Université Paris Descartes

Document de synthèse

présenté par

Olivier Bouaziz

En vue de l'obtention de

l'Habilitation à Diriger des Recherches
de l'Université Paris Descartes
Spécialité : MATHÉMATIQUES APPLIQUÉES

Contributions théoriques et appliquées à l'analyse de survie

Soutenue le 29 novembre 2018

devant le jury :

Jean-Francois Dupuy	Professeur (INSA de Rennes)
Agathe Guilloux	Professeur (Université d'Evry Val d'Essonne)
Pierre Joly	Maître de conférence (Université de Bordeaux), *rapporteur*
Aurélien Latouche	Professeur (CNAM)
Jean-Christophe Thalabard	PUPH (Université Paris Descartes)
Pascale Tubert-Bitter	Directrice de recherche (INSERM, UVSQ), *rapporteur*

Contents

CHAPTER 1

Publications and submitted papers

Publications

[P1] Olivier Bouaziz and Olivier Lopez. Conditional density estimation in a censored single-index regression model. *Bernoulli*, 16(2):514–542, 2010.

[P2] Olivier Bouaziz, Ségolen Geffray, and Olivier Lopez. Semiparametric inference for the recurrent events process by means of a single-index model. *Statistics*, 49(2):361–385, 2015.

[P3] Olivier Bouaziz, Fabienne Comte, and Agathe Guilloux. Nonparametric estimation of the intensity function of a recurrent event process. *Statistica Sinica*, 23(2):635–665, 2013.

[P4] Olivier Bouaziz and Agathe Guilloux. A penalized algorithm for event-specific rate models for recurrent events. *Biostatistics*, 16(2):281–294, 2014.

[P5] Olivier Bouaziz and Grégory Nuel. A change-point model for detecting heterogeneity in ordered survival responses. *Statistical methods in medical research*, 27(12):3595–3611, 2017.

[P6] Olivier Bouaziz and Grégory Nuel. L0 regularization for the estimation of piecewise constant hazard rates in survival analysis. *Applied Mathematics*, 8(3), 2017.

[P7] Grégory Nuel, Alexandra Lefebvre, and Olivier Bouaziz. Computing individual risks based on family history in genetic disease in the presence of competing risks. *Computational and mathematical methods in medicine*, 2017.

[P8] Tine D Clausen, Thomas Bergholt, Olivier Bouaziz, Magnus Arpi, Frank Eriksson, Steen Rasmussen, Niels Keiding, and Ellen C Løkkegaard. Broad-spectrum antibiotic treatment and subsequent childhood type 1 diabetes: a nationwide danish cohort study. *PloS one*, 11, 2016.

[P9] Olivier Bouaziz, David Courtin, Gilles Cottrell, Jacqueline Milet, Grégory Nuel, and André Garcia. Is placental malaria a long-term risk factor for mild malaria attack in infancy? Revisiting a paradigm. *Clinical Infectious Diseases*, 66(6):930–935, 2017.

Submitted papers

[S1] Olivier Bouaziz, Elodie Brunel, and Fabienne Comte. Nonparametric survival function estimation for data subject to interval censoring case 2.

[S2] Vivien Goepp, Jean-Christophe Thalabard, Grégory Nuel, and Olivier Bouaziz. Regularized bidimensional estimation of the hazard rate.

[S3] Olivier Bouaziz, Eva Fejerskov Lauridsen, and Grégory Nuel. Regression modelling of interval censored data based on the adaptive ridge procedure.

[S4] Jakob Schroder, Olivier Bouaziz, Ross Agner Agner, Torben Martinussen, Per Lav Madsen, Dana Li, Fa Nedaei, and Ulrik Dixen. Atrial fibrillation predicts atrial fibrillation - a hypothesis revisited in a clinical setting.

CHAPTER 2

Introduction

2.1 Time to event analysis: a brief review

2.1.1 The general framework

In survival analysis the event of interest is denoted T^* and the observations are:

$$\begin{cases} T = T^* \wedge C \\ \Delta = I(T^* \leq C), \end{cases}$$

where C is a censoring variable assumed to be independent of T^* (independent censoring assumption) and $I(\cdot)$ represents the indicator function. With such types of data a key function of interest is the hazard rate defined as:

$$\lambda(t) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T^* < t + \triangle t | T^* \geq t]}{\triangle t}.$$

Then, it can be easily shown that under independent censoring, we have (see [ABGK93] for instance):

$$\lambda(t) = \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T < t + \triangle t, \Delta = 1 | T \geq t]}{\triangle t}.$$

This last equation implies that the hazard rate can be estimated using only the observed data. Many quantities of interest are derived from this relation, such as the well known Nelson-Aalen estimator (see [Aal75] or [Nel72]) of the cumulative hazard function $\Lambda(t) := \int_0^t \lambda(u)du$ and the Kaplan-Meier estimator (see [KM58]) of the survival function $S(t) := \mathbb{P}[T^* > t]$.

In a regression context one also observes an external d dimension covariate vector $X(\cdot)$ which is allowed to be time-dependent. In this setting, one of the most popular model is the Cox regression model (see [Cox72]):

$$\lambda(t|X(t)) = \lambda_0(t) \exp(\beta_0 X(t)), \tag{2.1}$$

where β_0 is an unknown row d dimensional parameter and λ_0 is an unknown function. An alternative model is the Aalen model (see [Aal80], [Aal89] or more recently [MS07]), defined as

$$\lambda(t|X(t)) = \lambda_0(t) + \beta_0 X(t). \tag{2.2}$$

In both the Cox and Aalen models, estimating the hazard function amounts to estimate the β regression parameter and the baseline function λ_0. In the regression framework, the independent censoring assumption corresponds to assuming T^* to be conditionally independent to C given $(X(s), s \leq T^*)$.

In order to perform estimation, one observes the i.i.d sample $(T_i, \Delta_i)_{i=1,\ldots,n}$ in the non-parametric context or $(T_i, \Delta_i, \{X_i(s), s \leq T_i\})_{i=1,\ldots,n}$ in the regression context. It can be shown that the likelihood function is equal to:

$$L(\beta, \lambda_0) = \prod_{i=1}^{n} \lambda(T_i|X_i(T_i))^{\Delta_i} \exp\left(-\int_0^{T_i} \lambda(u|X_i(u))du\right). \tag{2.3}$$

In the Cox model, under the independent censoring assumption, the following Cox partial likelihood can be used in order to estimate β:

$$L^{cox}(\beta) = \prod_{i=1}^{n} \frac{\exp(\beta X_i(T_i))}{\sum_j I(T_j \geq T_i)\exp(\beta X_j(T_i))}.$$

This expression is valid when assuming a non-parametric baseline, that is the baseline is defined as a function putting mass only at the observed times (the event times T_i such that $\Delta_i = 1$). Once the β term has been estimated, the Breslow estimator (see [Bre72]) can be used to estimate the cumulative baseline function:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{\Delta_i I(T_i \leq t)}{\sum_j I(T_j \geq T_i)\exp(\beta X_j(T_i))}.$$

In the rest of the manuscript we will use the notations: $F(t) = \mathbb{P}[T^* \leq t]$, $G(t) = \mathbb{P}[C \leq t]$ and $H(t) = \mathbb{P}[T \leq t]$. We will also introduce the endpoint of the study $\tau > 0$, which is usually defined such that $\tau < \inf\{t : H(t) = 1\}$.

2.1.2 The counting process notation

Survival analysis data can be nicely represented using counting process notations. This approach was developed from the mid's 1970 to the early 1990 and is described in the seminal book from [ABGK93]. All the theoretical development of survival analysis estimators such as the Kaplan-Meier estimator, the regression estimators derived from the Cox model and so on, are derived from this counting process approach. More precisely, from these counting approach notations, a martingale decomposition can be derived. Then, a law of large numbers type of result can be derived from the Lenglart inequality (see [Len77]) and a central limit theorem for local martingales is derived from Rebolledo's theorem (see [Reb78] and [Reb80]). In particular, in the Cox regression framework, consistency and asymptotic normality of β_0 and Λ_0 was proved in [AG82] using these two ingredients. See also [FH91] for applications of the Lenglart inequality and Rebolledo's theorem in a survival analysis context.

Introduce the counting process of interest $N^*(t) = I(T^* \leq t)$ and its at risk process $Y^*(t) = I(T^* \geq t)$ for $t \geq 0$. Introduce also the observed counting and at risk processes denoted respectively by $N(t) = I(T \leq t, \Delta = 1)$ and $Y(t) = I(T \geq t)$ and let τ be the endpoint of the study. It will generally be assumed that $\mathbb{P}[T > t] > 0$ for all t in $[0, \tau]$. In the regression context, the data now consist of n independent replications $(N_i(t), Y_i(t), \{X_i(s), s \leq t\})_{i=1,\ldots,n}$, for $t \in [0, \tau]$.

By definition of the hazard rate, we have:

$$\mathbb{E}[dN^*(t)|\mathcal{F}^*_{t-}] = Y^*(t)\lambda(t|X(t))dt, \tag{2.4}$$

where $dN^*(t)$ represents the jump size at time t of the process N^* and $\mathcal{F}^*_t = \sigma\{N^*(s), Y^*(s), X(s), s \leq t\}$ is a filtration. Furthermore, $X(t)$ is also assumed to be measurable with respect to $\mathcal{F}^*_{t-}$.

We will also assume independent censoring which can be expressed in its general definition as (see [ABGK93] or [MS07]):

$$\mathbb{E}[dN^*(t)|\mathcal{F}^*_{t-}] = \mathbb{E}[dN^*(t)|\mathcal{G}_{t-}],$$

where $\mathcal{G}_t = \mathcal{F}^*_t \cup \sigma\{I(s \le C), s \le t\}$ is an enlarged filtration. This equation implies that the censoring process does not convey any additional information on the probability of a jump of the counting process. As a sufficient condition of this equation to hold one can assume T^* to be conditionally independent of C given $(X(s), s \le T^*)$. Now, using the innovation theorem it can be proved that (see [ABGK93] or [MS07]):

$$\mathbb{E}[dN(t)|\mathcal{F}_{t-}] = Y(t)\lambda(t|X(t))dt,$$

where $\mathcal{F}_t = \sigma\{N(s), Y(s), X(s), s \le t\}$ represents the observed filtration. As previously, this result is crucial as it implies that the hazard rate can be estimated using only the observed data.

Under independent censoring, it can then be proved that one has the following martingale decomposition:

$$N(t) = \int_0^t Y(u)\lambda(u|X(u))du + M(t),$$

where M is a martingale with respect to the filtration $\mathcal{F}_t$. This representation gives a direct expression of the residuals as $N(t) - \int_0^t Y(u)\lambda(u|X(u))du$ and the Lenglart's inequality and Rebolledo theorem allow to prove consistency and asymptotic normality of quantities

$$\int_0^t h(u)dM(u),$$

for any $\mathcal{F}_{t-}$ measurable function h.

2.2 Multi-state and competing risk situations

A competing risk situation arises when individuals are at risk of experiencing different type of events and any of these events precludes the occurrence of the others. A multi-state situation occurs when individuals are at risk of experiencing different events but these events are not necessarily mutually exclusive. Both situations are described by a set of discrete states that the individuals might occupy. The hazard risks for moving from one state to another can be different which allows great flexibility in the modelling approach. Figure 2.1 describes a competing risk situation with two events and the illness death model which is a special case of multi-state models with only three possible states.

Taking into account competing-risks is essential as an individual that will experience an event should not be at risk of experiencing another event anymore. In the competing risk scenario of Figure 2.1, on the left side, the so-called cause specific hazards λ_1 and λ_2 are defined as:

$$\lambda_k(t) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \le T^{*k} < t + \triangle t | T^{*1} \ge t, T^{*2} \ge t, X(t)]}{\triangle t}, \tag{2.5}$$

where T^{*k} represents the true time to event of type k, $k = 1, 2$. These quantities can be estimated from the observed data, treating the other event as a censoring event. However, care must be taken when the interest lies in a cumulative function. Typically, the cumulative incidence function is written

$$\mathbb{P}[T^{*k} \le t | (X(s), s \le t)] = \int_0^t \lambda_k(u) S(u|X(u))du, \tag{2.6}$$

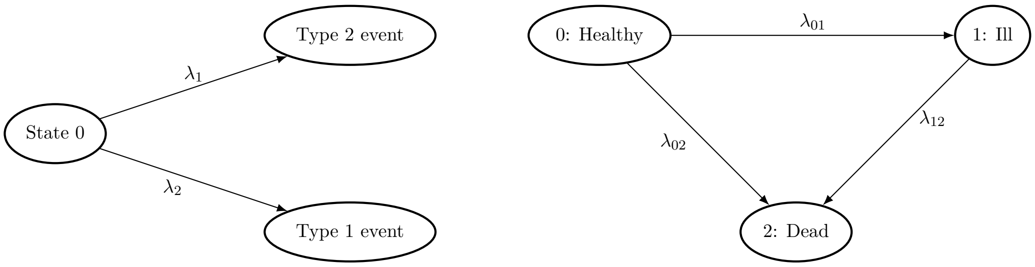


Figure 2.1: Examples of a competing risk situation on the left (with two types of events) and of the illness death model on the right.

where S is the conditional event-free survival function defined as $S(t|X(t)) = \exp(-\int_0^t(\lambda_1(u) + \lambda_2(u))du)$. See for example [MS07] for more details on competing risk models.

In the illness-death model, on the right panel of Figure 2.1, typical quantities of interest are the transition intensities. For two states $i, j \in \{0, 1, 2\}$, the transition intensity $P_{i,j}(x, y)$ is defined, for $x < y$, as the probability for a subject being in state i at time x to be in state j at time y. For example, in the homogeneous Markov illness-death model, the transition intensity for staying in the healthy state from time x to y, is equal to

$$P_{00}(x, y) = \exp\left(-\int_x^y (\lambda_{01}(u) + \lambda_{02}(u))du\right),$$

and the transition intensity for moving from the healthy state to the disease state between times x and y, and to be still occupying the disease state at time y, is equal to

$$P_{01}(x, y) = \int_x^y \exp\left(-\int_x^u (\lambda_{01}(v) + \lambda_{02}(v))dv\right) \lambda_{01}(u) \exp\left(-\int_u^y \lambda_{12}(w)dw\right) du. \tag{2.7}$$

All transition intensities can be computed in a similar way. For more complex multi-state designs, a transition probability matrix can be computed from a product-integral formula in a similar manner as the Kaplan-Meier estimator is derived in more classical contexts in survival analysis. This formula is typically resolved using a Kolmogorov forward differential equation (see [ABGK93]). For more details about multi-state models, see for instance [AK12].

2.3 Recurrent events with a terminal event

2.3.1 Modelling the rate function

Recurrent event data can be seen as an extension of standard survival data using the counting process approach. They occur when individuals may experience the same event several times. Typical examples in medical applications include the hospitalisations of a patient due to a specific disease, relapses from a disease, asthma attacks in respirology studies, epileptic seizures in neurology studies etc. See [CL07] for a thorough discussion of recurrent events models and applications. In many of these studies, the patients are also at risk of experiencing a terminal event, often death, which must be accounted for as a competing risk. We introduce the counting process of interest $N^*(t)$ which counts the number of recurrent events that have occurred before time t and the at risk process $Y^*(t) = I(T^* \geq t)$ where T^* represents the actual time of death. The recurrent event model is defined as:

$$\mathbb{E}[dN^*(t)|Y^*(t), X(t)] = Y^*(t)\lambda(t|X(t))dt, \tag{2.8}$$

where $\lambda(t|X(t))$ is called the rate function. Next this rate function can be modelled using the Cox model of Equation (2.1) or the Aalen model of Equation (2.2). In the absence of terminal event, the Cox rate model was introduced by [PC93], [LN95] and rigorous theoretical arguments were developed in [LWYY00]. The extension to the presence of terminal event in the Cox model is discussed in [LNC97] and [CL07]. Model (2.8) when the rate function is assumed to follow the Aalen model has been studied in great detail in [Sch02].

In Equation (2.8) it is important to stress that the rate function λ is defined by conditioning on $Y^*(t)$ and $X(t)$ in the left-hand side of the equation. Alternatively, one could condition on the entire history of the process N^*, namely $\mathcal{F}_t-$ as in Equation (2.4), in which case λ would represent the intensity of the recurrent event process N^*. However since λ in the right side of Equation (2.4), does not depend on the history of the process N^*, this would imply that all the influence of the prior events on the future recurrence, if there is any, is mediated through the time-varying covariate at time t. If X is time invariant, then this model would be equivalent to assuming an independent increments structure for N^* as in the Poisson process. In many medical applications, this independent increment assumption is not realistic and Model (2.8) should be used instead. This model is very general and as a matter of fact, it can be shown for instance that it encompasses the recurrent frailty model (see [LWYY00] for details).

As in classical survival analysis contexts, censoring will generally occur such that the observed recurrent event process is $N(t) = N^*(t \wedge C)$ and the observed at risk process is $Y(t) = I(T^* \wedge C \geq t)$. Under the following independent censoring assumption

$$\mathbb{E}[dN^*(t)|Y^*(t), X(t)] = \mathbb{E}[dN^*(t)|Y^*(t), I(C \geq t), X(t)],$$

it can be shown that Equation (2.8) holds with N^* and Y^* replaced by their observed counter parts N and Y. In other words, the relation $\mathbb{E}[dN(t)|Y(t), X(t)] = Y(t)\lambda(t|X(t))dt$ holds and inference can then be performed using the observed data.

Conditioning on the at-risk process at time t and not on the entire history also has implications on a theoretical point of view. In particular, martingale properties are no longer available and empirical process theory must be used instead. When modelling λ through a Cox or Aalen model, the asymptotic distributions of the regression parameters can be derived from the functional central limit theorem (see [Pol90]) or from central limit theorems for Donsker classes as in [VDVW96]. In the absence of terminal event, the theory for the regression parameters in the Cox model has been developed in [LWYY00].

In practice, the likelihood function is similar to the standard survival analysis context and the regression estimators are actually identical in the intensity model (when conditioning on the entire history) and in the rate model (2.8). However, the variance of the estimators is different in the rate model as it involves the covariance structure of the recurrent event increments, and a sandwich estimator is used to estimate this variance. As a consequence, assuming the Poisson assumption can substantially change statistical inference results in statistical tests or confidence intervals. Since only the variance of the estimators is changed in the rate model, the corresponding variance estimator is often called the robust variance estimator, in the sense that this estimator is robust to violation of the independent increment assumption. It should be noted that this robust variance estimator is equivalent to the one derived in cluster survival data as recurrent events can be considered as clustered data where each individual represents a different cluster. See [LWYY00] or [Wil00] for the explicit expression of this robust variance estimator.

2.3.2 Non-parametric estimation of the cumulative mean function

In a non-parametric setting, an alternative to the usual Kaplan-Meier survival estimator is to compute the average number of recurrent events experienced until any time point. From the independent censoring assumption, we have the two equations:

$$\begin{aligned}\mathbb{E}[dN^*(t)|Y^*(t)] &= Y^*(t)\lambda(t)dt\\ \mathbb{E}[dN(t)|Y(t)] &= Y(t)\lambda(t)dt.\end{aligned}$$

Taking the expectation in the second equation gives $\mathbb{E}[dN(t)] = H(t)\lambda(t)dt$, where $H(t) = \mathbb{P}[T^* \wedge C > t] = S(t)\cdot\mathbb{P}[C > t]$ and $S(t) = \mathbb{P}[T^* > t]$. An estimator of the cumulative rate function is then derived as

$$\widehat{\Lambda}(t) = \sum_{i=1}^{n}\int_0^t \frac{dN_i(u)}{\sum_{j=1}^{n} Y_j(u)}. \tag{2.9}$$

Taking now the expectation in the first equation gives $\mathbb{E}[dN^*(t)] = S(t)\lambda(t)dt$. It is then easily seen that

$$\mathbb{E}[N^*(t)] = \int_0^t \frac{S(u)\mathbb{E}[dN(u)]}{1 - H(u)} = \int_0^t \frac{\mathbb{E}[dN(u)]}{\mathbb{P}[C > u]}, \tag{2.10}$$

and the cumulative mean estimator is defined as:

$$\widehat{\mathbb{E}[N^*(t)]} = \sum_{i=1}^{n}\int_0^t \frac{\hat{S}(u)dN_i(u)}{\sum_{j=1}^{n} Y_j(u)}, \tag{2.11}$$

where $\hat{S}$ is the Kaplan-Meier estimator of S. This estimator was introduced by [GL00] in a different way and its theoretical derivations (such as the construction of confidence intervals) can be found in their paper.

Finally note that in the absence of a terminal event, an individual is always at risk of experiencing a recurrent event. In that case the counting process of interest verifies the equality $\mathbb{E}[dN^*(t)] = \lambda(t)dt$ and the observed counting process $N(t) = N^*(t\wedge C)$ verifies $\mathbb{E}[dN(t)|Y(t)] = Y(t)\lambda(t)dt$ where $Y(t) = I(C \geq t)$. Then the cumulative mean estimator is derived as

$$\widehat{\mathbb{E}[N^*(t)]} = \hat{\lambda}(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{dN_i(u)}{1 - \hat{G}(u-)}, \tag{2.12}$$

where $1 - \hat{G}$ is the Kaplan-Meier estimator of the censoring distribution.

2.3.3 Recurrent event models with dependence on prior recurrences

An alternative to Model (2.8) is to incorporate the effect of prior recurrences on the rate function. This is of interest when one suspects the rate to change as more recurrences occur. It is also a powerful tool for the purpose of prediction: knowing the past history of an individual (his number of previous recurrent events), the model will allow to predict the risk of experiencing a new recurrent event. These models are presented and discussed in great detail in [CL07]. As an illustration, we introduce the following recurrent event situation:

$$\begin{aligned}\mathbb{E}[dN(t)|Y_s(t), X(t)] &= Y_s(t)\lambda_s^E(t|X(t))dt,\\ \mathbb{E}[dN^T(t)|Y_s(t), X(t)] &= Y_s(t)\lambda_s^T(t|X(t))dt,\ s = 1,\dots,6,\end{aligned} \tag{2.13}$$

where $Y_s(t) = I(N(t-) = s-1, T \geq t)$, λ_s^E represent rate functions for the recurrent event process and λ_s^T represent hazard rates for the terminal event. In this model there are six different at-risk processes Y_s, corresponding to the situations where an individual has already experienced 0, 1, ..., or 5 and more events. Note that the rate functions λ^E and hazard rates

λ^T are allowed to change according to the number of previous recurrent events. For simplicity the model is written in terms of the observed at-risk and counting processes, since as previously, there is an equivalence between the model with the (unobserved) processes of interest and the observed ones under the independent censoring assumption. In our setting, the independent censoring assumption can be written as

$$\mathbb{E}[dN^*(t)|Y_s^*(t), X(t)] = \mathbb{E}[dN^*(t)|Y_s(t), X(t)], \tag{2.14}$$

where $Y_s^*(t) = I(N^*(t-) = s-1, T^* \geq t)$ are the true (unobserved) at risk processes.

A different and efficient way of characterising this model, is to use the multi-state representation. In Figure 2.2 below we see all the different possible states for an individual with the different hazard rates and rate functions. From this figure, it is also clear that the terminal event plays the role of a competing event and must be accounted for in the study as patients visiting this state are no longer at risk of experiencing a new event. The last state encompasses all the events equal to or greater than 5 and has a special status: individuals in this state are continuously at risk of experiencing an event with rate equal to λ_6^E. Transition intensities can then be computed using the multi-state approach of Section 2.2.

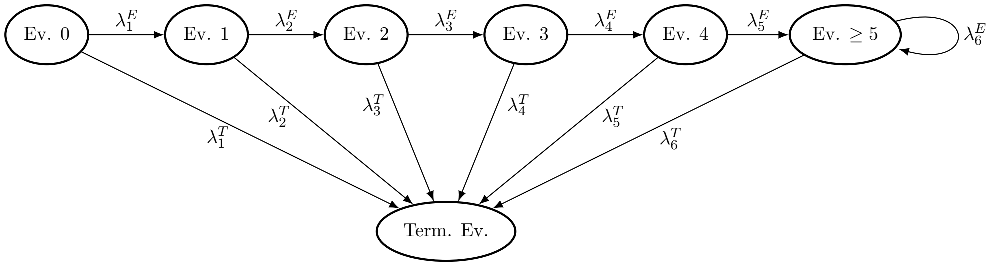


Figure 2.2: Illustration of a recurrent event model with dependence on prior events as a multi-state situation. Individuals start in the state Ev. 0 and can then move to the other states as time increases. The state Term. Ev. is an absorbing state.

2.4 Interval censoring

Interval censored data occur when the true time to event is only known to have occurred between two different times L and R. In other words, L and R are observed and we know that $\mathbb{P}(T^* \in [L, R]) = 1$. The type of data that we consider in this manuscript are called mixed interval censored and they include exact, right-censored, left-censored and interval-censored observations. They can be described as follows:

- Left censoring if $0 = L < R < \infty$
- Interval censoring if $0 < L < R < \infty$
- Exact observation if $L = R = T^*$
- Right censoring if $0 < L < R = \infty$.

For sake of simplicity we also introduce the notation $\delta \in \{0, 1\}$ which represents the right-censored status of the individuals, with $\delta = 1$ if the observation is left-censored, interval-censored, or exact and $\delta = 0$ if the observation is right-censored. The observations consist of the data $(L_i, R_i, X_i)_{i=1,\ldots,n}$ in the non-parametric context. In the regression context, a time independent covariate vector X_i is also observed.

2.4.1 Non-parametric analysis of interval-censored data

Interval-censored data are challenging to analyse. As a matter of fact, the non-parametric estimator of the survival function S is not explicit and iterative algorithms must be implemented. The asymptotic distribution of this estimator is not explicit and the estimator does not achieve the $n^{1/2}$ rate of convergence. The first algorithm for the survival function estimator was proposed by [Tur76]. This algorithm is based on the following Efron's self consistency equation (see [Efr67]):

$$S(t) = \mathbb{P}[L > t, \delta = 0] + \iint \frac{S(t) - S(r)}{S(l) - S(r)} I(l < t < r) dF_{L,R}(l, r),$$

where $F_{L,R}$ is the joint cumulative distribution function of (L, R). Since the L_is and R_is are observed, the empirical version of the previous equality holds when replacing S by $\hat{S}$ and $F_{L,R}$ by its empirical cumulative distribution function. This equation then needs to be solved with respect to $\hat{S}$. In Turnbull's algorithm, instead of directly solving the self consistency equation, one must first determine the innermost intervals, which are the intervals whose left and right end points are given by some of the L_is and R_is respectively and that contain no other L_is and R_is except at their end points. Once these are known, an iterative algorithm is used to determine the value of the survival estimator on these intervals. Using the notation of [Tur76], the m innermost intervals are defined as $[q_1, p_1], \dots, [q_m, p_m]$ and $s_j = F(p_j+) - F(q_j-)$ represents the contribution of the cumulative distribution function $F = 1 - S$ on the jth innermost interval. Turnbull proposed to estimate $\mathbf{s} = (s_1, \dots, s_m)$ using the self-consistent equation: for $j = 1, \dots, m$,

$$s_j = \pi_j(\mathbf{s}) := \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k}, \tag{2.15}$$

where $\alpha_{ij} = 1$ if $[q_j, p_j] \subset [L_i, R_i]$ and 0 otherwise, $\sum_j s_j = 1$ and $s_j \geq 0$. The Turnbull algorithm is an iterative algorithm where at the Kth step, $\hat{s}_j^{(K)} = \pi_j(\hat{\mathbf{s}}^{(K-1)})$ for $j = 1, \dots, m$ and $\hat{s}_j^{(0)} = 1/m$ for all j. At convergence, the $\hat{s}_j$s verify Equation (2.15). The final estimator of the survival function is then defined as

$$\hat{S}(t) = 1 - (\hat{s}_1 + \cdots + \hat{s}_j),$$

for $t \in (p_j, q_{j+1})$. Note that, as defined in [Tur76], some of the $\hat{s}_j$s can still be null on the innermost intervals. In other words, the innermost intervals do not necessarily refer to the support of $\hat{S}$. See also [Sun07] for more details about innermost intervals and the Turnbull's algorithm. Interestingly, it is possible to prove that this algorithm is actually equivalent to the EM algorithm developed by [DLR77] if one treats the time of interest T^* as an unobserved variable. See for example [GW92a] for the connection between Turnbull's estimator and the EM algorithm.

Several difficulties arise from Turnbull's estimator. First of all, the resulting survival estimator is not necessarily unique in the sense that more than one estimator may satisfy Equation (2.15). Secondly, from the definition of the estimator one can see that the estimator is actually not defined in the intervals (q_j, p_j). In [GW92a] and [Gro96], asymptotics of the Non Parametric Maximum Likelihood Estimator (NPMLE) are studied. Considering different scenarios depending on properties of the distribution of (L, R), it can be shown that the NPMLE has a $n^{1/3}$ rate of convergence or in the best case, a $(n \log(n))^{1/3}$ rate of convergence. Moreover, the asymptotic distribution of the NPMLE is not explicit. As a result, the variance of the NPMLE can be quite large and it is not directly possible to construct confidence intervals or statistical tests from the quantiles of the asymptotic distribution of the NPMLE. The slow convergence of the NPMLE is explained by the support of the survival estimator which is usually composed of a number of intervals that is much lower than the sample size. In other words, the support of the survival estimator consists of a much lower number of intervals than in a classical right-censored survival analysis. Note that, in the latter case, the union of innermost intervals is directly the support of $\hat{S}$ and it corresponds to the set of singletons composed of all non censored observations if the last time is observed. In the case where the last observation is right-censored, the interval whose left-endpoint is the last observation and right-endpoint infinity also corresponds to an innermost interval.

Finally, it should be noted that other algorithms exist to derive a non-parametric estimator of the survival function, such as the convex minorant algorithm. This algorithm was developed by [GW92a] and later by [Jon98] and is based on the isotonic regression. This method will not be used in this manuscript and shall not be discussed any further.

2.4.2 Regression modelling of interval-censored data

Assume that the observations consist of $(L_i, R_i, X_i)_{i=1,\dots,n}$, where X_i is a covariate vector. In a regression model, maximisation over the model parameters can be achieved using the likelihood of the observed data. Define $\boldsymbol{\theta}$ as the model parameter. In the mixed case of interval censored data, the contributions to the likelihood of exact observations can be separated from contributions of non-exact (interval censored, left censored or right censored) observations. The observed likelihood is equal to:

$$L^{\text{obs}}(\boldsymbol{\theta}) = \prod_{i \text{ not exact}} \left\{ \exp\Big(- \int_0^{L_i} \lambda(t|X_i)dt \Big) \Big(1 - \exp\Big(- \int_{L_i}^{R_i} \lambda(t|X_i)dt \Big) \Big) \right\}^{\delta_i} \\ \times \left\{ \exp\Big(- \int_0^{L_i} \lambda(t|X_i)dt \Big) \right\}^{1-\delta_i} \prod_{i \text{ exact}} \lambda(t|X_i) \exp\Big(- \int_0^{L_i} \lambda(t|X_i)dt \Big).$$

The hazard rate can be modelled by the Cox model (2.1) for example. In that case, a parametric baseline must be specified. The standard choices comprise the exponential, Weibull or piecewise constant baseline hazards. Using one of these specifications of the baseline will lead to a fully parametric model. Maximising $L^{\text{obs}}(\boldsymbol{\theta})$ can be performed using Newton-Raphson techniques and statistical inference and tests can be based on standard likelihood methods. See [Sun07] for more details about fully parametric regression models for interval censored data.

CHAPTER 3

Mathematical statistics

In this chapter, I describe my new statistical developments with focus on theoretical contributions. This chapter refers to the published papers [P1], [P2], [P3] and the submitted paper [S1]. The first two papers introduce single-index models as an alternative to the popular Cox model, one in the standard survival analysis context and the other one in the recurrent event context. The other two papers discuss non-asymptotic results for non-parametric estimators. Paper [P3] deals with the estimation of the rate function in the context of recurrent events using a kernel estimator with Lepski's method. Paper [S1] is concerned with the non-parametric estimation of the survival function for interval censored data.

3.1 Single-index model approaches for right-censored data

The Cox model is by far the most widely used model in survival analysis. Its popularity is mostly due to the nice interpretation of hazard ratios which are assumed to be constant over time: this allows to summarise the relative hazard risk of one covariate over another by a single number that is constant over time. However, this assumption is very strong and is violated in many real data applications. The aim of [P1] and [P2] were to generalise the Cox model using single-index models. See for instance [HHI93], [Ich93] and [XTLZ02] for a review of single-index model theory.

While the single-index model is usually defined in terms of its expectation, we will first focus our interest on the following alternative single-index model: let β_0 be a d dimensional row vector, we assume that

$$f(y|x) = f_{\beta_0}(y, \beta_0 x), \tag{3.1}$$

where $f(y|x)$ represents the conditional density of T^* given $X = x$ evaluated at y and $f_{\beta_0}(y|u)$ represents the conditional density of T^* given $\beta_0 X = u$ evaluated at y. For identifiability purposes, the first component of β_0 is assumed to be equal to one. It is straightforward to see that for a time independent covariate X the Cox model (2.1) satisfies this assumption as the hazard function completely specifies the distribution of the time variable T^*. As a matter of fact, our single-index model is very general: the class of proportional hazard models, the Aalen model, the Accelerated Failure Time model (see [BJ79]) or the proportional odds model (see [Ben83]) are all special cases of Model (3.1).

Model (3.1) was initially studied in the uncensored case by [DHH03]. However, their estimation procedure cannot be directly implemented in the censored framework since the response variables are not directly observed. A solution consists of using functionals of the Kaplan-Meier estimator. In order to define an analogue of this estimator to the bivariate case, we first rewrite the Kaplan-Meier estimator as a jump function: let $\hat{S}$ be the Kaplan-Meier estimator of the survival function, then we have:

$$1-\hat{S}(t)=\sum_{i=1}^{n}\Delta_i W_{in} I(T_i\leq t),$$

where W_{in} represent the jumps of $1-\hat{S}$. See [Efr67] or more recently [SD01] for this expression of the Kaplan-Meier estimator. The weights W_{in} can actually be expressed as a function of the Kaplan-Meier estimator of the censoring distribution G. Let $1-\hat{G}$ be this estimator, then $W_{in}=1/(1-\hat{G}(T_i-))$. In the bivariate context, [Stu93] proposed to extend the expression of $\hat{S}$ as a jump function to the estimation of the cumulative distribution function of (X,T^*) by:

$$\hat{F}_{X,T^*}(x,t)=\sum_{i=1}^{n}\Delta_i W_{in} I(X_i\leq x, T_i\leq t).$$

Since the Kaplan-Meier estimator is known to poorly behave in the tail of the distribution, our estimation method uses a truncation bound that can be adaptively chosen from the data. Finally, the truncation version of the conditional density of T^* given $\beta_0 X=u$ is computed using a non-parametric kernel estimator: for any parameter β,

$$\hat{f}_{\beta}^{h,\tau}(y|\beta x)=\frac{\int K_h(\beta x-\beta u)K_h(y-z)I(z\in A_\tau)d\hat{F}_{X,T^*}(u,z)}{\int K_h(\beta x-\beta u)I(z\in A_\tau)d\hat{F}_{X,T^*}(u,z)},$$

where K is a kernel, h a bandwidth, $K_h(\cdot)=K(\cdot/h)/h$ and A_τ is a sequence of compacts included in the set $\{t:0\leq t\leq\tau\}$, for $\tau\leq\tau_0$ where $\tau_0=\inf\{t:\mathbb{P}[T\leq t]=1\}$. Both the bandwidth of the kernel estimator and the truncation bounds are chosen from the data. The final estimator of β_0 is derived from likelihood arguments:

$$\hat{\beta}=\arg\max_{\beta}\sum_{i=1}^{n}\Delta_i W_{in}\log(\hat{f}_{\beta}^{h,\tau}(y|\beta x))I(T_i\in A_\tau).$$

In order to derive valid theoretical results of the proposed estimator, uniform convergence (with respect to the bandwidth and truncation bound) of the kernel estimator must be proposed. This is achieved by assuming the class of density functions to belong to some Donsker classes (see [VDVW96] for the definition of Donsker classes) and by using results for the uniform convergence of kernel estimators from [EM05].

In paper [P2] a similar single-index model is introduced in a recurrent event context with a terminal event. Using the notation of the introduction section, we assume that

$$\mathbb{E}[N^*(t)|X=x]=\mu_{\beta_0}(t,\beta_0 x),$$

where $\mu_{\beta_0}(t,u)=\mathbb{E}[N^*(t)|\beta_0 X=u]$ and β_0 is an unknown parameter vector. The estimation procedure then uses a least-square criterion based on a similar relation as the one derived in Equation (2.10) for the cumulative mean function. In our single-index model this can be written as $\mathbb{E}[N^*(t)|X]=\int_0^t\mathbb{E}[dN(u)|X]/\mathbb{P}[C>u]$ under independent censoring. This leads to the following least-square criterion that needs to be minimised:

$$\int\sum_{i=1}^{n}\left(\hat{\mu}_{\beta}(t,\beta X_i)-\int_0^t\frac{dN_i(u)}{1-\hat{G}(u-)}\right)^2 w(t)dt,$$

where $1 - \hat{G}(u)$ represents the Kaplan-Meier estimator of the censoring distribution and $w(t)$ represents some measure that is used to ensure the existence of the integral. The term $\hat{\mu}_\beta$ represents a kernel estimator of μ_β defined as

$$\hat{\mu}_\beta(t,u) := \int_0^t \frac{\sum_i K\left(\frac{\beta X_i - u}{h}\right) dN_i(s)}{\sum_j K\left(\frac{\beta X_j - u}{h}\right)(1 - \hat{G}(u-))},$$

where K is a kernel and h a bandwidth. As previously, theoretical properties of the estimator are derived uniformly on the bandwidth parameter and on the measure w using empirical processes methods. This allows to choose the bandwidth and the measure from the data in an efficient way. In particular the measure is optimally chosen such as to prevent estimation issues caused by large recurrent event values.

3.2 Non-asymptotic results for non-parametric estimators

In this section, I present the two papers [P3] and [S1] where non asymptotic results were proved in the context of non-parametric estimators. In the first of these papers we studied a kernel type estimator while in the second one we defined an estimator based on model selection theory. In both cases, the main results are oracle bounds for the proposed estimators. These kind of results are based on concentration inequalities due mainly to [Tal94] and [Tal95]. We recall below the concentration inequality that was used in both papers. This inequality can be found in [KR05].

Theorem 1 *Let $\xi_1, \dots, \xi_n$ be independent random variables and let $\nu_{n,\xi}(f) = \sum_{i=1}^n \{f(\xi_i) - \mathbb{E}[f(\xi_i)]\}/n$. Then, for a countable class of functions $\mathcal{F}$ uniformly bounded and $\alpha > 0$, we have*

$$\mathbb{E}\Big[\Big\{\sup_{f\in\mathcal{F}} \nu_{n,\xi}^2(f) - 2(1+2\alpha)A^2\Big\}_+\Big] \le \frac{4}{b}\left(\frac{W}{n} e^{-b\alpha\frac{nA^2}{W}} + \frac{49B^2}{bn^2\psi^2(\alpha)} e^{-\frac{\sqrt{2\alpha}b\psi(\alpha)}{7}\frac{nA}{B}}\right),$$

where $\psi(\alpha) = (\sqrt{1+\alpha} - 1) \vee 1$, $b = 1/6$ and

$$\sup_{f\in\mathcal{F}} \|f\|_\infty \le B, \quad \mathbb{E}\Big[\sup_{f\in\mathcal{F}} |\nu_{n,\xi}(f)|\Big] \le A, \quad \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbb{V}[f(\xi_i)] \le W.$$

3.2.1 Estimation of the rate function for recurrent event data with a terminal event

In the Introduction section, non-parametric estimation of the cumulative mean function $\mathbb{E}[N^*(t)]$ was discussed. The standard estimator of [GL00] defined in Equation (2.11) is a piecewise constant estimator with jumps at the observed recurrent events. Few other works using smoothing approach were introduced in this framework. In [BBM+81], the authors briefly presented a kernel estimator of the rate function when the recurrent events were supposed to be distributed according to a Poisson process and the censored times constant. Then, [CWH05] extended their results to a more general setting where no Poisson assumption is made, no terminal events are considered and the censoring variables are random, but observed. In our work [P3], we extended these results to a kernel estimator of the rate function in the context of non observed random censoring and in the presence of a terminal event. For this estimator, we developed an adaptive procedure to select the bandwidth, based on the work of [GL11]. We established oracle inequalities for the L2-risk and the integrated L2-risk of our estimator with a data-driven choice of the bandwidth. This was the first non-asymptotic result in this setting.

Our paper [P3] focused on the composite endpoint defined as a recurrent event or a terminal event. The rate function was thus defined in a slightly different way than in the Introduction section as for composite endpoints the recurrent event process should have jumps at any occurrence of either a recurrent event or a terminal event while in the introduction N^* will only jump at observed recurrent events. In order to be consistent with the notations of this manuscript, we present the kernel estimator as defined in the Introduction section. Let K be a kernel and h a bandwidth. Based on Equation (2.9) our smooth estimator of the rate function is defined by:

$$\hat{\lambda}_h(t) = \frac{1}{nh} \sum_{i=1}^{n} \int K\left(\frac{t-s}{h}\right) \frac{dN_i(s)}{1-\hat{H}(s-)},$$

where $\hat{H}(s) = \sum_i I(T_i \leq s)/n$ and $T_i = T_i^* \wedge C_i$. Introducing the pseudo version of this estimator as

$$\tilde{\lambda}_h(t) = \frac{1}{nh} \sum_{i=1}^{n} \int K\left(\frac{t-s}{h}\right) \frac{dN_i(s)}{1-H(s-)},$$

where $H(s) = \mathbb{P}[T \leq s]$, we can easily see the intuition behind this estimator. Using a change of variables, we have

$$\mathbb{E}[\tilde{\lambda}_h(t)] - \lambda(t) \leq \left(\int_{-1}^{1} K(u) \Big(\lambda(t+uh) - \lambda(t) \Big) \right) du,$$

and from a Taylor's expansion we can see that the bias term is of order h^b, where b represents the regularity of the rate function and the kernel is assumed to be of order b. Finally, in order to derive non asymptotic results on the rate function estimator $\hat{\lambda}$, the key tool is a concentration inequality due to Talagrand. See [P3] for the exact expression of this concentration inequality.

We finally provide the two oracle inequalities obtained for this estimator. Note that the kernel is assumed to be supported on $[-1, 1]$ such that the integral in the definition of $\hat{\lambda}_h(t)$ will vanish outside the interval $[t-h, t+h]$ and therefore estimation of λ is only performed for t such that $t \pm h \in [0, \tau]$, where τ is the endpoint of the study defined such that $\tau < \inf\{t : H(t) = 1\}$. We will assume that the survival functions of the censoring and time to event variables are bounded from below, that $N(t)$ is bounded from above for all t in $[0, \tau]$ and that $\sup_{t\in[0,\tau]} \lambda(t) < \infty$. For the L2-risk we proved that:

Theorem 2 *For $\mathcal{H}_n$ a finite discrete set of bandwidths such that Card$(\mathcal{H}_n) \leq n$,*

$$\forall h \in \mathcal{H}_n, nh \geq \kappa_1 \log(n), \text{ for some } \kappa_1 \geq 0,$$

and

$$\sum_{k:h_k \in \mathcal{H}_n} \frac{1}{nh_k} \lesssim \log^a(n), \text{ for some } a \geq 0,$$

the estimator $\hat{\lambda}_{\hat{h}}$ defined with the bandwidth $\hat{h}$ chosen by Goldenshluger and Lepski's method (see [GL11]) satisfies

$$\mathbb{E}\left[\left(\hat{\lambda}_{\hat{h}}(t_0) - \lambda(t_0) \right)^2 \right] \leq c(c_1^2 h^{2b} + V_0(h)) + c' \frac{\log^{(1+a)}(n)}{n},$$

where c is a positive constant and $V_0(h)$ is a quantity defined in Goldenshluger and Lepski's method for the choice of $\hat{h}$.

For the integrated L2-risk we proved that:

Theorem 3 *For $\mathcal{H}_n$ a finite discrete set of bandwidths such that Card$(\mathcal{H}_n) \leq n$,*

$$\sum_{k:h_k \in \mathcal{H}_n} \frac{1}{nh_k} \lesssim \log^a(n), \text{ for some } a \geq 0,$$

and

$$\sum_{k:h_k \in \mathcal{H}_n} \exp(-s/h_k) < \infty, \forall s \geq 0,$$

the estimator $\hat{\lambda}_{\hat{h}}$ defined with the bandwidth $\hat{h}$ chosen by Goldenshluger and Lepski's method (see [GL11]) satisfies

$$\int_h^{\tau-h} \mathbb{E}\left[\left(\hat{\lambda}_{\hat{h}}(t) - \lambda(t)\right)^2\right] dt \leq c(\tau c_1^2 h^{2b} + V(h)) + c' \frac{\log^{(1+a)}(n)}{n},$$

where c is a positive constant and $V(h)$ is a quantity defined in Goldenshluger and Lepski's method for the choice of $\hat{h}$.

3.2.2 Estimation of the survival function for interval censored data

We now consider the interval censored context presented in the Introduction section with no exact observations. Smooth estimators have already been proposed in the case 1 censoring, that is when a time variable is observed for all subjects and the true event time is known to have happened either before or after the observed time. For these types of data, [Yan00] studied the estimate of functionals of the survival function using locally linear smoothers and [BC09] proposed two adaptive estimators, one of quotient type and another one of regression type, using projection methods. For interval censored data with case 2, spline methods were introduced in [KS92] and a kernel method was studied in [BDS05] for the estimation of the density function. More recently, smooth alternatives to the NPMLE were proposed by using a kernel method in [GK11] and by introducing a log-concave constraint in the estimation procedure in [ABY16]. In the submitted paper [S1], we propose a new selection model estimator based on a least square criterion.

Introduce the variable ε which indicates if the observation is left-censored ($\varepsilon = -1$), interval-censored ($\varepsilon = 0$) or right-censored ($\varepsilon = 1$). Then the estimation procedure is based on the relations: $\mathbb{E}[1 - I(\varepsilon_i = -1)|L_i] = S(L_i)$ and $\mathbb{E}[I(\varepsilon_i = 1)|U_i] = S(U_i)$. The resulting estimator is somewhat complicated as it makes use of all types of observations (those for which $\varepsilon = -1$, $\varepsilon = 0$ and $\varepsilon = 1$) in order to obtain a large support of the estimator based on all observations. A key feature of this new estimator is that the basis of the model selection estimator does not need to be compactly supported. As a matter of fact, our results are valid for the Laguerre basis which is $\mathbb{R}_+$ supported. Elements of our theoretical results were borrowed from a recent work from [CGC18] to include this possibility in our results. We were then able to provide mean-square risk bounds for the resulting estimators, to compute general rates of convergence in the compactly supported case, and to propose a model selection device leading to an automatic bias variance trade-off.

We first consider a projection space $\Sigma_m(J) = \text{span}(\varphi_0, \dots, \varphi_{m-1})$ where $(\varphi_j)_{0 \leq j \leq m-1}$ constitutes an orthonormal basis $\langle \varphi_j, \varphi_k \rangle = \varepsilon_{j,k}$ with respect to the scalar product $\langle u, v \rangle = \int_J u(x)v(x)dx$. The domain J is the support of the basis and can be an interval $[a, b]$ if we consider histogram or trigonometric basis, or the interval $J = \mathbb{R}_+$ if we consider the Laguerre basis. An interesting property of these basis is to see that they all satisfy:

$$\forall m \in \mathbb{N} \setminus \{0\}, \quad \sup_{x \in I} \sum_{j=0}^{m-1} \varphi_j^2(x) := \Big\| \sum_{j=0}^{m-1} \varphi_j^2 \Big\|_\infty \leq c_\varphi^2 m,$$

for some constant $c_\varphi > 0$ depending on the basis only. Introduce the following matrices:

$$\begin{cases} \Phi_m^{(L)} & = (\varphi_j(L_i))_{1 \leq i \leq n, 1 \leq j \leq m}, \ \vec{\varepsilon}^{\,(L)} = (1 - I(\varepsilon_i = -1))_{1 \leq i \leq n} = (1 - I(X_i \leq L_i))_{1 \leq i \leq n}, \\ \Phi_m^{(R)} & = (\varphi_j(R_i))_{1 \leq i \leq n, 1 \leq j \leq m}, \ \vec{\varepsilon}^{\,(R)} = (I(\varepsilon_i = 1))_{1 \leq i \leq n} = (1 - I(X_i \leq R_i))_{1 \leq i \leq n}, \end{cases}$$

and

$$\Psi_{m,Z} = (\langle \varphi_j, \varphi_k \rangle_Z)_{1 \leq j,k \leq m}, \ \widehat{\Psi}_{m,Z} = (\langle \varphi_j, \varphi_k \rangle_{n,Z}) \text{ for } Z = L, R.$$

We have $\Psi_{m,Z} = \mathbb{E}[\widehat{\Psi}_{m,Z}]$ for $Z = L, R$ and

$$\widehat{\Psi}_{m,L} = \frac{1}{n} \Phi_m^{(L)\top} \Phi_m^{(L)}, \ \widehat{\Psi}_{m,R} = \frac{1}{n} \Phi_m^{(R)\top} \Phi_m^{(R)}.$$

Now, define the contrast

$$\gamma_n(t) = \|t\|_{n,R}^2 + \|t\|_{n,L}^2 - \frac{2}{n} \sum_{i=1}^n I(\varepsilon_i = 1) t(R_i) - \frac{2}{n} \sum_{i=1}^n I(\varepsilon_i \neq -1) t(L_i),$$

where for $Z = L, R$, $\|t\|_{n,Z}^2 = \sum_{i=1}^n t^2(Z_i)/n$. Our estimator is defined as:

$$\widehat{S}_m = \arg\min_{t \in \Sigma_m} \gamma_n(t).$$

In order to perform model selection, we first define the collection of models $\mathcal{M}_n$ by

$$\mathcal{M}_n = \left\{ m \in \mathbb{N} \setminus \{0\} : m(\|(\Psi_{m,L} + \Psi_{m,U})^{-1}\|_{\text{op}}^2 \vee 1) \leq \mathfrak{c} \frac{n}{\log(n)} \right\},$$

where

$$\mathfrak{c} = \left(6 \wedge \frac{1}{\|f_L + f_R\|_\infty} \right) \frac{1}{48 c_\varphi^2},$$

and f_L, f_R are the densities of L and R. The random set $\widehat{\mathcal{M}_n}$ is defined analogously to $\mathcal{M}_n$ but with $\Psi_{m,Z}$ for $Z = L, R$ replaced by $\widehat{\Psi}_{m,Z}$ and $\mathfrak{c}$ multiplied by 4. We propose to select our model in the following way:

$$\hat{m} = \arg\min_{m \in \widehat{\mathcal{M}_n}} [\gamma_n(\widehat{S}_m) + \text{pen}(m)],$$

with $\text{pen}(m) = \kappa m/n$ and κ is a numerical constant. The constant κ is calibrated on preliminary simulation experiments. Introduce the norm $\| \cdot \|_{L+R}$ such that $\|t\|_{L+R}^2 = \int t^2(x)(f_L(x) + f_R(x))dx$. Denote by S_J the survival function restricted on the domain J, that is $S_J(x) = 0$ for $x \notin J$. Based on results stated in [CGC18], we obtained the following oracle result.

Theorem 4 *Assume that $\int_J S^2(x)(f_L(x) + f_R(x))dx < +\infty$, and $\Phi_m^{(L)\top}\Phi_m^{(L)} + \Phi_m^{(R)\top}\Phi_m^{(R)}$ is invertible. We have*

$$\mathbb{E}[\|\widehat{S}_{\hat{m}} - S_J\|_n^2] \leq C \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in \Sigma_m} \|t - S_J\|_{L+R}^2 + \frac{m}{n} \right) + \frac{C'}{n},$$

where C is a numerical constants and C' is a constant depending on f_L, f_R, $\mathfrak{c}$.

3.3 A penalised event-specific rate model for recurrent events

In our work [P4] we considered a model for the recurrent event process where the rate function depends on prior recurrences. Similarly as in Section 2.3.3 of the Introduction section, the model is defined as:

$$\mathbb{E}[dN^*(t)|Y_s^*(t), X(t)] = Y_s^*(t)\lambda_s(t|X(t))dt, \quad s = 1, 2, \dots,$$

where $Y_s^*(t) = I(N^*(t-) = s-1, T^* \geq t)$ and T^* is a terminal event. We work under the independent censoring assumption of Equation (2.14) and we assume that a maximum of B events can be observed per individual. In practice, the value B must be chosen by the user and this value means that events after the Bth are removed from the study. The rate function is then modelled using either a multiplicative model based on the Cox model,

$$\lambda_s(t|X(t)) = \lambda_0(t, s)\exp(\beta_0(s)X(t)), \tag{3.2}$$

or an additive model based on the Aalen model,

$$\lambda_s(t|X(t)) = \lambda_0(t, s) + (\beta_0(s)X(t)). \tag{3.3}$$

In both models, λ_0 is an unknown baseline and β_0 an unknown parameter that both need to be estimated. Importantly, they both depend on the strata s representing the number of previous recurrent events already experienced by the individual. Model (3.2) was already introduced by [PWP81] while the Aalen model for event-specific data is new. The standard formula for the likelihood (2.3) can be easily extended to the event-specific context. The estimator in the multiplicative model, proposed by [PWP81] is defined as:

$$\begin{aligned}\hat{\beta}_{ES/mult} &\in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{d\times B}} L_n^{PL}(\beta)\\ &= \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{d\times B}} \left[-\frac{1}{n}\sum_{s=1}^{B}\sum_{i=1}^{n}\int_0^\tau \left\{\beta(s)X_i(t) - \log\left(\sum_{j=1}^{n} Y_j^s(t)\exp\left(\beta(s)X_j(t)\right)\right)\right\} Y_i^s(t)dN_i(t)\right].\end{aligned}$$

In the additive model, we can extend the work from [MS09a] and [MS09b] to define the following estimator:

$$\hat{\beta}_{ES/add} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{d\times B}} L_n^{PLS}(\beta) = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{d\times B}} \sum_{s=1}^{B}\left\{\beta(s)\mathbf{H}_n(s)\beta(s)^\top - 2\beta(s)\boldsymbol{h}_n(s)\right\},$$

where for all $s \in \{1, \dots, B\}$, $\mathbf{H}_n(s)$ are $d \times d$ symmetrical positive semidefinite matrices and $\boldsymbol{h}_n(s)$ are d-dimensional vectors equal to

$$\mathbf{H}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i^s(t)\Big(X_i(t)-\bar{X}^s(t)\Big)^{\otimes 2}dt \text{ and } \boldsymbol{h}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i^s(t)\Big(X_i(t)-\bar{X}^s(t)\Big)dN_i(t),$$

with $\bar{X}^s(t) = \sum_{i=1}^n X_i(t)Y_i^s(t)/\sum_{i=1}^n Y_i^s(t)$ and the convention that $0/0 = 0$.

The two estimators $\hat{\beta}_{ES/mult}$ and $\hat{\beta}_{ES/add}$ will be over-parametrised as soon as the number of covariates d and/or the number of strata B is large. If we take for example, as a rule of thumb, the criterion $\sqrt{n} < d \times B$ to determine that a problem is over-parametrised we see that in the Bladder tumour data example of [Bya80], there are four covariates with a maximum of 10 recurrences. Setting the parameter B to 5 gives $4 \times 5 = 20$ parameters that need to be estimated for a total of only 116 patients!

The idea of our penalisation method is to force the effect of a covariate on two consecutive values to be close to each other. This is achieved through a fused-lasso penalty (see for example [TSR+05]). For all $\beta = (\beta(s), s = 1, \dots, B)$ with $\beta(s) = (\beta^1(s), \dots, \beta^d(s))$, define for all $j = 1, \dots, d$

$$\beta^j = (\beta^j(1), \dots, \beta^j(B)) \text{ and } \operatorname{TV}(\beta^j) = \sum_{s=2}^{B} |\beta^j(s) - \beta^j(s-1)| = \sum_{s=2}^{B} |\Delta\beta^j(s)|.$$

We now consider the minimisers of the partial log-likelihood (respectively the partial least-squares) penalised with a covariate specific total variation. Define the penalised estimators in models (3.2) and (3.3) as:

$$\hat{\beta}_{\mathrm{TV}/mult} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{d\times B}} \left\{ L_n^{PL}(\beta) + \frac{\lambda_n}{n} \sum_{j=1}^{d} \operatorname{TV}(\beta^j) \right\} \text{ and}$$
$$\hat{\beta}_{\mathrm{TV}/add} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^{d\times B}} \left\{ L_n^{PLS}(\beta) + \frac{\lambda_n}{n} \sum_{j=1}^{d} \operatorname{TV}(\beta^j) \right\},$$

where λ_n is a tuning penalty parameter that needs to be chosen. It turns out that these penalised algorithms can be rewritten as lasso algorithms which facilitates their implementation. The multiplicative model is implemented from the **coxnet** function in the **glmnet** R package and the additive model is implemented from the **ahazpen** function in the **ahaz** R package. In these packages, the tuning parameter λ_n is chosen by 10-fold cross validation. A more efficient procedure, in terms of selection in consistency is the reweighted lasso (or two steps estimator) derived by [Zou06] or [CWB08]. We will use the two steps procedure in applications but theoretical results are proved only for the initial lasso type estimator. See the supplementary material of [P4] for more details regarding the implementation of our algorithm.

We provide hereafter, the obtained theoretical results for our penalised estimator in the multiplicative model. The results for the additive model are of similar nature and can be found in [P4]. Define first $A_s = \{t : \mathbb{P}[N^*(t-) = s-1, T^* \geq t] > 0\}$ and $\tau > 0$ such that $A_s^\tau = A_s \cap [0, \tau]$ and for all $s = 1, \dots, B$ and t in A_s^τ, suppose that $\mathbb{E}[Y^s(t) > 0]$ and $\mathbb{P}[E(B) \leq \tau] > 0$ where $E(B)$ represents the Bth recurrent event (not always observed). Define also for all $s = 1, \dots, B$ for all $t \geq 0$,

$$s^{(l)}(s, t, \beta) = \mathbb{E}[Y^s(t) X(t)^{\otimes l} \exp(\beta(s) X(t))], l = 0, 1, 2.$$

Introduce $\mathbf{e}(s, t, \beta) = s^{(1)}(s, t, \beta)/s^{(0)}(s, t, \beta)$, $\mathbf{v}(s, t, \beta) = s^{(2)}(s, t, \beta)/s^{(0)}(s, t, \beta) - \mathbf{e}(s, t, \beta)^{\otimes 2}$ and $\boldsymbol{\Sigma}(s, \beta) = \int_{A_s^\tau} \mathbf{v}(s, t, \beta) \mathbb{E}[Y^s(t) dN(t)]$. For any $s = 1, \dots, B$ and for any $t \geq 0$, the three functions $s^{(l)}(s, t, \beta_0)$ are bounded and $\mathbf{e}(s, t, \beta), \mathbf{v}(s, t, \beta)$ and $\boldsymbol{\Sigma}(s, \beta)$ are finite under classical assumptions (see [P4] for more details).

Theorem 5 *Assume that for each $s = 1, \dots, B$, $\boldsymbol{\Sigma}(s, \beta_0)$ is non-singular.*

1. *If $\lambda_n/n \to 0$ as $n \to \infty$ then $\hat{\beta}_{\mathrm{TV}/mult}$ converges to β_0 in probability.*

2. *If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ as $n \to \infty$ then $\sqrt{n}(\hat{\beta}_{\mathrm{TV}/mult} - \beta_0)$ converges in distribution to*

$$\operatorname*{arg\,min}_{u \in \mathbb{R}^{d\times B}} \Lambda_{mult}(u) = \operatorname*{arg\,min}_{u \in \mathbb{R}^{d\times B}} \Big[\sum_{s=1}^{B} \Big\{ \frac{1}{2} u(s)^\top \boldsymbol{\Sigma}(s, \beta_0) u(s) - u(s)^\top \xi_{mult}(s) \Big\}$$
$$+ \lambda_0 \sum_{j=1}^{d} \sum_{s=2}^{B} \Big\{ |\Delta u^j(s)| I(\Delta\beta_0^j(s) = 0) + sgn(\Delta\beta_0^j(s))(\Delta u^j(s)) I(\Delta\beta_0^j(s) \neq 0) \Big\} \Big],$$

and for each s, $\xi_{mult}(s)$ is a centred d-dimensional gaussian vector with covariance matrix equal to

$$\mathbb{E}\left[\left(\int_{A_s^\tau}(X(t)-\mathbf{e}(s,t,\beta_0))\,Y^s(t)dM^s(t)\right)^{\otimes 2}\right].$$

The process M^s in the theorem is centred and defined for all $s=1,\ldots,B$ and t in A_s^τ, by

$$M^s(t)=N(t)-\int_0^t \mathbb{E}\big[dN(r)|X(r),T\wedge C\geq r,N(r-)=s-1\big].$$

As explained in the introduction section, the process M^s is not a martingale due to the definition of the rate function which do not condition on all the entire history of the recurrent event process. Therefore empirical processes theory is needed. A class of function defined as integrals with respect to dM^s are shown to be Donsker using results from [VDVW96] and central limit theorem types are derived for this class of functions. Then, consistency and asymptotic normality are proved in a similar manner as in [KF00].

We finally illustrate the performance of our penalised algorithm on the bladder tumour cancer data of [Bya80]. The dataset is composed of 116 patients with 47 patients from the placebo group, 38 from the thiotepa group and 31 from the pyridoxine group. For interpretation purpose, the treatment variable is coded as two new binary variables, pyridoxine and thiotepa, making placebo the reference. On these patients, since 13.79% experienced at least five tumour recurrences and only 6.9% patients experienced six tumour recurrences or more, we set the parameter B to 5. In addition to these two treatment variables two supplementary covariates were recorded for each patient: the number of initial tumours and the size of the largest initial tumour. Figure 3.1 display the estimates obtained from the constant coefficient, unconstrained, total variation and two steps total variation estimators in the multiplicative model. The constant coefficient estimator is obtained in the simpler multiplicative model where the β parameter does not depends on the number of previous recurrences s. The unconstrained estimator shows very strong variations and is not interpretable as such. On the other hand, the constant coefficient estimator gives valuable information on the impact of each covariate, but in turn cannot detect a change in variation. Our two steps total-variation estimator reaches a compromise between interpretability and detection of variations in the effect of each covariate. Indeed, a very interesting result from our estimator on the Byar dataset comes from the effect of pyrodixine. While previous studies in the literature have resulted in contradictory conclusions about the efficiency of this treatment, our estimator shows that this treatment has actually a protective effect for the first three recurrences but then the risk of further recurrences is increased by this treatment. The same pattern is observed for the Aalen model, see [P4] for more details.

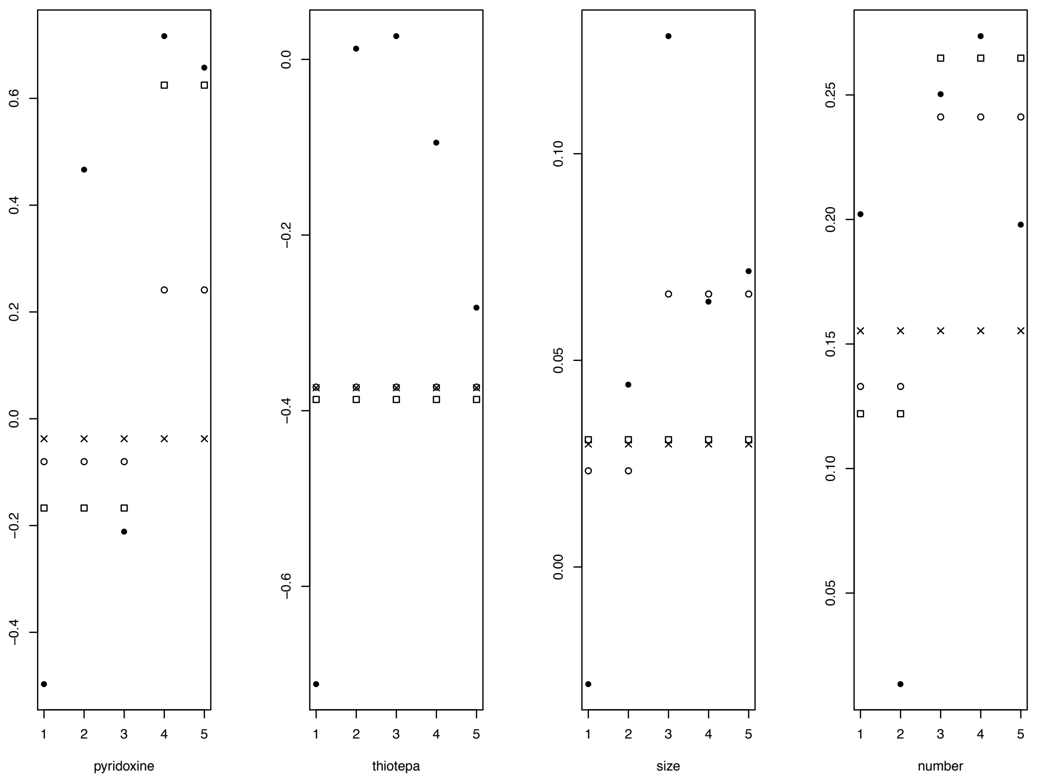


Figure 3.1: Estimates for the bladder data in the multiplicative model. The crosses represent the constant estimator, the filled circles the unconstrained estimator, the circles the total variation estimator and the squares the two steps total variation estimator.

CHAPTER 4

Computational statistical methods

In this chapter I present my contributions to new computational statistical methods. This chapter refers to the published papers [P5], [P6], [P7] and to the submitted papers [S2] and [S3]. In the first section, the paper [P5] presents a new method to detect heterogeneity for time to event data using breakpoint models. In paper [P6] the adaptive ridge algorithm of [RME12] and [FN16] is extended to the survival analysis context. This algorithm has then been used as a regularisation method in [S2] where the hazard is computed as a bi-dimensional function. When using the Cox model, it can also be used to model the baseline hazard function as a piecewise constant function with automatic choice of the cuts. This is especially useful when dealing with complex type of data. This is the purpose of [S3] which deals with interval censored data. Finally, the paper [P7] discusses risk predictions for developing genetic diseases when taking into account the family history. This article extends classical methods by taking into account the competing risk of death which is not negligible when dealing with diseases with possibly late age at onset such as cancer diseases.

4.1 Heterogeneity in survival analysis

Heterogeneity in survival analysis arises when the observed covariates do not properly account for all the variability in the survival distribution. This might typically be due to unobserved covariates or to individual specific variability such as it occurs in clustered data. In the latter case, a popular approach is to use frailty models which incorporate a random effect to capture individual variations. In this model, groups appartenance are known and the group effects are modelled through the random effect. A similar approach in an unsupervised context is the cure model, which assumes that the population is composed of two groups: the susceptibles which are at risk of developing the event and the non susceptibles which will never experience the event. While theses models have proved to be most useful, it is however likely that unaccounted latent heterogeneity remains in the survival signal. This might be due for example to an unknown interaction between a treatment and some exposure, or to some unaccounted heterogeneity of the disease itself (for example an unknown cancer sub-type). For instance, age at diagnosis might be associated with a higher chance to receive a new treatment or BMI might be associated with a specific exposure.

In [P5], we suggested a new approach considering survival heterogeneity as a breakpoint model in an ordered sequence of survival responses. The survival responses might be ordered according to any numerical covariate (ties are possible) like age at diagnosis, BMI, etc. The basic idea being that heterogeneity will be detected as soon as it is associated with the chosen covariate. From a statistical point of view we consider this situation as a change-point model where abrupt changes occur in terms of baseline hazard rates and/or in terms of proportional factors. In such a model, we aim at two objectives: first we want to estimate the hazard rates and the proportional factors in each homogeneous region through a Cox model considering parametric baseline hazards or a nonparametric baseline hazard. Secondly, we want to accurately provide the number and location of the breakpoints. A constrained Hidden Markov Model (HMM) method was suggested in the context of breakpoint analysis by [LRN13]. This method allows to perform a full change-point analysis in a segment-based model (one parameter by segment) providing linear EM (see [DLR77]) estimates of the parameter and a full specification of the posterior distribution of change points. In paper [P5] we adapted this method to the context of survival analysis with hazard rate estimates, where the estimation is performed through the EM algorithm to provide update of the estimates and the posterior distribution at each iteration step.

We suppose that the population is composed of K segments such that for $i = 1, \dots, n$, $R_i \in \{1, 2, \dots, K\}$ and R_i represents subject segment allocation. The R_is are unobserved meaning that we do not know in advance the segment allocations. Without loss of generality, we also assume that the R_is are ordered. For example, if the population is a mixture of three subpopulations such that we have $n = 10$ and two breakpoints occurring after positions 3 and 7 then $R_{1:10} = 1112222333$. Following the notations from the Introduction section, the model is defined as

$$\mathbb{E}[dN^*(t)|Y^*(t), X, R] = Y^*(t) \sum_{k=1}^{K} \lambda_k(t) \exp(X\beta_k) I(R = k) dt,$$

where the λ_k represent unknown baseline hazard functions and the β_k unknown regression parameters associated to each segment index. Therefore we assume Cox models on each unobserved segment, where baselines and/or regression coefficients can differ on each segment. Let $\Lambda_k(t) = \int_0^t \lambda_k(s) ds$ represents the cumulative baseline hazard function of the kth segment index. We denote by $\boldsymbol{\theta} = (\Lambda_1, \dots, \Lambda_K, \beta_1, \dots, \beta_K)$ the model parameter to be estimated.

In this model, the contribution of the ith individual to the likelihood $e_i(k; \boldsymbol{\theta})$ can be easily computed. From standard arguments on likelihood constructions in the context of survival analysis, see for instance [ABGK93], we have under independent and non informative censoring:

$$\log e_i(k; \boldsymbol{\theta}) = \int_0^\tau \{\log(\lambda_k(t)) + X_i\beta_k\} dN_i(t) - \int_0^\tau Y_i(t)\lambda_k(t) \exp(X_i\beta_k) dt,$$

where the equality holds true up to a constant that does not depend on the model parameter $\boldsymbol{\theta}$. Since the segment indexes are not observed, standard likelihood approaches cannot be directly implemented. To overcome this problem, an Expectation-Maximisation (EM) algorithm procedure is used. It consists in performing alternatively until convergence the following two-steps.

Expectation Step: compute the conditional expected log-likelihood,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \int_{R_{1:n}} \mathbb{P}(R_{1:n}|\text{data}; \boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(R_{1:n}, \text{data}; \boldsymbol{\theta}) dR_{1:n}$$

where $\boldsymbol{\theta}_{\text{old}}$ denote the previous value of the parameter and data $= (T_{1:n}, \Delta_{1:n}, X_{1:n})$.

Maximisation Step: update the parameter with

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}).$$

Assuming that the prior segmentation distribution $\mathbb{P}(R_{1:n};\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, it can be shown that:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_i(k;\boldsymbol{\theta}_{\text{old}}) \log e_i(k;\boldsymbol{\theta})$$

where for any $i \in \{1,\dots,n\}$, $k \in \{1,\dots,K\}$ and $\boldsymbol{\theta}$ we define:

$$w_i(k;\boldsymbol{\theta}) = \mathbb{P}(R_i = k|\text{data};\boldsymbol{\theta}).$$

In order to perform the E step we therefore need to compute the weights $w_i(k;\boldsymbol{\theta})$. This is achieved using Hidden Markov Models theory with the additional constraint that $R_n = K$. We start by choosing a prior: $\eta_i(k) = \mathbb{P}(R_i = k+1|R_{i-1} = k)$, such as a uniform prior. Introduce for all $i \in \{1,\dots,n\}$ and $k \in \{1,\dots,K\}$, $F_i(k;\boldsymbol{\theta}) = \mathbb{P}(\text{data}_{1:i}, R_i = k;\boldsymbol{\theta})$ and $B_i(k;\boldsymbol{\theta}) = \mathbb{P}(\text{data}_{i+1:n}, R_n = K|R_i = k;\boldsymbol{\theta})$ the forward and backward quantities. These quantities can be computed recursively using the following formulas:

$$F_i(k;\boldsymbol{\theta}) = F_{i-1}(k-1;\boldsymbol{\theta})\eta_i(k-1)e_i(k;\boldsymbol{\theta}) + F_{i-1}(k;\boldsymbol{\theta})(1-\eta_i(k))e_i(k;\boldsymbol{\theta}), \tag{4.1}$$

$$B_{i-1}(k;\boldsymbol{\theta}) = (1-\eta_i(k))e_i(k;\boldsymbol{\theta})B_i(k;\boldsymbol{\theta}) + \eta_i(k)e_{i+1}(k+1;\boldsymbol{\theta})B_i(k+1;\boldsymbol{\theta}), \tag{4.2}$$

and we can derive from them posterior distributions of interest:

$$\mathbb{P}(R_i = k|\text{data};\boldsymbol{\theta}) = w_i(k;\boldsymbol{\theta}) \propto F_i(k;\boldsymbol{\theta})B_i(k;\boldsymbol{\theta}), \tag{4.3}$$

$$\mathbb{P}(\text{BP}_k = i|\text{data};\boldsymbol{\theta}) \propto F_i(k;\boldsymbol{\theta})\eta_{i+1}(k)e_{i+1}(k+1;\boldsymbol{\theta})B_{i+1}(k+1;\boldsymbol{\theta}), \tag{4.4}$$

where $\{\text{BP}_k = i\} = \{R_i = k, R_{i+1} = k+1\}$. It is hence clear that Equation (4.3) allows to compute the marginal weights used in the EM algorithm while Equation (4.4) gives the marginal distribution of the k^{th} breakpoint. Note that the full posterior segmentation distribution can be proved to be an heterogeneous Markov chain which transition can be derived immediately from Equations (4.3) and (4.4) (see [LRN13] for more details).

Next, by observing that the quantity $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ corresponds to a weighted Cox likelihood, the maximisation step can be performed using Newton-Raphson algorithm in the usual way.

Finally, the number of segments is determined from a Bayesian Information Criteria (BIC). Note that the likelihood can also be derived from the forward-backward quantities and for any $i \in \{1,\dots,n\}$ as:

$$\mathbb{P}(\text{data}|\boldsymbol{\theta}) = \frac{\sum_{R_{1:n}} \mathbb{P}(\text{data}, R_{1:n}, R_n = K|\boldsymbol{\theta})}{\sum_{R_{1:n}} \mathbb{P}(R_{1:n}, R_n = K|\boldsymbol{\theta})} = \frac{\sum_{k=1}^{K} F_i(k;\boldsymbol{\theta})B_i(k;\boldsymbol{\theta})}{\sum_{k=1}^{K} F_i^0(k)B_i^0(k)},$$

where F^0 and B^0 are obtained through recursions (4.1) and (4.2) by replacing all $e_i(k;\boldsymbol{\theta})$ by 1. These quantities depend only on η, n and K, thus they do not need to be updated during the EM algorithm. The BIC is computed using the likelihood from the previous formula.

The method is illustrated on the Steno Memorial hospital dataset from [ABGK93]. These data concern diabetic patients and the time to event of interest is the time from diagnosis to death. These data are left-truncated as individuals did not reach the hospital directly after being diagnosed of diabetes and a diabetic patient that died before being included in the hospital dataset will never be observed. Nevertheless, our method easily accommodates for left truncation by modifying the at risk process in the weighted log-likelihood function Q. From the BIC the model with two breakpoints is chosen. The marginal distribution of the breakpoints were computed using Formula (4.4); the modes are located at years 1948 and 1962. Weighted Kaplan-Meier curves were also plotted, using the weights $w_i(k;\hat{\boldsymbol{\theta}})$ obtained at convergence of the EM algorithm. All these plots are displayed in Figure 4.1.

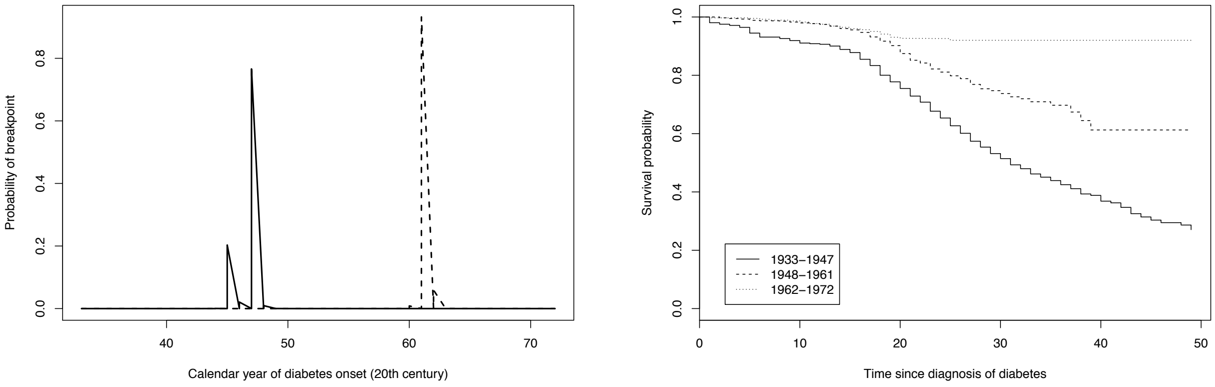


Figure 4.1: Marginal distribution of breakpoints in the Danish diabetes dataset on the left panel. The modes are located in 1948 and 1962. Weighted Kaplan-Meier estimators in the models with two breakpoints in the Danish diabetes dataset on the right panel.

4.2 The adaptive ridge method for piecewise constant hazard in survival analysis

In survival analysis, when interest lies on the estimation of the hazard rate, an attractive and popular model is the piecewise constant hazard (pch) model. This model is easy to interpret as the hazard rate is supposed to be constant on some pre-defined time intervals and plotting the hazard rate gives a quick sense of the evolution of the event of interest through time. Many epidemiological studies use this model to represent the hazard rate function either because it provides an interesting way to fit the hazard function or because the data are not available on the individual level.

While this model can be used in a nonparametric setting, it is often used in combination with covariates effects. This is the case for instance for the so called Poisson regression model (see [CH93] or [ABG08]) which assumes a proportional effect on the covariates and a piecewise constant hazard model for the baseline hazard. This model is widely used in practice typically when dealing with register data.

When modelling covariates effect through a proportional hazard model, the Cox model allows the baseline to stay unspecified. Through the Cox partial likelihood the regression effect can be estimated separately from the baseline. While this is a very interesting aspect of the Cox model, this nice separation between baseline estimation and regression effect estimation does not hold anymore in many extensions of this model. For instance, in frailty models (see among many other authors [Cla78], [Hou95], [TG00] and [RP00]) keeping a non-parametric baseline makes the estimation method much more complicated since baseline and regression parameters must be estimated simultaneously. In the joint modelling framework where one wants to model the association between a longitudinal variable and a time to event response through a random effect (see [TD04], [Riz12]), only parametric baseline functions are implemented in the widely used **jm** R package (see [Riz10]). As a matter of fact, the author in [Riz12] recommends either to use the piecewise constant baseline hazard or a spline basis baseline hazard which he says "*often work quite satisfactorily in practice*" (see page 53 of the book). The **frailtypack** R package (see [RMG12]) deals with more survival analysis situations involving a random effect such as nested frailty models (see [RFJ06]) or joint inference of recurrent and terminal events (see [RMPJG$^+$07]). In this package, the possible baseline hazard functions are the piecewise constant hazard, Weibull hazard and spline functions. However the use of spline baseline functions requires to specify in advance the number of knots used in the estimation and therefore can be seen as a smoothed version of the piecewise constant hazard functions where one must choose in advance the number of cuts.

Other contexts where the partial likelihood approach does not work anymore include the cure models framework (see for instance [Far82a] and [ST00]) and the analysis of interval-censoring data (see [Sun07] for instance). In the latter case, the nonparametric maximum likelihood estimator for the cumulative hazard or the survival function is known to be slow with a convergence rate of order $n^{-1/3}$ and the limiting distribution is not Gaussian (see [GW92b] for current status data and [Gro96] for case II intervals censored data). This problem pertains in the regression framework (see sections 5.2.3 and 6.2.2 in [Sun07] for instance). On the other hand, using parametric baseline functions such as the piecewise hazard functions allows to obtain classical parametric rate of convergence and makes the estimation procedure much more stable.

In our paper [P6], we only considered a setting without covariates and the aim was to estimate the baseline hazard function in a piecewise constant hazard model in the situation of right-censored data. We proposed a new method to automatically find the appropriate number and location of the cuts used in this model. Our algorithm is based on the works from [RME12] and [FN16] where starting from a large set of possible cut points an L0 penalty on the likelihood of the model forces many successive cuts to be equal providing a parsimonious estimate of the hazard function. The procedure is data-driven and inference taking into account both the variability from the estimates and the cut points positions can be derived. This penalised algorithm has also been applied to the context of age-period-cohort estimation in [S2] and to the interval-censoring problem in [S3].

4.2.1 The adaptive ridge algorithm in the absence of covariates

In this section we briefly explain how this adaptive ridge estimator works when there are no covariates available. More details can be found in [P6].

First, the hazard function is assumed to be piecewise constant on K cuts represented by $c_0, c_1, \ldots, c_K$, with the convention that $c_0 = 0$ and $c_K = +\infty$. Let $I_k(t) = I(c_{k-1} < t \leq c_k)$. We suppose that

$$\lambda(t) = \sum_{k=1}^{K} I_k(t) \exp(a_k),$$

for $k = 1, \ldots, K$. Note that the exponential baseline hazard is obtained from $K = 1$ in the piecewise constant hazard family. Denote by $\boldsymbol{a} = (a_1, \ldots, a_K)$ the model parameter we aim to estimate and let $L_n(\boldsymbol{a}) = \log \prod_{i=1}^n \mathbb{P}[T_i, \Delta_i; \boldsymbol{a}]$ represents the log-likelihood of the model. We have:

$$L_n(\boldsymbol{a}) = \sum_{i=1}^{n} \left\{ \log\left(\lambda(T_i)\right)\Delta_i - \int_0^{T_i} \lambda(t)dt \right\},$$

where the equality holds true up to a constant that does not depend on the model parameter $\boldsymbol{a}$. For computational purpose, it is interesting to note that the log-likelihood can be written in a Poisson regression form. Introduce $R_{i,k} = I(T_i \geq c_{k-1})(c_k \wedge T_i - c_{k-1})$, the total time individual i is at risk in the kth interval $(c_{k-1}, c_k]$, $O_{i,k} = I_k(T_i)\Delta_i$, the number of events for individual i in the kth interval. Also $R_k = \sum_{i=1}^n R_{i,k}$ and $O_k = \sum_{i=1}^n O_{i,k}$ are sufficient statistics and estimation can be carried out using only these two statistics. The log-likelihood can then be written again as (see [ABG08] p.223-225 for more details):

$$L_n(\boldsymbol{a}) = \sum_{k=1}^{K} \{O_k a_k - \exp(a_k) R_k\}. \tag{4.5}$$

Since L_n is concave, the maximum likelihood estimator has an explicit solution, obtained by maximisation of the log-likelihood: for $l = 1, \dots, L$,

$$\hat{a}_k = \log\left(\frac{O_k}{R_k}\right) \cdot \tag{4.6}$$

Now, we aim at using this estimator in the case where the number of cuts and their locations are unknown. We start with a large grid of cuts and we propose a penalised version of the piecewise constant hazard (pch) model which allows to simultaneously determine the locations of the cuts and the estimated values of the $\hat{a}_k$s. Based on the work from [FN16], we propose the following penalised log-likelihood:

$$L_n^{\text{pen}}(\boldsymbol{a}, \boldsymbol{w}) = \sum_{l=1}^{K} \{O_k a_l - \exp(a_k) R_k\} - \frac{\text{pen}}{2} \sum_{l=1}^{K-1} w_k (a_{k+1} - a_k)^2,$$

where $\boldsymbol{w} = (w_1, \dots, w_{K-1})$ are non-negative weights that will be iteratively updated in order for the weighted ridge penalty term to approximate the L0 penalty. The penalisation term is designed to force consecutive values of the a_ks to be close to each other. 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 Equation (4.5) and to the exponential model.

The score vector is denoted $U(\boldsymbol{a}, \boldsymbol{w}) = \partial L_n^{\text{pen}}(\boldsymbol{a}, \boldsymbol{w}) / \partial \boldsymbol{a}$ and its kth component, $k \in \{1, \dots, K\}$, is equal to:

$$O_k - R_k \exp(a_k) + (w_{k-1} a_{k-1} - (w_{k-1} + w_k) a_k + w_k a_{k+1}) \text{pen},$$

with the convention $w_0 = w_K = a_0 = a_{K+1} = 0$. Now introduce $I(\boldsymbol{a}, \boldsymbol{w}) = -\partial U(\boldsymbol{a}, \boldsymbol{w}) / \partial \boldsymbol{a}^T$, the opposite of the Hessian matrix. $I(\boldsymbol{a}, \boldsymbol{w})$ is a $K \times K$ non-negative definite band matrix whose bandwidth equals 1. Its diagonal elements are equal to

$$I(\boldsymbol{a}, \boldsymbol{w})_{k,k} = R_k \exp(a_k) + (w_{k-1} + w_k) \,\text{pen},$$

other elements next to the diagonal are defined for $k = 1, \dots, K-1$ by

$$I(\boldsymbol{a}, \boldsymbol{w})_{k,k+1} = I(\boldsymbol{a}, \boldsymbol{w})_{k+1,k} = -w_k \,\text{pen},$$

and all other elements are equal to zero, that is for k, k' such that $|k - k'| \geq 2$, $I(\boldsymbol{a}, \boldsymbol{w})_{k,k'} = 0$.

The vector parameter $\boldsymbol{a}$ is updated using the Newton-Raphson algorithm. For a given sequence of weights $\boldsymbol{w}^{(m-1)}$ obtained at the $(m-1)$th step, the mth Newton Raphson iteration step is obtained from the equation

$$\boldsymbol{a}^{(m)} = \boldsymbol{a}^{(m-1)} + I(\boldsymbol{a}^{(m-1)}, \boldsymbol{w}^{(m-1)})^{-1} U(\boldsymbol{a}^{(m-1)}, \boldsymbol{w}^{(m-1)}). \tag{4.7}$$

The inversion of the band matrix is performed through a fast (linear complexity) C++ implementation of the well-known LDL algorithm (variant of the LU decomposition for symmetric matrices). The complexity of this inversion is $\mathcal{O}(K)$. Initialisation of the Newton-Raphson algorithm can be obtained from the classical unpenalised estimator of the piecewise constant hazard model, that is $\boldsymbol{a}^{(0)} = \arg\max_a L_n(\boldsymbol{a})$.

Once the Newton-Raphson algorithm has reached convergence, the weights are updated at the mth step from the equation

$$w_l^{(m)} = \left((a_{k+1}^{(m)} - a_k^{(m)})^2 + \varepsilon^2\right)^{-1},$$

for $k = 1, \dots, K-1$ with $\varepsilon = 10^{-5}$ and where the $a_k^{(m)}$s represent the estimates of the a_ks 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 $|a_{k+1} - a_k| < \varepsilon$ and close to 1 when $|a_{k+1} - a_k| > \varepsilon$. Hence the penalty term tends to approximate the L0 norm. The weights are initialised by $w_k^{(0)} = 1$, 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 a_ks verifying $|a_{k+1} - a_k| > \varepsilon$. The non-penalised log-likelihood L_n is then maximised using this set of cuts and the final estimate is derived from Equation (4.6). It is important to stress that the penalised likelihood is used only to select a set of cuts. Reimplementing the non-penalised estimator in the final step enables to reduce the bias classically induced by penalised maximisation techniques (for more details about the adaptive ridge procedure see [RME12] or [FN16]).

For a given penalty value, the algorithm can be summarised by the following steps:

Step 0. Initialise the weights to $w_k^{(0)} = 1$, and initialise the hazard values from the unpenalised estimator, $\boldsymbol{a}^{(0)} = \arg\max_a L_n(\boldsymbol{a})$. Set $m = 1$ and $\text{sel}_k^1 = 0$.

Step 1. Compute the penalised estimator $\boldsymbol{a}^{(m)}$ from the Newton-Raphson algorithm (4.7). After convergence go to the next step.

Step 2. Update the weights: $w_k^{(m)} = \left((a_{k+1}^{(m)} - a_k^{(m)})^2 + \varepsilon^2\right)^{-1}$, $k = 1, \dots, K-1$. Define $\text{sel}_k^{(m)} = w_k^{(m)}(a_{k+1}^{(m)} - a_k^{(m)})^2$:

- If $\max_k |\text{sel}_k^{(m)} - \text{sel}_k^{(m-1)}| > 10^{-5}$, set $m = m+1$ and go to Step 1.
- If $\max_k |\text{sel}_k^{(m)} - \text{sel}_k^{(m-1)}| < 10^{-5}$, select all the c_ks for which $\text{sel}_k^{(m)} > 0.99$. Exit the algorithm.

This algorithm provides a selection of the cuts. From these cuts, an unpenalised estimator $\hat{\boldsymbol{a}}$ is implemented from Equation (4.6). Then a Bayesian Information Criteria (BIC) is calculated using the following formula:

$$\text{BIC} = -2L_n(\hat{\boldsymbol{a}}) + K^* \log(n),$$

where $K^* = \sum_k I(\text{sel}_k^{(m)} > 0.99)$ represents the dimension of the selected model. From a large grid of penalty values, the procedure is iterated and the penalty is chosen as the one that minimises the BIC. An important feature of the procedure is to use a warm start when the penalty value changes. At the beginning of the algorithm, the term $\boldsymbol{a}^{(0)}$ should be initialised from the estimator obtained at the previous value of the penalty. Using this warm start ensures the algorithm to converge very quickly when the number of selected cuts does not change for a new penalty. As a result, the global algorithm takes most of its time for the first penalty value.

Finally, we briefly illustrate how to take into account the uncertainty in the choice of the cut points and in the estimated values in the construction of pointwise confidence intervals. This methodology is performed for the estimation of the survival function on a single data example of size $n = 100$ with 38% of censoring. We use a resampling technique where for each sample a different penalty term can be chosen from the BIC. This provides a new hazard and survival estimates with a different set of cut points for each sample. Taking the quantiles of order 0.025 and 0.975 at each time point allows us to obtain 95% pointwise confidence intervals for the survival function. Interestingly, this resampling technique also allows us to compute an alternative pointwise estimate of the survival function by taking the pointwise medians of each bootstrap sample. This provides a very smooth estimate function and, in that sense, this kind of estimate can be seen as a smooth non-parametric estimate of the survival function. Following this methodology, the survival curve is plotted in Figure 4.2 along with its 95% pointwise confidence interval from 100 bootstrap samples. Our method shows very little difference from the classical Kaplan-Meier estimate and its pointwise confidence interval. Interestingly, our survival estimator and its pointwise confidence intervals have a smooth shape in contrast with the stepwise shape of the Kaplan-Meier estimator.

The regularisation path of the adaptive ridge estimator is illustrated on the more specific setting of age-period-cohort analysis (see [S2]). In this work, a ridge alternative of the algorithm is also implemented which is based on the adaptive ridge algorithm with the weights w_k simply equal to 1.

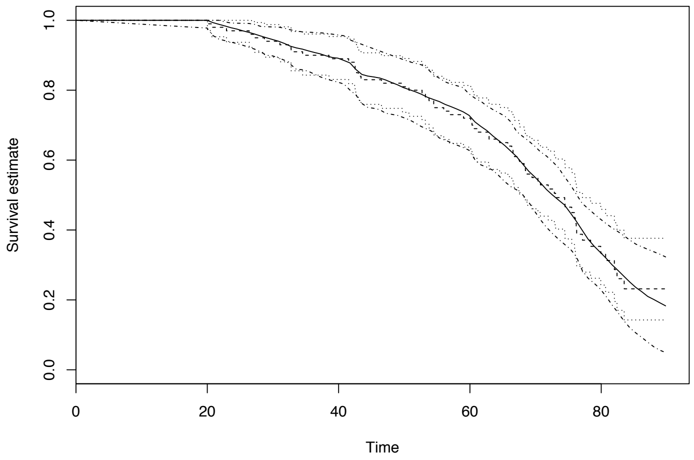


Figure 4.2: Estimates of the survival function. Dashed line: Kaplan Meier estimator along with its 95% pointwise confidence interval (dotted lines). Solid line: bootstrapped adaptive ridge estimator along with its 95% pointwise confidence interval (dot dash lines).

4.2.2 Extensions of age-period-cohort models

In epidemiological or demographic studies, with variable age at onset, individuals are recruited and followed-up during a long period of time, usually from birth. The data are then reported in the form of registers which contain the number of observed cases and the number of individuals at risk to contract a disease. These types of studies are of great interest for the statistician, especially when the event of interest will tend to occur at late ages, such as in cancer studies. However, these data are usually highly heterogeneous in terms of dates of birth and with respect to the calendar time. In such cases, it is therefore very important to take into account the variability of the age, the cohort effect (date of birth) and the period effect (the calendar time) in the hazard rate estimation. This is usually done using age-period-cohort estimation methods (see [YL13] and citations therein).

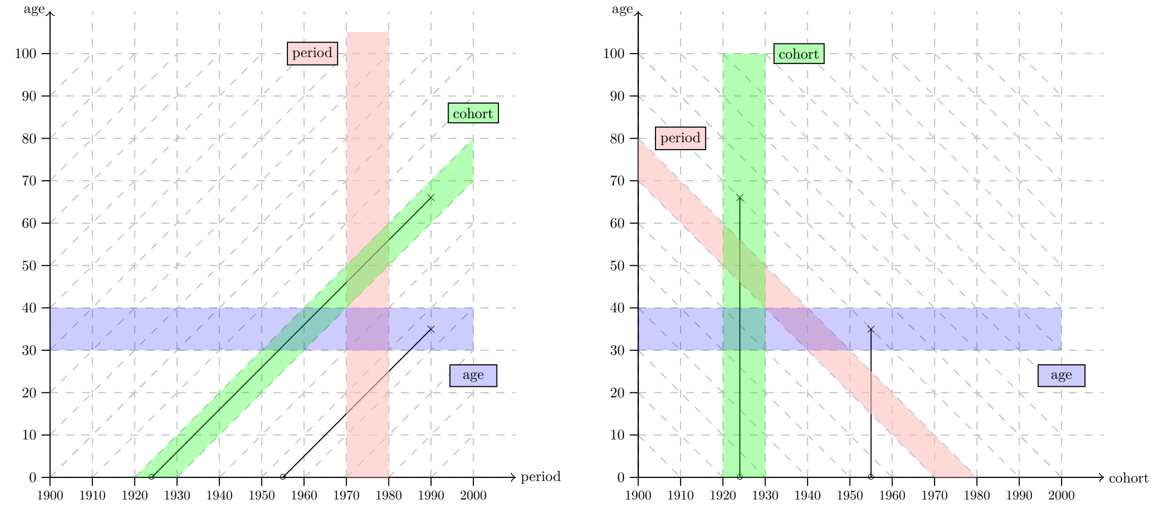


Figure 4.3: Lexis diagram. Age-Period diagram on the left panel and Age-Cohort diagram on the right panel.

In Figure 4.3 we see a typical Lexis diagram in the age-period or age-cohort plane. From this figure, one can discretise the hazard rate in JK intervals for example as

$$\log\left(\lambda(\text{age}, \text{cohort})\right) = \sum_{j=1}^{J}\sum_{k=1}^{K} \eta_{j,k} I(c_{j-1} \le \text{age} < c_j, d_{k-1} \le \text{cohort} < d_k), \tag{4.8}$$

where the c_js are the age intervals, the d_ks the cohort intervals and the $\eta_{j,k}$ are parameters to be estimated. The formula could be alternatively written as a function of age and period or as a function of cohort and period. In standard age-period-cohort analysis, there are three different effects of interest: the age effect α_j, the cohort effect β_k and the period effect γ_{j+k-1}. However fitting a model with all three effects such as

$$\eta_{j,k} = \mu + \alpha_j + \beta_k + \gamma_{j+k-1},$$

will clearly induce identifiability issues due to the relationship: period=age+cohort. As a consequence, [OG82] have proposed to fit each sub-model (age-cohort, age-period and period-cohort) and to use a weighting procedure to combine the three models. Different constraints have also been proposed to make the age-period-cohort model identifiable. However, as noticed by [Heu97], the obtained estimates highly depend on the choice of the constraints. [Hol83] proposed to directly estimate the linear trends of each effect but this procedure leads to results that are difficult to interpret. See [Car07] for a detailed discussion of the identifiability problem of the age-period-cohort model. Another method where the second order derivatives of the three effects are estimated is implemented in the **apc** package from [Nie15]. Finally, in the **epi** R package the method from [CPLH18] is implemented, where one submodel (say age-cohort) is fit and the period effect over the residuals of the first model is then implemented. Again, all these approaches lead to results that can be hard to interpret.

In [S2] we considered the nonparametric approach where the hazard rate is simply a bivariate function of either age-cohort, age-period or period-cohort with no specific structure of the hazard being assumed. Inference is made in two dimensions, but through the linear relationship period = age + cohort, the hazard rate can be represented as a function of any two of the three variables. Of course, for moderate sample sizes, these kind of approaches will suffer from overparametrisation. As a consequence, regularised methods have been proposed in order to avoid overfitting in this non-parametric context. A kernel-type estimator was proposed by [Ber81] and [MU90] where the cumulative hazard is smoothed using a kernel function. See [Kei90] for a thorough discussion of methods for hazard inference in age-period-cohort analysis. In our work, the adaptive ridge algorithm presented in Section 4.2.1 is used to take into account the issue of overfitting. This method results in a nice segmentation of the hazard rate into constant areas.

By convention, we consider the age-cohort Model (4.8). As previously, we introduce the sufficient statistics $O_{j,k}$ which represents the number of observed events in the rectangle (j,k) and $R_{j,k}$ which represents the total time at risk spent in the rectangle (j,k). The log-likelihood is then equal to:

$$\ell_n(\boldsymbol{\eta}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\left\{O_{j,k}\eta_{j,k} - \exp(\eta_{j,k})R_{j,k}\right\},$$

and the maximum likelihood estimator is equal to

$$\eta_{j,k}^{\text{mle}} = \log\left(\frac{O_{j,k}}{R_{j,k}}\right).$$

In this model there are JK parameters that need to be estimated which will usually lead to overfitting issues. We therefore consider the penalised log-likelihood

$$\ell_n^{\text{pen}}(\boldsymbol{\eta}) = \ell_n(\boldsymbol{\eta}) - \frac{\text{pen}}{2}\left\{\sum_{j,k} v_{j,k}\left(\eta_{j+1,k}-\eta_{j,k}\right)^2 + w_{j,k}\left(\eta_{j,k+1}-\eta_{j,k}\right)^2\right\},$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ represent weights and pen is a tuning penalty term. Again, maximisation of the penalised likelihood is performed using the Newton-Raphson algorithm but this time the complexity for the inversion of the Hessian matrix is $\mathcal{O}(\min(J,K))$. This should be compared to a full rank Hessian matrix whose complexity would be $\mathcal{O}(J^2K^2)$. In the algorithm, the bi-dimensional segmentation is performed by representing the connex components induced by the values of $v_{j,k}\left(\eta_{j+1,k}-\eta_{j,k}\right)^2$ and $w_{j,k}\left(\eta_{j,k+1}-\eta_{j,k}\right)^2$. Figure 4.4 shows how the connex components graph is constructed.

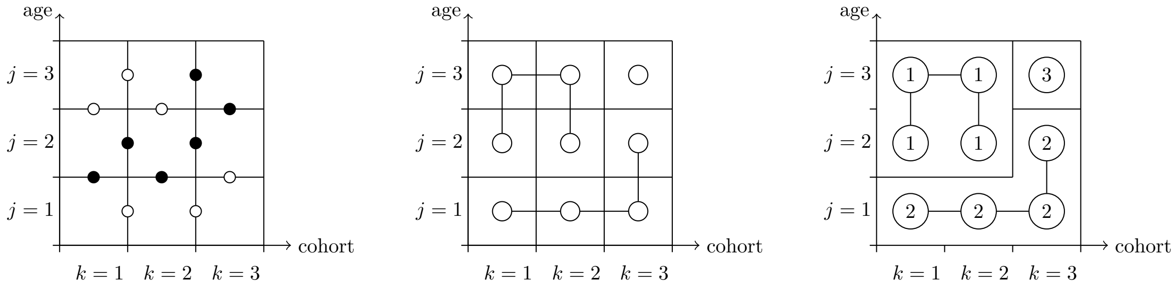


Figure 4.4: Representation of $v_{j,k}\left(\eta_{j+1,k}-\eta_{j,k}\right)^2$ and $w_{j,k}\left(\eta_{j,k+1}-\eta_{j,k}\right)^2$ on the left panel where empty circles correspond to the value 0 and filled circles correspond to the value 1; corresponding graph on the middle panel and segmentation through connected components on the right panel.

Once the segmentation is chosen, the unpenalised estimator is implemented using the chosen segmentation. The two regularisation paths for the whole algorithmic procedure (choice of the segmentation and computation of the final estimator) are illustrated on Figure 4.5 for the ridge estimator (with weights $v_{j,k}$ and $w_{j,k}$ equal to 1) and for the adaptive ridge estimator. We see in particular that, as the value of pen increases, the number of constant areas for the adaptive ridge decreases. Both methods converge to the same constant estimator as pen tends to infinity.

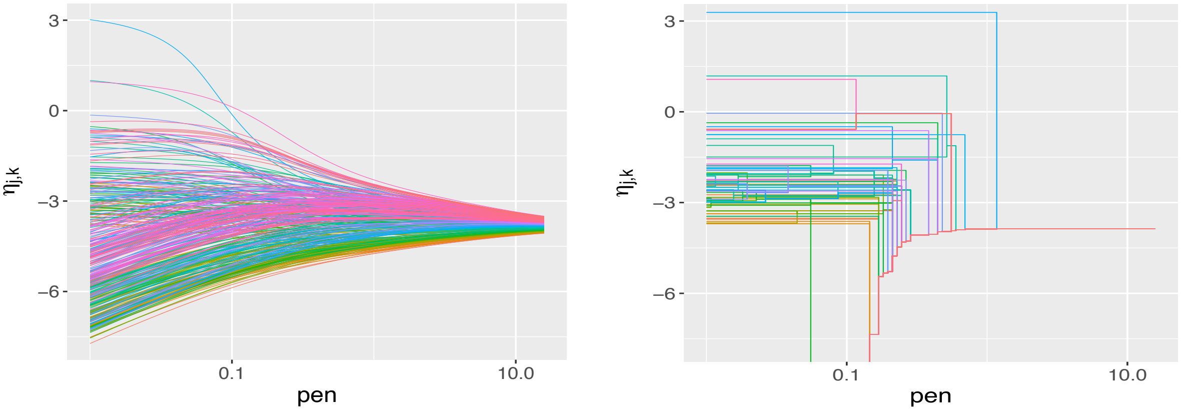


Figure 4.5: Regularisation path for the ridge algorithm on the left panel and for the adaptive ridge algorithm on the right panel.

Finally the penalty term is chosen using the Extended Bayesian Information Criterion (EBIC). This criterion is more parsimonious than the standard BIC, see [CC08] for more details about the EBIC.

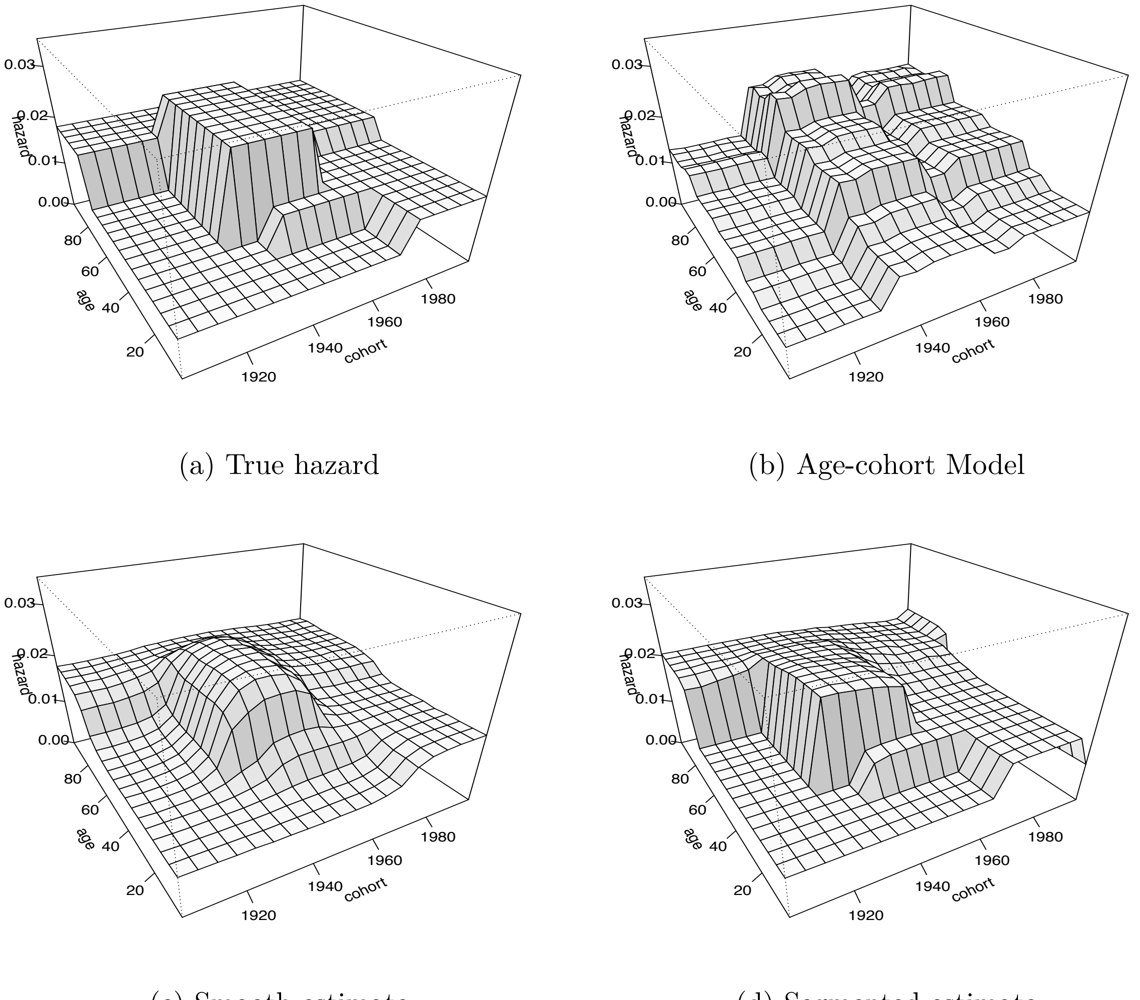
The method is illustrated on simulated data according to a piecewise constant hazard rate. The left top panel of Figure 4.6 shows the true hazard in the age-cohort pane, with four distinct regions. We considered three different estimators of this hazard from samples of size 4 000. The estimators were replicated 500 times and the median of these replications are displayed in Figure 4.6. The age-cohort model

$$\eta_{j,k} = \mu + \alpha_j + \beta_k,$$

is implemented on the top right panel. This estimator is shown to be highly biased due to the constraints imposed by the model. The lower left and right panels respectively show the ridge (with $v_{j,k} = w_{j,k} = 1$ for all j, k) and adaptive ridge estimates. They both show a good performance of the estimators, in particular the shape of the hazard is correctly captured by the adaptive ridge estimator.

The method is also illustrated on the *Surveillance, Epidemiology, and End Results* (SEER) dataset which is publicly available from the US National Cancer Institute. SEER collects medical data of cancers (including stage of cancer at diagnosis and the type of tumour) and follow-up data of patients in the form of a registry. Around 28 percent of the US population is covered by the program. The registry started in February 1973 and the available current dataset includes follow-up data until January 2015. In this study the duration of interest T^* is the time from breast cancer diagnosis to death in years, the cohort is the date of diagnosis (in years) and the period is the calendar time (in years). Patients continuously entered the study between 1973 and 2015 and right-censoring occurred for patients that were still alive at the end of follow-up or for those that were lost to follow-up. For the sake of comparison, the subsample of malignant, non-bilateral breast tumour cancers was extracted from the dataset, such that the data comprises 1 265 277 individuals with 60 percent of censored individuals. Observed times from diagnosis to death vary between 0 and 41 years, and the dates of cancer diagnosis vary between 1973 and 2015.

The adaptive ridge estimates for the whole sample and for each cancer stage are displayed in Figure 4.7. We can see that the different stages of cancer at diagnosis have a considerable impact on the survival times. For stage 1 cancers, the hazard is low few years after diagnosis, and steadily increases with time. There seems to be no effect of the date of diagnosis. On the opposite, for stage 2 cancers, a strong effect of the date of diagnosis can be noticed. Around the years 1995 – 1997, the hazard rate is seen to considerably decrease which could correspond to a



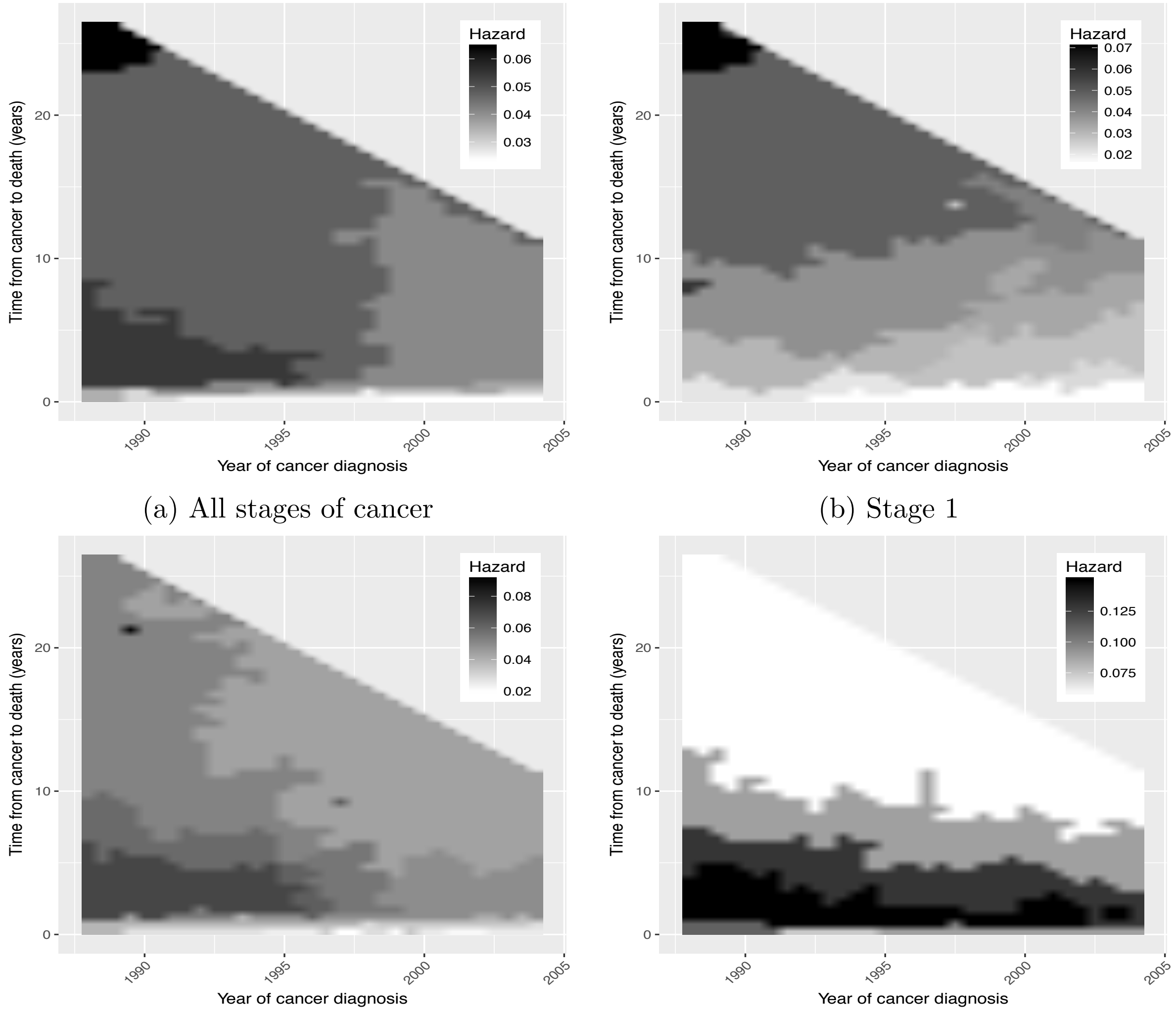
(a) True hazard

(b) Age-cohort Model

(c) Smooth estimate

(d) Segmented estimate

Figure 4.6: Piecewise constant true hazard and corresponding estimates. The sample size is 4 000 and the median hazard is taken over 500 simulations. The estimations are performed in the age-cohort plane and with different methods. Figure (a) represents the true hazard used to generate the data, Figure (b) represents the hazard estimated using the age-cohort model, Figure (c) represents the ridge estimate and Figure (d) represents the adaptive ridge estimate obtained using the EBIC criterion.



(a) All stages of cancer

(b) Stage 1

(c) Stage 2

(d) Stage 3

Figure 4.7: Estimated hazard of survival time after diagnosis of breast cancer for different stages of cancer using the adaptive ridge estimator.

medical improvement for the treatment of breast cancer in the United States. Finally, the figure for stage 3 cancers displays a very high hazard rate across all dates of diagnosis. From these plots we can conclude that the evolution in treatments of breast cancer had a significant impact on the survival times after diagnosis, but almost exclusively when cancers were diagnosed at stage 2.

4.2.3 Interval censoring with a cure fraction

In our work [S3] we studied interval-censored data such as defined in Section 2.4.1. We present a model that also allows for a cure fraction, that is there exists a subpopulation that cannot experience the event of interest. Cure models were initially introduced by [Far82b] and then further studied by [ST00] and [PD00]. They attempt to address estimation issues when the data have typically heavy censoring at the end of the follow-up period. In [ST00] and [PD00], the cure model was extended to a Cox proportional modelling of the latency part, with a nonparametric baseline. However, as explained in [ST00] for instance, identifiability issues can occur which are circumvented by imposing the survival function of the susceptible group to reach 0 at the last observed failure time. This condition is arbitrary and unnecessary for parametric baselines. In [S3] we consider a cure model for the mixed case of exact, left censored, right censored and interval censored data and we avoid the zero-tail constraint by using the piecewise constant baseline hazard and the adaptive ridge algorithm presented in Section 4.2.1.

We introduce the susceptibility status as the variable Y which equals 1 for patients that will eventually experience the event and 0 for patients that will never experience the event. The probability of being susceptible is equal to $p = \mathbb{P}[Y = 1]$. For a right censored individual, Y is not observed. The marginal survival function of T^* is $S(t) = (1-p) + pS(t|Y = 1)$ for $t < \infty$, where $S(t|Y = 1)$ is the survival function of the susceptibles. Note that $S(t) \to 1-p$ as $t \to \infty$. We assume an independent, non informative, random censoring model and that censoring is statistically independent of Y. We then assume the following Cox proportional hazard model for the time variable T^*:

$$\lambda(t|Y = 1, Z) = \lambda_0(t)\exp(\beta_0 Z),$$

where Z is a covariate vector and β_0 an unknown row parameter vector, both of dimension d_Z. The PH cure model specifies the hazard, conditional on Y and Z, to be equal to $\lambda(t|Y, Z) = Y\lambda(t|Y = 1, Z)$. As in Section 4.2.1, we model the baseline function through a piecewise constant hazard function. For $k = 1, \dots, K$,

$$\lambda_0(t) = \sum_{l=1}^{K} I_k(t)\exp(a_k),$$

with $I_k(t) = I(c_{k-1} < t \leq c_k)$ defined as previously. Under this model, note that the survival and density of the susceptible individuals are respectively equal to:

$$S(t|Y = 1, Z) = \exp\Big(-\sum_{l=1}^{K} e^{a_k+\beta_0 Z}(t \wedge c_k - c_{k-1})I(c_{k-1} \leq t)\Big),$$

$$f(t|Y = 1, Z) = \sum_{l=1}^{K} I_k(t)\exp\Big(a_k + \beta_0 Z - \sum_{j=1}^{k} e^{a_j+\beta_0 Z}(t \wedge c_j - c_{j-1})\Big).$$

The nonparametric situation is encompassed in our modelling approach as the special case where $Z = 0$. The case of only susceptible individuals corresponds to $p = 1$.

In the regression framework, a logistic link is used to model the probability of being susceptible with respect to some covariates X. Let

$$p(X) = \mathbb{P}[Y = 1|X] = \frac{\exp(\gamma_0 X)}{1 + \exp(\gamma_0 X)},$$

where X is a covariate vector including the intercept and γ_0 is a row parameter vector, both of dimension d_X.

The observed data consist of $(L_i, R_i, \delta_i, Z_i, X_i)$ for $i = 1, \ldots, n$ while T_i^* and Y_i are respectively incompletely observed and non observed data. We set $\boldsymbol{\theta}$ the model parameters we aim to estimate. In the following we will either study the non-parametric context when there are no covariates, in which case $\boldsymbol{\theta} = (a_1, \ldots, a_L, \beta, p)$, or the regression context in which case $\boldsymbol{\theta} = (a_1, \ldots, a_L, \beta, \gamma)$. In the regression context, we introduce the notations $p_i = \mathbb{P}[Y_i = 1|X_i]$ and $a_{i,k} = a_k + \beta Z_i$.

As presented in Section 2.4.2, there is an explicit representation of the observed likelihood that can be maximised with respect to the model parameters using the Newton-Raphson algorithm. 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 is therefore to use the EM algorithm based on the complete likelihood of the unobserved true event times and susceptibility status. This algorithm will result into a diagonal block matrix of the baseline coefficients. Combined with the adaptive ridge algorithm, this will lead to a tractable maximisation problem where the cuts will be automatically chosen by the procedure and estimation performed with the chosen cuts.

The complete likelihood is then defined by:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i} \prod_{i=1}^{n} \{f(T_i^*|Y_i = 1, Z_i; \boldsymbol{\theta})\}^{Y_i}.$$

Expectation Step:

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 and Y_is, given the L_is, R_is, δ_is, Z_is, X_is and $\boldsymbol{\theta}_{\text{old}}$. Let $\pi_i^{\text{old}} = \mathbb{E}[Y_i|\text{data}, \boldsymbol{\theta}_{\text{old}}]$. We have:

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

Then note that

$$\mathbb{E}[\log(f(T_i^*|Y_i = 1, Z_i; \boldsymbol{\theta}))|\text{data}, \boldsymbol{\theta}_{\text{old}}] = \int f(t|\text{data}, \boldsymbol{\theta}_{\text{old}}) \log f(t|Y_i = 1, Z_i; \boldsymbol{\theta}) dt$$

and under the assumption $\mathbb{P}(T^* \in [L, R]) = 1$,

$$f(t|\text{data}, \boldsymbol{\theta}_{\text{old}}) = \frac{f(t|Y_i = 1, Z_i; \boldsymbol{\theta}_{\text{old}}) I(L_i < t < R_i)}{S(L_i|Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i|Y_i = 1, Z_i, \boldsymbol{\theta}_{\text{old}})}.$$

Maximisation Step:

The M-step consists of maximising the quantity $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \mathbb{E}_{T^*_{1:n}, Y_{1:n}|\text{data}, \boldsymbol{\theta}_{\text{old}}}[\log(L(\boldsymbol{\theta}))]$

with respect to $\boldsymbol{\theta}$. In the absence of exact observations, we have:

$$\begin{aligned}
&Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})\\
&= \sum_{i=1}^{n} \Big\{ \pi_i^{\text{old}} \log(p_i) + (1-\pi_i^{\text{old}}) \log(1-p_i)\\
&\quad + \frac{\pi_i^{\text{old}} \int_{L_i}^{R_i} f(t|Y_i=1, Z_i; \boldsymbol{\theta}_{\text{old}}) \log f(t|Y_i=1, Z_i; \boldsymbol{\theta}) dt}{S(L_i|Y_i=1, Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i|Y_i=1, Z_i, \boldsymbol{\theta}_{\text{old}})} \Big\}\\
&= \sum_{i=1}^{n} \Big\{ \pi_i^{\text{old}} \log(p_i) + (1-\pi_i^{\text{old}}) \log(1-p_i) \Big\}\\
&\quad + \sum_{i=1}^{n} \Bigg\{ \frac{\pi_i^{\text{old}}}{S(L_i|Y_i=1, Z_i, \boldsymbol{\theta}_{\text{old}}) - S(R_i|Y_i=1, Z_i, \boldsymbol{\theta}_{\text{old}})}\\
&\quad\quad \times \sum_{k=1}^{K} J_{k,i} \int_{c_{k-1} \vee L_i}^{c_k \wedge R_i} \exp\Big(a_{i,k}^{\text{old}} - \sum_{j=1}^{k} e^{a_{i,j}^{\text{old}}} (t \wedge c_j - c_{j-1})\Big) \Big(a_{i,k} - \sum_{j=1}^{k} e^{a_{j,k}} (t \wedge c_j - c_{j-1})\Big) dt \Bigg\},
\end{aligned}$$

where $J_{k,i}$ is the indicator $I\{(L_i, R_i) \cap (c_{k-1}, c_k) \neq \emptyset\}$. If we include exact data, these observations will contribute in the same way as in Equation (4.5) for standard pch model with exact and right-censored data. We will therefore separate the contributions of exact observations to the contributions of left, interval and right censored data in the expression of Q. For $k = 1, \ldots, K$, introduce the quantities

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

and

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

For mixed case data, the M-step corresponds of maximising the quantity

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

One should note that the function Q separates the terms with γ (first line of the previous equation) and the terms involving $(a_1, \ldots, a_L, \beta)$ (second and last line of the previous equation) such that maximisation of these terms can be performed separately. In the absence of covariates, explicit estimators can be derived by maximising the function Q. In the regression framework, a Newton-Raphson procedure must be used. The Hessian of Q with respect to $\boldsymbol{\theta}$ is composed of four block matrices. The block corresponding to the second order derivatives with respect to the a_ks is diagonal and the other blocks are of full rank. The adaptive ridge is implemented by penalising the function Q:

$$l(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) - \frac{\text{pen}}{2} \sum_{k=1}^{K-1} w_k (a_{k+1} - a_k)^2,$$

where $\boldsymbol{w} = (w_1, \ldots, w_{K-1})$ are the adaptive weights as defined in Section 4.2.1. The block matrix of the Hessian of l corresponding to the second order derivatives with respect to the a_ks is, this time, tri-diagonal. As a result, the total complexity for the inversion of the Hessian of l is of order $\mathcal{O}(K)$. Nevertheless, for a given penalty, it should be noted that the global algorithm consists of an EM algorithm with a Newton-Raphson procedure at each step. As a consequence, a Generalised Expectation Maximisation (GEM) algorithm (see [DLR77]) is used instead of the standard EM where, as soon as the value of Q 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 is also used to accelerate the EM algorithm (see for instance [VR08]). Finally, a BIC is implemented from the observed likelihood and the penalty term is chosen as the one that minimises the BIC.

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. 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 regular visits to the dentist. In this study, we focused on a complication called ankylosis such that the variable of interest T^* is the time from replantation of the tooth to ankylosis. 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). All the covariates were included in a Cox model. Since age shows little variation for immature teeth, this last variable was only included in interaction with the stage of root formation. The data are described in great details in [ABJA95]. There is no need for a cure fraction in the model, as when implemented on the dataset, the cure fraction is estimated to 0%. This means that all patients will eventually develop ankylosis, a result that is supported by clinicians experience. The results are shown in Figure 4.8. The method found four cuts for the baseline hazard at time points 100, 500, 800 and 900. Statistical tests of the significance of the variables were implemented from log-ratio tests when assuming the cuts to be pre determined. It can be seen that the stage of root formation is highly significant with twice more 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. Taking only the significant covariates, survival curves can be plotted to illustrate the evolution of the risk with respect to time, as shown on the right panel of Figure 4.8 for mature and immature teeth with 20 minutes of storage time.

Covariates	HR	p-value
Mature	2.00	$1.89{\cdot}10^{-5}$
Storage time (hours)	1.23	0.0017
Physiologic	0.93	0.6980
Age>20 (mature teeth)	1.27	0.1272

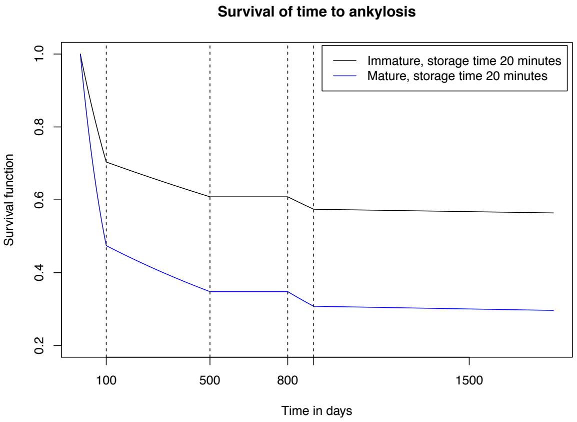


Figure 4.8: On the left panel, regression modelling of time to ankylosis on the dental dataset (HR: Hazard Ratio). The adaptive ridge found four cuts for the baseline hazard at times 100, 500, 800 and 900. On the right panel, estimates of the survival function from the reduced regression model with covariates: stage of root formation (mature/immature) and storage time.

This plot illustrates another 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. The survival after 100 days is estimated to 70.4% for the immature teeth and to 47.4% for the mature teeth. Then the risk decreases from 100 days to 500 days, with the survival after 500 days estimated to 60.8% for the immature teeth and to 34.8% for the mature teeth and finally the risk gets very low after 500 days.

4.3 Accounting for the competing risk of death in genetic studies

Complex diseases with variable age at onset typically have many interacting factors such as the age, lifestyle, environmental factors, treatments, genetic inherited components. The genetic component is generally composed of one or several genes including major genes for which a deleterious mutation rises significantly the risk of the disease and/or minor genes which participation in the disease is moderate by itself.

The mode of inheritance can be monogenic if a mutation in a single gene is transmitted or polygenic if mutations in several genes are transmitted. As an example of a major gene in a complex disease, the BRCA1 gene is well known to be strongly correlated with ovarian and breast cancer since the 90s [HNM+90] ,[CRT94]. Carriers of a deleterious mutation in BRCA1 gene have a much higher risk to be affected with relative risks ranging from 20 to 80 but deleterious mutations in BRCA1 gene only explain 5 to 10 % of the disease [MA16] as many other implicated known or unknown genes exist along with sporadic cases (cases with no inherited component).

The family history (FH) of such diseases is often the first tool for clinicians to detect a family of carriers of a deleterious mutation as any unusual accumulation of cases in relatives leads to suspect a deleterious allele in the family. With the appropriate model and computation, the FH can be used to better target the most appropriate individuals for a genetic testing and/or to identify high-risk individuals who require special attention (monitoring and/or treatments).

The first challenge to compute such a model comes from the fact that genotypes are mostly (if not totally) unobserved and that posterior carrier probability computations must sum over a large number of familial founders' genotypes configurations. Once such computations are carried out, deriving posterior individual disease risk is also a challenging task since the posterior carrier distribution changes over time. Finally, for diseases with possibly late age at onset (*e.g.* cancer), the competing risk of death is not negligible and must be accounted for. Classical familial risk models such as Claus-Easton (see [CRT91], [EBFC93]), BOADICEA (see [APSE04]), or the BayesMendel models (BRCAPRO, MMRpro, PancPRO and MelaPRO, see [CWL+06]) do not take into account the competing event of death. As a result, it is likely that individual predictions will tend to be overestimated from these models (see [DP12]). The main result of our work [P7] is that we show how to derive individual risk predictions from the family history while taking into account the competing risk of death.

More precisely, we place ourselves in an illness death situation (since death precludes the occurrence of the disease but the opposite is not true) such as described in Section 2.2 of the Introduction. In our context, the event of interest is the age of cancer diagnosis for individual i, denoted $T_i^{*\text{dis}}$ while the competing risk of death is denoted $T_i^{*\text{death}}$. We note FH the family history of a given patient, which includes the personal history and the genotype of all related individuals. The personal history comprises information on whether or not the relatives have yet developed the disease or died. The genotype of the jth relative is represented by $X_j \in \{00, 01, 10, 11\}$. We work under the Claus-Easton model (see [CRT91], [EBFC93]) which assumes an autosomal dominant mode of inheritance, such that a carrier and a non carrier have genotypes respectively equal to $X_j \neq 00$ and $X_j = 00$. Even with genetic testing, it is essential to understand that the X_js are, at best, partially observed. Indeed, even with a (hypothetical and unrealistic) 100% specificity/sensitivity test, a positive heterozygous carrier status cannot distinguish between genotypes 01 and 10. Moreover, genetic tests are in general only available for few individuals in the whole pedigree. Accounting for the unobserved genotypes is therefore of utmost importance. This is performed using Bayesian network and sum-product algorithms (see for instance [LS03] or [KF09]). This aspect of our work will not be developed here and we will only focus on how to perform individual risk predictions from the family history.

Let $t_0 > 0$ be a given time point for which it is known that patient i has not yet developed the disease and is still alive. We introduce $\pi = \mathbb{P}[X_i \neq 00|\text{FH}]$, where in the notation FH we have also included the information that $T_i^{*\text{dis}} > t_0$ and $T_i^{*\text{death}} > t_0$. For any $t > t_0$ we aim at computing the cumulative incidence of the disease given family history $\mathbb{P}[T_i^{*\text{dis}} \leq t|\text{FH}]$. We denote the cause specific hazard of the disease in a similar way as in Equation (2.5) but by also conditioning on the family history:

$$\lambda_{\text{dis}}(t) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T_i^{*\text{dis}} < t + \triangle t | T_i^{*\text{dis}} \geq t, T_i^{*\text{death}} \geq t, \text{FH}]}{\triangle t}.$$

We then model this hazard using a piecewise constant model on the cuts $0 = c_0, c_1, \ldots, c_K = +\infty$ such as in Section 4.2.1 and we denote by α_k the value of the hazard of the disease on the kth cut $(c_{k-1}, c_k]$. We also introduce $\beta_k := \alpha_k + \lambda_{\text{death}}(c_k)$ where λ_{death} is the hazard of death, which is obtained through register population data. Even though these register data usually contain informations only on death from all cause, we assume that death without cancer and death from all cause have a similar hazard. We also assume that the hazard for non carriers individuals (that is with $X_i = 00$) and for carriers (that is with $X_i \neq 00$) are known. These are respectively denoted λ_0 and λ_1 and their all cause survival functions are noted S_0 and S_1:

$$\lambda_0(t) := \lim_{\triangle t \to 0} \frac{\mathbb{P}[t \leq T_i^{*\text{dis}} < t + \triangle t | T_i^{*\text{dis}} \geq t, T_i^{*\text{death}} \geq t, X_i = 00]}{\triangle t},$$

$$S_0(t) := \mathbb{P}[T_i^{*\text{dis}} \wedge T_i^{*\text{death}} \geq t | X_i = 00] = \exp\left(-\int_0^t (\lambda_0(u) + \lambda_{\text{death}}(u))du\right),$$

and λ_1 and S_1 are defined similarly by conditioning on $X_i \neq 00$. Finally, we also note

$$S(t) := \mathbb{P}[T_i^{*\text{dis}} \wedge T_i^{*\text{death}} \geq t|\text{FH}] = \exp\left(-\int_{t_0}^{t} (\lambda_{\text{dis}}(u) + \lambda_{\text{death}}(u))du\right).$$

In [P7] we showed that given the family history, the cumulative incidence function of the disease can be computed through the following steps:

1. compute $\alpha_j = \lambda_{\text{dis}}(c_j)$ using the equation

$$\lambda_{\text{dis}}(t) = \frac{1}{S(t)}\left[\pi\frac{\left(S_1(t)\right)^2}{S_1(t_0)}\lambda_1(t) + (1-\pi)\frac{\left(S_0(t)\right)^2}{S_0(t_0)}\lambda_0(t)\right].$$

2. compute $\beta_j = \alpha_j + \lambda_{\text{death}}(c_j)$ and $S_\beta(c_j) = \exp(-\sum_{l=1}^{j}(c_l - c_{l-1})\beta_l)$.

3. then the marginal posterior probability of being diagnosed with the disease before age c_k, in the presence of death as a competing risk, is given for $k = 1, \ldots, K$ by:

$$\mathbb{P}[T_i^{dis} \leqslant c_k|\text{FH}] = \sum_{j=1}^{k}\frac{\alpha_j}{\beta_j}\Big(S_\beta(c_{j-1}) - S_\beta(c_j)\Big).$$

Finally the difference in risk prediction when accounting for the competing risk of death or when ignoring it is illustrated on an hypothetical example. Figure 4.9 represents an example of a moderate size (hypothetical) family with a severe history of breast and ovarian cancer. This family has a total of $n = 12$ individuals with the set of founders $\mathcal{F} = \{1, 2, 3, 4\}$ and the set of nonfounders $\mathcal{I}\backslash\mathcal{F} = \{5, 6, 7, 8, 9, 10, 11, 12\}$. There is no inbreeding (mating between individuals with a common ancestor) in this family but a mating loop (two families joined more than once by mating) due to the two brothers of the first nuclear family having children with two sisters of the second nuclear family. The individual risk of breast cancer for individuals 7 and 12 are represented on Figure 4.10 where we set $\pi = 0.553\%$ and $t_0 = 62$ years for individual 7 and $\pi = 44.6\%$ and $t_0 = 37$ years for individual 12. The risks are plotted for t ranging from t_0 to 100 years with and without taking into account the competing risk of death. We can see that the difference between the two curves for each individual is increasing with the age. We also observe that the individual risk of breast cancer eventually reaches a plateau which corresponds to the point where the incidence of breast cancer becomes negligible compared to the incidence of death in the elderly.

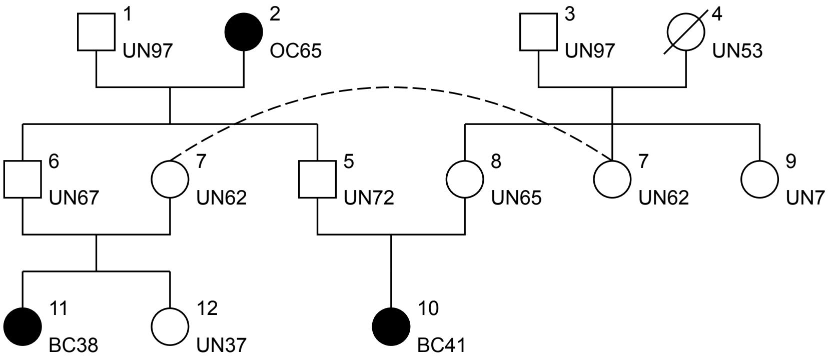


Figure 4.9: An hypothetical family with a severe FH of cancer. Squares correspond to males, circles to females, and affected individual are filled in black. Individual id on the top-right of the nodes, personal history of cancer (UN=UNaffected; BC=Breast Cancer; OC=Ovarian Cancer) on the bottom-right. The dashed line represents an identity link used to represent the mating loop (due to the mating between individuals 5/8 and 6/7) between brothers 5 and 6, and sisters 7 and 8.

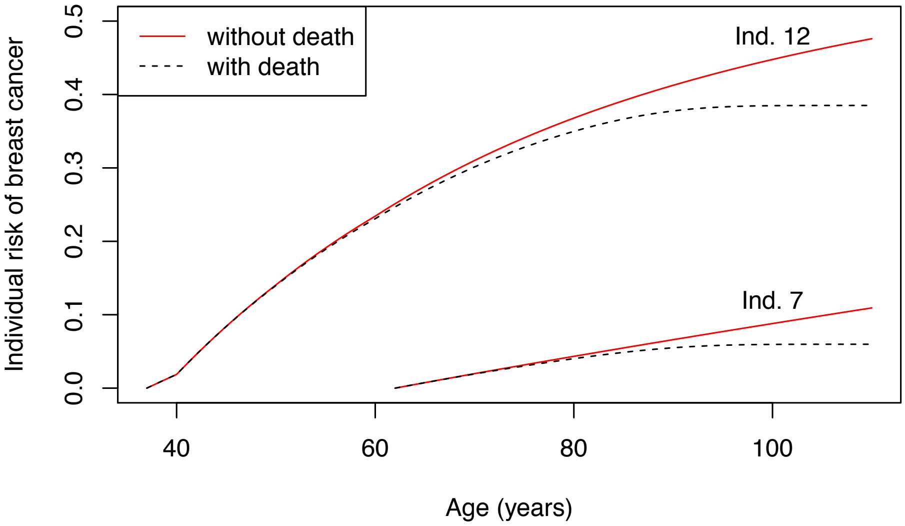


Figure 4.10: Individual risk of breast cancer with and without the competing risk of death for individual 7 and 12 of our hypothetical family from τ to 100 years with and without the competing risk of death.

CHAPTER 5

Applications to medical studies

In this chapter I present my statistical contributions to medical studies. The work in [P8] started while I was visiting the Biostatistics section at the University of Copenhagen. This work was in collaboration with the department of Gynecology and Obstetrics at the Hillerød hospital in Denmark. The aim of the study was to evaluate the effect of broad spectrum antibiotics on the risk of developing type 1 diabetes for children. This was performed from Danish register data using survival analysis theory. The work in [S4] is also a collaboration with the Biostatistics section of Copenhagen and it involves the Department of Cardiology at Hvidovre University Hospital in Copenhagen. The aim was to predict risks of future hospitalisations related to atrial fibrillation for patients suffering from this cardiac disease using recurrent event methods. Finally, in a work with the biology team at Institut de Recherche pour le Développement (IRD) at the Faculté de Pharmacie, Université Paris Descartes, we studied the susceptibility to malaria attacks for children born to mothers with placental malaria using recurrent event techniques. This work was published in [P9].

5.1 Modelling the effect of broad-spectrum antibiotic treatment on childhood type 1 diabetes onset from a nationwide Danish cohort study

Infectious morbidity and mortality have been reduced dramatically since the introduction of penicillin and other antibiotics. However, expanding use of antibiotics has unwanted ecological side-effects. Recent studies indicate that antibiotic treatment may influence the human organism in a long-term perspective and increase the risk of chronic diseases (see [WLV$^+$08], [VWGK15], [BFZ$^+$14]). It is suggested that the hypothetical disease-causing effect of antibiotics works through changes of the gut-microbiota, and broad-spectrum antibiotics are thought to have the most prominent effect (see [VWGK15]). Since 1997, the incidence rate of type 1 diabetes in 0-4 year old children in Denmark has stabilised at 14 new cases per 100,000 person-years, whereas the incidence rates among 5–9 and 10–14 year old children have been steadily increasing (see [CBE$^+$16] and [Vaa12]). Studies have linked microbiomic changes to an increased risk of type 1 diabetes, through a complex disturbance of the maturation of the immune system and an increased vulnerability to environmental triggers of autoimmunity (see [VWGK15]). As the microbiota is affected for several months following a broad-spectrum antibiotic treatment (see [JLEJ10]) it has been intriguing to link the increasing incidence of childhood type 1 diabetes to the rising use of broad-spectrum antibiotics. Published studies show inconsistent results (see [Vaa12], [BKV$^+$06] and [HS09] among others). Mode of delivery has also been linked to an increased risk of type 1 diabetes, possibly due to microbiomic changes, but findings from studies are conflicting (see [CBE$^+$16] and [CSJ$^+$08] among others). According to the "hygiene hypothesis" exposure to microbes, including those in the genital tract during birth, increases the microbial biodiversity and is thought to be beneficial for the immune system maturation and protective for later development of a variety of diseases. Therefore, harmful effects of broad-spectrum antibiotics would be expected to be most pronounced among children delivered by prelabor cesarean section, whom had not been exposed to the maternal vaginal flora (see [WLV$^+$08], [VWGK15]). In a previous study [CBE$^+$16] the authors studied the effect of mode of delivery on the onset of type 1 diabetes. In our work [P8], we aimed at evaluating the association of broad-spectrum antibiotic treatment during the first two years of life with subsequent onset of childhood type 1 diabetes and at exploring potential effect-modification by mode of delivery.

The data are based on four Danish nationwide registers: the Medical Birth Registry (1997 to 2010), the Fertility Database (1997 to 2010), the National Patient Registry (1977 to 2011), the Register of Medicinal Product Statistics (1997 to 2012) and additional information from the national Statistics Denmark registry (1997 to 2012). Linkage between registers and between children and their parents was performed (when possible) from the unique Danish personal identification number. All live-born children in Denmark from 1 January 1997 through 31 December 2010 ($n = 912,797$) were identified in the Medical Birth Registry. We excluded 38, 218 children from multiple pregnancies ($n = 37,895$) and pregnancies with errors in the personal identification number (n = 323). Furthermore, we excluded 16, 378 children with events before their two years birthday due to either death ($n = 3,412$), emigration ($n = 12,790$) or diagnosis of type 1 diabetes ($n = 176$). The final population included 858, 201 live-born singleton children born to 527, 927 mothers.

The children were followed from age two until their fifteenth birthday or end of follow-up which corresponded to December 2012. It is clear that all children had a different follow-up time since children born for example in January 2010 were included in the study at age two in January 2012 and could be followed a maximum of one year only. Therefore, a survival analysis is needed to take into account the differences in follow-up times. Moreover, drop-off could occur due to immigration. In this study, the time-scale is the age and the end of follow-up in December 2012 (for 773, 359 children), immigration of the child (for 21, 787 children) or turning 15 (for 60, 934 children) all correspond to censoring. Death of the child (for 618 children) corresponds to a competing risk since the child cannot develop type 1 diabetes after death. Type 1 diabetes diagnosis was only observed on 1, 503 children.

Outpatient redemptions of antibiotic prescriptions for the child during the first two years of life were classified into either: any type of antibiotics (yes or no), narrow-spectrum antibiotics (yes or no) or broad-spectrum antibiotics (yes or no), classified in accordance with the Danish Integrated Antimicrobial Monitoring and Research Program 2013.

Descriptive statistics of the study describing in particular the consumption of antibiotics (narrow, broad or any of the two) can be found in our paper [P8]. A first result concern the implementation of three Cox models, each of them including either the covariate any antibiotics, narrow antibiotics or broad antibiotics. They were all adjusted along with the following covariates: mode of delivery (vaginal, intrapartum cesarean section or prelabor cesarean section), sex, parity, paternal age (three groups), maternal age (three groups), paternal education (three groups), maternal education (three groups), paternal type 1 diabetes (yes or no) and maternal type 1 diabetes (yes or no). Since some of the children included in the study were born from the same mother, the data must be considered as clustered data. As mentioned in the Introduction section 2.3.1, a robust sandwich estimator exists for clustered data which accommodates for correlated times for children born to the same mother. Finally, missing covariates were handled by complete case analysis. The results of these three models were that children who had redeemed prescriptions on any type of antibiotics during the first two years of life had a comparable rate of childhood type 1 diabetes, to children without redemptions of antibiotics

(HR = 1.06, 95% CI = $[0.94, 1.19]$). The same pattern was found regarding narrow-spectrum antibiotics. In contrast, children who had redeemed prescriptions on broad-spectrum antibiotics had a higher rate of childhood type 1 diabetes as compared to children who did not (HR = 1.13, 95% CI = $[1.02, 1.25]$). Apart from broad-spectrum antibiotics the following predictors of type 1 diabetes was found: primiparity (HR = 1.12; 95% CI = $[1.00, 1.26]$), paternal type 1 diabetes diagnosed before childbirth (HR = 11.21; 95% CI = $[8.85, 14.20]$) and maternal type 1 diabetes diagnosed before childbirth (HR= 6.19; 95% CI = $[4.31, 8.91]$).

In a second part, the following interaction model was studied. We introduce the dichotomous covariates AB representing the redemption of antibiotics (yes or no for any, broad or narrow types), intra for intrapartum cesarean section as mode of delivery, and prelabor for prelabor cesarean section as mode of delivery. All the other covariates are represented by X_j, for $j = 1, \dots, 12$. The interaction Cox model is defined as:

$$\lambda_0(t) \exp\left(\beta_{01}\text{AB} + \beta_{02}\text{intra} + \beta_{03}\text{prelabor} + \beta_{04}\text{AB} \cdot \text{intra} + \beta_{05}\text{AB} \cdot \text{prelabor} + \sum_{j=1}^{12} \beta_j X_j\right).$$

As a consequence, the hazard ratio of antibiotic redemption and vaginal delivery versus no antibiotic redemption and vaginal delivery is equal to: $e^{\beta_{01}}$. The hazard ratio of antibiotic redemption and intrapartum cesarean section versus no antibiotic redemption and intrapartum cesarean section is equal to: $e^{\beta_{01}+\beta_{04}}$. The hazard ratio of antibiotic redemption and prelabor cesarean section versus no antibiotic redemption and prelabor cesarean section is equal to: $e^{\beta_{01}+\beta_{05}}$. Estimates of the parameters are performed in the usual way for Cox models. However, confidence intervals of the last two types of hazard ratios require the knowledge of the multidimensional distribution of the vector parameters. This is well known in survival analysis theory. In practice, confidence intervals and tests of functionals of the parameter estimate are implemented in the **lava** package [HBJ13] through the **estimate** function.

A global test for an interaction between antibiotic redemption and mode of delivery corresponds to the test:

$$(\text{H}_0) : \beta_{04} = \beta_{05} = 0 \quad \text{vs} \quad (\text{H}_1) : \beta_{04} \neq 0 \text{ or } \beta_{05} \neq 0,$$

and is again implemented from the **lava** package. The p-value from this test was equal to 0.0023 which indicates a strong association between antibiotics redemption and mode of delivery. In vaginally delivered children, redemption of antibiotics, regardless of type, was not associated with an increased rate of childhood type 1 diabetes (for example, in the case of broad spectrum antibiotics redemption, the confidence interval for $e^{\beta_{01}}$ is equal to $[0.94, 1.18]$). In contrast, an association with broad-spectrum antibiotics was found among children delivered by intrapartum cesarean section (HR = 1.70, 95% CI = $[1.15, 2.51]$) as well as by prelabor cesarean section (HR = 1.63, 95% CI = $[1.11, 2.39]$). In children delivered by prelabor cesarean section redemption of antibiotics of any type was associated with a two-fold increased rate of childhood type 1 diabetes (HR = 1.91, 95% CI= $[1.14, 3.20]$). Regardless of mode of delivery, narrow-spectrum antibiotics were not associated with childhood type 1 diabetes.

As defined by [AA99], a Number Needed to Treat/Harm (NNT or NNH) was calculated. If the antibiotic redemption was the only covariate, the NNH would correspond to:

$$\text{NNH} = \left(\hat{S}(\tau|\text{AB} = 1) - \hat{S}(\tau|\text{AB} = 0)\right)^{-1},$$

where $\hat{S}(t|\text{AB} = 1)$ and $\hat{S}(t|\text{AB} = 0)$ represent the Kaplan-Meier estimators of the survival function evaluated at time t respectively for children that were treated with antibiotics and for children that were never treated with antibiotics during their first two years of life. The time

τ corresponds to the end-point of the study which is 15 years in our case. Then, we need to take into account the competing risk of death and to adjust for all the other covariates when computing the NNH. The competing risk of death is dealt with using Formula (2.6) of Section 2.2 of the Introduction and adjusting for the other covariates is performed by taking the average over all the covariates values. Let $\hat{S}(t|\text{AB}, \boldsymbol{X})$ be the estimated event-free survival function defined conditionally on antibiotics redemption and the covariate vector $\boldsymbol{X}$ (which comprises all the covariates except antibiotics redemption). Let $\hat{\lambda}(t|\text{AB}, \boldsymbol{X})$ be the estimated hazard derived from the Cox model. We then compute the adjusted cumulative incidence

$$\hat{F}(t|\text{AB}) = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{t}\hat{\lambda}(u|\text{AB}, \boldsymbol{X}_i)\hat{S}(u|\text{AB}, \boldsymbol{X}_i)du,$$

and the NNH is equal to

$$\text{NNH} = \Big(\hat{F}(\tau|\text{AB} = 0) - \hat{F}(\tau|\text{AB} = 1)\Big)^{-1}.$$

In our study, the NNH were estimated to $2,218$ in all children regardless of delivery mode, 433 in children delivered by intrapartum cesarean section and 562 in children delivered by prelabor cesarean section. Confidence intervals can also be computed, typically using a bootstrap approach but they were not calculated in the present study.

5.2 Assessing the effect of placental malaria on malaria attacks for Beninese children

Placental malaria (PM) due to P. falciparum is estimated to cause up to $200,000$ infant deaths every year [SNPM01]. Among the consequences that have been attributed to PM, children born to a mother with P. falciparum infected placenta seem more susceptible to malaria. This potential susceptibility, defined as a shorter delay of occurrence of the first malaria infection, seems to be due to a phenomenon of immune tolerance (IT) which is only partially understood. This would involve the transfer of parasite proteins from mother with PM to foetus leading to a modification of immune development of the foetus [BBEL07]. To date, only five studies have reported an association between PM and the delay of first malaria infection and the existence of IT, but without precisely accounting for local variability in exposure to vector bites (see for example [LHCP+97] or [LPWC+11]). However, recently [LPCMP+12] showed that adequately taking into account variability in exposure to malaria, by means of a predictive statistical model (see [CKP+12] for the development of this predictive model), reinforced the association between PM and the time to first occurrence of malaria infection [LPCC+13]. Due to the protocol used in these previous studies (*i.e.* children follow-up of from birth, to survey the occurrence of infections, which are prone to censoring due to drop-off or end of the study) survival analyses based on Cox models are particularly suited. However, these studies only considered the first malaria infection, and were therefore unable to explore whether or not such children remain more susceptible to malaria after the first infection. We believe this question is important from a public health point of view, in order to potentially provide more suitable prevention strategies. Here, we propose to explore this question by means of a recurrent events model allowing the analysis of not only the first malaria attack, but all malaria attacks occurring during the follow-up. The aim of the study was to assess the impact of PM on the overall risk of malaria attacks from birth to 18 months of life in a Beninese cohort. More precisely, the main objective was to determine whether PM is a risk factor for only the first malaria attack occurring after birth or also for all subsequent attacks occurring during the first 18 months of life.

The work in [P9] is based on a cohort study conducted in Benin whose protocol is detailed in [LPCMP+12]. The study included nine villages and three health centres: Tori Avame, Tori Cada and Tori Gare, providing primary healthcare as well as a maternity ward for antenatal care and childbirth. Malaria is endemic in the study region and transmitted mainly by *Anopheles gambiae ss* and *Anopheles funestus* species. More than 600 pregnant women (from 9 villages), visiting one of the three health centres for antenatal care (ANC) and having no intention to move out of the region, were included in this study at delivery from June 2007 to July 2008. Twins, stillbirths and HIV-positive women were excluded. Two months before the beginning of the study, study supervisors and community health workers informed women about the study. Midwives were told to present the study to all women frequenting the ANC from the 7th month of pregnancy.

The infants were visited weekly at home from birth to 18 months and axillary temperature was measured by a community health worker. In case of a temperature higher than 37.5°C, or a history of fever within the last 24 hours, mothers were told to bring their children to the health centre where a questionnaire was filled up and a Parascreen rapid diagnostic test (RDT) was made, to obtain an immediate diagnosis of symptomatic malaria infection. A thick blood smear (TBS) was performed to provide a later confirmation of the RDT result. Following a positive RDT, the infant was treated by an artemisinin-based combination (arthemeter and lumefantrine) as recommended by the Beninese National Malaria Control Program. A systematic TBS was performed monthly, to detect asymptomatic infections. Mothers were also invited to bring their infants to the health centre at any time, for free attendance in case of fever or any clinical signs, and the same procedure was applied. A malaria attack was defined as an axillary temperature higher than 37.5° (or a history of fever within the last 24 hours) and a positive RDT and/or TBS diagnosed during weekly home visit or during an unscheduled visit of the child and his mother at the health centre. An asymptomatic infection was defined as a systematic monthly TBS positive without fever, history of fever or any other clinical sign.

550 live-birth singletons (among 646 initially selected newborns included in the cohort) were included in the analyses. The events of interest are thus the malaria attacks of the Beninese children. They will be studied using models for recurrent events as the ones presented in Section 2.3. The study period was from birth until age 18 month and 827 malaria attacks were observed during this period. On the overall follow-up, we observed 201 children who did not experience any malaria attack, 133 children who experienced exactly one, and 216 children who experienced two or more. It should be noted that due to censoring, the observed malaria attacks do not correspond to the malaria attacks of interest which are the malaria attacks that one would observe in case of a complete follow-up from birth to age 18 month. The average follow-up time was 16.8 months. Children contributed to a total of $281,903$ person-days at risk. Very few children died during the follow-up (17 in total) and although death precludes the occurrence of further malaria attacks we did not take into account death as a terminal event in this study. All the results were also computed considering death as a terminal event and were very similar.

Besides placental malaria, the covariates included in Cox models for recurrent events are the gravidity status (primigravid vs multigravid), the low birth weight (defined as birth weight < 2500, yes or no), the age (4 groups), the use of bed-net (yes or no) and the environmental risk defined as a score assessed from the predictive model of [CKP+12]. The information about the maternity ward (three different health centres has previously described) was also taken into account, either as a stratification for the baseline hazard in the Cox model or as a random effect in a Cox frailty model. For descriptive statistics about all the covariates, see our paper [P9]. In particular, only 11% of the children were born to a mother having placental malaria.

First, we implemented the estimator (2.12) as presented in the Introduction section for estimating the average number of recurrent events $\widehat{\mathbb{E}[N^*(t)]}$. For t corresponding to the end of follow-up (18 months), we obtained an estimated average of malaria attacks equal to 1.636 for the whole population, equal to 1.626 for children born to mother without placental malaria and equal to 1.728 for children born to mother with placental malaria. These estimates can be interpreted as average numbers of recurrent events in the hypothetical situation where all children were followed up from birth until 18 months without censoring. Then we implemented a model depending on prior recurrences as presented in Section 2.3.3. This model can be seen as a multi-state model (see Figure 5.1) with only three possible states: 0 recurrent event, 1 recurrent event and 2 or more recurrent events (modelling further recurrences leads to similar rate functions).

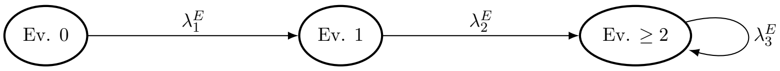


Figure 5.1: The recurrent event model with dependence on prior events seen as a multi-state situation for the Benin study. All individuals start in the state Ev.0. In the regression setting, all the rate functions for the recurrent events are assumed to be proportional.

In this multi-state scenario all the different hazard rates can be easily estimated. Then, a simulated method was developed in order to simulate recurrent events in an hypothetical context where all individuals would be followed-up from birth to 18 months without censoring. We denote by Λ_k^E the cumulative hazard rate functions and we use the notations $\lambda_k^E = \lambda_3^E$ for $k \geq 3$. The simulation scheme is motivated by the fact that if an individual has already experienced $E_1, \ldots, E_k$ events whose realisations are denoted $e_1, \ldots, e_k$, then for $t > e_k$, $\mathbb{P}[E_{k+1} > t | E_{k+1} > e_k] = \exp(-(\Lambda_{k+1}^E(t) - \Lambda_{k+1}^E(e_k)))$ which is distributed as a uniform distribution $[0, 1]$ when evaluated at $t = E_{k+1}$. As a consequence, $\Lambda_{k+1}^E(E_{k+1}) - \Lambda_{k+1}^E(e_k)$ is distributed as an exponential distribution with parameter 1. For a given value of τ, the simulations were performed from the following algorithm:

Step 1. Initialise $e_0 = 0$ and $k = 0$.

Step 2. Repeat

- draw $\mathcal{E}$ following an exponential distribution with parameter 1.
- if $\widehat{\Lambda_{k+1}^E}(\tau) - \widehat{\Lambda_{k+1}^E}(e_k) < \mathcal{E}$ exit the algorithm.
- else solve $\widehat{\Lambda_{k+1}^E}(t) - \widehat{\Lambda_{k+1}^E}(e_k) = \mathcal{E}$ for $t \in [e_k, \tau]$.
- set $e_{k+1} = t$ and $k = k + 1$.

Step 3. Return k and $e_1, \ldots, e_k$.

The algorithm returns the value k corresponding to the number of recurrent events experienced by an individual before time τ and the times $e_1, \ldots, e_k$ representing his/her corresponding recurrent event times. See also the supplementary material of [P4] for a theoretical justification of this type of accept-reject algorithm for simulating recurrent events. Note that the algorithm could be easily extended to take into account the competing risk of death by computing cumulative incidence functions. We took τ equal to 18 month. Repeating a large number of times the procedure allows to obtain simulated data with complete follow-up from birth to 18 month. This allows to compute the probability distribution of the number of recurrent events experienced by a child. The method was also applied from stratified samples with respect to the placental malaria status of the mother. The results are presented in Table 5.1. We can see that the probability of getting exactly one malaria attack is much greater for children born to mother with placental malaria. However, the placental malaria of the mother seems to have no more impacts on further malaria attacks of the children.

Number	0	1	2	3	4	5	6	7+	Mean
All	33.90	22.30	18.40	10.90	6.80	3.90	2.40	1.40	1.636
PM= 0	34.21	21.75	18.10	12.03	6.71	3.79	2.09	1.32	1.626
PM= 1	25.13	29.25	17.70	11.40	7.94	4.80	2.23	1.55	1.728

Table 5.1: Estimated probability distribution of the number of malaria attacks. These estimations were obtained through a simulation method. They correspond to the distribution of malaria attacks in the hypothetical situation of complete follow-up from birth to 18 month.

In order to take into account potential confounders we used a Cox model for recurrent events adjusted with respect to all the covariates as previously described. The model is defined with dependence on prior recurrent events and with placental malaria entered as an interaction term with respect to the number of previous recurrent events. We introduce the dichotomous covariate PM representing the placental malaria status of the mother (yes or no) and all the other covariates are represented by X_j, for $j = 1, \dots, 7$. Following the notations of Section 2.3.3, the model is defined as follows:

$$\mathbb{E}[dN(t)|Y_s(t), X] = Y_s(t)\lambda_s^E(t)dt,$$

where $Y_s(t) = I(N(t-) = s-1)$, $s = 1, 2, 3, \dots$ and the $\lambda_s^E(t)$ correspond to the rate functions in the multi-state model of Figure 5.1 with the convention $\lambda_s^E = \lambda_3^E$ for $s \geq 3$. The $\lambda_s^E(t)$ are then modelled as:

$$\lambda_1^E(t) = \lambda_0(t)\exp\left(\beta_{01}\text{PM} + \sum_{j=1}^{7}\beta_j X_j\right),$$

$$\lambda_2^E(t) = \lambda_0(t)\exp\left(\beta_{02}\text{PM} + \alpha_{01} + \sum_{j=1}^{7}\beta_j X_j\right),$$

$$\lambda_3^E(t) = \lambda_0(t)\exp\left(\beta_{02}\text{PM} + \alpha_{02} + \sum_{j=1}^{7}\beta_j X_j\right).$$

Note that there are only two different values for β_{0s}, for $s = 1$ and $s = 2$ since we only want to separate the effect of the first malaria attack to the effect of all other malaria attacks. In this model the parameter $e^{\beta_{0s}}$ corresponds to the hazard ratio for placental malaria for the first malaria attack when $s = 1$ and for any other malaria attack when $s = 2$. The term $e^{\alpha_{01}}$ corresponds to the hazard ratio for experiencing a second malaria attack versus experiencing a first malaria attack for a child whose mother did not have placental malaria while $e^{\alpha_{02}}$ corresponds to the hazard ratio for experiencing a new malaria attack knowing that the child has already experienced at least two versus experiencing a first malaria attack (irrespective of the placental malaria status). We found that $\widehat{e^{\beta_{01}}} = 1.33$ with 95% CI $= [1.00, 1.76]$, $\widehat{e^{\beta_{02}}} = 0.90$ with 95% CI $= [0.66, 1.22]$, $\widehat{e^{\alpha_{01}}} = 1.72$ with 95% CI $= [1.41, 2.09]$ and $\widehat{e^{\alpha_{02}}} = 2.09$ with 95% CI $= [1.70, 2.58]$. The results are of great interest since they show an effect of placental malaria on the first malaria attack but not on subsequent ones! The huge effects of the $\widehat{e^{\alpha_{0s}}}$ seem to suggest unaccounted heterogeneity in the data. Frailty models were also examined using the **coxme** package that fits the method from [TG00] and [RP00] and using the package **frailtypack** that fits the the method developed in [RMG12]. They allow to include a random effect accounting for individual heterogeneity or to include nested random effect in order to account for both heterogeneity on the village and on the individual levels. All these models showed similar results as our general model and are therefore omitted.

5.3 Prediction of future risk of hospitalisations in paroxysmal and persistent atrial fibrillation patients using recurrent event methods

Atrial fibrillation (AF) is a cardiac disease that is characterised by irregular or abnormal heart rate. The duration of abnormal beating can go from brief episodes to long-time lasting and can even become constant over time. AF is defined as paroxysmal AF (PAF) if the episode terminates spontaneously in less then seven days. It is defined as persistent AF (PeAF) if the episode lasts longer than seven days and does not stop without treatment. Finally, it is defined as permanent (PermAF) if the patient is experiencing an ongoing episode that cannot be stopped even with medication. This disease mainly concerns patients over 50 years old and can results in various complications such as stroke, embolism or heart failure.

From January 1st 2008 to December 1st 2012, patients with AF were enrolled in the “Atrial Fibrillation Survey – Copenhagen (ATLAS-CPH)” from both the in- and outpatient clinics at the Department of Cardiology at University Hospital Copenhagen, Hvidovre, Denmark. Inclusion criteria were age > 18 years, recent (< 1 month) AF documented via either standard 12-lead electrocardiogram (ECG) or home monitoring and ability to give oral and written consent. PAF was defined as at least one recorded AF episode with spontaneous conversion to sinus rhythm, no documentation or suspicion of a reversible primary cause, and excluding other forms of AF. PeAF was defined as at least one recorded episode of AF lasting > 7 days, or where either medical or electrical cardioversion was needed to restore sinus rhythm (in accordance with the Danish Cardiology Society AF guidelines at this time). Patients were excluded if AF type was PermAF, defined as AF that was accepted by both the patient and physician, and accordingly rhythm control interventions were not pursued. Patients were also excluded if they had previously been, or were at any time during the follow up period, treated for AF with an invasive ablation procedure or anti arrhythmic surgery, or if estimated survival was < 1 year from inclusion date. Patients undergoing treatment with sodium or potassium channel blocking anti arrhythmic drugs and patients with bradycardia pacemakers or implantable cardioverter defibrillators were not excluded. The aim of [S4] was to check if AFHs may be a strong predictor of future risk of AF symptoms in PAF and PeAF patients, and further to build a predictive model of future AFH risk in individual patients using recurrent event models.

During enrolment, numerous baseline covariates were recorded through completion of an extensive questionnaire, supplemented with data from the patient’s comprehensive digitalised medical record. The variables encompassed general informations on the patient as well as non-cardiac comorbidity: age (treated as continuous), gender, diabetes mellitus status (yes or no), hypertension (yes or no), heart failure (yes or no), valvular heart disease (yes or no), ischemic heart disease (yes or no), chronic obstructive pulmonary disease (COPD, yes or no), alcohol consumption (> 5 or < 5 units/day).

The follow-up ended on March 1st 2014 and recurrent events were defined as hospitalisations directly related to a new episode of AF, with a severity or duration of symptoms leading to hospital contact and ensuing admittance to the cardiology ward following evaluation by the cardiologist on duty. The AF diagnosis was confirmed by ECG and possible electrical or pharmacological cardioversion treatment was confirmed in the medical records. Only hospitalisations where symptomatic AF with or without cardioversion treatment was the primary reason for admittance to the ward were classified as recurrent events. There was no loss to follow-up in this dataset, but censoring still occurred at the end of follow-up for all patients. Due to censoring, the follow-up times vary according to the inclusion dates (it ranges from 1 year and 2 month to 6 years and 2 month) and dedicated survival analysis methods must be applied. A terminal event (TE) was defined as either a) progression to PermAF, defined as the date on which patient and physician agreed on accepting the presence of permanent or very frequent recurrent AF, or b) death, in which case the date of death was available in the medical record and used as the terminal date. All non-AF hospitalisations and visits to the outpatient clinic were scrutinised for potential progression to PermAF in the follow-up period.

A total of 174 patients were enrolled in the study. The mean follow up duration was $1,279$ days, and the patients all contributed to a total of $222,459$ person-days. There were 325 observed AFHs in the follow-up period, divided among 84 patients (ranging from 1 to 17 events per patient). 89% of all patient experienced from 1 to 7 events. A TE was experienced by 45 patients prior to the study end date, 18 due to death and 27 due to disease progression to PermAF. Baseline characteristics for the covariates are shown on Table 5.2. No patients in our sample had thyroid or renal disease of any kind (therefore not shown).

Variables	Levels	Freq. (Perc.)
AF type	paroxysmal	50(28.6)
	persistent	125(71.4)
gender	male	125(71.4)
	female	50(28.6)
age	median {iqr}	$63.0\{52.5, 68.0\}$
alcohol	$0-5$	93(56.4)
	5+	72(43.6)
	missing	10
tobacco	never	88(53.3)
	ex smoker	46(27.9)
	smoking	31(18.8)
	missing	10
hypertension	yes	82(46.9)
	no	93(53.1)
heart failure	yes	14(8.0)
	no	161(92.0)
heart valv dis	yes	12(6.9)
	no	163(93.1)
isch heart dis	yes	23(13.1)
	no	152(86.9)
diabetes	no	151(86.3)
	yes	24(13.7)
copd	yes	11(6.3)
	no	164(93.7)

Table 5.2: Descriptive statistics of the baseline covariates with frequencies and percentages for the categorical variables. For age, the median and interval quantile range (iqr) are provided.

We first start with a non-parametric analysis. The terminal event must be accounted in our analysis (45 observed TE for 174 patients) and we use the estimator (2.11) presented in the Introduction section to compute the estimated cumulative mean number of AFHs over time. This estimator is further computed on the subsamples of PAF and PeAF and the results are displayed on Figure 5.2. On average, a patient will experience one AFH after 635 days and two AFHs after 1 613 days. As shown in the plots, persistent patients have a much worse condition than paroxysmal patients. For example, if the patient has persistent AF then he/she will experience one AFH after 485 days on average and if the patient has paroxysmal AF then he/she will experience one AFH after after 1 146 days on average.

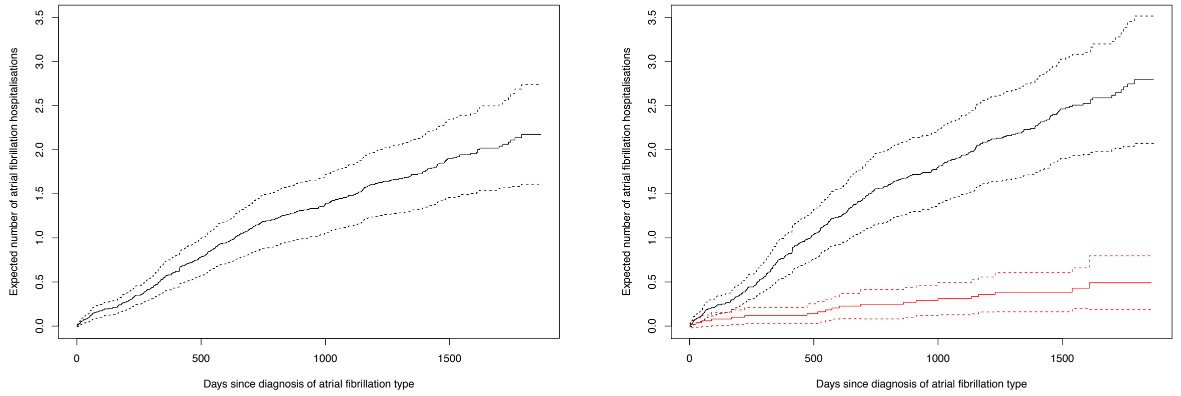


Figure 5.2: Estimated cumulated mean number of AF hospitalisations over time along with 95% confidence intervals. The left panel is for the whole population while the right panel is for the subsamples of PAF and PeAF.

Next a Cox model for the recurrent event process and a Cox model for the terminal event were implemented, both including all the covariates. The recurrent event model was also implemented with dependence on prior recurrent events as defined in Equation (2.13) (note in particular that both models for the recurrent event process and the terminal event are implemented). This multi-state situation is also described by Figure 5.3 with 3 possible states for the AFHs and one absorbing state corresponding to the terminal event. For simplicity, the hazard rates for the terminal event were all assumed equal such that λ^T does not depend on prior recurrences. Also $\lambda_1^E, \lambda_2^E, \lambda_3^E$ are assumed to be proportional. This model is summarised as follows. We first introduce the dichotomous covariate AFtype representing the type of AF and all the other covariates are represented by X_j, for $j = 1, \dots, 9$.

$$\mathbb{E}[dN(t)|Y_s(t), X] = Y_s(t)\lambda_s^E(t)dt,$$
$$\mathbb{E}[dN^T(t)|Y_s(t), X(t)] = Y_s(t)\lambda^T(t|X(t))dt,$$

where $Y_s(t) = I(N(t-) = s - 1)$, $s = 1, 2, 3, \dots$ and the $\lambda_s^E(t)$ correspond to the rate functions in the multi-state model of Figure 5.3 with the convention $\lambda_s^E = \lambda_3^E$ for $s \geq 3$. The $\lambda_s^E(t)$ are then modelled as:

$$\lambda_1^E(t) = \lambda_0(t) \exp\left(\beta_0 \text{AFtype} + \sum_{j=1}^{9} \beta_j X_j\right),$$
$$\lambda_2^E(t) = \lambda_0(t) \exp\left(\beta_0 \text{AFtype} + \alpha_{01} + \sum_{j=1}^{9} \beta_j X_j\right),$$
$$\lambda_3^E(t) = \lambda_0(t) \exp\left(\beta_0 \text{AFtype} + \alpha_{02} + \sum_{j=1}^{9} \beta_j X_j\right).$$

The parameter β_0 corresponds to the effect of AF type on the risk of further recurrences, the β_js, $j = 1, \dots, 9$, correspond to the effect of all other covariates and the α_{0s} $(s = 1, 2)$ correspond to the effects of prior recurrences on the risk to experience new ones. The $e^{\alpha_{0s}}$ should be interpreted as hazard ratio for the risk of experiencing a new recurrent event knowing the patient has already experienced s AFHs as compared to the risk of experiencing a first AFH.

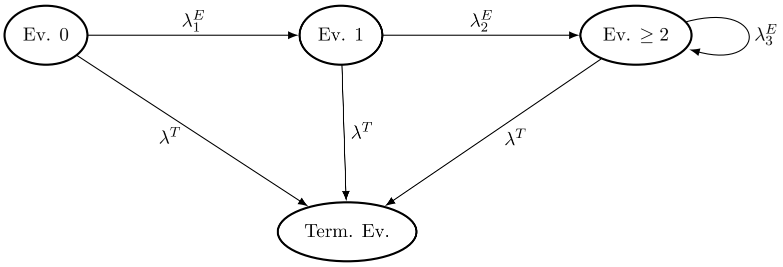


Figure 5.3: The recurrent event model with dependence on prior events seen as a multi-state situation for the atrial fibrillation study. All individuals start in the state Ev. 0. The death hazard rates λ^D are all identical. In the regression setting, all the rate functions for the recurrent events are assumed to be proportional.

The results for this model are shown in Table 5.3. In the last column, the pvalues for the standard Cox model (without dependence on prior recurrences) are also displayed. Both models show a very significant effect of AF type with a 3.64 fold increase in risk for persistent patients as compared to paroxysmal patients in the multi-state model. Both models disagree on the effect of age (borderline significant for the multi-state model and highly significant in the standard Cox model), diabetes (borderline significant for the multi-state model and highly significant in the standard Cox model) and of alcohol consumption (pvalue= 18% for the multi-state model and pvalue= 5% in the standard Cox model). The multi-state model exhibits strong and very significant effects of previous AFHs with a risk 2.6 times higher for patients that have already experienced one AFH and a risk 6.5 times higher for patients that have already experienced two AFH as compared to patients that have not yet experienced any AFH. These strong effects of previous recurrences highlight heterogeneity in the model that is not taken into account in the standard Cox model. The Cox model for the terminal event shows non significant effects of all covariates except for age which is highly significant (results not shown).

	Hazard ratio	2.5 %	97.5 %	p-value	p*
AF type (persistent)	3.64	2.41	5.50	0.0000	0.0000
gender (female)	1.11	0.81	1.52	0.5249	0.9596
age	0.99	0.97	1.00	**0.0546**	**0.0243**
hypertension (no)	0.85	0.59	1.21	0.3650	0.3854
heart fail. (no)	0.90	0.51	1.59	0.7274	0.9539
heart valv. dis. (no)	1.19	0.80	1.77	0.4030	0.5127
isch. heart dis. (no)	0.96	0.49	1.86	0.9037	0.7070
diabetes (yes)	0.48	0.21	1.12	**0.0910**	**0.0090**
copd (no)	0.76	0.46	1.25	0.2810	0.2810
alcohol (5+)	0.78	0.54	1.13	**0.1828**	**0.0507**
AF 1	2.62	1.76	3.89	0.0000	
AF 2+	6.50	4.42	9.56	0.0000	

Table 5.3: Results from the Cox model with dependence on prior recurrences. In the last column, the p* represent the p-values obtained from the previous model without adjustment with respect to the number of previous AFHs.

In order to perform individual predictions taking into account AFH history, we select a reduced model from a stepwise variable selection procedure at the 15% level. However, the diabetes covariate was omitted from the final model. This decision was based both on the relatively small proportion of patients with diabetes in our patient sample (only 14%) and the enigmatic nature of AF symptomatology in diabetic patients; diabetes is a risk factor for AF but may also contribute to a larger proportion of silent AF (see [SPAC12] and [RSM$^+$15]) that could not be ascertained in our study set-up, as we did not have access to continuous patient ECG monitoring. Other methods, such as Lasso methods were also implemented which selected the same model. The resulting model is shown in Table 5.4. The model for the terminal event only includes the effect of age, its corresponding hazard ratio is equal to 1.05 and is highly significant (pvalue$< 10^{-4}$).

	Hazard ratio	2.5 %	97.5 %	p-value
AF type (persistent)	3.20	2.01	5.11	0.0000
age	0.99	0.98	1.00	0.0909
AF 1	2.97	2.04	4.32	0.0000
AF 2+	7.54	5.47	10.40	0.0000

Table 5.4: Final Cox model with dependence on prior recurrences used for the individual predictions of future AFHs.

Finally, we want to use this model in order to provide predictions on the risk of future AFHs of a patient knowing his/her history of previous AFHs, his/her age and his/her type of AF. As presented in Section 2.2 of the Introduction, transition intensities could be plotted but these quantities only consider the probability of experiencing exactly one new event and do not consider all the other events. As a matter of fact, it turns out that it is often more useful in practice to compute the probability of visiting the next state (that is of experiencing a new event) in the future knowing the actual state (that is the number of recurrent events already experienced) of the individual. For prediction purposes, this probability provides an estimation of the risk of health deterioration and take into account all the future events. This probability is also increasing with respect to time while transition intensities are not necessarily monotone. We explain, in what follows, how to compute these prediction curves.

Let x and y be two time points such that $x < y$, where x corresponds to the current time and y is the time for prediction. We use the multi-state framework, such as in Section 2.2 of the Introduction. Given that the individual is in state s at time x, the probability of staying in this state between times x and y is equal to

$$\tilde{P}^E_{s,s}(x,y) = \exp\left(-\int_x^y (\lambda^E_{s+1}(u) + \lambda^T(u))du\right), \ s = 0, 1.$$

Given that the individual is in state s at time x, the probability that the individual will pass through the next state during the time interval $[x, y]$ is equal to

$$\tilde{P}^E_{s,s+1}(x,y) = \int_x^y \exp\left(-\int_x^u (\lambda^E_{s+1}(v) + \lambda^T(v))dv\right)\lambda^E_{s+1}(u)du, \ s = 0, 1.$$

Note the difference between the expressions of $\tilde{P}^E_{s,s+1}$ and $P^E_{s,s+1}$ from Equation (2.7). In $\tilde{P}^E_{s,s+1}$, the individual is allowed to experience any new events (terminal event or recurrent events) between the times u and y in the integral.

The last state corresponding to $s = 2$ must be dealt with in a different manner, since when an in individual is in the state Ev. ≥ 2, he/she is continuously at risk of experiencing new events with rate equal to λ^E_3. Given that the individual is in state 2 at time x, the probability of staying in this state between times x and y is equal to

$$\tilde{P}^E_{2,2}(x,y) = \exp\left(-\int_x^y \lambda^T(u)du\right).$$

This probability is just the individual's survival probability as the only risk that he/she is exposed is the terminal event. Given that the individual is in state 2 at time x, the probability that the individual will experience at least one more event during the time interval $[x, y]$, is equal to

$$\tilde{P}^E_{2,2+}(x,y) = 1 - \exp\left(-\int_x^y \exp\left(-\int_x^u \lambda^T(v)dv\right)\lambda^E_3(u)du\right).$$

Finally, since the terminal event is a competing risk, it is often important to also reports risks associated with the terminal event. Indeed, these previous curves can be misleading in case the hazard rate of the terminal event is high as compared to the recurrent event rates. Low values of the $\tilde{P}_{s,s+1}$ could then be solely due to high risk of experiencing a terminal event: for example, an individual that is at very high of dying will be at very low risk of experiencing an hospitalisation since it is very likely that the patient will die before the hospitalisation occurs.

Given that the individual is in state s at time x, the probability of experience a terminal event between times x and y is equal to

$$\tilde{P}_{s,T}(x,y) = 1 - \exp\left(-\int_x^y \lambda^T(u)du\right), \quad s = 0, \ldots, 2.$$

Given that the individual is in state s at time x, the probability that the individual will either pass through the next recurrent event state or will experience a terminal event during the time interval $[x,y]$ is equal to

$$\tilde{P}_{s,ET}(x,y) = 1 - \exp\left(-\int_x^y \left(\lambda^T_{s+1}(u)du + \lambda^E_{s+1}(u)du\right)\right), \quad s = 0, \ldots, 2.$$

All these prediction estimators are estimated using plug-in estimators $\widehat{\lambda^E_s}$ and $\widehat{\lambda^T}$. In the AF study, the risk of experiencing a terminal event is very low, so for sake of simplicity we only shows the curves of the type $\widehat{\tilde{P}^E_{s,s+1}}$ and the curves for $\widehat{\tilde{P}^E_{s,ET}}$ are omitted since they are very similar. The estimated curve $\widehat{\tilde{P}^E_{s,s+1}}$ is shown on Figure 5.4 for a hypothetical 60 year old patient with persistent AF with $s = 180$ days. We can see that previous AF hospitalisations dramatically increase the risk of future AF hospitalisations. Figure 5.4 also shows the estimated curve $\widehat{\tilde{P}_{s,T}}$ for the same patient, this curve does not depend on the number of previous AFHs since the hazard rate for the terminal event is a function of age only.

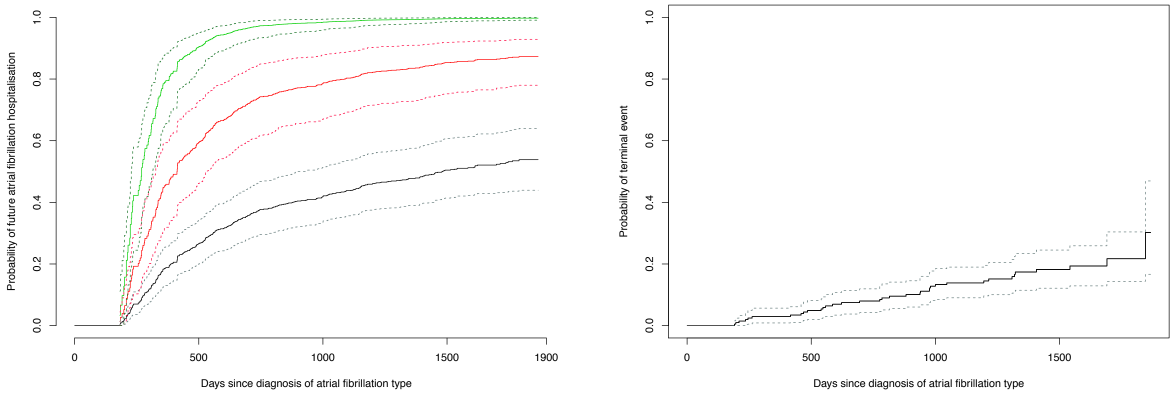


Figure 5.4: Predictions of future AF hospitalisation (left panel) and future terminal event (right panel) for a 60 years old patient with persistent AF, after 180 days since study entry. On the left panel, the gray curve represents the risk for a patient that has never experienced AF hospitalisation, the red curve represents the risk for a patient that has already experienced one AF hospitalisation and the the green curve represents the risk for a patient that has already experienced two AF hospitalisations. 95% confidence intervals are plotted along the curves.

CHAPTER 6

Perspectives

6.1 Detecting heterogeneity in survival analysis in a multidimensional space

The breakpoint model presented in the article [P5] could be extended to detect heterogeneity along several covariates instead of just one. This would allow to build a powerful data mining tool for survival data. In order to use the method developed in [P5] in a multi-dimensional setting, the main key is to find an adequate way of ordering the data according to several continuous covariates. Let X be a d dimensional covariate vector. Each individual can be characterised by its coordinate X_i in a d dimensional space. Then, after performing a Principal Component Analysis (PCA) we can project each point on the space of the principal components (let us say three for instance), and then fit a "principal curve" from a standard smoother such as local polynomial smoother. Finally, projecting the points of each individual on the smoothed curve allows to define a natural ordering of the points. Our breakpoint model can then be directly applied on these ordered survival data. The R package **princurve** fits a principal curve on a data matrix of arbitrary dimension and could be combined with our algorithm to allow for heterogeneity detection in a multidimensional space.

6.2 Extension of classical age-period-cohort models

In [S2], we studied the age-period-cohort model by considering the hazard as a bivariate function due to the relationship period=age+cohort. However, standard APC models fit an age effect, a period effect and a cohort effect in a regression model and one of the main goal is to retrieve these estimated effects. These models are very constrained as they impose the same age effect for every cohort and every period and vice versa. An extension of these models could be to consider the model:

$$\log \lambda_{j,k} = \mu + \alpha_j + \beta_k + \delta_{j,k},$$

where μ is an intercept term, α_j represents the age effect, β_k the cohort effect and $\delta_{j,k}$ is an interaction term. This model allows a great flexibility by including an interaction term and it encompasses the age-cohort model in the situation $\delta_{j,k} = 0$. However, the model will be over-parametrised even for small sample sizes. To remedy to this problem, we propose to use a likelihood based method penalised with respect to $|\delta_{j,k}|$ using our adaptive ridge procedure. This will result in the selection of few areas where the interaction term is non-null which correspond to areas where the standard age-cohort model is not appropriate. As a result the model provides interpretable age and cohort effects and also takes into account an interaction term which generalises the standard age-cohort model.

6.3 A Brier Score for evaluating the prediction performance of recurrent event models

In the Atrial Fibrillation study [S4], we produced predictive models but the prediction performance of the final model was not assessed. In survival analysis, standard tools such as the ROC or AUC curves cannot be directly used since the event of interest is not always observed. New tools, dedicated to deal with time to event variables have been developed. Among them, the Brier score [GS06] is a useful one to assess the predictive ability of a model over time. This score is based on estimated conditional survival functions and works for time to event data but has not yet been generalised to multiple events such as recurrent events. A modification of the Brier score could be defined by evaluating the cumulative mean function for recurrent events. Let N^* be the (unobserved) recurrent event of interest. In the absence of terminal event, $\int_0^t \mathbb{E}[dN(u)]/(1-G(u)) = \mathbb{E}[N^*(t)]$, where $N_i(t) = N_i^*(t \wedge C_i)$ and G is the cumulative distribution function of the censoring variable C. Now, using our prediction model for $\hat{\lambda}(\cdot|X)$ the quantity $\int_0^t \hat{\lambda}(u|X)du$ should be a good predictor of $\mathbb{E}[N^*(t)]$. As a consequence, our predictive score could be defined as:

$$\sum_{i=1}^{n} \int \left(\int_0^t \hat{\lambda}(u|X_i)du - \int_0^t \frac{dN_i(u)}{1-\hat{G}(u)} \right)^2 dt,$$

where $1-\hat{G}(\cdot)$ is the Kaplan-Meier estimator of the censoring distribution. To take into account a terminal event, the previous formula can be modified by

$$\sum_{i=1}^{n} \int \left(\int_0^t \hat{S}(u)\hat{\lambda}(u|X_i)du - \int_0^t \hat{S}(u)\frac{dN_i(u)}{\sum_j Y_j(u)} \right)^2 dt,$$

where $\hat{S}$ is the Kaplan-Meier estimator of the terminal event, $N_i(t) = N_i^*(t \wedge C_i \wedge T_i^*)$ and $Y_i(t) = I(T_i^* \wedge C_i \geq t)$. This predictive score can be evaluated for our prediction model developed in [S4]. Our model could then be compared with other methods to predict future atrial fibrillation hospitalisations. For example [GL02] proposed a Cox model that works when censoring times are known. Since censoring is only due to the end of study in the atrial fibrillation dataset, their estimation method could also be implemented. Finally, our single-index modelling approach introduced in [P2] could provide an alternative model. The single-index model is more general than the Cox model and it is very likely that we could obtain more performant predictions using this model.

6.4 A fast algorithm for regression modelling of interval censored data

In [S3] we provided a new regression modelling of interval censored data using a piecewise constant baseline. This method uses the adaptive ridge algorithm and is a performant estimation method in case of interval censored data. An alternative way to fit regression modelling with interval censored while keeping a flexible baseline hazard is to use pseudo-regression. This method was developed by [AKR03] and allows to perform regression modelling from a non parametric estimator of the survival function. A Jackknife technique is used to generate pseudo-observations and the regression estimates are derived from a generalised linear model with an adequate link function.

Therefore, our idea is to use a non-parametric estimator of the survival function such as the Turnbull estimator. Using the notations of Section 2.4.1 of the Introduction, we define the lth leave one out estimate $\hat{s}_j^{-l}$ by omitting the lth observation when computing the estimator. It is defined as:

$$\hat{s}_j^{-l} = \frac{1}{n-1} \sum_{i \neq l} \frac{\alpha_{ij} \hat{s}_j^{-l}}{\sum_k \alpha_{ik} \hat{s}_k^{-l}}.$$

However, computing all pseudo-estimates can be time consuming since the Turnbull estimator itself is slow to compute. We propose to compute the alternative estimator:

$$\tilde{s}_j^{-l} = \frac{1}{n-1} \sum_{i \neq l} \frac{\alpha_{ij} \hat{s}_j}{\sum_k \alpha_{ik} \hat{s}_k}.$$

and to use the approximation $\hat{s}_j^{-l} \approx \tilde{s}_j^{-l}$. Straightforward calculations then give:

$$\tilde{s}_j^{-l} = \hat{s}_j + \frac{\hat{s}_j}{n-1} \left(1 - \frac{\alpha_{lj}}{\sum_k \alpha_{lk} \hat{s}_k} \right).$$

These approximated pseudo-estimates can be easily computed for all j and l. Then, the lth *pseudo-value* of $1 - \hat{S}$ is defined as:

$$1 - \hat{S}_{(l)}(t) = m(\hat{s}_1 + \cdots + \hat{s}_j) - (m-1)(\tilde{s}_1^{-l} + \cdots + \tilde{s}_j^{-l}),$$

for $t \in (p_j, q_{j+1})$. These approximated pseudo-estimates were used to perform a Cox regression and lead to very performant estimations of the regression coefficients. They are fast to compute since the approximated pseudo-estimates can be directly computed from the initial estimator $\hat{s}_j$ and do not need to resample the data. On the opposite, for the standard leave-one out estimates the Turnbull estimator needs to be computed n times.

Finally, the idea of providing fast pseudo-estimates is also of interest for right-censored data. In that case, Von Mises expansion (see for example [VDVW96]) can be used in order to approximate the difference between the Kaplan-Meier estimator and its leave one out estimate. This would also provide direct formulas for computing pseudo-estimates and should lead to a faster algorithm than the standard Jackknife method where the Kaplan-Meier estimator is computed n times.

Bibliography

[AA99] Douglas G Altman and Per K Andersen. Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ*, 319(7223):1492–1495, 1999.

[Aal75] Odd O Aalen. *Statistical inference for a family of counting processes.* PhD thesis, University of California, Berkeley, 1975.

[Aal80] Odd O Aalen. A model for non-parametric regression analysis of counting processes. In Kozek A. Klonecki W. and J. Rosinski, editors, *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory*, pages 1–25, New-York, 1980. Springer-Verlag.

[Aal89] Odd O Aalen. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989.

[ABG08] Odd O Aalen, Ørnulf Borgan, and Hakon Gjessing. *Survival and Event History Analysis.* Statistics for Biology and Health. Springer, 2008.

[ABGK93] Per K Andersen, Ørnulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York, 1993.

[ABJA95] JO Andreasen, MK Borum, HL Jacobsen, and FM Andreasen. Replantation of 400 avulsed permanent incisors. 1. diagnosis of healing complications. *Dental Traumatology*, 11(2):51–58, 1995.

[ABY16] Clifford Anderson-Bergman and Yaming Yu. Computing the log concave npmle for interval censored data. *Statistics and Computing*, 26(4):813–826, 2016.

[AG82] Per K Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[AK12] Per K Andersen and Niels Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31(11-12):1074–1088, 2012.

[AKR03] Per K Andersen, John P Klein, and Susanne Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

[APSE04] Antonis C Antoniou, Paul Pharoah, Paula Smith, and Douglas F Easton. The boadicea model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer*, 91(8):1580–1590, 2004.

[BBEL07] Kelly Broen, Kim Brustoski, Ilka Engelmann, and Adrian JF Luty. Placental plasmodium falciparum infection: causes and consequences of in utero sensitization to parasite antigens. *Molecular and biochemical parasitology*, 151(1):1–8, 2007.

[BBM$^+$81] Robert Bartoszynski, Barry W Brown, Charles M McBride, James R Thompson, et al. Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary poisson process. *The Annals of Statistics*, 9(5):1050–1060, 1981.

[BC09] Elodie Brunel and Fabienne Comte. Cumulative distribution function estimation under interval censoring case 1. *Electronic journal of statistics*, 3:1–24, 2009.

[BDS05] John Braun, Thierry Duchesne, and James E Stafford. Local likelihood density estimation for interval censored data. *Canadian Journal of Statistics*, 33(1):39–60, 2005.

[Ben83] Steve Bennett. Analysis of survival data by the proportional odds model. *Statistics in medicine*, 2(2):273–277, 1983.

[Ber81] Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, Technical Report, Univ. California, Berkeley, 1981.

[BFZ$^+$14] Charles L Bailey, Christopher B Forrest, Peixin Zhang, Thomas M Richards, Alice Livshits, and Patricia A DeRusso. Association of antibiotics in infancy with early childhood obesity. *JAMA pediatrics*, 168(11):1063–1069, 2014.

[BJ79] Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

[BKV$^+$06] S Brugman, FA Klatter, JTJ Visser, ACM Wildeboer-Veloo, HJM Harmsen, J Rozing, and NA Bos. Antibiotic treatment partially protects against type 1 diabetes in the bio-breeding diabetes-prone rat. is the gut flora involved in the development of type 1 diabetes? *Diabetologia*, 49(9):2105–2108, 2006.

[Bre72] Norman E Breslow. Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

[Bya80] DP Byar. The veterans administration study of chemoprophylaxis for recurrent stage 1 bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa. In *Bladder tumors and other topics in urological oncology*, pages 363–370. Springer, 1980.

[Car07] Bendix Carstensen. Age–period–cohort models for the lexis diagram. *Statistics in medicine*, 26(15):3018–3045, 2007.

[CBE$^+$16] Tine Dalsgaard Clausen, Thomas Bergholt, Frank Eriksson, Steen Rasmussen, Niels Keiding, and Ellen C Løkkegaard. Prelabor cesarean section and risk of childhood type 1 diabetes. *Epidemiology*, 27(4):547–555, 2016.

[CC08] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[CGC18] Fabienne Comte and Valentine Genon-Catalot. Nonparametric regression with non compactly supported bases. Submitted, 2018.

[CH93] David G Clayton and Michael Hills. *Statistical models in epidemiology.* OUP Oxford, 1993.

[CKP$^+$12] Gilles Cottrell, Bienvenue Kouwaye, Charlotte Pierrat, Agnès Le Port, Aziz Bouraïma, Noël Fonton, Mahouton Norbert Hounkonnou, Achille Massougbodji, Vincent Corbel, and André Garcia. Modeling the influence of local environmental factors on malaria transmission in benin and its implications for cohort study. *PLoS One*, 7(1):e28812, 2012.

[CL07] Richard J Cook and Jerald Lawless. *The statistical analysis of recurrent events.* Springer Science & Business Media, 2007.

[Cla78] David G Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[Cox72] David R Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

[CPLH18] Bendix Carstensen, Martyn Plummer, Esa Laara, and Michael Hills. *Epi: A Package for Statistical Analysis in Epidemiology*, 2018. R package version 2.30.

[CRT91] Elisabeth B Claus, Neil Risch, and Douglas W Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232, 1991.

[CRT94] Elizabeth B Claus, Neil Risch, and W Douglas Thompson. Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction. *Cancer*, 73(3):643–651, 1994.

[CSJ$^+$08] Christopher R Cardwell, LC Stene, G Joner, O Cinek, J Svensson, MJ Goldacre, RC Parslow, P Pozzilli, G Brigis, D Stoyanov, et al. Caesarean section is associated with an increased risk of childhood-onset type 1 diabetes mellitus: a meta-analysis of observational studies. *Diabetologia*, 51(5):726–735, May 2008.

[CWB08] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[CWH05] Chin-Tsang Chiang, Mei-Cheng Wang, and Chiung-Yu HUANG. Kernel estimation of rate function for recurrent event data. *Scandinavian journal of statistics*, 32(1):77–91, 2005.

[CWL$^+$06] Sining Chen, Wenyi Wang, Shing Lee, Khedoudja Nafa, Johanna Lee, Kathy Romans, Patrice Watson, Stephen B Gruber, David Euhus, and Kenneth W Kinzler. Prediction of germline mutations and cancer risk in the lynch syndrome. *Jama*, 296(12):1479–1487, 2006.

[DHH03] Michel Delecroix, Wolfgang Härdle, and Marian Hristache. Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86(2):213–226, 2003.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[DP12] Antoine De Pauw. *Estimation des risques de cancer du sein et de l'ovaire des femmes sans mutation des gènes BRCA1et BRCA2: apport des modèles de calcul de risque*. PhD thesis, Paris 7, 2012.

[EBFC93] Douglas F Easton, DT Bishop, D Ford, and GP Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. the breast cancer linkage consortium. *American journal of human genetics*, 52(4):678, 1993.

[Efr67] Bradley Efron. The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853. University of California Press, Berkeley, CA, 1967.

[EM05] Uwe Einmahl and David M Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

[Far82a] Vern T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

[Far82b] Vern T Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

[FH91] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley and Sons Inc., New York, 1991.

[FN16] Florian Frommlet and Grégory Nuel. An adaptive ridge procedure for L0 regularization. *PLoS ONE*, 11(2):1–23, 02 2016.

[GK11] Piet Groeneboom and Tom Ketelaars. Estimators for the interval censoring problem. *Electronic Journal of Statistics*, 5:1797–1845, 2011.

[GL00] Debashis Ghosh and DY Lin. Nonparametric analysis of recurrent events and death. *Biometrics*, 56(2):554–562, 2000.

[GL02] Debashis Ghosh and Danyu Y Lin. Marginal regression models for recurrent and terminal events. *Statistica Sinica*, pages 663–688, 2002.

[GL11] Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.

[Gro96] Piet Groeneboom. Lectures on inverse problems. In *Lectures on probability theory and statistics*, pages 67–164. Springer, 1996.

[GS06] Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

[GW92a] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media, 1992.

[GW92b] Piet Groeneboom and Jon A Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science and Business Media, 1992.

[HBJ13] Klaus K Holst and Esben Budtz-Jørgensen. Linear latent variable models: the lava-package. *Computational Statistics*, 28(4):1385–1452, 2013.

[Heu97] Carsten Heuer. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, pages 161–177, 1997.

[HHI93] Wolfgang Hardle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, pages 157–178, 1993.

[HNM$^+$90] MK Hall, JMand Lee, B Newman, J Morrow, LA Anderson, B Huey, and MC King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684, 1990.

[Hol83] Theodore R Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, pages 311–324, 1983.

[Hou95] Philip Hougaard. Frailty models for survival data. *Lifetime data analysis*, 1(3):255–273, 1995.

[HS09] Anders Hviid and Henrik Svanström. Antibiotic use and type 1 diabetes in childhood. *American journal of epidemiology*, 169(9):1079–1084, 2009.

[Ich93] Hidehiko Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

[JLEJ10] Cecilia Jernberg, Sonja Löfmark, Charlotta Edlund, and Janet K Jansson. Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology*, 156(11):3216–3223, 2010.

[Jon98] Geurt Jongbloed. The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7(3):310–321, 1998.

[Kei90] Niels Keiding. Statistical inference in the lexis diagram. *Philosophical Transactions: Physical Sciences and Engineering*, 332(1627):487–509, 1990.

[KF00] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KM58] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[KR05] Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[KS92] Charles Kooperberg and Charles J Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328, 1992.

[Len77] E Lenglart. Relation de domination entre deux processus. *Ann. Inst. H. Poincaré Sect. B (NS)*, 13(2):171–179, 1977.

[LHCP+97] Jean Yves Le Hesran, Michel Cot, Philippe Personne, Nadine Fievet, Béatrice Dubois, Mathilde Beyeme, Christian Boudin, and Philippe Deloron. Maternal placental infection with plasmodium falciparum and malaria morbidity during the first 2 years of life. *American journal of epidemiology*, 146(10):826–831, 1997.

[LN95] Jerald F Lawless and Claude Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168, 1995.

[LNC97] Jerald F Lawless, Claude Nadeau, and Richard J Cook. Analysis of mean and rate functions for recurrent events. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 37–49. Springer, 1997.

[LPCC+13] Agnès Le Port, Gilles Cottrell, Fabrice Chandre, Michel Cot, Achille Massougbodji, and André Garcia. Importance of adequate local spatiotemporal transmission measures in malaria cohort studies: application to the relation between placental malaria and first malaria infection in infants. *American journal of epidemiology*, 178(1):136–143, 2013.

[LPCMP+12] Agnès Le Port, Gilles Cottrell, Yves Martin-Prevel, Florence Migot-Nabias, Michel Cot, and André Garcia. First malaria infections in a cohort of infants in benin: biological, environmental and genetic determinants. description of the study site, population methods and preliminary results. *BMJ open*, 2(2):e000342, 2012.

[LPWC+11] Agnès Le Port, Laurence Watier, Gilles Cottrell, Smaila Ouédraogo, Célia Dechavanne, Charlotte Pierrat, Antoine Rachas, Julie Bouscaillou, Aziz Bouraima, and Achille Massougbodji. Infections in infants during the first 12 months of life: role of placental malaria and environmental factors. *PloS one*, 6(11):e27516, 2011.

[LRN13] The Minh Luong, Yves Rozenholc, and Grégory Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. *Computational Statistics & Data Analysis*, 68:129–140, 2013.

[LS03] Steffen L Lauritzen and Nuala A Sheehan. Graphical models for genetic analyses. *Statistical Science*, pages 489–514, 2003.

[LWYY00] Danyu Y Lin, Lee-Jen Wei, I Yang, and Z Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730, 2000.

[MA16] Amir Mehrgou and Mansoureh Akouchekian. The importance of brca1 and brca2 genes mutations in breast cancer development. *Medical Journal of the Islamic Republic of Iran*, 30:369, 2016.

[MS07] Torben Martinussen and Thomas H Scheike. *Dynamic regression models for survival data.* Springer Science & Business Media, 2007.

[MS09a] Torben Martinussen and Thomas H Scheike. The additive hazards model with high-dimensional regressors. *Lifetime data analysis*, 15(3):330–342, 2009.

[MS09b] Torben Martinussen and Thomas H Scheike. Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, 36(4):602–619, 2009.

[MU90] Ian W McKeague and Klaus J Utikal. Identifying nonlinear covariate effects in semimartingale regression models. *Probability theory and related fields*, 87(1):1–25, 1990.

[Nel72] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[Nie15] Bent Nielsen. apc: An r package for age-period-cohort analysis. *R Journal*, 7(2), 2015.

[OG82] Clive Osmond and MJ Gardner. Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1(3):245–259, 1982.

[PC93] Margaret S Pepe and Jianwen Cai. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American statistical Association*, 88(423):811–820, 1993.

[PD00] Yingwei Peng and Keith BG Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

[Pol90] David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

[PWP81] Ross L Prentice, Benjamin J Williams, and Arthur V Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

[Reb78] Rolando Rebolledo. Sur les applications de la théorie des martingales à l'étude statistique d'une famille de processus ponctuels. In *Journées de Statistique des Processus Stochastiques*, pages 27–70. Springer, 1978.

[Reb80] Rolando Rebolledo. Central limit theorems for local martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 51(3):269–286, 1980.

[RFJ06] Virginie Rondeau, Laurent Filleul, and Pierre Joly. Nested frailty models using maximum penalized likelihood estimation. *Statistics in medicine*, 25(23):4036–4052, 2006.

[Riz10] Dimitris Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[Riz12] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R.* CRC Press, 2012.

[RME12] Ralph CA Rippe, Jacqueline J Meulman, and Paul HC Eilers. Visualization of genomic changes by segmented smoothing using an L0 penalty. *PloS one*, 7(6):e38230, 2012.

[RMG12] Virginie Rondeau, Yassin Mazroui, and Juan R Gonzalez. Frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, 47(4):1–28, 2012.

[RMPJG$^+$07] Virginie Rondeau, Simone Mathoulin-Pelissier, Hélène Jacqmin-Gadda, Véronique Brouste, and Pierre Soubeyran. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4):708–721, 2007.

[RP00] Samuli Ripatti and Juni Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, 2000.

[RSM$^+$15] Maria R Rizzo, Ferdinando C Sasso, Raffaele Marfella, Mario Siniscalchi, Pasquale Paolisso, Ornella Carbonara, Maria Carmela Capoluongo, Nadia Lascar, Caterina Pace, and Celestino Sardu. Autonomic dysfunction is associated with brief episodes of atrial fibrillation in type 2 diabetes. *Journal of Diabetes and its Complications*, 29(1):88–92, 2015.

[Sch02] Thomas H Scheike. The additive nonparametric and semiparametric aalen model as the rate function for a counting process. *Lifetime Data Analysis*, 8(3):247–262, 2002.

[SD01] Glen A Satten and Somnath Datta. The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210, 2001.

[SNPM01] Richard W Steketee, Bernard L Nahlen, Monica E Parise, and Clara Menendez. The burden of malaria in pregnancy in malaria-endemic areas. *The American journal of tropical medicine and hygiene*, 64(1_suppl):28–35, 2001.

[SPAC12] Tobias Schoen, Aruna D Pradhan, Christine M Albert, and David Conen. Type 2 diabetes mellitus and risk of incident atrial fibrillation in women. *Journal of the American College of Cardiology*, 60(15):1421–1428, 2012.

[ST00] Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[Stu93] Winfried Stute. Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1):89–103, 1993.

[Sun07] Jianguo Sun. *The statistical analysis of interval-censored failure time data.* Springer Science & Business Media, 2007.

[Tal94] Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

[Tal95] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

[TD04] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[TG00] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model.* Statistics for Biology and Health. Springer Science and Business Media, New York, 2000.

[TSR$^+$05] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[Tur76] Bruce W Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

[Vaa12] Outi Vaarala. Gut microbiota and type 1 diabetes. *The review of diabetic studies: RDS*, 9(4):251, 2012.

[VDVW96] Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[VR08] Ravi Varadhan and Christophe Roland. Simple and globally convergent methods for accelerating the convergence of any em algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353, 2008.

[VWGK15] Pajau Vangay, Tonya Ward, Jeffrey S Gerber, and Dan Knights. Antibiotics, pediatric dysbiosis, and disease. *Cell host & microbe*, 17(5):553–564, 2015.

[Wil00] Rick L Williams. A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56(2):645–646, 2000.

[WLV$^+$08] Li Wen, Ruth E Ley, Pavel Yu Volchkov, Peter B Stranges, Lia Avanesyan, Austin C Stonebraker, Changyun Hu, F Susan Wong, Gregory L Szot, Jeffrey A Bluestone, et al. Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature*, 455(7216):1109, 2008.

[XTLZ02] Yingcun Xia, Howell Tong, WK Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[Yan00] Song Yang. Functional estimation under interval censoring case 1. *Journal of statistical planning and inference*, 89(1):135–144, 2000.

[YL13] Yang Yang and Kenneth C Land. *Age-period-cohort analysis: New models, methods, and empirical applications*. Chapman and Hall/CRC, 2013.

[Zou06] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.