Biometrika (2020), **xx**, x, *pp*. 1–16
Printed in Great Britain

Non-Smooth Backfitting for Excess Risk Additive Regression Model with Two Survival Time-Scales

BY M. HIABU

School of Mathematics and Statistics, University of Sydney
Camperdown NSW 2006, Australia
munir.hiabu@sydney.edu.au

J. P. NIELSEN

Cass Business School, City, University of London
106 Bunhill Row, London, EC1Y 8TZ, U.K.
Jens.Nielsen.1@city.ac.uk

T. H. SCHEIKE

Section of Biostatistics, Department of Public Health, University of Copenhagen
Øster Farimagsgade 5B, 1014 Copenhagen K, Denmark
thsc@sund.ku.dk

SUMMARY

We consider an extension of Aalen's additive regression model allowing covariates to have effects that vary on two different time-scales. The two time-scales considered are equal up to a constant that varies for each individual, such as for example follow-up time and age in medical studies or calendar time and age in longitudinal studies. The model has been introduced in Scheike (2001) where it was solved via smoothing techniques. We present a new backfitting algorithm estimating the structured model without having to use smoothing. Estimators of the cumulative regression functions on the two time-scales are suggested by solving local estimating equations jointly on the two time-scales. We provide large sample properties and simultaneous confidence bands. The model is applied to data on myocardial infarction providing a separation of the two effects stemming from time since diagnosis and age.

Some key words: Aalen model, counting process, disability model, illness-death model, generalized additive models, multiple time-scales, non-parametric estimation, varying-coefficient models.

1. INTRODUCTION

In many bio-medical applications in survival analysis it is of interest and needed to use multiple time-scales. A medical study will often have a follow-up time, for example time since diagnosis, for patients of different ages. In this case, both time-scales will contain important but different information about how the risk of, for example, dying is changing. We therefore consider the situation with two time-scales that are equivalent up to a constant for each individual, such as for example follow-up time and age. Specifically, if a patient is included in a study at age, a_0, then the age of the patient at follow-up time t is $a_o + t$. One may see this as arising from the the illness-death model, or the disability model, where the additional time-scale may be duration in the illness state of the model; see Keiding (1991) for a general discussion of these models. There is rather limited work on how to deal with multiple time-scales in a biomedical context, see for example Oakes (1995); Duchesne & Lawless (2000) and Iacobelli & Carstensen (2013); Lee et al. (2017) and references therein. The first two references deal with choosing a relevant time-scale via transformations, thus considering and finding one useful timescale, an aim rather different from ours, whereas Iacobelli & Carstensen (2013) an Lee et al. (2017) consider semi-parametric models for dealing with time-scales such as age and follow-up time in a multiplicative hazard setting. These models are quite flexible and easy to fit as pointed by Iacobelli & Carstensen (2013).

The aim of our approach is to consider the time-scales jointly and provide a simple non-parametric regression approach where each time-scale contributes additively to the mortality. The regression setting models the effect of covariates by additive Aalen models on each time-scale (Aalen, 1989; Huffer & McKeague, 1991; Andersen et al., 1993; Martinussen & Scheike, 2006). This allows covariates to have effects that change over two different time-scales. We are able to interpret that change of effect for each of the two time-scale separately.

We consider an extension of the additive Aalen structure in this paper. Whether one uses an additive or a multiplicative structure, like the Cox proportional hazard, to model the effects of the time-scales, and what model that fits best depends on the setting, but we here use the additive structure because the estimation turns out to be simpler on this scale. A structured approach, e.g., additive or multiplicative assuming a suitable fit, has the advantage that the number of covariates considered effects the estimation performance only linearly. A fully nonparametric approach is subject to the curse of dimensionality, i.e., exponentially deteriorating estimation performance in the number of covariates. Additionally, a structured model enables interpretation of the effects for each time scale by visualizing the one-dimensional components.

A popular setting were main effects of several time-scales is considered is the age-period-cohort model, see e.g., (Kuang et al., 2008a), where interest is on describing how the different time-scales, i.e., age, period and cohort, affect the hazard. Our model is a simper because we only allow for two time-scales but also more complicated because we allow for additional covariates to affect the hazard.

In an illustrative example, we consider patients that experience acute myocardial infarction (AMI), and aim at predicting the intensity considering the two time-scales age and time since myocardial infarction. In a first analysis we do not consider additional covariates. Here as expected we find a strong effect of the duration time-scale, with a much increased mortality just after the AMI. This duration effect is visible because our model automatically adjusts for possible age effects. As a consequence, we can make survival predictions for patients given their age at diagnosis. These predictions are direct functions of the mortality components on either time-scale. Second, we consider a structured regression approach of the same data where we study the importance of different factors on each time-scale. Here we see that only two covariates seem to be important for the duration time-scale and in addition we get a quantification of their importance.

The model of this paper was previously considered in Scheike (2001) where estimation was based on smoothing for one of the time-scales. A study closely related to ours is Kauermann & Khomski (2006) who studied the two most common time-scales: age and duration. They consider a multiplicative hazard model without covariates that is estimated via splines. In contrast our approach is an additive hazard model including covariates and estimating without smoothing. Alternative smoothing methodologies to multiplicative hazard estimation includes Linton et al. (2003); Huang (1999b); Hastie & Tibshirani (1986); Lin et al. (2016). None of the known multiplicative hazard approaches including the ones mentioned above are able to estimate without smoothing, include time varying covariate-effects, or are able to provide simultaneous confidence bands as the additive approach of this paper does provide. We do know that smoothing improves efficiencies of cumulatively estimated quantities, see Guillen et al. (2007) for the simplest possible case. However, smoothing is also a complexity and experts applying survival analysis have developed a practical way of smoothing by eye the underlying rough non-parametric estimators of Kaplan & Meier (1958); Nelson (1972). The advantage of providing estimators without smoothing is that there can be no confusion from the complicated process of picking the smoothing procedure first and the amount of smoothing after that. Even if a smoothing approach is eventually used, then the smoothing free procedure would always count as a benchmark approach to check whether something went wrong during the smoothing. Our backfitting approach is different from standard backfitting in regression (Mammen et al., 1999). In the backfitting approach of this paper, the non-parametric dynamics is only taking place in the two time directions, and the end result is therefore closer to the classical approach of Nelson (1972) with a non-smooth estimator of the dynamics in the one-dimensional time axis. What is obtained through Aalen's additive hazard regression model on two time axes is that the dynamics of the two time effects are adjusted for covariates in a way that keep the one-dimensional structure of the non-parametric dynamics. The expert user of survival methodology can therefore use the well developed intuition from looking at Nelson-Aalen estimators and Kaplan-Meier estimators when interpreting the empirical results based on the new methodology of this paper. Another advantage of estimating directly the cumulative functions is that we are able to obtain a simple uniform asymptotic description of our estimators. We are thus able to construct confidence bands and intervals, that are based on bootstrapping the underlying martingales.

2. Aalen's Additive Hazard Model for Two Time-Scales

Let $N_i(t)\ \ (i = 1, ..., n)$ be n independent counting processes that do not have common jumps and are adapted to a filtration that satisfy the usual conditions (Andersen et al., 1993). We are interested in hazard models where the hazard can be written as a sum of components of excess risk terms additively on two time-scales. In the simplest case with no additional covariates, the intensity of the counting process is $\lambda_i(t) = Y_i(t)\{\alpha_1(t) + \beta_1(t + a_i)\}$, where $Y_i(t)$ is the at-risk indicator, $\alpha_1(t)$ is the hazard, or excess hazard, related to the duration time-scale t and $\beta(t + a_i)$ gives the hazard on the age time-scale. In the more general case we consider a regression formulation of this model that makes it possible, for example, to compare the components on each time-scale for males and females, or when additivity is not perfectly satisfied one might allow different components for the duration time-scale depending on strata defined from a_i.

We assume that the counting processes have intensities given by

$$\begin{aligned}\lambda_i(t) &= \sum_{j=1}^{p} X_{ij}(t)\alpha_j(t) + \sum_{k=1}^{q} Z_{ik}(t)\beta_k(t + a_i)\\ &= X_i(t)\alpha(t) + Z_i(t)\beta(t + a_i), \quad (0 \leq t \leq t_{max}), \qquad (1)\end{aligned}$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)^T$ and $\beta = (\beta_1, \ldots, \beta_q)^T$ are vectors of unknown one-dimensional deterministic functions. We do not impose any structural assumption on α, β. The vectors $X_i^T(t) \in \Re^p$ and $Z_i^T(t) \in \Re^q$ are predictable cadlag covariates with $X(t)$ and $Z(t)$ having almost surely full rank, and a_i is a real-valued random variable observed at time $t = 0$. No indicator variables are introduced but are absorbed in the covariates.

The model is the sum of two Additive Aalen Models running on two different time-scales, see also Scheike (2001). The two time-scales are t and $a = t + a_i \in [a_0, a_{max}]$ where the latter time-scale is specific to each individual and a_0 is some lower-limit that depends on the possible range of the second time-scale. The left summand of (1) captures via α how the effect that X has on the intensity λ varies over t and the right summand captures via β how the effect that Z has on λ varies over a. The case that X and Z have some or all columns equal is generally allowed but comes with identification issues discussed in the next section.

In the illness-death model, say, t might be time since diagnosis (duration) among subjects that have entered the illness stage of the model and a_i could be the age when the transition to the illness stage occurred, such that $t + a_i$ is the age of the subject.

After introducing some notation we present an estimation procedure that leads to explicit estimators of $A(t) = \int_0^t \alpha(s)ds = (\int_0^t \alpha_1(s)ds, \dots, \int_0^t \alpha_p(s)ds)^T$ and $B(a) = \int_{a_0}^a \beta(u)du = (\int_{a_0}^a \beta_1(u)du, \dots, \int_{a_0}^a \beta_q(u)du)^T$. The cumulative functions have the advantage compared to $\alpha(s)$ and $\beta(a)$ that they may be used for inferential purposes since a more satisfactory simultaneous convergence can be established for these processes. We derive the asymptotic distribution for these estimators and a bootstrapping procedure quantifying the estimation uncertainty. Based on the cumulative functions $A(t)$ or $B(a)$ one may estimate the intensity $\alpha(t)$ or $\beta(a)$ by smoothing techniques.

We introduce the following notation. Let $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$ such that $M_i(t) = N_i(t) - \Lambda_i(t)$ are martingales. Let further $N(t) = \{N_1(t), ..., N_n(t)\}^T$ be the n-dimensional counting process, $\Lambda(t) = \{\Lambda_1(t), ..., \Lambda_n(t)\}^T$ is its compensator, such that $M(t) = \{M_1(t), ..., M_n(t)\}^T$ is an n-dimensional martingale, and define matrices $X(t) = (X_1(t), \dots, X_n(t))^T$ and $Z(t) = \{Z_1(t), \dots, Z_n(t)\}^T$, with dimensions $n \times p$ and $n \times q$, respectively. The individual entry times are summarized in one vector $a_\bullet = (a_1, \dots, a_n)$. A superscript $a > 0$ denotes a shift in the argument, i.e, for a generic function f, $f^a(y) = f(y + a)$. For a generic matrix $C(t)$, with n rows $C_i(t)$, and a n-dimensional vector v, $C^v(t)$ is defined through shifting the rows: $C_i^v(t) = C_i(t + v_i)$. For a generic matrix C, a minus superscript, C^-, denotes the Moore-Penrose inverse. An integral, $\int$, with no limits denotes integration over the whole range.

3. Identification of the parameters

One needs to be careful when covariates are part of both the X and the Z design, such that their risk contribution is changing with respect to both time-scales. This is for example the case for the simple additive model where the hazard is given as $\alpha_1(t) + \beta_1(t + a_i)$, or when such models are considered jointly for males and females. For some covariates, we might find that the risk is well described using only one time-scale, for example, in the simple additive model when $\alpha_1(t) \equiv \alpha_1$, and then we can describe the hazard of the event of interest by one function on the age time-scale.

In the general case, the parameters are not uniquely determined without some constraints when they enter both the X and the Z design. Specifically, we can add and subtract a constant to all components that enter both the X and the Z design. If a covariate enters the model only on one time-scale its identification is solely a matter of whether the design matrix is invertible as for the additive hazards model, see (Martinussen & Scheike, 2006, Section 5). To estimate the components that enter both time-scales we therefore need to choose some identifiability constraint. We have chosen to subtract a constant from the time-component such that it integrates to 0 over the considered time-scale. Almost equivalent, we might also require that one of the components integrate to some specific constant, such as the background population mortality on the age time-scale. It is typically easy reparametrising the functions once they are estimated, to go between different parametrizations depending on the constraint that is imposed. In addition to be able to learn about their effects, and how they change over the time-scales we need to estimate them, even if, for example, survival predictions using the model are not influenced by how we specifically identify the parameters of the model.

Without loss of generality we assume that X and Z share the first d $(0 \leq d \leq \min(p, q))$ columns, i.e., for all $i = 1, \ldots, n$,

$$X_{il} = Z_{il}, \quad l = 1, ..., d.$$

In the sequel we resolve the identification issue by adding the constraint that

$$A_l(t_{max}) = \int_0^{t_{max}} \alpha_l(s)\mathrm{d}s = 0, \quad (l = 1, \ldots, d). \tag{2}$$

As noted above it is easy to move between different solutions defined by different constraints, by simply reparametrizing a given solution.

4. Least squares minimisation ignoring the identification of the nonparametric parameters

In this section we show how the standard least squares estimates for the additive hazards models, (Martinussen & Scheike, 2006, Section 5), can be adapted to work for the two time-scale model. This leads to a set of backfitting equations that we then subsequently solve under identifiability constraint. In this section we ignore the identification problem keeping in mind that the solutions below are not unique. We motivate our estimator $(\widehat{A}, \widehat{B})$ as the limit of the following least squares criteria.

$$\arg\min_{\overline{A},\overline{B}} \sum_i \frac{1}{\varepsilon} \int \left\{ \int_t^{t+\varepsilon} \mathrm{d}N_i(s) - \sum_j \int_t^{t+\varepsilon} X_{ij}(s) d\overline{A}_j(s) - \sum_k \int_t^{t+\varepsilon} Z_{ik}(s) d\overline{B}_k^{a_i}(s) \right\}^2 \mathrm{d}t,$$

for ε converging to zero, and where the integrals can be understood as Stieltjes integrals, noting that X_i and Z_i are left continuous. Minimisation runs over all possible integrators. We note that the minimizer, if it exists, will be a right continuous step-function, since $\int_0^t \mathrm{d}N_i(s)$ is a right continuous step function. To simplify notation we will generally work in matrix notation so that above minimisation criteria can also be written as

$$\arg\min_{\overline{A},\overline{B}} \sum_i \frac{1}{\varepsilon} \int \left\{ \int_t^{t+\varepsilon} \mathrm{d}N_i(s) - \int_t^{t+\varepsilon} X_i(s) d\overline{A}(s) - \int_t^{t+\varepsilon} Z_i(s) d\overline{B}^{a_i}(s) \right\}^2 \mathrm{d}t.$$

Computations using calculus of variations lead to $(\widehat{A}, \widehat{B})$ solving the following first order conditions for all $t \in [0, t_{max}]$, $a \in [a_0, a_{max}]$:

$$\sum_i X_i(t)^T \left\{ dN_i(t) - X_i(t) d\widehat{A}(t) - Z_i(t)\mathrm{d}\widehat{B}^{a_i}(t)\mathrm{d}t \right\} = 0,$$
$$\sum_i Z_i^{-a_i}(a)^T \left\{ dN_i^{-a_i}(a) - Z_i^{-a_i}(a) d\widehat{B}(a) - X_i^{-a_i}(a)\mathrm{d}\widehat{A}^{-a_i}(a) \right\} = 0.$$

Rearranging yields

$$\begin{aligned}
&\sum_i X_i(t)^T dN_i(t) - \sum_i X_i(t)^T Z_i(t)\mathrm{d}\widehat{B}^{a_i}(t) = X(t)^T X(t)\mathrm{d}\widehat{A}(t),\\
&\sum_i Z_i^{-a_i}(a)^T dN_i^{-a_i}(a) - \sum_i Z_i^{-a_i}(a)^T X_i^{-a_i}(a)\mathrm{d}\widehat{A}^{-a_i}(a) = Z^{-a_\bullet}(a)^T Z^{-a_\bullet}(a) d\widehat{B}(a).
\end{aligned}$$

The last set of equations can be further rewritten as

$$\widehat{A}(t) = \int_0^t X(s)^- dN(s) - \int E_1(t|u) d\widehat{B}(u) \tag{3}$$

$$\widehat{B}(a) = \int_{a_0}^a Z^{-a_\bullet}(u)^- dN^{-a_\bullet}(u) - \int E_2(a|s) d\widehat{A}(s), \tag{4}$$

where

$$\begin{aligned}
E_1(s,u) &= \sum_i \{X^T(u-a_i)X(u-a_i)\}^{-1} X_i^{-a_i,T}(u) Z_i^{-a_i}(u) I(a_i \le u \le a_i + s),\\
E_2(u,s) &= \sum_i \{Z^{-a_\bullet,T}(s+a_i) Z^{-a_\bullet}(s+a_i)\}^{-1} Z_i^T(s) X_i(s) I(a_0 - a_i \le s \le u - a_i).
\end{aligned}$$

This last set of equations shows how to compute a solution when the component on the other time-scale is known, and we therefore denote these as backfitting equations, Breiman & Friedman (1985); Buja et al. (1989). The functions E_1 and E_2 keep track of how much adjustment is needed from the other time-scale. Specifically, $E_1(t|u)$, shows for time t on the follow-up time-scale, how much adjustment is needed from the age time-scale at age u. The solutions provided by the last set of equations do not guarantee that individual cumulative hazards will be increasing everywhere on the time-scales, but enforcing a positivity constraint will make the estimation of the components much more complicated, and is therefore not considered further in analogy with how the standard additive hazards model is fitted, Aalen (1989); Huffer & McKeague (1991).

Remark 1. In the case with no covariates, i.e., $d = p = q = 1$, with

$$\lambda_i(t) = Y_i(t)\{\alpha(t) + \beta(a_i + t)\},$$

with $X_i(s) = Z_i(s) = Y_i(s) \in \Re$, the adjustment functions are

$$\begin{aligned}
E_1(s,u) &= \sum_i \frac{1}{\sum_{i'} Y_{i'}^{-a_i}(u)} Y_i^{-a_i}(u) I(a_i \le u \le a_i + s),\\
E_2(u,s) &= \sum_i \frac{1}{\sum_{i'} Y_{i'}^{-a_{i'}}(s + a_i)} Y_i(s) I(a_0 - a_i \le s \le u - a_i).
\end{aligned}$$

5. Establishing existence, identification and uniqueness of the estimator

In §3 we outlined the identification problem but ignored it when establishing the estimator in the previous section. In this section we provide a fully identified estimator of our problem. When aiming to solve equations (3) and (4) the identification problem can no longer be ignored. In order to get a better grip of the situation we will now rewrite the backfitting equations as a linear operator equation. We can write the equations (3) and (4) in matrix notation:

$$\begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix} = \begin{pmatrix} \int_0^t X(s)^- dN(s) \\ \int_{a_0}^a Z^{-a_\bullet}(u)^- dN^{-a_\bullet}(u) \end{pmatrix} + \begin{pmatrix} 0 & -E_1 \\ -E_2 & 0 \end{pmatrix} \times \begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix},$$

where with some miss-use of notation $E_l f(\cdot) = \int E_l(\cdot, y) f(y) \mathrm{d}x, (l = 1, 2)$. Or equivalently

$$\widehat{\theta} = \widehat{m} + E\widehat{\theta}, \tag{5}$$

$$\widehat{\theta} = \begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix}, \quad \widehat{m} = \begin{pmatrix} \int_0^t X(s)^- dN(s) \\ \int_{a_0}^a Z^{-a_\bullet}(u)^- dN^{-a_\bullet}(u) \end{pmatrix}, \quad E = \begin{pmatrix} 0 & -E_1 \\ -E_2 & 0 \end{pmatrix}.$$

Note that $\widehat{m}$ is composed of the marginal Aalen estimators of the two time-scales, t and a. Additionally, the linear operator E is compact because it is the composition of an integral operator, which is compact, and a derivative operator, which is bounded. The operator E being compact means that it can be arbitrarily close approximated by a finite dimensional matrix which simplifies both the numerical and theoretical considerations. If the eigenvalues of E are bounded away from one, then, $(I - E)$ is invertible and we have

$$\widehat{\theta} = (I - E)^{-1}\widehat{m}.$$

Hence existence and uniqueness of our proposed estimator can be translated to properties of the eigenvalues of E. One can for instance easily verify that if some covariates are both in the X and the Z design, then E will have an eigenvalue equal to one - as discussed in the following remark.

Remark 2. Consider the most simple case $1 = d = p = q$, i.e., $\lambda_i(t) = Y_i(t)\{\alpha(t) + \beta(a_i + t)\}$. Given a constant $c \in \Re$, consider the pair of linear function $f_1 = (f_{11}, f_{12})^T$ with $f_{11}(s) = cs, \; f_{12}(u) = -c(u - a_0)$. Assuming that $\sum Y_i(s)$ and $\sum Y_i(u - a_i)$ are bounded away from zero on the whole range $s \in [0, t_{max}], u \in [a_0, a_{max}]$, one can easily verify that $E_2 f_{11}(u) = c \int E_2(u|s)\mathrm{d}s = c(u - a_0)$, $E_1 f_{12}(s) = -c \int E_1(s|u)\mathrm{d}u = -cs$. To see this, e.g., for the second equation, note

$$\int E_1(s|u)\mathrm{d}u = \sum_i \int_{a_i}^{a_i+s} \frac{1}{\sum_{i'} Y_{i'}(u - a_i)} Y_i^{-a_i}(u)\mathrm{d}u = \int_0^s \frac{\sum_i Y_i(t)}{\sum_{i'} Y_{i'}(t)} \mathrm{d}t = s.$$

Hence, we have

$$E \begin{pmatrix} f_{11} \\ f_{12} \end{pmatrix} = \begin{pmatrix} -E_1 f_{12} \\ -E_2 f_{12} \end{pmatrix} = \begin{pmatrix} f_{11} \\ f_{12} \end{pmatrix}.$$

Or equivalently, 1 is an eigenvalue of E with corresponding eigenfunction $f_1 = (f_{11}, f_{12})^T$. The identification issue of the model carries over to the estimator. For the general case, fix constants $c_1, \ldots, c_d$ and define f_l as a $\Re^{p+q}$ valued function having all entries but the $l'th$ and the $(d + l)'th$ equal zero:

$$f_l(s, u) = \{0, \cdots, 0, c_l s, 0, \cdots, 0, -c_l(u - a_0), 0, \cdots, 0)\}^T, \; (l = 1, \ldots, d).$$

With analogue arguments one can show that the eigenspace corresponding to eigenvalue equal to one includes the functions in $Lin(f_1, \ldots f_d)$.

We now utilize constraint (2) and incorporate it into new backfitting equations:

$$\widehat{A}(t) = \int_0^t X(s)^- dN(s) - \int E_1(t|u) d\widehat{B}(u), \tag{6}$$

$$\widehat{B}(a) = \int_{a_0}^a Z^{-a_\bullet}(u)^- dN^{-a_\bullet}(u) - \int E_2(a|s) d\widehat{A}(s) + \frac{\widehat{A}^{d_q}(t_{max})}{t_{max}}(a - a_0), \tag{7}$$

where $\widehat{A}^{d_q}$ is the q-dimensional vector $\widehat{A}^{d_q} = (A_1, \ldots, A_d, 0, \ldots, 0)^T$. This translates to the new operator equation

$$\widehat{\theta} = \widehat{m} + \overline{E}\widehat{\theta}, \quad \overline{E} = \begin{pmatrix} 0 & -E_1 \\ -\overline{E}_2 & 0 \end{pmatrix}, \tag{8}$$

where $\overline{E}_2 h(a) = \int E_2(a|s)dh(s) - (a - a_0)h^{d_q}(t_{max})t_{max}^{-1}$. The next proposition states that the solutions of (8) include all relevant solutions of (5) and that every solution of (8) is a solution of (5).

PROPOSITION 1. *(Reparameterisation of a given solution) For every solution $\widehat{\theta}$ of* (5)*, define*

$$\widehat{\theta}_0 = (I - \Pi)\widehat{\theta},$$

where

$$\Pi \begin{pmatrix} h_1(t) \\ h_2(a) \end{pmatrix} = \begin{pmatrix} \Pi_1 h_1(t) \\ \Pi_2 h_2(a) \end{pmatrix} = \begin{pmatrix} t h_1^{d_p}(t_{max})t_{max}^{-1} \\ -(a - a_0)h_1^{d_q}(t_{max})t_{max}^{-1} \end{pmatrix}.$$

The function $(I - \Pi)$ is a projection onto the space of functions $\{(A, B)^T : \Re^2 \mapsto \Re^{p+q} |\ A_l(t_{max}) = 0,\ l = 1 \ldots, d\}$, i.e, fulfilling the identification constraint (2)*. Hence, $\widehat{\theta}_0$ is a solution of* (8) *and*

$$\widehat{\theta}_0 + Lin(f_1, \ldots f_d), \tag{9}$$

are further solutions of (5)*. Conversely, for every solution $\widehat{\theta}_0$ of* (8)*, all functions of the form* (9) *are solutions of* (5)*.*

The proof can be found in the Supplementary Material.

With Proposition 1 at hand it is justified to define our estimator as the solution of (8). We will now discuss existence and uniqueness of the solution of (8).

Note that E is known and hence one can calculate a numerical approximation of its eigenvalues by working on a grid. Consider the sub-space

$$K = \{h = (h_1, \ldots, h_d, 0, \ldots, 0)|\ h_l : \Re \to \Re,\ x \mapsto c_l x,\ c_l \in \Re,\ l = 1, \ldots, d\}.$$

It holds that $\overline{E}_2 = E_2(I - \Pi_2)$, where Π_2, as defined in Proposition 1 is a projection into K. We have $K \subseteq ker(I - E_2)$. We can check whether K equals $ker(I - E_2)$. This can be done by calculating the dimension of the eigenspace of E_2 corresponding to an eigenvalue equal to one. The dimension will be at least d. If it is exactly d, then $K = ker(I - E_2)$.

The next proposition states that if $ker(I - E_2) = K$, and $ker(I - E) = Lin(f_1, \ldots, f_d)$, then both $I - \overline{E}_2$ and $I - \overline{E}$ are bijective.

PROPOSITION 2. *Assume that E_2 has eigenvalue 1 with multiplicity d. Then, $(I - \overline{E}_2)$ will be bijective. If furthermore E has Eigenvalue 1 with multiplicity d, then $(I - \overline{E})$ is bijective and hence invertible. In particular a solution of equations* (8) *exists and it is unique.*

The proof can be found in the Supplementary Material.

We thus have demonstrated that a way to find all solutions of the backfittingequation (5) is to solve equation (8) first. Under the assumptions of Proposition 2, that solution exists and is unique, hence allowing for straightforward numerical calculations as outlined in the Supplementary Material. Other specific solutions of (5) under specific constraints can be computed following Proposition 1.

6. Asymptotics

We have

$$\theta = m + \overline{E}\theta, \tag{10}$$

where m arises from $\widehat{m}$ by replacing N by Λ. This is seen by direct calculations when dN is replaced by $dM + d\Lambda$ with its intensity $d\Lambda_i = X_i(t)\alpha(t)dt + Z_i(t+a_i)\beta(a_i+t)da$. This equation holds when all inverses exist everywhere, which is required in the asymptotic conditions. Importantly, $\overline{E}$ is the observable operator from the previous sections and not some asymptotic limit.

We conclude that the least square solution (6) and (7) is a plug-in estimator of (10). The estimation error is given as

$$\widehat{\theta} - \theta = \widehat{m} - m + \overline{E}(\widehat{\theta} - \theta). \tag{11}$$

As in the last section, If $\overline{E}$ has eigenvalues all bounded away from one, then

$$\widehat{\theta} - \theta = (I - \overline{E})^{-1}(\widehat{m} - m).$$

So the asymptotic behaviour of $\widehat{\theta} - \theta$ can be deduced from the asymptotic behaviour of $(I - \overline{E})^{-1}$ and $\widehat{m} - m$, with the former being observed and the latter being the compound estimation error of two additive Aalen models on different time-scales.

Theorem 1. *Under assumptions (A)–(E), given in the Supplementary Material, the estimator $\widehat{\theta}$ exists. Furthermore the estimator $\widehat{\theta}$ is $n^{1/2}$ consistent:*

$$n^{-1/2}(\widehat{\theta} - \theta) \to (I - \widetilde{E})^{-1}U,$$

in Skorohod space $D^{p+q}[0, a_{max}]$. Here, $(\widehat{\theta} - \theta)$ is treated as one stochastic process defined on $[0, a_{max}]$ by setting for $j = 1, \ldots, p$ and $\nu \in [t_{max}, a_{max}]$, $(\widehat{\theta} - \theta)_j(\nu) = (\widehat{\theta} - \theta)_j(t_{max})$. And similarly, for $j = p+1, \ldots, p+q$ and $\nu \in [0, a_0]$, $(\widehat{\theta} - \theta)_j(\nu) = 0$. The process U is a $p+q$ dimensional mean-zero Gaussian process with covariation matrix $\Sigma(\nu_1, \nu_2)$ described in the Supplementary Material, and $\widetilde{E}$ is the limit of $\overline{E}$.

The proof can be found in the Supplementary Material.

7. Confidence Bands

While we could use the central limit theorem of the previous section to construct confidence bands, it is often easier computationally and often also leads to better small sample performance to use some bootstrapping procedure (Lin et al., 1994; Bluhmki et al., 2018; Beyersmann et al., 2013). Following these authors, we propose a wild bootstrap approach based on the relationship

$$\widehat{\theta} - \theta = (I - \overline{E})^{-1}(\widehat{m} - m) = (I - \overline{E})^{-1}\begin{pmatrix} \int_0^t X(s)^- dM(s) \\ \int_{a_0}^a Z^{-a_\bullet}(u)^- dM^{-a_\bullet}(u) \end{pmatrix} = (I - \overline{E})^{-1}\begin{pmatrix} \mathcal{M}_1 \\ \mathcal{M}_2 \end{pmatrix}.$$

Since $(I - \overline{E})^{-1}$ is known, it is enough to only approximate $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2)^T$. We propose two possible wild bootstrap approximations:

$$\widehat{\mathcal{M}}^{(1)} = \begin{pmatrix} \int_0^t X(s)^- d\widetilde{N}(s) \\ \int_{a_0}^a Z^{-a_\bullet}(u)^- d\widetilde{N}^{-a_\bullet}(u) \end{pmatrix}, \quad \widetilde{N}_i(s) = G_i N_i(s),$$

or alternatively

$$\widehat{\mathcal{M}}^{(2)} = \begin{pmatrix} \int_0^t X(s)^{-} d\widetilde{M}(s) \\ \int_{a_0}^a Z^{-a_\bullet}(u)^{-} d\widetilde{M}^{-a_\bullet}(u) \end{pmatrix},$$

$$\int_0^t \widetilde{M}_i(s)\mathrm{d}s = G_i \left[\int_0^t N_i(s)\mathrm{d}s - \left\{ \int_0^t (X_i(s)\mathrm{d}\widehat{A}(s) + \int_0^t Z_i(s)\mathrm{d}\widehat{B}(s+a_i) \right\} \right],$$

where G_i is a mean zero random variable with unit variance. The random variable G_i is generated such that for fixed i, it is independent of all other variables. It is straight forward to confirm that for any of the two choices, $\widehat{\mathcal{M}}^{(r)}$ ($r = 1, 2$), is a mean zero process that has the same covariance as $\mathcal{M}$; the covariance of $\mathcal{M}$ is given in the Supplementary Material. Hence, we directly derive the following proposition.

PROPOSITION 3. *Under assumptions (A)–(E), given in the Supplementary Material, the bootstrapped estimation error is uniformly consistent, i.e., for* $r = 1, 2$

$$n^{-1/2}\left\{(I-\overline{E})^{-1}\widehat{\mathcal{M}}^{(r)}\right\} \to (I-\widetilde{E})^{-1}U,$$

in Skorohod space $D^{p+q}[0,T]$*, where* U *is is described in Theorem 1.*

The proof can be found in the Supplementary Material.

One useful consequence of this is that we can estimate standard errors of our estimator $\hat{\theta}$ based on the approximation from the bootstrap. We denote these estimators as $\hat{\sigma}_r(t)$ for the two possibilities $r = 1, 2$.

COROLLARY 1. *Under assumptions (A)–(E), the bootstrapped errors lead to confidence bands* $CB^{(r)}$ *for* $\theta(\nu)$ *over* $\nu \in [\nu_1, \nu_2]$ *providing an asymptotic coverage probability of* $1-\alpha$*, where* $CB^{(r)}(\nu) = \theta(\nu) \pm c_{1-\alpha}\hat{\sigma}_r(\nu)$*, and*

$$c_{1-\alpha} = (1-\alpha) \quad \textit{quantile of} \quad \mathcal{L}\left\{ \sup_{[\nu_1,\nu_2]} n^{-1/2} \frac{\left|(I-\overline{E})^{-1}\widehat{\mathcal{M}}^{(r)}\right|}{\hat{\sigma}_r} \mid X, Z, N \right\}$$

We explore the performance of the estimator of the standard error and the uniform bands in the next section.

8. SIMULATIONS

We generated data from the simple two-time-scale model with age and duration that resemble the data we consider in a worked example in the next section. We assume that the hazard for those under risk is given as $\beta(t+a_i) + \alpha(t)$, where $\beta(a) \equiv 0.067$ and the entry ages were drawn uniformly from $[0, 25]$ but with a point-mass at 10 % in 0 (to avoid difficulties with left truncation in the estimation). The $\alpha(t)$ component was piecewise constant with rate 0.32 in the time-interval $[0, 0.25]$, then 0.48 in $(0.25, 0.5]$ and then finally to satisfy our constraint -0.044 in $(0.5, 5]$, so that $\int_0^5 \alpha(s)ds = 0$. All subjects were censored after 5 years of follow up. In all simulations we used a discrete approximation based on a time-grid of either 100 points in both the age direction $[0, 30]$ and on the duration time-scale $[0, 5]$. We considered sample sizes 100, 200 and 400 to estimate the time component $A(t)$ and age component $B(a)$. In Table 1, based on 1000 realizations, we show a) the bias, b) the point-wise mean standard error. Additionally based on 100 bootstrap samples in each run, using $G_i dN_i$, we report c) the bootstrap estimate of mean standard error and d) coverage of the bootstrapped confidence interval.

n	age	bias	mean se	sd	cov	time	bias	mean se	sd	cov
100	6.7	-0.001	0.224	0.231	0.912	1.0	0.018	0.044	0.045	0.954
100	13.8	0.009	0.297	0.298	0.935	2.0	0.015	0.039	0.04	0.946
100	20.9	0.018	0.351	0.357	0.943	3.0	0.009	0.032	0.034	0.951
100	27.9	0.027	0.391	0.402	0.938	4.0	0.005	0.024	0.024	0.966
100	35.0	0.078	0.460	0.464	0.932	5.0	0.000	0.016	0.017	0.874
200	6.7	0.006	0.158	0.155	0.94	1.0	0.009	0.031	0.031	0.951
200	13.8	0.003	0.207	0.206	0.942	2.0	0.007	0.027	0.027	0.960
200	20.9	0.001	0.243	0.237	0.948	3.0	0.005	0.022	0.022	0.966
200	27.9	0.004	0.271	0.262	0.945	4.0	0.002	0.017	0.017	0.972
200	35.0	0.006	0.328	0.329	0.933	5.0	0.000	0.011	0.012	0.933
400	6.7	−0.004	0.114	0.118	0.948	1.0	0.006	0.022	0.022	0.951
400	13.8	−0.006	0.148	0.153	0.946	2.0	0.005	0.019	0.019	0.957
400	20.9	0.002	0.173	0.18	0.937	3.0	0.003	0.015	0.015	0.960
400	27.9	0.010	0.192	0.196	0.943	4.0	0.002	0.012	0.012	0.970
400	35.0	0.013	0.235	0.245	0.934	5.0	0.000	0.008	0.008	0.950

Table 1. *Estimation performance and uncertainty estimated from bootstrap for sample sizes* $n = 100, 200, 400$ *for the age component* $B(a)$ *and time component* $A(t)$ *for selected ages and time points. Results are based on 1000 realisations and a bootstrap with 100 repetitions. We report bias of the estimates (bias), mean of estimated standard errors (mean se), standard deviation of bootstrap estimates (sd) and 95 % pointwise coverage (cov).*

We note that the backfitting algorithm is almost unbiased across all sample size and improves as the sample size increases. This is despite the fact that the component in the time-direction is increasing very rapidly, and thus would be hard to estimate based on smoothing based techniques.

We note that the standard error is well estimated by the bootstrapped standard deviation across all sample sizes and for both components. In addition the pointwise coverage is close to the nominal 95 % level for the larger sample sizes. But even for $n = 100$ the coverage is reasonable for most time-points for the two components.

Finally, we also considered the performance of the confidence bands based on our bootstrap approach.

n	coverage (age)	coverage (time)
100	0.797	0.792
200	0.912	0.915
400	0.952	0.939

Table 2. *Coverage of confidence bands estimated from bootstrap for sample sizes* $n = 100, 200, 400$ *for the age and time component. Based on 1000 realisations and a boostrap with 100 repetitions.*

When n gets larger these bands are quite close to the nominal 95 % level, but for $n = 100$ the asymptotics have not quite set in to make the entire band work well.

9. Applications

9.1. *The TRACE study*

The TRACE study group (see e.g. Jensen et al. (1997)) has collected information on more than 4000 consecutive patients with acute myocardial infarction (AMI) with the aim of studying the prognostic importance of various risk factors on mortality. We here consider a subset of 1878 of these patients that are available in the timereg R package. At the age at entry, i.e., age of diagnosis, the patients had various risk factors recorded. We will first consider the simple model with the only effects of the two-time-scales age and duration. Afterwards we will incorporate additional risk factors. It is expected that the duration time-scale has a strong initial effect on dying that then disappears when patients survive the first period right after their AMI. We will see that the duration effect is primarily due to two risk factors, and we will quantify their effects on the duration time-scale.

9.2. *The simple model without additional covariates*

We estimate the two-time-scale model $\lambda(t) = Y_i(t)\{\alpha(t) + \beta(t + a_i)\}$ under the identifiability condition $\int_0^5 \alpha(s)ds = 0$. We use only the subset of patients that were more than 40 years of age, and only consider the first five years of follow-up time after the diagnosis.

For comparison, we also estimate the mortality on the two time-scales separately, i.e., we also consider the models $\lambda(t) = Y_i(t)\alpha(t)$ and $\lambda(t) = Y_i(t)\beta(t + a_i)$. In the sequel we call these two models marginals, and the estimates are given in Figure 1. Panel (a) shows the cumulative function on the age time-scale, i.e, $B(a)$, as the marginal estimate (full line) and as component in the two-time-scale model, i.e., adjusted for duration effects (broken line). Panel (b) shows the mortality on the duration time-scale, i.e, A(t), as the marginal estimate (full line) and as component in the two-time-scale model, i.e., adjusted for age effects (broken line). In addition we show 95% confidence bands based on our bootstrap (regions), and the pointwise confidence intervals (dotted line).

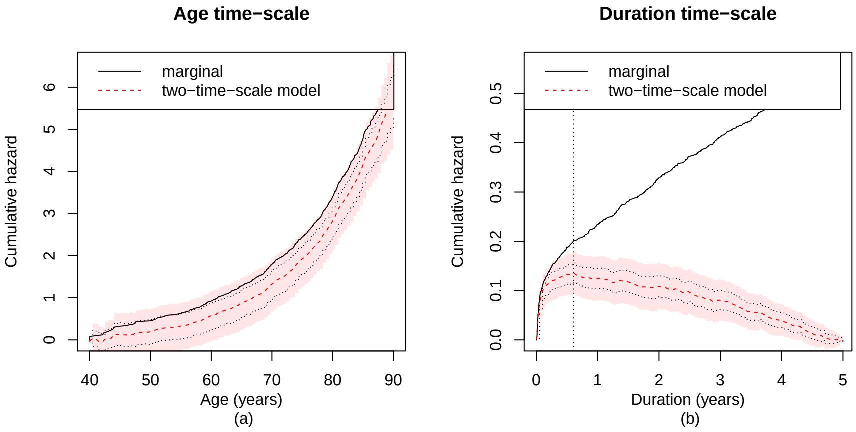

Fig. 1. Cumulative baseline on the two time-scales estimated marginally (full line) and in the two-time-scale model (broken line). Confidence bands (regions) and pointwise confidence intervals (dotted lines).

Taking out the duration effect slightly alters the estimates on the age time-scale. In contrast, on the duration time-scale, the time effect is strongly affected by adjustment of the age effect estimates, and here the two-time-scale model more clearly demonstrates what is going on on the duration time-scale. Before we come to this, we stress that the interpretation of the cumulative functions on the two-time-scales is not straight forward, due to the constraint that needs to be imposed to identify a specific solution. While straight lines can be added and subtracted without altering the fit, differences in the slope, i.e., second derivatives, do not depend on a specific solution. Back to our results, the duration effect has a very steep slope initially and then after surviving the first 220 days we see a protective effect (dotted vertical line). After those 220 days there is a constant, i.e., not much changing, negative slope, hinting that the duration effect vanishes after 220 days.

The two time-scale components that jointly make up the hazard for an individual, can also be used for prediction purpose. Within the additive structure, the duration effect can be interpreted as giving relative survival due to the duration time-scale. In addition we conclude that there is important variation along the duration time-scale. This is formally tested by using the confidence bands.

9.3. *Regression modelling of effects*

We now illustrate a more detailed regression analysis where we study the importance of the important risk-predictors, VF (ventricular fibrilation, yes/no), CHF (clinical heart pump failure, yes/no) and diabetes (yes/no).

We start by considering the following intensity model

$$\begin{aligned}\lambda_i(t) = {} & \alpha_1(t) + \alpha_2(t)\text{VF}_i + \alpha_3(t)\text{CHF}_i + \alpha_4(t)\text{diabetes}_i \\ & +\beta_1(t+a_i) + \beta_3(t+a_i)\text{VF}_i + \beta_4(t+a_i)\text{CHF}_i + \beta_4(t+a_i)\text{diabetes}_i\end{aligned}$$

where a_i is the age of the ith subject at the time of entry. To identify the model we assume that $\int_0^5 \alpha_j(s)ds = 0$ for all j.

We estimated the model using our backfitting equations and estimated the uncertainty using our bootstrap approach.

First considering the duration effects that are identified up to a slope within the first 5 years after the AMI, we note that VF and CHF are important risk predictors on this time-scale. The presence of CHF or VF will increase the mortality quite notably right after the AMI. Taking out effects of VF and CHF there is only a rather small effect left for the intercept (subjects without diabetes, CHF and VF), with the cumulative being at most 0.04 within the first time-period. In addition we note that diabetes does not seem to interact with the duration time-scale.

On the age time-scale VF, CHF and diabetes are not significantly different from a straight line. Hence, in terms of describing the mortality all effects of VF, CHF and diabetes seems rather consistent with constant additive effects, and the mortality might therefore also be described with intercept and diabetes to represent the age time-scale and effects of CHF and VF together with intercept on the duration time-scale, that is

$$\lambda_i(t) = \beta_1(t+a_i) + \beta_2(t+a_i)\text{diabetes}_i + \alpha_1(t) + \alpha_2(t)\text{VF}_i + \alpha_3(t)\text{CHF}_i.$$

In this model, β_2, α_2 and α_3 can be fitted without identifiability constraints which makes the interpretation of the effects simpler.

Indeed the structured model, shows the effects of VF and CHF are strongly time-varying on duration time-scale, and increases the mortality significantly right after the AMI. There is only a rather small duration effect left for the intercept when the VF and CHF duration effects are taken out. It is also interesting that VF leads to a highly increased hazard only in the time right after

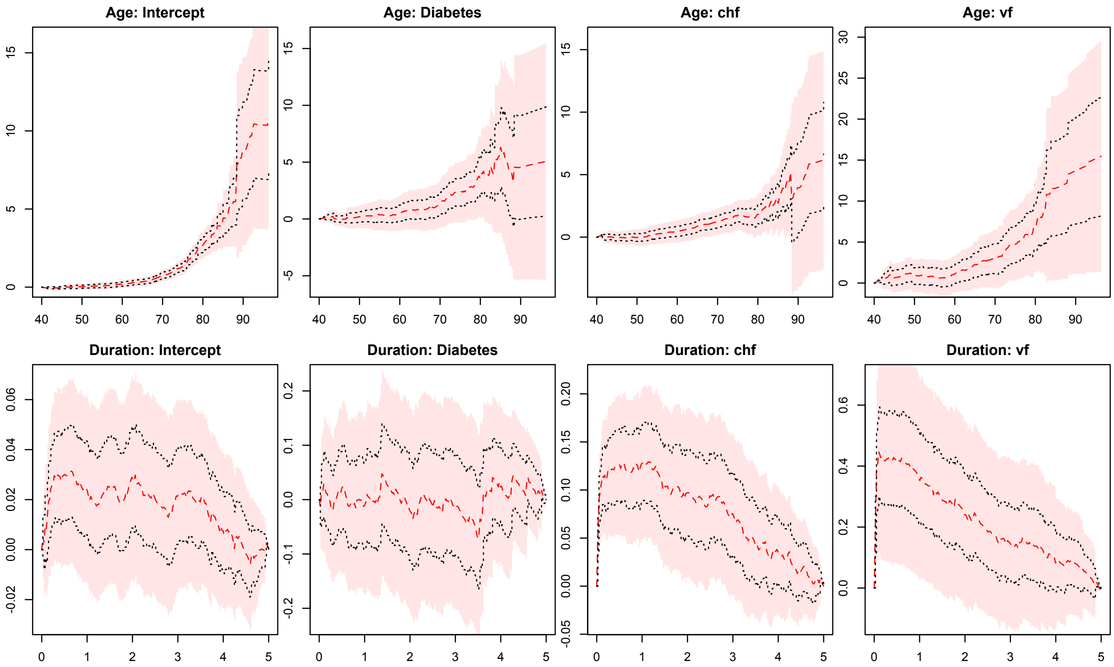

Fig. 2. Estimates (broken line). Confidence bands (regions) and pointwise confidence intervals (dotted lines).

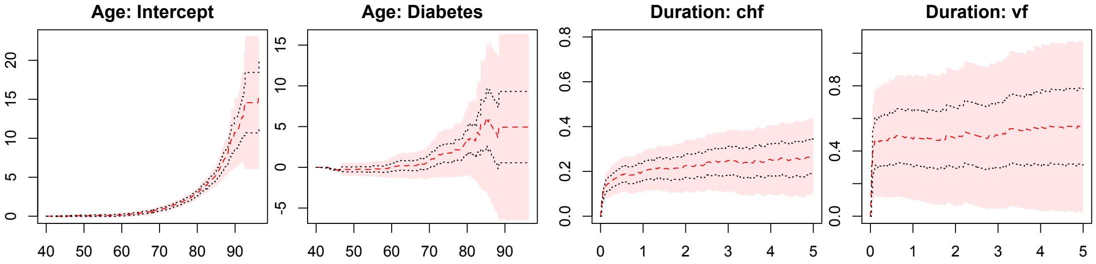

Fig. 3. Estimates (broken line). Confidence bands (regions) and pointwise confidence intervals (dotted lines).

AMI, note that this is in contrast to the overall effect on the duration time-scale considered in the previous section. Further we note that CHF has an effect on the duration time-scale that is less local than VF, but still with a strongly increased risk right the AMI.

10. Discussion

By using the additive structure we have demonstrated that one can estimate and separate the effects of two time-scales directly via a backfitting algorithm that does not involve smoothing. Our model includes a regression setting with multiple covariates using both time-scales simultaneously. In the example of the TRACE data-set, we find that VF and CHF are primarily acting through the duration time-scale, while diabetes acts via the age-time scale.

An advantage of working only on the cumulative scale is that we achieve a uniform asymptotic description leading to a simple bootstrap procedure for getting estimates of the uncertainty and for constructing for example confidence intervals. These cumulative hazards may form the basis for smoothing based estimates when the hazard function itself is of interest, but often the cumulative hazards are the quantities of key interest for example when interest is on survival predictions.

Our model has two key assumptions. (1) Covariates have a linear effect on survival and (2) additivity between the two time-scales and between the different covariate effects. With regard to the first point, the assumption on linearity did not have an impact in our particular application because we only studied the effect of binary variables on survival. If continuous variables are to be included that do not act linearly, then one can extend our model easily via splines, e.g. p-splines (Eilers & Marx, 1996), while obeying the framework set in this paper. Alternatively, a more complicated nonparametric approach could be employed by replacing $\alpha_j(t)x_j(t)$ by $\alpha_j(t, x_j(t))$. This latter approach changes the problem from a one-dimensional to a two-dimensional problem and would make smoothing necessary. We now discuss the second point, i.e., the additivity assumption. Assuming some structure is necessary so that the estimation performance does not deteriorate exponentially with the number of covariates. Additionally, if not prediction but interpretation is the goal, i.e., understanding how different time-scales and covariates running on these different time-scales affect survival, then a model that separates the different effects must be employed. An alternative to the additive structure considered in this paper is the multiplicative, i.e., proportional hazard, structure, employed by Iacobelli & Carstensen (2013). A useful extension in both the additive and the multiplicative structure is the manual inclusion of interaction effects between the time-scales. To allow for interactions without changing the framework of our paper one can, for example, consider to divide the age-interval in two disjoint subsets: I_1 and I_2:

$$\beta(t + a_i) + I(a_i \in I_1)\alpha_1(t) + I(a_i \in I_2)\alpha_2(t).$$

Another variation could bring our new methodology closer to the type of approach of Iacobelli & Carstensen (2013), for example,

$$\beta(t + a_i) + \alpha_1(t)\exp\{\beta(a_i + t)\},$$

providing a new multiplicative time-age effect. This latter model has been considered in Lee et al. (2017) and is beyond the implemented framework of this paper. It would be interesting to see how an extension of our model would compare to the solution provided by Lee et al. (2017). It would also be interesting for the future to extend our new model to incorporate both our new additive modelling and the multiplicative hazard approaches of Iacobelli & Carstensen (2013) and Lee et al. (2017) in one single model.

R code for fitting the simple model is available at
`https://github.com/MHiabu/Two_Timescale_Aalen`.

References

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statist. Med.* **8**, 907–925.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

Beyersmann, J., Termini, S. D. & Pauly, M. (2013). Weak convergence of the wild bootstrap for the aalen–johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics* **40**, 387–402.

Bluhmki, T., Schmoor, C., Dobler, D., Pauly, M., Finke, J., Schumacher, M. & Beyersmann, J. (2018). A wild bootstrap approach for the aalen–johansen estimator. *Biometrics* **74**, 977–985.

BREIMAN, L. & FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* **80**, 580–598.

BUJA, A., HASTIE, T., TIBSHIRANI, R. et al. (1989). Linear smoothers and additive models. *The Annals of Statistics* **17**, 453–510.

CARSTENSEN, B. (2007). Age–period–cohort models for the lexis diagram. *Statist. Med.* **26**, 3018–3045.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187–220.

DUCHESNE, T. & LAWLESS, J. (2000). Alternative time scales and failure time models. *Lifetime Data Anal.* **6**, 157–179.

EILERS, P. H. & MARX, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science* **11**, 89–102.

GRAMBSCH, P. M., THERNEAU, T. M. & FLEMING, T. R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **51**, 1469–1482.

GUILLEN, M., NIELSEN, J. P. & PEREZ-MARIN, A. M. (2007). Improving the efficiency of the nelson–aalen estimator: The naive local constant estimator. *Scand. J. Statist.* **34**, 419–431.

HASTIE, T. & TIBSHIRANI, R. (1986). Generalized additive models. *J. Amer. Statist. Assoc.* **81**, 297–318.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

HUANG, J. (1999a). Efficient estimation of the partly linear additive cox model. *The Annals of Statistics* **27**, 1536–1563.

HUANG, J. (1999b). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27**, 1536–1563.

HUFFER, F. W. & MCKEAGUE, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Assoc.* **86**, 114–129.

IACOBELLI, S. & CARSTENSEN, B. (2013). Multiple time scales in multi-state models. *Statist. Med.* **32**, 5315–5327.

JENSEN, G. V., TORP-PEDERSEN, C., HILDEBRANDT, P., KOBER, L., NIELSEN, F. E., MELCHIOR, T., JOEN, T. & ANDERSEN, P. K. (1997). Does in-hospital ventricular fibrillation affect prognosis after myocardial infarction? *European Heart Journal* **18**, 919–924.

KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

KAUERMANN, G. & KHOMSKI, P. (2006). Additive two-way hazards model with varying coefficients. *Comput. Statist. Data Anal.* **51**, 1944–1956.

KEIDING, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *J. Roy. Statist. Soc. Ser. A* , 371–412.

KUANG, D., NIELSEN, B. & NIELSEN, J. P. (2008a). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95**, 979–986.

KUANG, D., NIELSEN, B. & NIELSEN, J. P. (2008b). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95**, 979–986.

LEE, M., GOUSKOVA, N. A., FEUER, E. J. & FINE, J. P. (2017). On the choice of time scales in competing risks predictions. *Biostatistics* **18**, 15–31.

LIN, D. Y., FLEMING, T. & WEI, L. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.

LIN, H., HE, Y. & HUANG, J. (2016). A global partial likelihood estimation in the additive cox proportional hazards model. *J. Statist. Plann. Inference* **169**, 71–87.

LINTON, O. B., NIELSEN, J. P. & VAN DE GEER, S. (2003). Estimating multiplicative and additive hazard functions by kernel methods. *Ann. Statist.* **31**, 464–492.

MAMMEN, E., LINTON, O. B. & NIELSEN, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Stat* **27**, 1443–1490.

MARTINUSSEN, T. & SCHEIKE, T. (2006). *Dynamic Regression Models for Survival Data*. Springer-Verlag New York.

NELSON, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* **14**, 945–966.

OAKES, D. (1995). Multiple time scales in survival analysis. *Lifetime Data Anal.* **1**, 7–18.

O'SULLIVAN, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* **9**, 531–542.

O'ULLIVAN, F. (1993). Nonparametric estimation in the cox model. *The Annals of Statistics* **21**, 124–145.

SCHEIKE, T. (2001). A generalized additive regression model for survival times. *Ann. Statist.* , 1344–1360.

SLEEPER, L. A. & HARRINGTON, D. P. (1990). Regression splines in the cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* **85**, 941–949.

THERNEAU, T. M., GRAMBSCH, P. M. & FLEMING, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.