldsymbol{\theta}) = \log(\tilde{\mathrm{L}}_n^{\text{obs}}(\boldsymbol{\theta}))/n$. We introduce $\ell(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}^\dagger}[I(L_i \neq R_i)\log(g_{\boldsymbol{\theta}}(L_i,R_i,Z_i))] + \mathbb{E}_{\boldsymbol{\theta}^\dagger}[I(L_i = R_i)\log(f(L_i \mid Z_i,\boldsymbol{\theta}))]$ and we write:

$$|\ell_n^{\text{pen}}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \leq |\ell_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| + \frac{\text{pen}}{2n} \sum_{k=1}^{K-1} \hat{w}_k^{(1)} (a_{k+1} - a_k)^2.$$

The two terms on the right-hand side of the equation converge toward 0 in probability: the first one from the law of large numbers, and the second one from the consistency of $\hat{w}_k^{(1)}$ and the condition $\text{pen}/n \to 0$.

Then, from Jensen inequality,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}^\dagger}\left[-I(L_i \neq R_i)\log\left(\frac{g_{\boldsymbol{\theta}}(L_i,R_i,Z_i)}{g_{\boldsymbol{\theta}^\dagger}(L_i,R_i,Z_i)}\right)\right] &\geq -\log\left(\mathbb{E}_{\boldsymbol{\theta}^\dagger}\left[I(L_i \neq R_i)\frac{g_{\boldsymbol{\theta}}(L_i,R_i,Z_i)}{g_{\boldsymbol{\theta}^\dagger}(L_i,R_i,Z_i)}\right]\right) \\ &\geq -\log\left(\iiint_{l\neq r} \frac{g_{\boldsymbol{\theta}}(l,r,z)}{g_{\boldsymbol{\theta}^\dagger}(l,r,z)} g_{\boldsymbol{\theta}^\dagger}(l,r,z)dldrdz\right) = 0.\end{aligned}$$

The same reasoning applies to $\mathbb{E}_{\boldsymbol{\theta}^\dagger}[I(L_i = R_i)\log(f(L_i \mid Z_i,\boldsymbol{\theta})/f(L_i \mid Z_i,\boldsymbol{\theta}^\dagger))]$ which proves that $\ell(\boldsymbol{\theta}) \leq \ell(\boldsymbol{\theta}^\dagger)$ for all $\boldsymbol{\theta}$. To conclude, we have proved that $|\ell_n(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \to 0$ in probability, with $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_n^{\text{pen}}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^\dagger = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$. The concavity of $\ell_n^{\text{pen}}(\boldsymbol{\theta})$ yields that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^\dagger$ in probability.

Proof of 2. and 3.

We start by working on the true set of cuts $\mathcal{A}^*$. We need to define the estimator $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*}$, that is our estimator using the true set of cuts. In particular we need to define the value of $\hat{\hat{a}}_{k,\mathcal{A}^*}$ on each interval $c_{k-1}^* < t \leq c_k^*$. As a matter of fact, for a given $n$ the sets $\mathcal{A}_n$ and $\mathcal{A}^*$ might be different and therefore some $\hat{\hat{a}}_{k,\mathcal{A}^*}$ might not exist. We set:

$$\exp(\hat{\hat{a}}_{k,\mathcal{A}^*}) = \hat{\hat{\lambda}}_{0,\mathcal{A}_n}(c_{k-1}^*).$$

This definition is arbitrary and any value of $t \in (c_{k-1}^*, c_k^*]$ could be taken for $\hat{\hat{\lambda}}_{0,\mathcal{A}_n}(t)$. We now also define $\ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}) = \log(\mathrm{L}_{n,\mathcal{A}^*}^{\text{obs}}(\boldsymbol{\theta}))$ the observed log-likelihood defined using the true set of cuts $\mathcal{A}^*$. From a Taylor expansion, we have:

$$\nabla_{\boldsymbol{\theta}} \ell_{n,\mathcal{A}^*}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*}) = \nabla_{\boldsymbol{\theta}} \ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}^*) + (\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*} - \boldsymbol{\theta}^*)^t \nabla^2_{\boldsymbol{\theta}} \ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*}),$$

where $\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*}$ is on the line segment between $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*}$ and $\boldsymbol{\theta}^*$. As a consequence,

$$\sqrt{n}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*} - \boldsymbol{\theta}^*)^t = -(\nabla^2_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*})/n)^{-1}(\nabla_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}^*) - \nabla_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*}))\frac{1}{\sqrt{n}}\cdot \tag{7}$$

From the result in 1. of this theorem, $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*} \to \boldsymbol{\theta}^*$ in probability, and thus $\nabla^2_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*})/n - \nabla^2_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}^*)/n$ converges to 0 in probability and $-\nabla^2_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*})/n \to -\mathbb{E}[\nabla^2_{\boldsymbol{\theta}}h^*_{\boldsymbol{\theta}}(L_i,R_i,Z_i))|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}] = \Sigma$ in probability.

The key to the proof is now to show that $\nabla_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*})/\sqrt{n}$ converges to 0 in probability. We denote by $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*}$ the estimator that maximises $\ell_{n,\mathcal{A}^*}(\boldsymbol{\theta})$. Noticing that $\nabla_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\hat{\boldsymbol{\theta}}_{\mathcal{A}^*}) = 0$ we have

$$\nabla_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*})/\sqrt{n} = \sqrt{n}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*} - \hat{\boldsymbol{\theta}}_{\mathcal{A}^*})^t\nabla^2_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*})/n, \tag{8}$$

where $\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*}$ is on the line segment between $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*}$ and $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*}$. Since $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*} \to \boldsymbol{\theta}^*$ and $\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*} - \hat{\boldsymbol{\theta}}_{\mathcal{A}^*} \to 0$ in probability, we can prove as previously that $\nabla^2_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*})/n \to \Sigma$ in probability.

We now work on the initial set of cuts $\{c_1,\dots,c_K\}$ and we define $\hat{\boldsymbol{\theta}}^\dagger$, the estimator $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*}$ that is defined on $\{c_1,\dots,c_K\}$ (this is always possible since $\mathcal{A}^* \subset \{c_1,\dots,c_K\}$). We need to prove that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^\dagger)^t$ converges to 0 in probability which will imply that $\sqrt{n}(\hat{\hat{\boldsymbol{\theta}}}_{\mathcal{A}^*} - \hat{\boldsymbol{\theta}}_{\mathcal{A}^*})^t$ converges to 0 in probability. Introduce the function:

$$\psi_n(u,v) := \ell_n(\hat{\boldsymbol{\theta}}^\dagger + (u,v)/\sqrt{n}) - \ell_n(\hat{\boldsymbol{\theta}}^\dagger) - \frac{\text{pen}}{2n}\sum_{k=1}^{K-1}\hat{w}_k^{(1)}(V(\hat{a}_k^\dagger + u_k/\sqrt{n}) - V(\hat{a}_k^\dagger)),$$

where $(u,v) = (u_1,\dots,u_K,v_1,\dots,v_{d_Z})$ is a row vector of dimension $(K+d_Z)$ and $V(a_k) = (a_{k+1}-a_k)^2$. For

$$(\hat{u},\hat{v}) = \arg\min_{u,v}\psi_n(u,v),$$

we have $\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}^\dagger + \hat{u}/\sqrt{n}$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\dagger + \hat{v}/\sqrt{n}$, that is $\hat{u} = \sqrt{n}(\hat{\boldsymbol{a}} - \hat{\boldsymbol{a}}^\dagger)$ and $\hat{v} = \sqrt{n}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^\dagger)$. We now study the limit of $\psi_n$. First of all,

$$\ell_n(\hat{\boldsymbol{\theta}}^\dagger + (u,v)/\sqrt{n}) - \ell_n(\hat{\boldsymbol{\theta}}^\dagger) = \frac{(u,v)}{\sqrt{n}}\nabla_\theta\ell_n(\hat{\boldsymbol{\theta}}^\dagger) + \frac{1}{2n}(u,v)\nabla^2_\theta\ell_n(\hat{\boldsymbol{\theta}}^\dagger)(u,v)^t + o_{\mathbb{P}}(1),$$

where the $o_{\mathbb{P}}(1)$ is obtained from the law of large numbers applied to the partial derivatives of order three of $\ell_n(\tilde{\boldsymbol{\theta}}_n)$, for a $\tilde{\boldsymbol{\theta}}_n$ on the line segment between $\hat{\boldsymbol{\theta}}^\dagger$ and $(u,v)/\sqrt{n}$. By definition, $\hat{\boldsymbol{\theta}}^\dagger$ maximises $\ell_n$ and therefore $\nabla_\theta\ell_n(\hat{\boldsymbol{\theta}}^\dagger) = 0$. By the law of large numbers, $\frac{1}{2n}(u,v)\nabla^2_\theta\ell_n(\hat{\boldsymbol{\theta}}^\dagger)(u,v)^t$ converges in probability toward $\frac{1}{2}(u,v)\nabla^2_\theta\ell(\boldsymbol{\theta}^\dagger)(u,v)^t = -\frac{1}{2}(u,v)\Sigma(u,v)^t$. Secondly,

$$V(\hat{a}_k^\dagger + u_k/\sqrt{n}) - V(\hat{a}_k^\dagger) = \frac{2}{\sqrt{n}}(\hat{a}_{k+1}^\dagger - \hat{a}_k^\dagger)(u_{k+1}-u_k) + \frac{(u_{k+1}-u_k)^2}{n}.$$

Since $\hat{w}_k^{(1)} \to ((a_{k+1}^\dagger - a_k^\dagger)^2 + \varepsilon^2)^{-1}$, $\hat{a}_{k+1}^\dagger - \hat{a}_k^\dagger \to a_{k+1}^\dagger - a_k^\dagger$ in probability and

$$\left|\frac{a_{k+1}^\dagger - a_k^\dagger}{(a_{k+1}^\dagger - a_k^\dagger)^2 + \varepsilon^2}\right| < 1,$$

we see that $V(\hat{a}_k^\dagger + u_k/\sqrt{n}) - V(\hat{a}_k^\dagger) \to 0$ in probability. To summarise we have shown that $\psi_n(u,v) \to -\frac{1}{2}(u,v)\Sigma(u,v)^t$ in probability. Since $\Sigma$ is a positive definite matrix, $-\frac{1}{2}(u,v)\Sigma(u,v)^t$ is minimal for $(u,v) = (0,0)$. This proves that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^\dagger)^t$ converges to 0 in probability.

Going back to Equations (7) and (8), and from the asymptotic normality of $\nabla_{\boldsymbol{\theta}}\ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}^*)/\sqrt{n}$

using the Central Limit Theorem, we finally obtain:

$$\sqrt{n}(\hat{\tilde{\boldsymbol{\theta}}}_{\mathcal{A}^*} - \boldsymbol{\theta}^*)^t = -(\nabla^2_{\boldsymbol{\theta}} \ell_{n,\mathcal{A}^*}(\tilde{\boldsymbol{\theta}}_{\mathcal{A}^*})/n)^{-1}(\nabla_{\boldsymbol{\theta}} \ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}^*))\frac{1}{\sqrt{n}} + o_{\mathbb{P}}(1) \longrightarrow \Sigma^{-1}\mathcal{N}(0,\Sigma),$$

in distribution. This concludes the proof.

## A.6 Extended simulation study for the piecewise constant hazard model: two scenarios that include exact observations and a cure fraction

We consider two new scenarios which include a proportion of non-susceptible individuals. For the susceptibles, the data include left, interval and right-censored observations along with a proportion of exact observations. The model is defined by Equations (2) and (3) of the main paper with a logistic link for the probability of being cured. In both scenarios, the $Z$ covariate, $\beta$ coefficient and $\lambda_0$ baseline function are all generated as in the simulation section of the main paper. The $X$ covariate is of dimension $d_X = 2$ (including the intercept) and follows a Bernoulli distribution with parameter 0.8. In Scenario S3, $\gamma = (\log(2.35), \log(2))^t$ and in Scenario S4, $\gamma = (\log(0.8), \log(2))^t$. These values yield an average number of susceptible individuals $\mathbb{E}[p(X)]$ respectively equal to 80% and 58%. Among the susceptibles, both scenarios correspond to a proportion of 18% of exact observations, 19% of left observations, 40% of interval-censored observations and 23% of right-censored observations. The results are presented in Table 9. Only our adaptive ridge estimator has been implemented for these two scenarios. The $\gamma$ estimator is initialised to 0 in the EM algorithm.

A slight deterioration of the variance estimation of $\hat{\beta}$ and $\hat{\lambda}_0$ is seen when a cure fraction is included and the degree of deterioration increases as the proportion of cured gets bigger. On the other hand the bias of the parameter estimates is similar with or without the cure fraction. In the presence of a cure fraction, the $\gamma$ parameter is less accurately estimated as compared to the $\beta$ parameter both in terms of bias and variance. Nevertheless the results show that as the sample size increases the bias and variance of $\hat{\gamma}$ get smaller with a bias very close to 0 for a sample size equal to 1 000. The estimation performance of $\mathbb{E}[p(X)]$ was also investigated by computing the average value of $\sum_i \hat{p}(X_i)/n$ for all generated samples where $\hat{p}(X)$ is defined as in Equation (3) of the main paper with $\gamma$ replaced by $\hat{\gamma}$. For example, in Scenario S4 we found a bias and empirical standard error (SE) equal for $n = 200$ to 0.057 (SE = 0.064), for $n = 400$ to 0.046 (SE = 0.044) and for $n = 1\,000$ to 0.033 (SE = 0.028).

More simulations were conducted. In particular, the cure model without covariates for the cure fraction was also implemented in Scenario S1, Model M1 of the main paper such that the parameters to be estimated are $\boldsymbol{\theta} = (a_1, \ldots, a_L, \beta, p)$ with the true value of $p$ equal to 1. In replications of samples of size 400, it was seen that the model estimated the proportion of susceptibles $p$ to a value greater than 0.99 in 98% of cases and the lowest value on the 500 replications for the estimation of $p$ was equal to 0.95. This highlights the very high specificity of our model in terms of detecting a cure fraction. It shows that our model does not tend to overestimate the proportion of cured when the population is homogeneous, which is a very important feature of the estimation method. On the other hand, a scenario identical to Scenario S1, Model M1 but with a true proportion of susceptibles equal to $p = 0.7$ was also considered. In replications of samples of size 400, the estimator of $p$ was equal to 0.712 on average and only 0.5% of the estimates where greater than 0.99. This suggests in turn a high sensitivity of our model to detect heterogeneity in interval censored data.

## A.7 Computational cost of the adaptive ridge algorithm

The complexity for the inversion of the Hessian of $\ell$ is of order $\mathcal{O}(K)$, in the case $K >> d_X + d_Z$ (see Section A.2 in the Supporting Information about the Schurr complement). However, for a given penalty, it should be noted that the global algorithm for maximising $Q$ or $\ell^{\text{pen}}$ consists of an EM algorithm with a Newton-Raphson procedure at each step. As a consequence, in the simulations and for the dental dataset a Generalised Expectation Maximisation (GEM) algorithm (see [1]) is used instead of the standard EM where, as soon as the value of $Q$ or $\ell^{\text{pen}}$ increases, the Newton-Raphson procedure is stopped. This results in computing only a few steps of the Newton-Raphson algorithm (very often only one step is needed). As the EM algorithm is usually very slow to reach convergence the `turboEM` R package with the `squareEM` option is used to accelerate the procedure (see for instance [3]). Finally, the algorithm must be iterated for the whole sequence of penalties. In order to evaluate the global computational cost, numerical experiments were conducted which showed that, for a maximum of $K_{\max}$ initial cuts, the total complexity of the whole procedure is of order $\mathcal{O}(nK_{\max}^{1/2})$.

More specifically, the computation time for the method was evaluated on replicated samples for the three sample sizes $n = 200, 400, 1\,000$ and for different values of the maximal number of initial cuts: $K_{\max} = 18, 40, 80$. We estimated the implementation of the whole method with 200 penalty values to $0.0016 \times nK_{\max}^{1/2}$ minutes. For example, for $n = 400, K_{\max} = 40$ the whole program takes 4 minutes, for $n = 400, K_{\max} = 80$ it takes 5.7 minutes, for $n = 1\,000, K_{\max} = 40$ it takes 10.12 minutes and for $n = 1\,000, K_{\max} = 80$ it takes 14.3 minutes. These values are given as an indication of the algorithmic complexity and should be considered with caution as the implementation has not been optimised. In particular, computation of the $A_{k,i}^{\text{old}}$ and $B_{k,i}^{\text{old}}$ terms could be improved by computing the set of values $(c_k \wedge R_i, c_{k-1} \vee L_i)$ such that $(L_i, R_i) \cap (c_{k-1}, c_k) \neq \emptyset$ more efficiently in C++. Also the non-penalised MLE is implemented for each selection of cuts. For small penalty values, the set of selected cuts can be quite large and the `turboEM` R package has trouble to converge in these cases. For very large set of selected cuts it often does not converge at all and the algorithm is stopped after 200 iterations. This procedure could be greatly improved by only implementing the MLE for reasonable sets of cuts.

Finally, it should be noted that the adaptive ridge procedure needs only to be implemented once on the dataset, in order to detect the set of cuts. Then given this set of cuts, the piecewise-constant hazard model is much faster to compute. For example in Scenario S1 from the main paper with three cuts, the computation time of the piecewise-constant hazard maximum likelihood model is on average respectively equal to 1.13, 1.80 and 3.33 seconds for $n = 200, 400, 1\,000$.

## A.8 The likelihood ratio approach to construct confidence intervals

As shown in Section 5, statistical inference in our model reduces to a fully parametric problem since, after selection of the cuts, one can consider these cuts as fixed and the asymptotic distribution of the final estimator is identical to the asymptotic distribution one would get if the true cuts were initially provided.

Statistical tests are implemented from the likelihood ratio test which is based on the observed likelihood $\mathrm{L}_n^{\text{obs}}$. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with $\theta_1$ of dimension $d$. To test the null hypothesis $\mathrm{H}_0 : \theta_1 = \theta_0$, with $\theta_0$ known, one can use the test statistic $-2\log(\mathrm{L}_n^{\text{obs}}(\theta_0, \hat{\theta}_2)/\mathrm{L}_n^{\text{obs}}(\hat{\theta}_1, \hat{\theta}_2))$ which follows a chi-squared distribution with $d$ degrees of freedom from standard likelihood theory. Confidence intervals can also be constructed from the likelihood ratio statistic. Let us assume that $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with $\theta_1$ of dimension 1 and consider the test $\mathrm{H}_0 : \theta_1 = \theta_0$ versus $\mathrm{H}_1 : \theta_1 \neq \theta_0$. The $1 - \alpha$ confidence interval level of the parameter $\theta_1$ will be determined by the set of values $\theta_0$ such that the previous test is not significant at the significance level $\alpha$. Note that the p-value of the test

is defined by (with a slight abuse of notation for the realisation of the test statistic)

$$\mathbb{P}\left[\chi^2(1) > -2\log\left(\frac{\mathrm{L}_n^{\mathrm{obs}}(\theta_0,\hat{\theta}_2)}{\mathrm{L}_n^{\mathrm{obs}}(\hat{\theta}_1,\hat{\theta}_2)}\right)\right],$$

and the test is non-significant if this value is greater than $\alpha$. Let $q_{\chi^2}^{1-\alpha}$ be the $1-\alpha$ quantile of the $\chi^2(1)$ distribution. The bounds of the confidence intervals can therefore be determined by resolving the equation

$$\log(\mathrm{L}_n^{\mathrm{obs}}(\theta_0,\hat{\theta}_2)) + \frac{1}{2}q_{\chi^2}^{1-\alpha} - \log(\mathrm{L}_n^{\mathrm{obs}}(\hat{\theta}_1,\hat{\theta}_2)) = 0, \tag{9}$$

with respect to $\theta_0$. This equation has two solutions and since it is clear that $\theta_0 = \hat{\theta}_1$ is part of the confidence interval (the p-value equals one for this value), a grid search can be performed using for example the `uniroot` package with the two starting intervals $[\hat{\theta}_1 - c; \hat{\theta}_1]$ and $[\hat{\theta}_1; \hat{\theta}_1 + c]$, where $c$ is a positive constant. This constant can be chosen arbitrarily large and should satisfy that the left-hand side of Equation (9) is of opposite sign for $\theta_0 = \hat{\theta}_1 - c$ and $\theta_0 = \hat{\theta}_1 + c$. See [4] for more details about the likelihood ratio test approach for constructing confidence intervals.

A more classical method for deriving confidence intervals can be based on the normal approximation of the model parameter obtained from Theorem 5.1. It requires to compute the Hessian matrix of the observed log-likelihood. The details for this approach are given in the next section.

## A.9 Score vector and Hessian matrix for the observed log-likelihood

Computation of the Hessian matrix of the observed log-likelihood $\partial^2 \log(\mathrm{L}_n^{\mathrm{obs}}(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}^2$ evaluated at $\boldsymbol{\theta} = \hat{\theta}$ can be done by direct calculation or by using the following relationship which makes use of the complete likelihood $\mathrm{L}_n$ (see [2]):

$$\frac{\partial \log\left(\mathrm{L}_n^{\mathrm{obs}}(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}} = \mathbb{E}\left[\frac{\partial \log\left(\mathrm{L}_n(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}}\,\middle|\,\text{data}, \boldsymbol{\theta}\right]. \tag{10}$$

In the above equation, the Hessian can be computed based on the complete likelihood by taking the derivative of the right-hand side of the equation with respect to $\boldsymbol{\theta}$. For simplicity, we assume that all individuals are susceptibles. Then,

$$\begin{aligned}
\log\left(\mathrm{L}_n(\boldsymbol{\theta})\right) = &\sum_{i \text{ not exact}} \sum_{k=1}^{K} I(c_{k-1} < T_i \le c_k)\Big(a_{i,k} - \sum_{j=1}^{k} e^{a_{i,j}}(T_i \wedge c_j - c_{j-1})\Big),\\
&+ \sum_{i \text{ exact}} \sum_{k=1}^{K} \big\{O_{i,k}a_{i,k} - \exp(a_{i,k})R_{i,k}\big\}\\
\frac{\partial \log\left(\mathrm{L}_n(\boldsymbol{\theta})\right)}{\partial a_k} = &\sum_{i \text{ not exact}}^{n} \Big\{I(c_{k-1} < T_i \le c_k) - \sum_{l=k}^{K} I(c_{l-1} < T_i \le c_l)e^{a_{i,k}}(T_i \wedge c_k - c_{k-1})\Big\},\\
&+ \sum_{i \text{ exact}} \big\{O_{i,k} - \exp(a_{i,k})R_{i,k}\big\}\\
\frac{\partial \log\left(\mathrm{L}_n(\boldsymbol{\theta})\right)}{\partial \beta} = &\sum_{i=1}^{n}\sum_{l=1}^{K} I(c_{l-1} < T_i \le c_l)Z_i\Big(1 - \sum_{j=1}^{l} e^{a_{i,j}}(T_i \wedge c_j - c_{j-1})\Big)\\
&+ \sum_{i \text{ exact}} \sum_{l=1}^{K} Z_i\big\{O_{i,l} - \exp(a_{i,l})R_{i,l}\big\}.
\end{aligned}$$

We now need to take the expectation conditionally on the data of the last two equations. This will involve the quantities

$$\mathbb{P}[c_{k-1} < T_i \leq c_k \mid \text{data}, \boldsymbol{\theta}] = \frac{S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta}) - S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta})}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})},$$

and

$$\begin{aligned}
&\mathbb{E}[I(c_{k-1} < T_i \leq c_k) T_i \mid \text{data}, \boldsymbol{\theta})] \\
&= J_{k,i} \int_{c_{k-1} \vee L_i}^{c_k \wedge R_i} t \exp\Big(a_{i,k} - \sum_{j=1}^{k} e^{a_{i,j}} (t \wedge c_j - c_{j-1})\Big) dt \times \frac{1}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})},
\end{aligned}$$

$$\begin{aligned}
&= \big\{\big(\exp(-a_{i,k}) + c_{k-1} \vee L_i\big) \exp(-e^{a_{i,k}} c_{k-1} \vee L_i) - \big(\exp(-a_{i,k}) + c_k \wedge R_i\big) \exp(-e^{a_{i,k}} c_k \wedge R_i)\big\} \\
&\quad \times \frac{\exp\big(e^{a_{i,k}} c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}} (c_j - c_{j-1})\big) J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}.
\end{aligned}$$

Calculation of the right-hand side of Equation (10) is now straightforward. We first separate exact and non exact observations in the following way:

$$\frac{\partial \log(\mathrm{L}_n^{\text{obs}}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \sum_{i \text{ not exact}} \frac{\partial \mathrm{L}_{i,1}^{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \sum_{i \text{ exact}} \frac{\partial \mathrm{L}_{i,2}^{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

For the non-exact observations, we introduce

$$\begin{aligned}
C_{i,k}(\boldsymbol{\theta}) &= \frac{S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta}) - S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta})}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}, \\
D_{i,k}(\boldsymbol{\theta}) &= J_{k,i} \big\{\big(\exp(-a_{i,k}) + c_{k-1} \vee L_i\big) \exp(-e^{a_{i,k}} c_{k-1} \vee L_i) \\
&\qquad - \big(\exp(-a_{i,k}) + c_k \wedge R_i\big) \exp(-e^{a_{i,k}} c_k \wedge R_i)\big\} \frac{\exp\big(e^{a_{i,k}} c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}} (c_j - c_{j-1})\big)}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})},
\end{aligned}$$

such that

$$\begin{aligned}
\frac{\partial \mathrm{L}_{i,1}^{\text{obs}}(\boldsymbol{\theta})}{\partial a_k} &= C_{i,k}(\boldsymbol{\theta}) - e^{a_{i,k}} \Big(D_{i,k}(\boldsymbol{\theta}) - c_{k-1} C_{i,k}(\boldsymbol{\theta})\Big) - e^{a_{i,k}} (c_k - c_{k-1}) \sum_{l=k+1}^{K} C_{i,l}(\boldsymbol{\theta}), \\
\frac{\partial \mathrm{L}_{i,1}^{\text{obs}}(\boldsymbol{\theta})}{\partial \beta} &= Z_i \Big\{ C_{i,k}(\boldsymbol{\theta}) - C_{i,k}(\boldsymbol{\theta}) \sum_{j=1}^{k-1} e^{a_{i,j}} (c_j - c_{j-1}) - e^{a_{i,k}} \Big(D_{i,k}(\boldsymbol{\theta}) - c_{k-1} C_{i,k}(\boldsymbol{\theta})\Big) \Big\}.
\end{aligned}$$

For the exact observations we have

$$\begin{aligned}
\frac{\partial \mathrm{L}_{i,2}^{\text{obs}}(\boldsymbol{\theta})}{\partial a_k} &= O_{i,k} - \exp(a_k + \beta Z_i) R_{i,k}, \\
\frac{\partial \mathrm{L}_{i,2}^{\text{obs}}(\boldsymbol{\theta})}{\partial \beta} &= Z_i \sum_{l=1}^{K} \Big\{ O_{i,l} - \exp(a_l + \beta Z_i) R_{i,l} \Big\}.
\end{aligned}$$

For the Hessian matrix $\partial^2 \log(\mathrm{L}_n^{\text{obs}}(\boldsymbol{\theta}))/\partial \boldsymbol{\theta}^2$, we first compute

$$
\begin{aligned}
\frac{\partial S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta})}{\partial a_k} &= -\left(L_i I_k(L_i) + c_k I(L_i > c_k)\right) e^{a_{i,k}} S(L_i \mid Z_i, \boldsymbol{\theta}), \\
\frac{\partial S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta})}{\partial \beta} &= -Z_i \sum_{l=1}^{K} (c_l \wedge c_{k-1} \vee L_i - c_{l-1}) I(c_{l-1} \leq c_{k-1} \vee L_i) e^{a_{i,k}} S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta}), \\
\frac{\partial S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta})}{\partial a_k} &= -(c_k \wedge R_i - c_{k-1}) e^{a_{i,k}} S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta}) I(R_i \geq c_{k-1}), \\
\frac{\partial S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta})}{\partial \beta} &= -Z_i \sum_{l=1}^{K} (c_l \wedge c_k \wedge R_i - c_{l-1}) I(c_{l-1} \leq c_k \wedge R_i) e^{a_{i,k}} S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta}), \\
\frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})}{\partial a_k} &= -(c_k \wedge L_i - c_{k-1}) e^{a_{i,k}} S(L_i \mid Z_i, \boldsymbol{\theta}) I(L_i \geq c_{k-1}), \\
\frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})}{\partial \beta} &= -Z_i \sum_{l=1}^{K} (c_l \wedge L_i - c_{l-1}) e^{a_{i,l}} S(L_i \mid Z_i, \boldsymbol{\theta}) I(L_i \geq c_{l-1}), \\
\frac{\partial S(R_i \mid Z_i, \boldsymbol{\theta})}{\partial a_k} &= -(c_k \wedge R_i - c_{k-1}) e^{a_{i,k}} S(R_i \mid Z_i, \boldsymbol{\theta}) I(R_i \geq c_{k-1}), \\
\frac{\partial S(R_i \mid Z_i, \boldsymbol{\theta})}{\partial \beta} &= -Z_i \sum_{l=1}^{K} (c_l \wedge R_i - c_{l-1}) e^{a_{i,l}} S(R_i \mid Z_i, \boldsymbol{\theta}) I(R_i \geq c_{l-1}),
\end{aligned}
$$

such that calculation of the partial derivatives of $C_{i,k}(\boldsymbol{\theta})$ are calculated from the formulas

$$
\begin{aligned}
\frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial a_k} &= \frac{\partial S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta})/\partial a_k - \partial S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta})/\partial a_k}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})} \\
&\quad - C_{i,k}(\boldsymbol{\theta}) \frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})/\partial a_k - \partial S(R_i \mid Z_i, \boldsymbol{\theta})/\partial a_k}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}, \\
\frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial \beta} &= \frac{\partial S(c_{k-1} \vee L_i \mid Z_i, \boldsymbol{\theta})/\partial \beta - \partial S(c_k \wedge R_i \mid Z_i, \boldsymbol{\theta})/\partial \beta}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})} \\
&\quad - C_{i,k}(\boldsymbol{\theta}) \frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})/\partial \beta - \partial S(R_i \mid Z_i, \boldsymbol{\theta})/\partial \beta}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}.
\end{aligned}
$$

Then, we can show that

$$
\begin{aligned}
\frac{\partial}{\partial a_k} \sum_{l=k+1}^{K} C_{i,l}(\boldsymbol{\theta}) &= \frac{(c_k \vee L_i - c_{k-1}) e^{a_{i,k}} \sum_{l=k}^{K} S(c_l \vee L_i \mid Z_i, \boldsymbol{\theta})}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})} \\
&\quad - \frac{(c_k \wedge R_i - c_{k-1}) e^{a_{i,k}} I(R_i \geq c_{k-1}) \sum_{l=k+1}^{K} S(c_l \vee R_i \mid Z_i, \boldsymbol{\theta})}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})} \\
&\quad - \sum_{l=k+1}^{K} C_{i,l}(\boldsymbol{\theta}) \frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})/\partial a_k - \partial S(R_i \mid Z_i, \boldsymbol{\theta})/\partial a_k}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}.
\end{aligned}
$$

We now introduce:

$$
\begin{aligned}
E_{i,k} = \exp(-a_{i,k} - e^{a_{i,k}} c_{k-1} \vee L_i) &+ \big(\exp(-a_{i,k}) + c_{k-1} \vee L_i\big)\big(\exp(a_{i,k} - e^{a_{i,k}} c_{k-1} \vee L_i) c_{k-1} \vee L_i\big) \\
&+ \exp(-a_{i,k} - e^{a_{i,k}} c_{k-1} \vee L_i) + \big(\exp(-a_{i,k}) + c_k \wedge R_i\big)\big(\exp(a_{i,k} - e^{a_{i,k}} c_k \wedge R_i) c_k \vee R_i\big),
\end{aligned}
$$

such that

$$\begin{aligned}
\frac{\partial D_{i,k}(\boldsymbol{\theta})}{\partial a_k} &= -\frac{E_{i,k}\exp\left(e^{a_{i,k}}c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}}(c_j - c_{j-1})\right)J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})} + D_{i,k}(\boldsymbol{\theta})e^{a_{i,k}}c_{k-1}J_{k,i} \\
&\quad - D_{i,k}(\boldsymbol{\theta})\frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})/\partial a_k - \partial S(R_i \mid Z_i, \boldsymbol{\theta})/\partial a_k}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}J_{k,i}, \\
\frac{\partial D_{i,k}(\boldsymbol{\theta})}{\partial \beta} &= -Z_i\frac{E_{i,k}\exp\left(e^{a_{i,k}}c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}}(c_j - c_{j-1})\right)J_{k,i}}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})} \\
&\quad + Z_i D_{i,k}(\boldsymbol{\theta})(e^{a_{i,k}}c_{k-1} - \sum_{j=1}^{k-1} e^{a_{i,j}}(c_j - c_{j-1}))J_{k,i} \\
&\quad - D_{i,k}(\boldsymbol{\theta})J_{k,i}\frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})/\partial \beta - \partial S(R_i \mid Z_i, \boldsymbol{\theta})/\partial \beta}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\frac{\partial^2 \mathrm{L}_1^{\mathrm{obs}}(\boldsymbol{\theta})}{\partial a_k^2} &= \frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial a_k} - e^{a_{i,k}}\Big(D_{i,k}(\boldsymbol{\theta}) - c_{k-1}C_{i,k}(\boldsymbol{\theta}) + \frac{\partial D_{i,k}(\boldsymbol{\theta})}{\partial a_k} - c_{k-1}\frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial a_k}\Big) \\
&\quad - e^{a_{i,k}}(c_k - c_{k-1})\left(\sum_{l=k+1}^{K} C_{i,l}(\boldsymbol{\theta}) + \frac{\partial}{\partial a_k}\sum_{l=k+1}^{K} C_{i,l}(\boldsymbol{\theta})\right), \\
\frac{\partial^2 \mathrm{L}_1^{\mathrm{obs}}(\boldsymbol{\theta})}{\partial a_k \partial \beta} &= Z_i\Big\{\frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial a_k} - \frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial a_k}\sum_{j=1}^{k-1} e^{a_{i,j}}(c_j - c_{j-1}) \\
&\quad - e^{a_{i,k}}\Big(D_{i,k}(\boldsymbol{\theta}) - c_{k-1}C_{i,k}(\boldsymbol{\theta}) + \frac{\partial D_{i,k}(\boldsymbol{\theta})}{\partial a_k} - c_{k-1}\frac{\partial C_{i,k}(\boldsymbol{\theta})}{\partial a_k}\Big)\Big\}, \\
\frac{\partial^2 \mathrm{L}_1^{\mathrm{obs}}(\boldsymbol{\theta})}{\partial \beta^2} &= Z_i\Big\{\frac{\partial C_{i,k}(\boldsymbol{\theta})^t}{\partial \beta} - \frac{\partial C_{i,k}(\boldsymbol{\theta})^t}{\partial \beta}\sum_{j=1}^{k-1} e^{a_{i,j}}(c_j - c_{j-1}) \\
&\quad - e^{a_{i,k}}\Big(Z_i^t D_{i,k}(\boldsymbol{\theta}) - c_{k-1}Z_i^t C_{i,k}(\boldsymbol{\theta}) + \frac{\partial D_{i,k}(\boldsymbol{\theta})^t}{\partial \beta} - c_{k-1}\frac{\partial C_{i,k}(\boldsymbol{\theta})^t}{\partial \beta}\Big)\Big\},
\end{aligned}$$

and for the exact observations

$$\begin{aligned}
\frac{\partial^2 \mathrm{L}_2^{\mathrm{obs}}(\boldsymbol{\theta})}{\partial a_k^2} &= -\exp(a_k + \beta Z_i)R_{i,k}, \\
\frac{\partial^2 \mathrm{L}_2^{\mathrm{obs}}(\boldsymbol{\theta})}{\partial a_k \partial \beta} &= -Z_i\exp(a_k + \beta Z_i)R_{i,k}, \\
\frac{\partial^2 \mathrm{L}_2^{\mathrm{obs}}(\boldsymbol{\theta})}{\partial \beta^2} &= -Z_i Z_i^t \sum_{l=1}^{K}\Big\{\exp(a_l + \beta Z_i)R_{i,l}\Big\}.
\end{aligned}$$

# References


[1] Dempster, Arthur P, Laird, Nan M and Rubin, Donald B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

[2] Louis, Thomas A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

[3] Varadhan, Ravi and Roland, Christophe. (2008). Simple and globally convergent

methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics* **35**(2), 335–353.

[4] Zhou, Mai. (2015). *Empirical likelihood method in survival analysis*. Chapman and Hall/CRC.

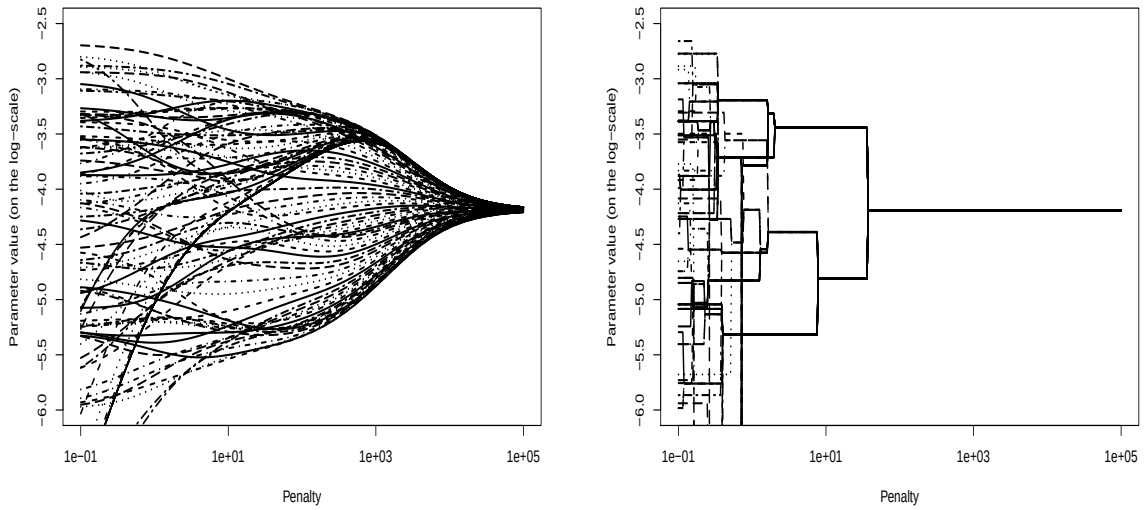


Figure 2: Regularization path for the ridge on the left panel and for the adaptive ridge on the right panel. The $x$-axis represents the penalty value and the $y$-axis represents the estimated values of the $a_k$'s.

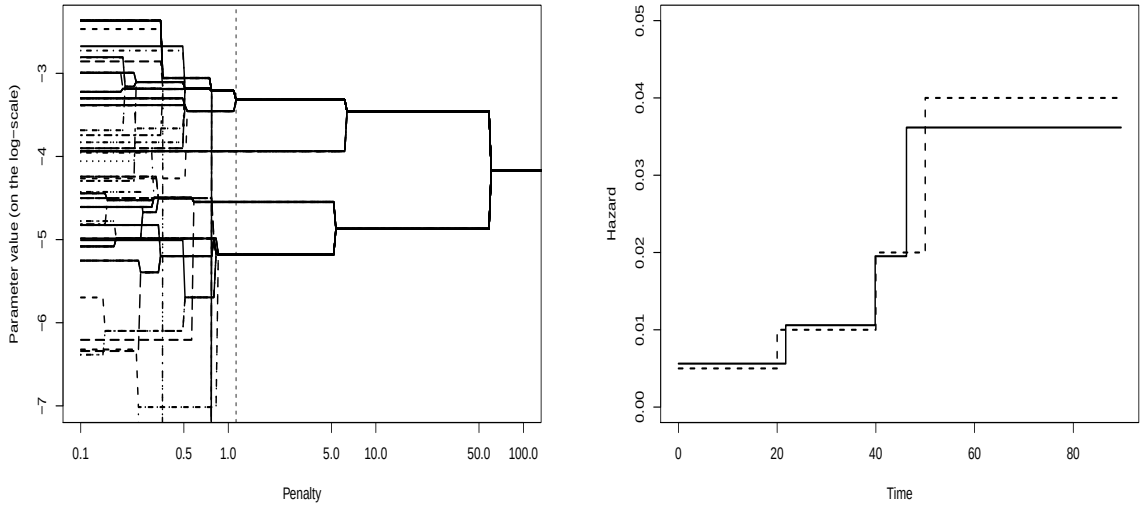


Figure 3: Regularization path for the adaptive-ridge on the left panel. The estimated set of cuts using the BIC is shown as a vertical dotted line. The resulting piecewise constant hazard estimator is shown on the right panel as a solid line. The dotted line represents the true hazard.

Table 9: Simulation results for the estimation of $\beta$ and $S_0$ in Scenarios S3 and S4. S3: 80% of susceptible individuals. S4: 58% of susceptible individuals. Among the susceptible individuals, 18% of exact data, 19% of left-censoring, 40% of interval-censoring, 23% of right-censoring.

| | | Adaptive Ridge estimate | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | Bias($\hat{\beta}$) | SE($\hat{\beta}$) | MSE($\hat{\beta}$) | Bias($\hat{\gamma}$) | SE($\hat{\gamma}$) | MSE($\hat{\gamma}$) | IBias$^2$($\hat{S}_0$) | IVar($\hat{S}_0$) | TV($\hat{\lambda}_0$) |
| S3 | 200 | −0.015 | 0.291 | 0.085 | 0.102 | 0.498 | 0.259 | 0.004 | 0.324 | 0.840 |
| | | 0.003 | 0.236 | 0.056 | 0.011 | 0.630 | 0.398 | | | |
| | 400 | −0.017 | 0.207 | 0.043 | 0.075 | 0.356 | 0.132 | 0.002 | 0.160 | 0.659 |
| | | −0.005 | 0.162 | 0.026 | 0.027 | 0.433 | 0.189 | | | |
| | 1 000 | 0.006 | 0.127 | 0.016 | 0.025 | 0.184 | 0.035 | 0.001 | 0.059 | 0.414 |
| | | 0.006 | 0.094 | 0.009 | 0.012 | 0.198 | 0.039 | | | |
| S4 | 200 | −0.021 | 0.387 | 0.150 | 0.077 | 0.479 | 0.235 | 0.005 | 0.563 | 1.195 |
| | | −0.010 | 0.310 | 0.096 | 0.038 | 0.511 | 0.262 | | | |
| | 400 | −0.023 | 0.255 | 0.066 | 0.048 | 0.296 | 0.090 | 0.003 | 0.255 | 0.810 |
| | | 0.003 | 0.209 | 0.044 | 0.016 | 0.309 | 0.096 | | | |
| | 1 000 | −0.009 | 0.150 | 0.023 | 0.032 | 0.186 | 0.036 | 0.001 | 0.096 | 0.530 |
| | | 0.008 | 0.124 | 0.015 | 0.004 | 0.205 | 0.042 | | | |