Supplementary Material for: Regression modelling of interval censored data based on the adaptive ridge procedure

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

[1]MAP5 (UMR CNRS 8145), Université de Paris
[2]Ressource Center for Rare Oral Diseases, Copenhagen University Hospital, Rigshospitalet, Denmark
[3]LPSM, CNRS 7599, 4 place Jussieu, Paris, France

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

For $k = 1, \ldots, K$, $i = 1, \ldots, n$, 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).

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 ℓ^{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_ks which is either diagonal (for Q) or tri-diagonal (for ℓ^{pen}).

Let $\mathcal{I}(a, \beta)$ be minus the Hessian matrix of Q or ℓ^{pen} for the maximisation problem with respect to $a_1, \ldots, a_L$ and $\beta_1, \ldots, \beta_{d_Z}$. 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 ℓ^{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 ℓ^{pen}, the $K \times K$ matrix A is either diagonal (for Q) or a band matrix of bandwidth equal to 1 (for ℓ^{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 $K >> d_Z$.

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:

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

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

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 1 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 2 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.

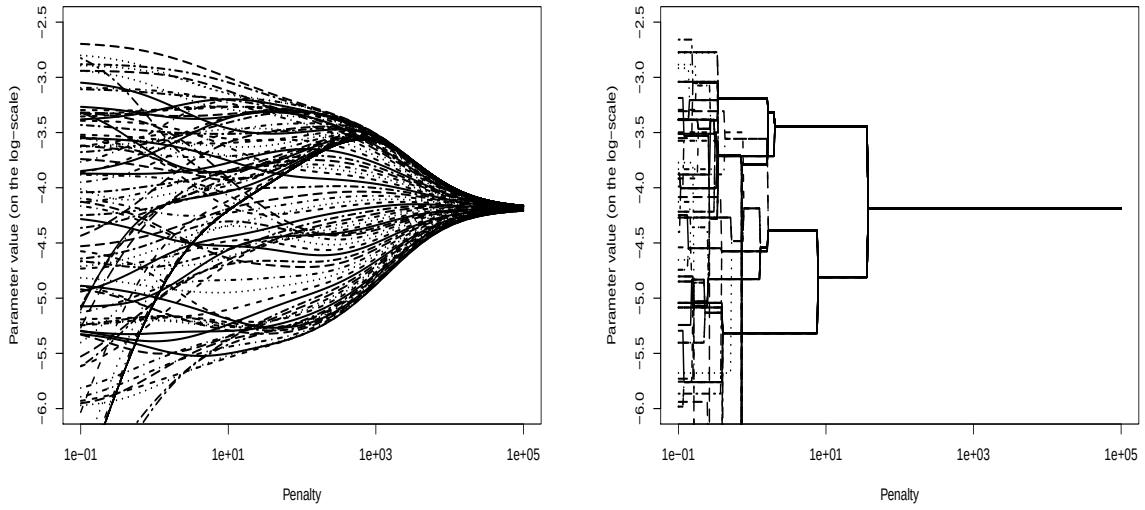

Figure 1: Regularization path for the ridge on the left panel and for the adaptive ridge on the right panel. The x-axis represents the penalty value and the y-axis represents the estimated values of the a_k's.

5 Proof of Theorem 4.1 of the main document

The proof follows two steps.

1. First of all, we prove that, at a given step of the adaptive-ridge algorithm, maximising Equation (4) using the EM algorithm is equivalent to maximising

$$\ell_*^{\text{pen}}(\boldsymbol{\theta}) = \log(L_n^{\text{obs}}(\boldsymbol{\theta})) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k^{(l-1)} (a_{k+1} - a_k)^2.$$

   Let $\boldsymbol{\theta}_{\text{old}}$ be the current parameter value of the EM algorithm. In the following we use the notation $f_{L,R,Z,T}(\text{data}_{1:n}, T_{1:n}, \boldsymbol{\theta})$ to represent the joint density of $(L_1, \ldots, L_n, R_1, \ldots, R_n,$

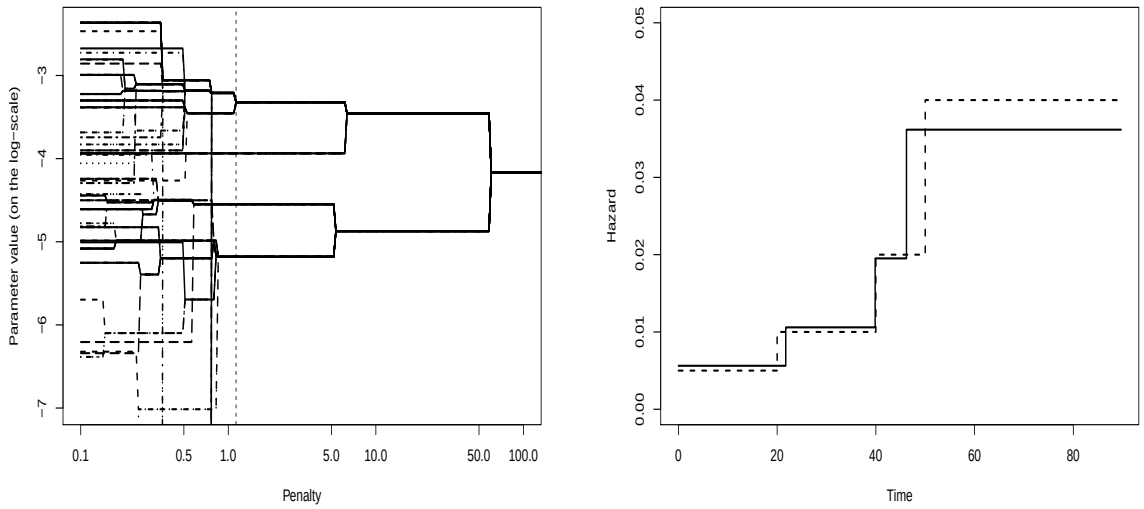

Figure 2: Regularization path for the adaptive-ridge on the left panel. The estimated set of cuts using the BIC is shown as a vertical dotted line. The resulting piecewise constant hazard estimator is shown on the right panel as a solid line. The dotted line represents the true hazard.

$Z_1, \ldots, Z_n, T_1, \ldots, T_n)$ evaluated at the same observations with parameter $\boldsymbol{\theta}$. We have:

$$\begin{aligned}
\log(L_n^{\text{obs}}(\boldsymbol{\theta})) &= \log\left(\int f_{L,R,Z,T}(\text{data}_{1:n}, T_{1:n}, \boldsymbol{\theta}) dT_{1:n}\right) \\
&= \log\left(\int f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}}) \frac{f_{L,R,Z|T}(\text{data}_{1:n} \mid T_{1:n}, \boldsymbol{\theta}) f_T(T_{1:n}, \boldsymbol{\theta})}{f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}})} dT_{1:n}\right) \\
&\geq \int f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}}) \log\left(\frac{f_{L,R,Z|T}(\text{data}_{1:n} \mid T_{1:n}, \boldsymbol{\theta}) f_T(T_{1:n}, \boldsymbol{\theta})}{f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}})} dT_{1:n}\right),
\end{aligned}$$

where the last inequality was obtained from Jensen inequality and the fact that $\int f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}}) dT_{1:n} = 1$. Then, define

$$\begin{aligned}
&\ell_1(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) \\
&= \int f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}}) \log\left(\frac{f_{L,R,Z|T}(\text{data}_{1:n} \mid T_{1:n}, \boldsymbol{\theta}) f_T(T_{1:n}, \boldsymbol{\theta})}{f_{T|L,R,Z}(T_{1:n} \mid \text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}}) f_{L,R,Z}(\text{data}_{1:n}, \boldsymbol{\theta}_{\text{old}})} dT_{1:n}\right),
\end{aligned}$$

we have

$$\ell_*^{\text{pen}}(\boldsymbol{\theta}) - \ell_*^{\text{pen}}(\boldsymbol{\theta}_{\text{old}}) \geq \ell_1(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k^{(l-1)} \left((a_{k+1} - a_k)^2 - (a_{k+1}^{\text{old}} - a_k^{\text{old}})^2\right).$$

By defining this time

$$\ell_2(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \ell_*^{\text{pen}}(\boldsymbol{\theta}_{\text{old}}) + \ell_1(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} \hat{w}_k^{(l-1)} \left((a_{k+1} - a_k)^2 - (a_{k+1}^{\text{old}} - a_k^{\text{old}})^2\right),$$

we directly see that $\ell_*^{\text{pen}}(\boldsymbol{\theta}) \geq \ell_2(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})$ and $\ell_2(\boldsymbol{\theta}_{\text{old}} \mid \boldsymbol{\theta}_{\text{old}}) = \ell_*^{\text{pen}}(\boldsymbol{\theta}_{\text{old}})$. Finally, we notice that maximising $\ell_2(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}})$ is equivalent to maximising Equation (4). We note $\hat{\boldsymbol{\theta}}$ such a maximiser:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_2(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}) = \arg\max_{\boldsymbol{\theta}} \ell^{\text{pen}}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{old}}),$$

and $\ell_*^{\text{pen}}(\hat{\boldsymbol{\theta}}) \geq \ell_2(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}_{\text{old}}) \geq \ell_2(\boldsymbol{\theta}_{\text{old}} \mid \boldsymbol{\theta}_{\text{old}}) = \ell_*^{\text{pen}}(\boldsymbol{\theta}_{\text{old}})$.

2. The second step consists in showing that maximising ℓ_*^{pen} with the iterative adaptive ridge penalisation is equivalent to maximising $\log(\mathrm{L}_n^{\text{obs}}(\boldsymbol{\theta})) - \kappa \sum_{k=1}^{K-1} p(\Delta a_k)$, for some $\kappa > 0$. To that purpose, we use a Local Quadratic Approximation (see Fan and Li (2001) and Hunter and Li (2005)) of $p(\beta)$. For all $\beta^{(l)} \in \mathbb{R}$, for all $\beta \in \mathbb{R}$, one can easily show that $p(\beta) \leq q(\beta \mid \beta^{(l)})$ with

$$q(\beta \mid \beta^{(l)}) = \frac{\log(1 + (\beta^{(l)})^2/\varepsilon^2)}{\log(1 + 1/\varepsilon^2)} + \frac{\beta^2 - (\beta^{(l)})^2}{\varepsilon^2 + (\beta^{(l)})^2} \cdot \frac{1}{\log(1 + 1/\varepsilon^2)}.$$

It is also directly seen that $q(\beta^{(l)} \mid \beta^{(l)}) = p(\beta^{(l)})$. Now, define

$$g(\boldsymbol{\theta} \mid \boldsymbol{\theta^{(l)}}) = \log(L_n^{\text{obs}}(\boldsymbol{\theta})) - \kappa \sum_{k=1}^{K-1} q(\Delta a_k \mid \Delta a_k^{(l)}).$$

Notice that $g(\hat{\boldsymbol{\theta}}^{(l)} \mid \hat{\boldsymbol{\theta}}^{(l)}) = \log(\mathrm{L}_n^{\text{obs}}(\hat{\boldsymbol{\theta}}^{(l)})) - \kappa \sum_{k=1}^{K-1} p(\Delta \hat{a}_k^{(l)})$ and let $\hat{\boldsymbol{\theta}}^{(l+1)} = \arg\max_{\boldsymbol{\theta}} g(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(l)})$. Then:

$$\begin{aligned}\log(\mathrm{L}_n^{\text{obs}}(\hat{\boldsymbol{\theta}}^{(l+1)})) - \kappa \sum_{k=1}^{K-1} p(\Delta \hat{a}_k^{(l+1)}) &\geq g(\hat{\boldsymbol{\theta}}^{(l+1)} \mid \hat{\boldsymbol{\theta}}^{(l)}) \geq g(\hat{\boldsymbol{\theta}}^{(l)} \mid \hat{\boldsymbol{\theta}}^{(l)}) \\ &\geq \log(\mathrm{L}_n^{\text{obs}}(\hat{\boldsymbol{\theta}}^{(l)})) - \kappa \sum_{k=1}^{K-1} p(\Delta \hat{a}_k^{(l)}).\end{aligned}$$

We conclude the proof by noticing that maximising $g(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(l)})$ is equivalent to maximising $\ell_*^{\text{pen}}(\boldsymbol{\theta})$ with pen $= 2\kappa/\log(1 + 1/\varepsilon^2)$ and $\hat{w}_k^{(l-1)} = \left((\hat{a}_{k+1}^{(l-1)} - \hat{a}_k^{(l-1)})^2 + \varepsilon^2\right)^{-1}$, which is the adaptive-ridge algorithm.

6 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

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

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, δ) 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

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

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 $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_n^{\mathrm{pen}}(\boldsymbol{\theta})$, where

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

with $\ell_n(\boldsymbol{\theta}) = \log(\tilde{\mathrm{L}}_n^{\mathrm{obs}}(\boldsymbol{\theta}))/n$. 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:

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

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 $\mathrm{pen}/n \to 0$.

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 $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_n^{\mathrm{pen}}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^\dagger = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$. The concavity of $\ell_n^{\mathrm{pen}}(\boldsymbol{\theta})$ yields that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^\dagger$ in probability.

Proofs 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 $\ell_{n,\mathcal{A}^*}(\boldsymbol{\theta}) = \log(\mathrm{L}^{\mathrm{obs}}_{n,\mathcal{A}^*}(\boldsymbol{\theta}))$ 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,

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

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

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

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 $\{c_1, \ldots, c_K\}$ and we define $\hat{\boldsymbol{\theta}}^\dagger$, the estimator $\hat{\boldsymbol{\theta}}_{\mathcal{A}^*}$ that is defined on $\{c_1, \ldots, c_K\}$ (this is always possible since $\mathcal{A}^* \subset \{c_1, \ldots, c_K\}$). 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:

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

where $(u, v) = (u_1, \ldots, u_K, v_1, \ldots, v_{d_Z})$ 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 ψ_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 ℓ_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,

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

Since $\hat{w}^{(1)}_k \to ((a^\dagger_{k+1} - a^\dagger_k)^2 + \varepsilon^2)^{-1}$, $\hat{a}^\dagger_{k+1} - \hat{a}^\dagger_k \to a^\dagger_{k+1} - a^\dagger_k$ in probability and

$$\left| \frac{a^\dagger_{k+1} - a^\dagger_k}{(a^\dagger_{k+1} - a^\dagger_k)^2 + \varepsilon^2} \right| < 1,$$

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 Σ 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 (1) and (2), 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.

7 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, β coefficient and λ_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 1. Only our adaptive ridge estimator has been implemented for these two scenarios. The γ estimator is initialised to 0 in the EM algorithm.

Table 1: Simulation results for the estimation of β 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			

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 γ parameter is less accurately estimated as compared

to the β 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 γ 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.

8 Computational cost of the adaptive ridge algorithm

The complexity for the inversion of the Hessian of ℓ is of order $\mathcal{O}(K)$, in the case $K >> d_X + d_Z$ (see Section 2 in the Supplementary Material about the Schurr complement). However, for a given penalty, it should be noted that the global algorithm for maximising Q or ℓ^{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 Dempster *and others* (1977)) is used instead of the standard EM where, as soon as the value of Q or ℓ^{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 Varadhan and Roland (2008)). 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 $n = 200, 400, 1\,000$.

9 The likelihood ratio approach to construct confidence intervals

As shown in Section 5 of the main document, 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 $\mathrm{L}_n^{\mathrm{obs}}$. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ with θ_1 of dimension d. To test the null hypothesis $\mathrm{H}_0 : \theta_1 = \theta_0$, with θ_0 known, one can use the test statistic $-2\log(\mathrm{L}_n^{\mathrm{obs}}(\theta_0, \hat{\theta}_2)/\mathrm{L}_n^{\mathrm{obs}}(\hat{\theta}_1, \hat{\theta}_2))$ 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 θ_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 θ_1 will be determined by the set of values θ_0 such that the previous test is not significant at the significance level α. 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 α. 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

$$\log(\mathrm{L}_n^{\mathrm{obs}}(\theta_0, \hat{\theta}_2)) + \frac{1}{2}q_{\chi^2}^{1-\alpha} - \log(\mathrm{L}_n^{\mathrm{obs}}(\hat{\theta}_1, \hat{\theta}_2)) = 0, \tag{3}$$

with respect to θ_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 (3) is of opposite sign for $\theta_0 = \hat{\theta}_1 - c$ and $\theta_0 = \hat{\theta}_1 + c$. See Zhou (2015) 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 of the main document. It requires to compute the Hessian matrix of the observed log-likelihood. The details for this approach are given in the next section.

10 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 L_n (see Louis (1982)):

$$\frac{\partial \log\left(\mathrm{L}_n^{\mathrm{obs}}(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}} = \mathbb{E}\left[\frac{\partial \log\left(\mathrm{L}_n(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}} \,\middle|\, \text{data}, \boldsymbol{\theta}\right]. \tag{4}$$

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,

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

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

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

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

Calculation of the right-hand side of Equation (4) 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

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

such that

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

For the exact observations we have

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

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

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

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

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

Then, we can show that

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

$$- \sum_{l=k+1}^{K} C_{i,l}(\boldsymbol{\theta})\frac{\partial S(L_i \mid Z_i, \boldsymbol{\theta})/\partial a_k - \partial S(R_i \mid Z_i, \boldsymbol{\theta})/\partial a_k}{S(L_i \mid Z_i, \boldsymbol{\theta}) - S(R_i \mid Z_i, \boldsymbol{\theta})}.$$

We now introduce:

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

such that

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

Finally, we have

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

and for the exact observations

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

References

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.

Fan, Jianqing and Li, Runze. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**(456), 1348–1360.

Hunter, David R and Li, Runze. (2005). Variable selection using mm algorithms. *Annals of statistics* **33**(4), 1617.

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.

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.

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