Structured Fusion Lasso Penalised Multi-state Models

Holger Reulen* and Thomas Kneib*

*Chair of Statistics, University of Göttingen

April 13, 2015

Abstract Multi-state models generalize survival or duration time analysis to the estimation of transition-type specific hazard rate functions for multiple transition-types. When each of the transition-type specific risk functions is parametrized with several distinct covariate effect coefficients, this leads to a high dimensional model which is hard to interpret. Aiming for the estimation of covariate effect coefficients equal to zero, or an equal value for two coefficients that belong to the same covariate but to two different transition-types, are two options to decrease the parameter space dimensionality and to work out a clearer image of the underlying multi-state model structure. The first issue can be approached by the penalisation of the absolute values of the covariate coefficients during the estimation. The penalisation of absolute differences between coefficients of effects of the same covariate on different transition-types leads to sparse competing risk relations within a multi-state model, and concerns the second issue of equality of covariate effect coefficients. Using piece-wise exponential models, this concept is expandable to baseline hazard rate functions. Throughout this article, a new estimation approach providing sparse multi-state modelling by the above principles is established, based on the estimation of multi-state models and simultaneously penalising the L_1-norm of covariate coefficients and their differences in a structured way. The new multi-state modelling approach is illustrated on peritoneal dialysis study data and the implemented R package `penMSM` is briefly described.

Keywords *Multi-state models, Regularisation, Structured fusion Lasso penalty, Cross-transition effects*

1 Introduction

Multi-state models are a general model class for the timing of events and have a wide range of applications in medicine, for example in epidemiology or clinical trials, where individuals progress through the different states of a disease. In a general definition, a multi-state model is characterized as a system of multivariate survival data where the individuals under study may experience a sequence of transitions across time. Each transition is characterized by an entry and an exit state-type, the time when the entry state-type is reached and the duration of the sojourn time until the transition is either observed or censored. We will address this in more detail in Section 2. The durations of the sojourn times are influenced by transition-type specific covariate effects, where in the most prominent model class for survival and multi-state models, the Cox proportional hazards models (Cox, 1972), the transition-type specific covariate effects are linked to the transition-type specific baseline hazard rate functions by the exponential function:

$$\lambda_{q,i}\left(t\right) = \lambda_{q,0}\left(t\right) \cdot \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}_q\right),$$

with transition-type set $q \in \{1, \ldots, Q\}$, individuals $i = 1, \ldots, N$, time t, transition-type specific baseline hazard rate function $\lambda_{q,0}\left(t\right)$, individual specific covariate vectors $\mathbf{x}_i$ as a collection of covariate observations $x_{p,i}$, $p = 1, \ldots, P$, and corresponding transition-type specific covariate coefficient vectors $\boldsymbol{\beta}_q$. The magnitudes of the transition-type specific covariate coefficients result in a reduction of the expected duration of the sojourn times for negative coefficients and a prolongation for positive coefficients. The product $\mathbf{x}_i^\top \boldsymbol{\beta}_q$ results in the individual and transition-type specific linear predictor $\eta_{q,i}$.

Each transition sequence is characterized by a series of distinct entry and exit state-types following certain paths of possible transition-types. This system of paths is illustrated by a multi-state model state-chart, where distinct state-types are treated as nodes and possible transition-types are considered using directed arrows. Here, the transitions between two state-types may be reversible or irreversible: in the first case one arrow exists between two state-types, in the second case two arrows exist between two state-types. The state-types themselves can be either absorbing or transient: in the first case there are only arrows that are directed to the respective node, in the second case there is at least one arrow leaving the respective node. Figure 1 shows the state-chart for an illness-death model with recovery. This is a three state-type model with the transition-types between the state-types healthy (H) and illness (I) being reversible transition-types, while the transition-types to the state-type death (D) are considered as being irreversible transition-types. Consequently, H and I are transient state-types, while D is an absorbing state-type.

In general, the transition-types of a multi-state model are all representing distinct transition-types, but some of them have a closer relation with respect to their practical

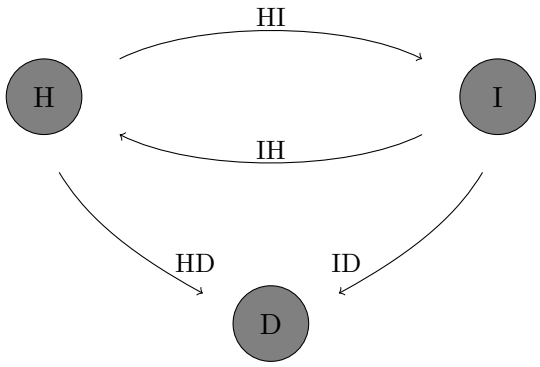

Figure 1: Illness-death model with recovery: State-chart illustrating the underlying process that leads to the sequences of events. State-type H denotes healthy, state-type I denotes illness, and state-type D denotes death. Transition-type IH is the representative for a recovery.

interpretation. In the illness-death model, the transition-types that point from H to I (HI) or from I to D (ID) have a major common attribute: they both lead to the aggravation of a patients health situation. While some of the risk factors that are associated with the sojourn times in the entry state-types of these transition-types, may have different strengths of effects, there may be others that can be described with the same effect magnitude on both transition-types, i.e. $\beta_{\text{HI}} = \beta_{\text{ID}} \neq \beta_{\text{IH}}$ for covariate x_p and transition-type I to H (IH). In the underlying data generating process of this model, this latter class of equal effects is mainly controlled by the aggravation component, while the signal of the single transition-type specific components is within the range of the noise. The covariate effects are equal across two, or in other situations even more transition-types and will be referred to as *cross-transition-type effects*. The value of cross-transition-type effects with respect to the interpretation of results gets clearest when we look on the relative transition-type hazard rate function for two transition-types q, q':

$$\frac{\lambda_{q,i}\left(t\right)}{\lambda_{q',i}\left(t\right)} = \frac{\lambda_{q,0}\left(t\right) \cdot \exp\left(x_{1,i}\beta_{1.q}\right) \cdot \ldots \cdot \exp\left(x_{P,i}\beta_{P.q}\right)}{\lambda_{q',0}\left(t\right) \cdot \exp\left(x_{1,i}\beta_{1.q'}\right) \cdot \ldots \cdot \exp\left(x_{P,i}\beta_{P.q'}\right)}.$$

If $\beta_{p.q} = \beta_{p.q'}$, the respective term will cancel out for any value of $x_{p,i}$. If $\lambda_{q,0}\left(t\right) = \lambda_{q',0}\left(t\right)$, i.e. the transition-type specific baseline hazard rate functions are equal across time, the relative transition-type baseline hazard rate function can be expressed by a single proportionality factor:

$$\frac{\lambda_{q,0}\left(t\right)}{\lambda_{q',0}\left(t\right)} = \exp\left(\gamma_{q.q'}\right).$$

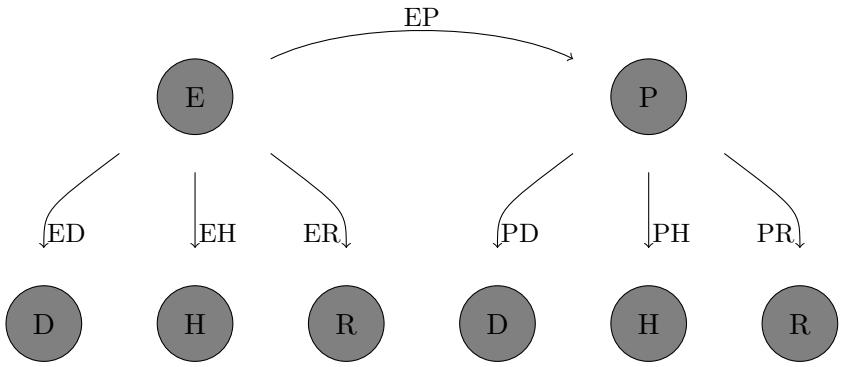

Figure 2: State-chart for the multi-state model on the peritoneal dialysis program data. State-type E represents the entrance to the peritoneal dialysis program, state-type P represents an affection with peritonitis, state-type D represents the death of a patient, state-type H represents a transfer to haemodialysis, and R represents a renal transplantation.

Analyses of this type are realizable using the piece-wise exponential model approach. For partial likelihood analyses, the latter investigation of proportionality of baseline hazard rates is not possible since the baseline hazard rate functions stay unspecified in their functional forms.

Another type of grouping structure for transition-types is present in the five state-type model for peritoneal dialysis program data which will be analysed during this article to illustrate the established multi-state modelling approach. In the peritoneal dialysis program data, patients with a chronic kidney disease participated in a peritoneal dialysis program at the Peritoneal Dialysis Unit, Nephrology Department, Hospital Geral de Santo António, Porto, Portugal, between 1980 and 2011. All of the 424 patients under study begin the dialysis program with the entrance (E) to the peritoneal dialysis program and are at-risk to transition into one of three absorbing state-types: death (D), transfer to haemodialysis (H), and renal transplantation (R), or into the transient state-type peritonitis (P). If a patient has reached the transient state-type P, she or he is immediately at risk again to reach one of the three absorbing state-types D, H, or R. The state-chart for this special multi-state model is presented in Figure 2. This data set has already been analysed in an un-penalised multi-state model Teixeira et al. (2015) and we refer to these results as a benchmark model in Section 3. A closer look on Teixeira et al. (2015) in combination with the model structure illustrated in Figure 2 suggests to take into account the possibility of cross-transition-type effects: While some covariates or risk factors have considerable differences between the estimated coefficients for the same absorbing exit state-type, others, like for example the presence of diabetes in connection with a transition to state-type D, share only small differences. These small differences

are possibly negligible from an applied perspective and a data-driven algorithm setting these to 0 may be of a high practical value.

Several general overview articles for multi-state models face the point of shared features across transition types, e.g.: "Of course it is not of much interest to make a joint model which gives the same results as the models fitted separately to each transition. The point comes from reducing this model to one which is more parsimonious yet sensible" (Carstensen and Plummer, 2011), "[...] go a step further in order to analyse more parsimonious models where some baseline intensities are proportional or where some covariates have the same effect on several transition intensities" (Andersen and Keiding, 2002), or "Interaction effects between covariates and strata may be used to assess whether covariate effects vary across competing outcomes [...]" (Clark et al., 2003). None of these articles or to our best knowledge any other publication gives a data-driven, and hence labour un-intensive, algorithm to solve this problem and answering the question about the occurrence of shared features across transition types.

In general, the number of degrees of freedom of a multi-state model, i.e. the number of regression parameters to be estimated, is the number of all transition-type specific effect parameters. This quantity increases as a product of two numbers, the number of covariates to be considered and the number of transition-types. Consequently, in a multi-state model with a large number of transition-types and/or potential covariates, the maximum model is quite complex and hard to interpret. To address this difficulty and also to alleviate over-fitting, a practical solution is to regularize the number of degrees of freedom of the model and to assume that only a few of all the transition-type specific effects are actually relevant for prediction. That is, the vector of model coefficients is sparse, meaning that the transition-type specific covariates whose associated model coefficients take a value equal to zero do not contribute to the decisions made by the multi-state process and are hence considered to be irrelevant. If performed in a non-automated way, the search of this subset of relevant model coefficients is a very labour-intensive task. For an automated, i.e. data-driven, approach, different alternative strategies to estimate the model coefficients under the sparsity assumption are described in the literature for regression models, such as linear or generalized linear models, or the Cox proportional hazards model in event time analysis. The *Least angle shrinkage and selection operator* (*Lasso*) (Tibshirani, 1996) is one of the most prominent approaches with the central idea to regularize the absolute value of model coefficients with the consequence that single model coefficients may be estimated exactly to 0. The Lasso is successfully applied to Cox proportional hazards models (Tibshirani et al., 1997) and generalized to penalties with additional positivity constraints, constraints on the absolute differences between model coefficients and constraints on squared model coefficient values (Goeman, 2010). This is accompanied with an alternative estimation algorithm using a series of directional Taylor approximations and results in the very robust performance of the R (R Development Core Team, 2014) add-on package `penalised` (Goeman, 2012). However, the implementation

in `penalised` as well as the presentation in Goeman (2010) allows to imply constraints on the absolute differences between model coefficients in a structured manner only in a limited manner, i.e. as it is used for ordinal covariates or in feature selection. This does not allow to adequately use prior knowledge about a possible grouping structure within the multi-state model.

The process of inducing the covariate coefficients under the sparsity assumption can be facilitated when prior information is available about groups of features that are expected to be jointly relevant or jointly irrelevant for prediction (Huang and Zhang, 2010), i.e. when different groups of covariate coefficients are expected to be jointly equal to or jointly different from zero. Finding this type of information can be difficult in practice, but is in many practical multi-state models directly defined by the underlying state-chart as defined in the above examples of the illness-death model with recovery, or the peritoneal dialysis program multi-state model. Having this information at hand might be beneficial to improve the estimates of the covariate coefficients and to reduce the number of samples required to obtain a good generalization performance. As described by Puig et al. (2011), there is in general a very wide range of applications where sparsity at the group level is beneficial, including regression with grouped variables, source localization, or whole genome association mapping. As with the individual sparsity assumption, sparsity at the group level can be introduced in the estimation process of the model coefficients by considering specific regularization norms at the group level.

To sum up the central points: regularization is a natural task in multi-state modelling induced by the highly parametrized nature of this model class, and using additional information about the structure of the model can be beneficial for the estimation result with respect to interpretability and generalisability. Both points can be incorporated into a single regularized estimation approach for multi-state models based on penalties for absolute values of model coefficients in combination with penalties for particular selected pairwise differences of model coefficients. The selection of pairs can be gathered from the state-chart of the multi-state model. The theory and implementation of this approach is described in the following section, based on a model with proportional hazards assumption, or with a piece-wise exponential model formulation. The theoretical concepts and practical steps behind this special penalisation concept are described, i.e. how to set up a suitable penalty-term, and the performance is illustrated on a real data set in Section 3.

2 Penalised likelihood formulation of a multi-state model

This section describes the construction of the multi-state modelling approach proposed by this article. The basic concepts are very thoroughly described in Andersen and Keiding (2002) and Andersen et al. (1993), and the following introduction is in especially

closely based on Andersen and Keiding (2002).

2.1 Likelihood formulation for multi-state models

The underlying mathematical concept of how to treat a multi-state model is a multi-state process $Y(t)$, $t \in \mathcal{T}$, a stochastic process with a finite state-type space $\mathcal{K} = \{1, \ldots, k, \ldots, k', \ldots, K\}$ and right-continuous sample path $Y(t+) = Y(t)$, where $t+$ denotes a time point directly after t. The sample path depicts itself as a constant function between the times of transitions of the process, where the support of the time is a subset of the positive real values $\mathcal{T} = [0, t_{\max}]$, with $0 < t_{\max} < \infty$. Over the course of time, a multi-state process $Y(\cdot)$ generates a history $\mathcal{Y}_t$, which is the σ-algebra generated by the observed sample path in the interval $[0, t]$.

We may define transition-type specific transition probabilities:

$$P_q(s,t) = P_{k.k'}(s,t) = \mathbb{P}\left(Y(t) = k' \,|\, Y(s) = k, \mathcal{Y}_{s-}\right),$$

for $k, k' \in \mathcal{K}$, $s, t \in \mathcal{T}$, $s \leq t$, transition-types denoted by $q = k.k' \in \mathcal{Q}$ referring to direct transitions from the initial state-type k to the exit state-type k', and $s-$ defined in analogy to $t+$.

Upon this definition we may furthermore define transition-type specific transition intensities:

$$\lambda_q(t) = \lim_{\Delta_t \downarrow 0} \frac{P_q(t, t+\Delta_t)}{\Delta_t},$$

which we shall assume exist.

Of course, not all transition-types defined by combinations of k and k' will be meaningful in praxis. State-charts, as e.g. in Figures 1 and 2, are a useful tool for graphical representations of the transition-types of a multi-state model that are possibly observable. Let $\mathcal{Q} = \{1, \ldots, q, \ldots, q', \ldots, Q\}$ define the set of possibly observable transition-types, i.e. combinations as $q = k.k'$ with $k, k' \in \mathcal{K}$ for which one t exists such that $\lambda_q(t) > 0$.

Assume that multi-state processes $Y_i(t)$ are observed over intervals $[0, t_{\max,i}]$ for individuals i, $i = 1, \ldots, N$, where $t_{\max,i}$ is the time of termination of the observation for individual i. Since the individual processes are constant between observed transitions, it is equivalent to record the state-type at the origin $Y_i(0)$ and the counting processes:

$$C_{q,i}(t) = \text{number of observed transition with transition-type } q \text{ for } i \text{ in } [0,t],$$

described by the times $t_{q,i,c}$ of these transitions, with:

$$0 < t_{q,i,1} < \ldots < t_{q,i,C_{q,i}(t_{\max,i})} = t_{\max,i}.$$

We represent the overall transition counting process by $C_q(t) = \sum_{i=1}^{N} C_{q,i}(t)$. We will

further need an at-risk indicator process for transition-type $q = k.k'$ which we define by:

$$R_{q,i}(t) = \mathrm{I}_{\{Y_i(t-)=k\}},$$

and $R_q(t) = \sum_{i=1}^{N} R_{q,i}(t)$, with $C_{q,i}(t) = C_{q,i}(t_{\max,i})$ and $R_{q,i}(t) = 0$ for $t > t_{\max,i}$. The at-risk processes $R_q(t)$ and $R_{q,i}(t)$ are identical across $t \in \mathcal{T}$ for all $q \in \mathcal{Q}$ with the same entry state-type $k \in \mathcal{K}$.

The likelihood L conditional on the initial distribution of the multi-state process and the density of covariates is (Andersen et al., 1993):

$$L = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \left(\prod_{q=1}^{Q} \left[\exp\left(- \int_0^{t_{\max,i}} \lambda_{q,i}(t) R_{q,i}(t) \, \mathrm{d}t \right) \prod_{c=1}^{C_{q,i}(t_{\max,i})} \lambda_{q,i}(t_{q,i,c}) \right] \right).$$

We may rewrite each individual contribution L_i as:

$$L_i = \prod_{q=1}^{Q} \left[\lambda_{q,i}(t_{q,i,1}) \exp\left(- \int_0^{t_{q,i,1}} \lambda_{q,i}(t) R_{q,i}(t) \, \mathrm{d}t \right) \prod_{c=2}^{C_{q,i}(t_{\max,i})} \lambda_{q,i}(t_{q,i,c}) \exp\left(- \int_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}(t) R_{q,i}(t) \, \mathrm{d}t \right) \right].$$

As indicated by this formulation, it is very natural to include left-truncation, also known as late entry, and right-censoring into this framework. For left-truncation, only the lower integral boundary has to be changed from the value 0 to the time of left-truncation which we represent by $t_{q,i,0}$:

$$L_i = \prod_{q=1}^{Q} \left[\lambda_{q,i}(t_{q,i,1}) \exp\left(- \int_{t_{q,i,0}}^{t_{q,i,1}} \lambda_{q,i}(t) R_{q,i}(t) \, \mathrm{d}t \right) \prod_{c=2}^{C_{q,i}(t_{\max,i})} \lambda_{q,i}(t_{q,i,c}) \exp\left(- \int_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}(t) R_{q,i}(t) \, \mathrm{d}t \right) \right].$$

For right censoring, a non-censoring indicator δ_i has to be included for the potential last

jump of the counting processes at $t_{\max,i}$:

$$L_i = \prod_{q=1}^{Q} \left[\lambda_{q,i}\left(t_{q,i,1}\right) \exp\left(- \int_{t_{q,i,0}}^{t_{q,i,1}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \cdot \right.$$
$$\cdot \left(\prod_{c=2}^{C_{q,i}(t_{\max,i})-1} \lambda_{q,i}\left(t_{q,i,c}\right) \exp\left(- \int_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \right) \cdot$$
$$\left. \cdot \lambda_{q,i}\left(t_{\max,i}\right)^{\delta_i} \exp\left(- \int_{t_{q,i,C_{q,i}\left(t_{\max,i}\right)-1}}^{t_{\max,i}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \right].$$

So we can treat this as a combined product over c where a transition count specific non-censoring indicator $\delta_{q,i,c}$ is always equal to one despite $\delta_{q,i,C_{q,i}(t_{\max,i})}$ which may also be equal to zero:

$$L_i = \prod_{q=1}^{Q} \left[\prod_{c=1}^{C_{q,i}(t_{\max,i})} \left(\lambda_{q,i}\left(t_{q,i,c}\right)^{\delta_{q,i,c}} \exp\left(- \int_{t_{q,i,c-1}}^{t_{q,i,c}} \lambda_{q,i}\left(t\right) R_{q,i}\left(t\right) \mathrm{d}t \right) \right) \right]. \qquad (1)$$

2.2 Parametrization of the transition-type specific hazard rate functions

In event-time analysis, statistical models are often obtained by specifying transition-type specific hazard rate functions $\lambda_{q,i}\left(t\right)$ for each individual i. The most frequently used regression models have a multiplicative structure with a baseline transition hazard rate function $\lambda_{q,0}\left(t\right)$ assumed common for all individuals. For an individual i the transition-type specific baseline hazard rate is modelled by:

$$\lambda_{q,i}\left(t\right) = \lambda_{q,0}\left(t\right) \exp\left(\mathbf{x}_i^{\top} \boldsymbol{\beta}_q \right),$$

with time-fixed covariates $\mathbf{x}_i = \left(x_{1,i}, \ldots, x_{P,i}\right)^{\top}$ and respective effects $\boldsymbol{\beta}_q = \left(\beta_{q,1}, \ldots, \beta_{q,P}\right)^{\top}$ on transition-type q, which form together an transition-type specific linear predictor $\eta_{q,i} = \mathbf{x}_i^{\top} \boldsymbol{\beta}_q$. Hence, the effect of a covariate x_p is described by factors of proportionality $\exp\left(\beta_{q.p}\right)$ on the transition-type specific baseline-hazard rate function $\lambda_{q,0}\left(t\right)$. Furthermore, the major class of models is the continuous time Markov model where the multi-state process $Y\left(t\right)$ is a Markov process, i.e. takes the assumption that the dependence of the transition-type specific hazard rates functions $\lambda_q\left(t\right)$ on the history $\mathcal{Y}_t$ is only via the current state of $Y\left(t\right)$ and possibly via time-fixed covariates.

In the simplest model class the transition-type specific baseline hazard rates are kept

constant:

$$\lambda_{q,0}(t) = \lambda_{q,0},$$

or piece-wise constant:

$$\lambda_{q,0}(t) = \lambda_{q,0}^{(j)}, \quad t^{(j-1)} < t \leq t^{(j)},$$

for a time-axis decomposed into several sub-intervals by artificial time points $t^{(0)} = 0, t^{(1)}, t^{(2)}, \ldots, t^{(j)}, \ldots t^{(J)} = t_{\max}$.

The Cox partial likelihood model class (Cox, 1972) leaves the baseline hazard rate functions unspecified but assumes them to be equal across individuals. If one is not interested in the underlying functional form of baseline hazard rate function, the Cox partial likelihood model is a good choice since it leaves no room for functional misspecification.

We will use both, a piece-wise constant and an unspecified hazard rate function parametrisation to set up fusion Lasso penalised multi-state models. Of course, many parametric alternatives of the baseline hazard rate function specification exist, but will not be treated in this article. As a ground concept, all of these models need a pre-estimation data management procedure that is described in the following section.

2.3 Event history observations in the long format

Throughout this article, the transition-types of a multi-state model are represented by $q = 1, \ldots, Q$, and a transition-type q is composed by an entry state-type $\text{state}_{\text{entry},q}$ and an exit state-type $\text{state}_{\text{exit},q}$. Furthermore, entry time $t_{\text{entry},i}$ and exit time $t_{\text{exit},i}$ represent transition (or censoring) times, where the information about non-censoring of the event q is captured by a transition-type specific non-censoring indicator $\delta_{q,i}$. In combination, the difference $t_{\text{exit},i} - t_{\text{entry},i}$ between these two time points measures the length of duration of the at-risk spell i. It is important to note that an at-risk spell i here refers to one time interval for one observation unit and each i is represented by a corresponding line in the final dataset, a consequence of using the long format for multi-state model observations as introduced by Putter et al. (2007). The number of competing at-risk processes is visualized by arrows pointing away from one node in the state-chart of a multi-state model, as e.g. in Figures 1 and 2. For the respective entry state-type $\text{state}_{\text{entry}}$ of a specific observed (or censored) transition, the number of competing at-risk processes defines the number of at-risk spells that result for this (censored) observation.

In a competing risk setting with four competing exit states, for instance, one observed event or censoring time leads to four lines in the long format dataset, representing four at-risk spells. For the comprehensive non-censoring indicator δ as used in the long data format, only one line out of four has the capability to take the value 1 if it has been actually observed (else $\delta = 0$), the other three out of four lines strictly take value $\delta = 0$. In the exemplary illness-death model with recovery (state-chart in Figure 1) and one

time-constant covariate x_p, two exemplary original event history observations with the three state-types healthy, illness, and death represented by $\{\mathrm{H}, \mathrm{I}, \mathrm{D}\}$, transition-types by $q \in \{\mathrm{HI}, \mathrm{HD}, \mathrm{IH}, \mathrm{ID}\}$, might look like this:

patient id	$\text{state}_{\text{entry}}$	$\text{state}_{\text{exit}}$	t_{entry}	t_{exit}	δ_{HI}	δ_{HD}	δ_{IH}	δ_{ID}	x_p
1	H	D	0	2.28	0	1	0	0	$x_{p,1}$
2	H	I	0	1.5	1	0	0	0	$x_{p,2}$
2	I	H	1.5	7.89	0	0	1	0	$x_{p,2}$
2	H	I	7.89	9.15	1	0	0	0	$x_{p,2}$
2	I	NA	9.15	10	0	0	0	0	$x_{p,2}$

Here, we use transition-type specific non-censoring indicators δ_q. The last observation of patient 2 has been right-censored at time 10 and the exit state-type is consequentially not available (NA).

In the long format, these observations are assigned by the following exemplary data set, where we use a global non-censoring indicator δ and transition-type specific covariates $x_{p.\text{HI}}$, $x_{p.\text{HD}}$, $x_{p.\text{IH}}$, and $x_{p.\text{ID}}$:

patient id	$\text{state}_{\text{entry}}$	$\text{state}_{\text{exit}}$	t_{entry}	t_{exit}	δ	$x_{p.\text{HI}}$	$x_{p.\text{HD}}$	$x_{p.\text{IH}}$	$x_{p.\text{ID}}$
1	H	I	0	2.28	0	$x_{p,1}$	0	0	0
1	H	D	0	2.28	1	0	$x_{p,1}$	0	0
2	H	I	0	1.5	1	$x_{p,2}$	0	0	0
2	H	D	0	1.5	0	0	$x_{p,2}$	0	0
2	I	H	1.5	7.89	1	0	0	$x_{p,2}$	0
2	I	D	1.5	7.89	0	0	0	0	$x_{p,2}$
2	H	I	7.89	9.15	1	$x_{p,2}$	0	0	0
2	H	D	7.89	9.15	0	0	$x_{p,2}$	0	0
2	I	H	9.15	10	0	0	0	$x_{p,2}$	0
2	I	D	9.15	10	0	0	0	0	$x_{p,2}$

A data-set in the long-data format will consist of the extended number of n lines, with $n > N$.

The construction of transition-type specific covariate vector versions was performed using transition-type indicators:

$$\psi_{q,i} := \mathrm{I}_{\{\text{state}_{\text{entry},i}=\text{state}_{\text{entry},q}\}} \cdot \mathrm{I}_{\left\{\text{state}_{\text{exit},i}=\text{state}_{\text{exit},q}\right\}}.$$

It is now convenient to formulate a general linear predictor in the spirit of general

regression models:

$$\eta_i = (\psi_{1,i}x_{1,i}, \psi_{1,i}x_{2,i}, \ldots, \psi_{1,i}x_{P,i}, \psi_{2,i}x_{1,i}, \ldots, \psi_{Q,i}x_{P,i},) \cdot \begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \vdots \\ \beta_{P,1} \\ \beta_{1,2} \\ \vdots \\ \beta_{P,Q} \end{pmatrix},$$

which are equally introduced in each transition-type specific hazard rate formulation:

$$\lambda_{q,i}(t) = \lambda_{q,0}(t) \exp(\eta_i).$$

Here the index p, $p = 1, \ldots, P$, represents the P covariates used to model the transition-type specific factors of proportionality on the transition-type specific baseline hazard rate functions. The allocation of one spell i to its suitable transition-type q is described in a broader context in Andersen et al. (1993) and is a mandatory practice to use the long format. In the following, the vector collecting all transition-type specific covariate coefficients are represented by $\boldsymbol{\beta} := (\beta_{1,1}, \ldots, \beta_{P,Q})^\top$.

2.4 Stratified partial likelihood formulation for multi-state models

The Cox log partial likelihood can be derived as a profile likelihood from the full likelihood (Johansen, 1983) and is then given in the stratified log partial likelihood form by:

$$\begin{aligned} \mathrm{pl}(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \sum_{q=1}^{Q} \sum_{c=1}^{C_{q,i}(t_{\max,i})} \log \left(\left(\frac{\exp(\eta_{q,i})}{\sum_{j=1}^{N} \sum_{c=1}^{C_{q,j}(t_{\max,j})} R_{q,j}(t_{q,i,c}) \exp(\eta_{q,j})} \right)^{\delta_{q,i,c}} \right) \\ &= \sum_{i=1}^{N} \sum_{q=1}^{Q} \sum_{c=1}^{C_{q,i}(t_{\max,i})} \delta_{q,i,c} \left(\eta_{q,i} - \log \left(\sum_{j=1}^{N} \sum_{c=1}^{C_{q,j}(t_{\max,j})} R_{q,j}(t_{q,i,c}) \exp(\eta_{q,j}) \right) \right), \end{aligned}$$

where $\eta_{q,i} = \mathbf{x}_i^\top \boldsymbol{\beta}_q$. By the application of the long format given in Section 2.3, we may formulate this in a much more compact way:

$$\mathrm{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\eta_i - \log \left(\sum_{j \in \mathrm{R}_i} \exp(\eta_j) \right) \right),$$

where i and j now represent single lines in the long-format data, which allows to drop the index q from the risk-set formulation:

$$\mathrm{R}_i := \{j : t_{\mathrm{entry},j} < t_{\mathrm{exit},i} \leq t_{\mathrm{exit},j},\, \mathrm{state}_{\mathrm{entry},i} = \mathrm{state}_{\mathrm{entry},j},\, \mathrm{state}_{\mathrm{exit},i} = \mathrm{state}_{\mathrm{exit},j}\}\,.$$

This is an at-risk line index oriented notation of the at-risk process described in Section 2.1. The respective information about entry state-types/times and exit state-types/times is captured in the long format using vectors $\mathbf{t}_{\mathrm{entry}}$, $\mathbf{t}_{\mathrm{exit}}$, $\mathbf{state}_{\mathrm{entry}}$, $\mathbf{state}_{\mathrm{exit}}$, alike for the event indicator $\boldsymbol{\delta}$, and the matrix of transition-type specific covariate information $\mathbf{X}$. The abbreviation $\mathrm{pl}\,(\boldsymbol{\beta})$ is used to refer to this log partial likelihood in the following.

A general penalised negative log partial likelihood is now formulated as:

$$\mathrm{pnpl}\,(\boldsymbol{\beta}) = -\mathrm{pl}\,(\boldsymbol{\beta}) + \mathrm{pen}\,(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta})\,.$$

Here, the penalty can take several forms, depending on features of the covariate mechanisms that one wants to detect. This is technically achieved by a penalty structure matrix $\mathbf{D}$ which is described later in Section 2.6, accompanied by the introduction of different types of penalty structures. For the estimation of penalised models proposed in Section 2.7.1, we will need the score vector $\mathbf{s}\,(\boldsymbol{\beta}) = \dfrac{\partial\,\mathrm{pl}\,(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}}$ with components $s_p\,(\boldsymbol{\beta}) = \dfrac{\partial\,\mathrm{pl}\,(\boldsymbol{\beta})}{\partial\,\beta_p}$:

$$s_p\,(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\delta_i x_{i,p} - \delta_i \sum_{j \in R_i} \frac{\exp\,(\eta_j)\, x_{j,p}}{\sum\limits_{k \in R_i} \exp\,(\eta_k)} \right),$$

and the Fisher information matrix $\mathbf{F}\,(\boldsymbol{\beta}) = \dfrac{\partial^2\,\mathrm{pl}\,(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}\,\partial\,\boldsymbol{\beta}^\top}$ with components $F_{p,p'}\,(\boldsymbol{\beta}) = \dfrac{\partial^2\,\mathrm{pl}\,(\boldsymbol{\beta})}{\partial\,\beta_p\,\partial\,\beta_{p'}}$:

$$F_{p,p'}\,(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \frac{\sum\limits_{j \in R_i} \exp\,(\eta_j)\, x_{j,p} x_{j,p'}}{\sum\limits_{k \in R_i} \exp\,(\eta_k)} - \\ - \sum_{i=1}^{n} \delta_i \frac{\left(\sum\limits_{j \in R_i} \exp\,(\eta_j)\, x_{j,p} \right) \cdot \left(\sum\limits_{j \in R_i} \exp\,(\eta_j)\, x_{j,p'} \right)}{\left(\sum\limits_{k \in R_i} \exp\,(\eta_k) \right)^2}.$$

2.5 Piece-wise exponential model

The piece-wise exponential model for single transition-type survival models with the assumption of piece-wise constant baseline hazard rate functions is frequently used in the literature and e.g. described in Rodríguez-Girondo et al. (2013). The assumption of piece-wise constant baseline hazard rate functions needs the definition of an artificial decomposition of the time axis into several sub-intervals as for instance in Section 2.2. This is referred to as *data augmentation* (Rodríguez-Girondo et al., 2013). There are several possibilities to rely on already available software performing data augmentation, such as the function `Lexis` from the `R` add-on package `Epi` (Carstensen et al., 2014), (Carstensen and Plummer, 2011). By this representation of event-times via data augmentation, we are able to use a Poisson maximum likelihood estimation scheme, which is motivated in the following.

For piece-wise exponential models we define a measure $\Delta_{q,i,c}^{(j)}$ that specifies the time-length in the j-th time sub-interval in which individual i was at-risk for the c-th transition of transition-type q. Here, i represents one single event-history observation again as in Section 2.1. $\Delta_{q,i,c}^{(j)}$ will be equal to zero in most cases and may take a maximum value of $t^{(j)} - t^{(j-1)}$. Additionally we define $j_{q,i,c}$ which specifies the sub-interval in which the c-th transition of transition-type q for individual i occurred. We furthermore include the sub-interval specific constant baseline hazard rate $\lambda_{q,0}^{(j)}$ into the exponential function of the linear predictor by:

$$\exp\left(\log\left(\lambda_{q,0}^{(j)}\right) + \mathbf{x}_i^\top \boldsymbol{\beta}\right) =: \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right).$$

Using this, we can now characterize each survival function component by:

$$\exp\left(-\sum_{j=1}^{J} \Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right) = \prod_{j=1}^{J} \exp\left(-\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right),$$

and each hazard rate component by:

$$\exp\left(\alpha_q^{(j_{q,i,c})} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)^{\delta_{q,i,c}}.$$

The fully composed likelihood is then of the form:

$$L = \prod_{i=1}^{N} \left(\prod_{q=1}^{Q} \left[\prod_{c=1}^{C_{q,i}(T_i)} \left(\exp\left(\alpha_q^{(j_{q,i,c})} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)^{\delta_{q,i,c}} \cdot \right.\right.\right.$$
$$\left.\left.\left. \cdot \prod_{j=1}^{J} \exp\left(-\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right) \right) \right] \right).$$

We now specify sub-interval specific versions $\delta_{q,i,c}^{(j)}$ of $\delta_{q,i,c}$ which take always value zero, despite for the interval $j_{q,i,c}$:

$$L = \prod_{i=1}^{N} \left(\prod_{q=1}^{Q} \left(\prod_{c=1}^{C_{q,i}(T_i)} \left(\prod_{j=1}^{J} \left[\exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)^{\delta_{q,i,c}^{(j)}} \cdot \right. \right. \right. \right. \right.$$
$$\left. \left. \left. \left. \left. \cdot \exp\left(-\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right) \right] \right) \right) \right) \right).$$

Taking the log of this likelihood transforms each of the outer products to a sum with the same indices, and each of the inner factors changes to:

$$\delta_{q,i,c}^{(j)} \cdot \left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right).$$

This yields a likelihood that agrees, except for constants, with the likelihood one would obtain from Poisson distributed observations. For clarification, assume that some δ_i was Poisson distributed with mean $\mu_i = t_i \lambda_i$, and hence results in the following log likelihood contribution (with ignoring of constant factors):

$$\begin{aligned} l_i &\propto \delta_i \log(\mu_i) - \mu_i \\ &= \delta_i \log(t_i \lambda_i) - t_i \lambda_i \\ &= \delta_i \log(t_i) + \delta_i \log(\lambda_i) - t_i \lambda_i. \end{aligned}$$

Since $\delta_i \log(t_i)$ does not depend on any parameter in λ_i, it can be ignored from the point of view of estimation:

$$l_i \propto \delta_i \log(\lambda_i) - t_i \lambda_i.$$

We yield a likelihood that is proportional to the likelihood for Poisson distributed response observations $\delta_{q,i,c}^{(j)}$ with mean $\Delta_{q,i,c}^{(j)} \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)$, and so we get to the following log likelihood in the estimation of the piece-wise exponential model (n now represents the number of sub-interval at-risk observations according to the above data augmentation):

$$\mathrm{l}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(-\Delta t_i \cdot \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right) + \delta_i \cdot \log\left(\Delta t_i \cdot \exp\left(\alpha_q^{(j)} + \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right) \right),$$

where $\Delta t_i := t_i^{(j)} - t_i^{(j-1)}$ serves as offset. Here conditioning on $\{\mathbf{t}_{\text{entry}}, \mathbf{t}_{\text{exit}}, \mathbf{state}_{\text{entry}}, \mathbf{state}_{\text{exit}}, \boldsymbol{\delta}, \mathbf{X}\}$ is again suppressed for notational simplicity.

The linear predictor η_i^{pe} in the piece-wise exponential model is composed by:

$$\eta_i^{\text{pe}} = \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{q=1}^{Q} \psi_{q,i} \cdot \log\left(f_{t_q}\left(t_{q,i}\right)\right).$$

with $f_{t_q}\left(t_{q,i}\right)$ in the role of the transition-type specific baseline hazard rate function $\lambda_{q,0}\left(t\right) = \exp\left(\alpha_q^{(\bullet)}\right)$, and $\psi_{q,i}$ as defined in Section 2.3. A specification for modelling of continuous covariates that is often used throughout the literature is the class of *fractional polynomials* (Lambert et al., 2005). For the estimation of the baseline hazard rate function, we use a specification that was used in the modelling of this type of effect class before (Carstensen and Center, 2005), i.e.:

$$f_{t_q}\left(t_{q,i}\right) = \sum_m t_{q,i}^m \beta_{t_q,m},$$

with $m = \left\{\frac{1}{3}, \frac{1}{2}, 0, 1, \frac{3}{2}, 2\right\}$ and $t_{q,i}^0 := \log\left(t_{q,i}\right)$. This setup is furthermore applicable to any non-linear effect component in the model, such as the effect of age in the application described in Section 3.

Again, we need the score vector $\mathbf{s}\left(\boldsymbol{\beta}\right) = \dfrac{\partial\, \mathrm{l}\left(\boldsymbol{\beta}\right)}{\partial\, \boldsymbol{\beta}}$ with components $s_p\left(\boldsymbol{\beta}\right) = \dfrac{\partial\, \mathrm{l}\left(\boldsymbol{\beta}\right)}{\partial\, \beta_p}$:

$$\mathbf{s}\left(\boldsymbol{\beta}\right) = \mathbf{X}^\top\left(\boldsymbol{\delta} - \boldsymbol{\mu}\right),$$

where $\mu := \Delta t_i \cdot \exp\left(\eta_i^{\text{pe}}\right)$, and the Fisher information matrix $\mathbf{F}\left(\boldsymbol{\beta}\right) = \dfrac{\partial^2\, \mathrm{l}\left(\boldsymbol{\beta}\right)}{\partial\, \boldsymbol{\beta}\, \partial\, \boldsymbol{\beta}^\top}$ with components $F_{p,q}\left(\boldsymbol{\beta}\right) = \dfrac{\partial^2\, \mathrm{l}\left(\boldsymbol{\beta}\right)}{\partial\, \beta_p\, \partial\, \beta_q}$:

$$\mathbf{F}\left(\boldsymbol{\beta}\right) = \mathbf{X}^\top \mathrm{diag}\left(\boldsymbol{\mu}\right)\mathbf{X},$$

for the estimation algorithm that will be defined in Section 2.7.1.

2.6 Penalty types

The *least absolute shrinkage and selection operator*, short *Lasso*, introduced by Tibshirani (1996) maximizes a likelihood "subject to the sum of the absolute value of the coefficients being less than a constant" (Tibshirani, 1996). A Lasso type penalty term penalising the absolute value of all elements of the parameter vector $\boldsymbol{\beta}$ is constructed as:

$$\mathrm{pen}_{\mathrm{L}}\left(\lambda, \mathbf{D}, \boldsymbol{\beta}\right) = \lambda \sum_{p=1}^{P} \left|\beta_p\right| = \lambda \sum_{l=1}^{P} \mathbf{d}_l^\top \left|\boldsymbol{\beta}\right|,$$

using penalty parameter λ, parameter vector $\boldsymbol{\beta}$, and difference vectors $\mathbf{d}_l^\top = (0, \ldots, 0, 1, 0, \ldots, 0)$ taking value 1 in the l-th entry, and 0 else. The vectors $\mathbf{d}_l$ are then stored as lines in the penalty structure matrix $\mathbf{D}$, which is here equal to the $P \times P$ dimensional identity matrix. Matrix $\mathbf{D}$ is useful to incorporate several penalty types into a composite approach. This is made more clear by the following passage.

The fused Lasso (Tibshirani et al., 2005) was introduced with the intention to generalize the Lasso penalisation approach "for problems with features that can be ordered in some meaningful way" (Tibshirani et al., 2005). This is achieved by penalising the L_1-norm, i.e. the absolute value, of both the coefficients and their successive differences. By this, the fused Lasso leads to sparsity of the coefficients and also of the their differences between adjacent covariate level effects. A fused Lasso type penalty term penalising the absolute value of all elements of the parameter vector $\boldsymbol{\beta}$ and all of the successive differences is constructed as:

$$\text{pen}_{\text{FL}}(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}) = \text{pen}_{\text{L}}(\lambda_1, \mathbf{D}_{\text{L}}, \boldsymbol{\beta}) + \text{pen}_{\text{F}}(\lambda_2, \mathbf{D}_{\text{F}}, \boldsymbol{\beta}),$$

with $\text{pen}_{\text{F}}(\lambda_2, \mathbf{D}_{\text{F}}, \boldsymbol{\beta}) = \lambda_2 \sum_{p=1}^{P-1} |\beta_{p+1} - \beta_p| = \lambda \sum_{l=1}^{P-1} |\mathbf{d}_{\text{F},l}^\top \boldsymbol{\beta}|$, fusion difference vectors $\mathbf{d}_{\text{F},l}^\top = (0, \ldots, 0, -1, 1, 0, \ldots, 0)$, penalty parameter vector $\boldsymbol{\beta}$, and a combined penalty structure matrix $\mathbf{D} = [\mathbf{D}_{\text{L}}, \mathbf{D}_{\text{F}}]$, with the fusion difference vectors $\mathbf{d}_{\text{F},l}$ stored as lines in the penalty structure matrix $\mathbf{D}_{\text{F}}$.

The pairwise fused Lasso proposed by Petry et al. (2011) extends the fused Lasso (Tibshirani et al., 2005) to situations where the predictors have no natural ordering. In other words, not only next neighbour coefficient differences, but all pairwise coefficient differences are penalised. Here, the fusion difference vectors $\mathbf{d}_{\text{F},l}^\top = (0, \ldots, 0, -1, 1, 0, \ldots, 0)$ for differences between adjacent effects are supplemented by the remaining vectors $\mathbf{d}_{\text{all pairwise fusion},l}$ leading to all possible pairwise effect differences.

To penalise transition-type specific covariate coefficients and their pairwise differences in multi-state modelling, a fusion Lasso penalty term is introduced. This penalty term takes into account the information provided by the state-chart of the multi-state model and is of the general form:

$$\text{pen}_{\text{SFL}}(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}) = \lambda_1 \sum_{q=1}^{Q} \sum_{p=1}^{P} |\beta_{p,q}| + \lambda_2 \sum_{q,q'} \sum_{p=1}^{P} |\beta_{p,q} - \beta_{p,q'}|,$$

with $\lambda_1 \sum_{q=1}^{Q} \sum_{p=1}^{P} |\beta_{q,j}|$ as Lasso type and $\lambda_2 \sum_{q,q'} \sum_{p=1}^{P} |\beta_{p,q} - \beta_{p,q'}|$ as fusion type penalty term. Here, the indices q, q' represent suitable pairs of transitions that are to be extracted from the state-chart of the multi-state model.

One important aspect is that the solutions of penalisation approaches, such as the

Lasso, are not equivariant under scaling of the used covariates (Hastie et al., 2001). In other words, there is a dependency on the scales of the covariates with respect to the solution of the penalised estimation when unique penalty parameter values λ_1 and λ_2 are selected for both penalty term components. An often used solution to get rid of this general problem in penalisation algorithms is to use standardized covariate versions (Hastie et al., 2001), i.e. $\mathbf{x}_p^* := \frac{\mathbf{x}_p - \hat{\mu}_{\mathbf{x}_p}}{\hat{\sigma}_{\mathbf{x}_p}}$, with $\hat{\mu}_{\mathbf{x}_p}$ as the empirical mean of $\mathbf{x}_p$, and $\hat{\sigma}_{\mathbf{x}_p}$ as the empirical standard deviation of $\mathbf{x}_p$. For the interpretation, the results are back-transformed after the estimation is performed. It is important to note here that using fusion penalties for differences between transition-type specific covariates, this scaling has to be performed on combined or grouped measures $\hat{\mu}_{\mathbf{x}_{p,q.q'}}$ and $\hat{\sigma}_{\mathbf{x}_{p,q.q'}}$ to maintain the fusion feature for effect estimates after the back-transformation step. This has the consequence that not all covariates within the penalised estimation task are having the exactly same scaling. However, there is no escape from this problem: scaling each covariate independently leads to equal scales, group scaling leads to the preservation of the fusion feature. For the peritoneal dialysis program data that is analysed in Section 3, the range for standard deviations of the scaled covariates $\mathbf{x}_p^*$ ranges between 0.87 and 1.1, and we therefore consider these differences to be of only low relevance, all the more in consideration of the fact that the frequencies of transition-type observation are quite unbalanced in this example. A way to control whether these results lead to any false effect fusions, is to re-estimate the model with individually scaled covariates and check if any different effect fusions occur.

2.7 Estimation approaches

The approach introduced by this article may be performed on the log partial likelihood $\mathrm{lpl}(\boldsymbol{\beta})$ or the log likelihood $\mathrm{l}(\boldsymbol{\beta})$, where the conditioning on other quantities is still notationally suppressed as described in Sections 2.4 and 2.5. For a clear notation we use the abbreviation $\mathrm{l}(\boldsymbol{\beta})$, but always mean this as a representative for $\mathrm{lpl}(\boldsymbol{\beta})$, too.

A naive opportunity to get penalised estimates would be to directly minimize the penalised negative log partial likelihood $\mathrm{pnl}(\boldsymbol{\beta})$ with respect to the parameters $\boldsymbol{\beta}$, i.e. by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\mathrm{pnl}\left(\boldsymbol{\beta}\right) \right),$$

using direct numerical optimization techniques, such as the Nelder-Mead (Nelder and Mead, 1965) algorithm. Since the performance of this would not be optimal with respect to computational cost and instability, we can alternatively use a modified version of an algorithm that is based on a first order Taylor series expansion Tibshirani et al. (1997): $\left(\mathbf{z} - \boldsymbol{\eta}\right)^\top \mathbf{A} \left(\mathbf{z} - \boldsymbol{\eta}\right)$ for the log (partial) likelihood $\mathrm{l}(\boldsymbol{\beta})$, with $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{u} = \frac{\partial \mathrm{l}}{\partial \boldsymbol{\eta}}$, $\mathbf{A} = -\frac{\partial^2 \mathrm{l}}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^\top}$, and $\mathbf{z} = \boldsymbol{\eta} - \mathbf{A}^{-1}\mathbf{u}$. These first and second derivatives of the $\mathrm{l}(\boldsymbol{\beta})$ with respect to the linear predictor $\boldsymbol{\eta}$ are given for instance in Hastie and Tibshirani (1990).

If the calculation of $\mathbf{A}$ is computationally very burdensome – as in the partial likelihood case –, one is able to use a reduced version of $\mathbf{A}$ that contains the same diagonal elements $a_{i,i}$, but is equal to 0 in all other entries (Hastie and Tibshirani, 1990).

An iterative minimization is then performed by:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left((\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{A} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\lambda}, \mathbf{D}, \boldsymbol{\beta}) \right),$$

where $\mathbf{z}$ and $\mathbf{A}$ are calculated using the version of $\boldsymbol{\beta}$ from the respective previous iteration. Before the first iteration, we set $\hat{\boldsymbol{\beta}} = (0, \ldots, 0)^\top$. This algorithm is iteratively performed using numerical optimization techniques for the nested minimization in each step. The algorithm is pursued until an updated version of $\hat{\boldsymbol{\beta}}$ – for the first time – does not change any more when compared to its previous version (with respect to a certain tolerance, e.g. $\frac{\sum_j |\hat{\beta}_{\text{old}} - \hat{\beta}_{\text{new}}|}{\sum |\hat{\beta}_{\text{new}}|} < 10^{-5}$).

2.7.1 Penalised iterative re-weighted least squares algorithm

An alternative estimation approach is the penalised iteratively re-weighted least squares (PIRLS) algorithm (Oelker and Tutz, 2013). The iteratively re-weighted least squares (IRLS) algorithm is very familiar since it is used in generalized linear models to find the maximum likelihood estimates. The PIRLS approach provides an estimation framework that is build-up on the well established IRLS basis, gives a flexible possibility to incorporate penalties, and yields very stable results – equal to those of suitably specified reference software implementations.

To use the PIRLS algorithm for multi-state models, it is mandatory to calculate the score vector and Fisher information as described in Sections 2.4 and 2.5. Furthermore we need a local quadratic approximation of the penalty matrix $\mathbf{P}_\lambda$ (Oelker and Tutz, 2013):

$$\mathbf{P}_{\boldsymbol{\lambda}} = \sum_{l=1}^{L} \lambda_l \mathbf{d}_l = \sum_{l=1}^{L} \lambda_l \frac{\partial \xi\left(||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}\right)}{\partial ||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}} \frac{\mathcal{D}_l\left(\mathbf{d}_l^\top \boldsymbol{\beta}\right)}{\mathbf{d}_l^\top \boldsymbol{\beta}} \mathbf{d}_l \mathbf{d}_l^\top.$$

Here, a penalty function of the form $\xi\left(||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}\right) = ||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}$ is used. Consequently the derivative neutralises, since $\frac{\partial \xi\left(||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}\right)}{\partial ||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l}} = 1$. Without exception, penalty terms are taken into account that penalise the L_1-norm, i.e. $||\mathbf{d}_l^\top \boldsymbol{\beta}||_{N_l} = ||\mathbf{d}_l^\top \boldsymbol{\beta}||_1 = |\mathbf{d}_l^\top \boldsymbol{\beta}|$. The derivative of a quadratic approximation (Oelker and Tutz, 2013) to this L_1-norm is $\mathcal{D}_l\left(\mathbf{d}_l^\top \boldsymbol{\beta}\right) = \frac{\mathbf{d}_l^\top \boldsymbol{\beta}}{\sqrt{\left(\mathbf{d}_l^\top \boldsymbol{\beta}\right)^2 + c}}$, where c is a very small constant (we use $c = 10^{-8}$ as in recent literature (Petry et al., 2011)). This leads to the following form of the quadratic

approximation of the penalty matrix $\mathbf{P}_\lambda$:

$$\mathbf{P}_{\boldsymbol{\lambda}} = \sum_{l=1}^{L} \lambda_l \mathbf{P}_l = \sum_{l=1}^{L} \lambda_l \cdot \frac{\mathbf{d}_l^\top \boldsymbol{\beta}}{\sqrt{\left(\mathbf{d}_l^\top \boldsymbol{\beta}\right)^2 + c}} \mathbf{d}_l \mathbf{d}_l^\top.$$

The construction of the penalty structure vector $\mathbf{d}_l$ is of the type $(0, \ldots, 0, 1, 0, \ldots, 0)^\top$ to penalise a single effect (Lasso term) and of the type $(0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^\top$ to penalise the difference between two effects (fusion term), with Lasso terms and fusion terms as described in Section 2.6. The PIRLS algorithm is then composed in the following way (using a step-length factor $\nu \in (0, 1]$ and iteration counter h) (Oelker and Tutz, 2013):

$$\hat{\boldsymbol{\beta}}_{(h+1)} = \hat{\boldsymbol{\beta}}_{(h)} - \nu \cdot \left(-\mathbf{F}\left(\hat{\boldsymbol{\beta}}_{(h)}\right) - \mathbf{P}_{\boldsymbol{\lambda}}\right)^{-1} \left(\mathbf{s}\left(\hat{\boldsymbol{\beta}}_{(h)}\right) - \mathbf{P}_{\boldsymbol{\lambda}} \hat{\boldsymbol{\beta}}_{(h)}\right).$$

Again, this algorithm is determined when the relative successive differences between the estimated coefficients is smaller than a fixed convergence criterion. We define the starting vector as $\boldsymbol{\beta} = (0, \ldots, 0)^\top$ (Oelker and Tutz, 2013). Using several different starting values is a good way to control for the algorithm running into local optima, a problem that has never occurred during the research process leading to this article.

2.8 Selection of penalty parameters

There exist several alternative criteria for tuning parameter or model selection. One of the most often used criteria to select optimal penalty parameters $\boldsymbol{\lambda}$ is the *Akaike Information Criterion (AIC)* defined by:

$$\text{AIC} = -2 \cdot \text{lpl} + 2 \cdot \text{df}.$$

For the calculation of the AIC we need a measure for the model complexity, i.e. the model degrees of freedom (df). In analogy to an article on the estimation of non-linear covariate effects in a Cox PH models using penalised splines (Gray, 1992), the model degrees of freedom are calculated by:

$$\text{df} = \text{trace}\left((\mathbf{F} + \mathbf{P})(\mathbf{F} + \mathbf{P})^{-1} \mathbf{F} (\mathbf{F} + \mathbf{P})^{-1}\right) = \text{trace}\left(\mathbf{F} (\mathbf{F} + \mathbf{P})^{-1}\right),$$

where $\mathbf{F}$ is the unpenalised model Fisher information, and $\mathbf{P}$ is the second derivative matrix of the penalty function.

An alternative definition is established by Tibshirani et al. (2005) in the fused Lasso

context with:

$$\mathrm{df} = p - \#\left\{\beta_j = 0\right\} - \#\left\{\beta_j - \beta_{j-1} = 0; \beta_j, \beta_{j-1} \neq 0\right\}.$$

In other words, this definition is to "count a sequence of one or more consecutive non-zero and equal β_j-values as one degree of freedom" (Tibshirani et al., 2005). Gertheiss and Tutz (2010) slightly adapted this definition to ordinal penalties by counting the number of unique non-zero coefficients induced by the respective current parameter estimates. However, we prefer to use the approach using the penalised Fisher information matrix, since it has the advantage that the degrees of freedom formulation is a continuous function of $\boldsymbol{\lambda}$, while the definitions in Tibshirani et al. (2005) and Gertheiss and Tutz (2010) introduce discontinuous step functions. Moreover, it corresponds very naturally to the PIRLS approach introduced in Section 2.7.1.

Using a grid search is an often applied strategy for the selection of a tuning parameter couple, as e.g. in Tibshirani et al. (2005), or Zou and Hastie (2005). This is pursued by taking all pair-wise combinations constructed by candidate vectors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ and then choosing the best combination (λ_1, λ_2) amongst all pair-wise combinations of tuning parameter values with respect to a selection criterion. Typically one would set-up a candidate grid using a finder lower part and then checking on a coarser grid for larger values, e.g. by using candidate vectors $\boldsymbol{\lambda}_l = (0, 0.01, 0.1, 1, 2, 3, \ldots, 10, 50, 100)$, $l = 1, 2$.

2.9 Software implementation

The established methods are implemented in the R (R Development Core Team, 2014) add-on package `penMSM` (Reulen, 2015). The central function within this package is the same-named function `penMSM`, performing the PIRLS algorithm established in Section 2.7.1. The function returns a list with the first element `B` being a matrix with the estimated effects (row dimension) throughout the iterations (column dimension) of the iterative estimation procedure. The second list entry called `aic` contains the value of the Akaike Information Criterion as defined in Section 2.8. If explicitly wished by setting the logical object `diagnostics` to `TRUE` in the function call, the return list additionally contains the values of the Fisher matrix, score vector, and approximated penalty matrix values throughout the iterations of the PIRLS estimation. Besides other mandatory objects it is key to forward the penalty structure matrices `PSM1` (Lasso part) and `PSM2` (fusion part) as well as the vectors with the penalty parameters for the respective penalty components (`lambda1` for the Lasso part and `lambda2` for the fusion part) to the `penMSM` function. Examples for these object definitions are given in Section 3.

3 Structured fusion Lasso estimation of a four state-type model for peritoneal dialysis program data

The potential of the structured fusion Lasso estimation for multi-state models is demonstrated during this section by analysing peritoneal dialysis study data for chronic renal disease patients as an application example, with clearly declaring that the goal of this section is not to perform an analysis that is completely adequate from a strictly medical point of view. Peritoneal dialysis is a class of dialysis methods that has important advantages in comparison to other dialysis methods, such as haemodialysis. Some exemplary advantages are a longer salvage of the remaining renal function, less frequent complications with respect to the dialysis access, and greater independence of patients from dialysis centres – eligible patients can independently carry out the treatment at home which results in a boost of life quality, for example patients are still able to travel. However, a major disadvantage of the peritoneal dialysis is a higher risk that the abdominal cavity is infected with pathogenic bacteria when getting into contact with the environment, with peritonitis – an inflammation of the peritoneum, the thin tissue that lines the inner wall of the abdomen and covers most of the abdominal organs – as the possible consequence. Therefore, patients must work very carefully and as sterile as possible when changing the dialysis solutions. Moreover, since peritoneal dialysis uses sugar-based solutions to perform dialysis, patients affected by diabetes will have to additionally adapt their diabetic medication.

In general, end stage renal disease is a worldwide increasing health problem, with a considerable amount of patients in need of a renal replacement treatment, or having some degree of renal dysfunction (Parmar, 2002). Moreover, the complications of "diabetes and hypertension are the two most common causes of end stage renal disease and are associated with a higher risk of death from cardiovascular disease" (Parmar, 2002). Hence, increasing the knowledge about the underlying mechanisms that lead to different complications and complication sequences after starting a peritoneal dialysis program is of high interest. The following state-types of a four state-type model are present in the peritoneal dialysis program study and will be considered during the analysis:

- Entrance to the peritoneal dialysis program (E): initial state-type for all patients.
- Occurrence of peritonitis (P): possible transient state-type.
- Death of the patient (D): absorbing state-type.
- Transfer to haemodialysis (H): absorbing state-type.
- Renal transplantation (R): absorbing state-type.

Figure 2 shows the state-chart that illustrates the possible seven transition-types between the four state-types. The four state-type process splits up into two nested competing risk

models, with both containing transition-types to the absorbing state-types D, H, and R. The numbers of observed transitions with entrance to the study (E) as initial state-type are 214 for the transition to peritonitis (EP), 47 transitions to death (ED), 56 transitions to transfer to haemodialysis (EH), 67 transitions to renal transplantation (ER), and 40 right-censored observations. For the transition-types with peritonitis (P) as initial state-type, 47 transitions to death (PD), 94 transitions to transfer to haemodialysis (PH), and 48 transitions to renal transplantation (PR) have been observed, with 26 right-censored observations. The transition-type specific median sojourn times given in months since the entry to the study are illustrated as vertical lines in Figure 5.

Four personal or clinical characteristics are taken into account for possibly influencing the transition-type specific hazard rate functions:

- Age of the patients (*Age*, measured in years),
- Sex of the patients (*Sex*, male/female, with reference category defined as female),
- *Diabetes* (no diabetes as reference category), and
- Previous renal transplantation therapy (*PRRT*, no PRRT as reference category).

The age of the patients has been taken into account with a potentially non-linear effect based on a fractional polynomial modelling as introduced for the baseline hazard rate specification in Section 2.5.

Appendix A gives some code snippets on how to set up design and penalty matrices for the piece-wise exponential modelling in the peritoneal dialysis application example.

3.1 Piece-wise exponential model results

Using $\Delta_{(j)} = 1$ is an adequate choice for the length of time sub-intervals since it offers a fine partition of the time axis with observed transition times in the interval $[1, 118]$ on the one hand, and a data-frame with a 44786 lines on the other hand, meaning that the estimation can still be performed in an acceptable time, i.e. a number of 100 iterations of the introduced approach take less than one minute using an ordinary notebook (Intel Core i7 2640M @ 2.80GHz CPU) and the software implementation described in Section 2.9.

We will compare the results of the introduced approach with the results of applying the benchmark software BayesX (Belitz et al., 2012), (Kneib and Hennerfeind, 2008) and estimating a model that is constructed very similarly to an un-penalised version of our approach. Here, transition-type specific effects and 95% confidence intervals are specified for the baseline hazard rate functions, effects of age, sex, diabetes and PRRT. Furthermore, the non-linear functional form of the baseline hazard rate functions and the effects of age are estimated using penalised B-Splines (Eilers and Marx, 1996). This

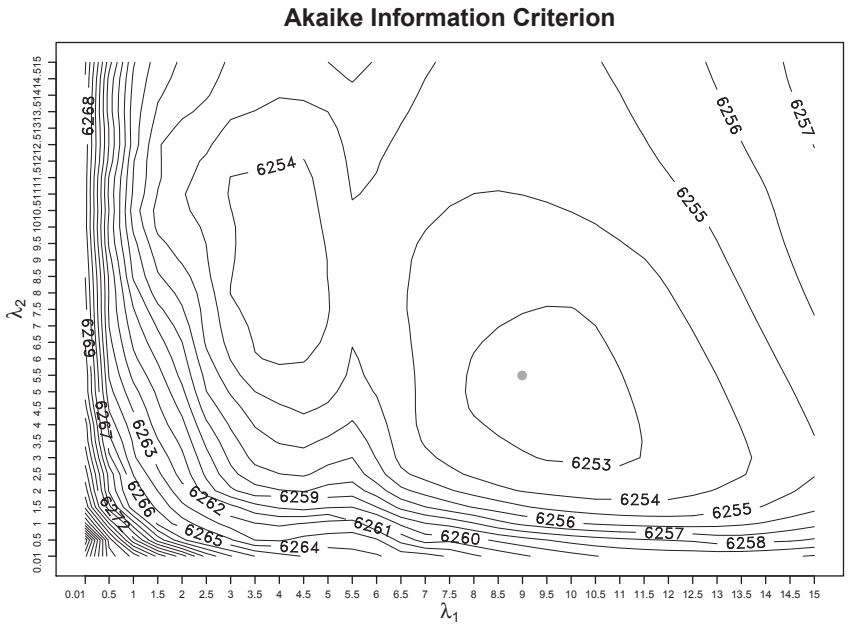

Figure 3: Contour plot of the Akaike Information Criterion values on the (31×31)-grid constructing combinations of the penalty parameters $\lambda_1, \lambda_2 = 0.01, 0.5, 1, 1.5, \ldots, 14.5, 15$. The black lines represent contour lines of the two dimensional AIC surface, the grey point indicates the $(\lambda_1, \lambda_2) = (9, 5.5)$ coordinate with the minimal AIC of 6252.54.

model is set up using additional time-varying effects and published before in Teixeira et al. (2015).

We estimate the model on a grid of (λ_1, λ_2) combinations constructed by all pairwise combinations for $\lambda_1, \lambda_2 = 0.01, 0.5, 1, 1.5, \ldots, 14.5, 15$. The resulting AIC values are illustrated in Figure 3. The minimal AIC with value 6252.54 is reached for the combination $(\lambda_1, \lambda_2) = (9, 5.5)$. Any value of (λ_1, λ_2) outside the illustrated region (results not shown) leads to a considerable increase of the AIC, and the minimum at $(9, 5.5)$ therefore stands for the global AIC minimum. The results for the minimum AIC model are illustrated in Figures 5 and 6, and described in the following.

Figure 4 shows the paths of the 112 coefficients across the penalty parameter ranges. The left illustration shows the regularisation of the coefficients towards the value 0, the right illustration shows the fusion of coefficients. Both of these regularisation features become stronger with increasing penalty parameters. Note here that, as in any fusion Lasso framework, the influences of the penalty parameters λ_1 and λ_2 on the model coefficients are not independent. For example by changing the Lasso penalty term, in many cases different coefficient levels are introduced. Since these changes will be different

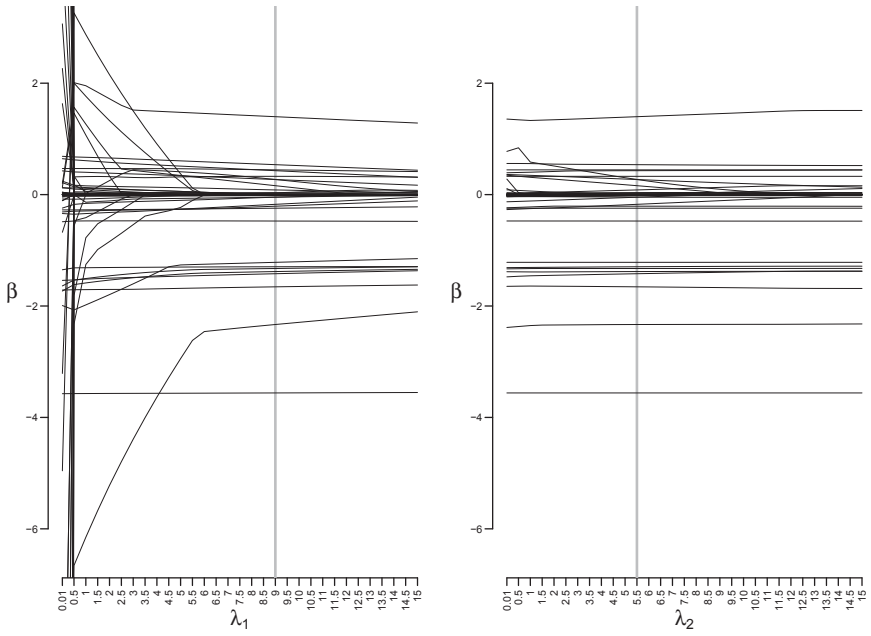

Figure 4: Paths of the penalised estimation regression coefficients in the peritoneal dialysis program data application. For fixed values of the respective other penalty parameter at the value for the minimum AIC model (9 for λ_1 and 5.5 for λ_2, see Figure 3), the penalty parameters λ_1 and λ_2 increase through the values $0.01, 0.5, 1, 1.5, \ldots, 14.5, 15$.

for different coefficients, the influence of the fusion penalty changes even if the fusion penalty parameter λ_2 is held constant.

As illustrated in Figure 5, we get constant baseline hazard rate function estimates for transition-types ER, PD, and PH, but observe no fusion of baseline hazard rate function estimates. However, we receive a cross-transition-type age effect, i.e. a fusion of transition-type specific age effects, for the transition-type combination ER with PR: older patients have to wait longer for a renal transplantation with and without prior occurrence of a peritonitis. As a result of the structured fusion Lasso penalised estimation we see that the effects for covariate sex are all equal to or smaller than 0. This yields a clearer image of the sex effect in comparison to the un-penalised estimation, where the effect was smaller than 0 for four transition-types, and larger than 0 for three other transition-types (denoted by the middle of the 95% confidence intervals). To be more precise, all of the effects larger than 0 in the un-penalised estimation are shrunken to 0, whereas all the effects smaller than 0 stay below and different to the value 0. Cross-transition-type effects are obtained for the transitions to death (for the effect of diabetes and PRRT) and for the transfer to renal transplantation (again for the effects of diabetes

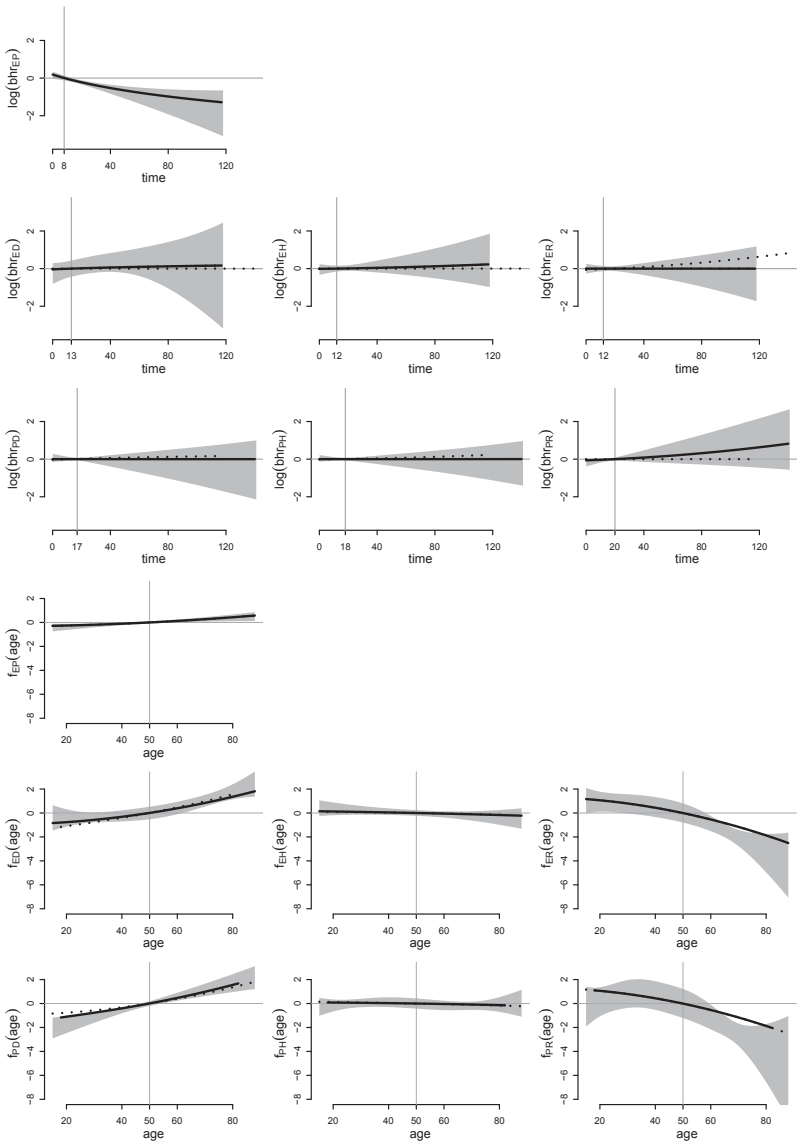

Figure 5: Estimated transition-type specific log baseline hazard rate functions (seven plots at the top) and effects of age (seven plots at the bottom) for the piece-wise exponential modelling in the peritoneal dialysis program data application with the minimum AIC penalty parameter combination $(\lambda_1, \lambda_2) = (9, 5.5)$. Solid black lines represent the estimated effects by the structured fusion Lasso penalised multi-state model using the fractional-polynomial set-up as described in Section 2.5. Dotted lines represent the estimate of the respective fusion penalised transition-type partner. Grey areas illustrate point-wise 95% confidence intervals of the benchmark model using the software BayesX, where smooth effect estimates are obtained using penalised B-Splines. The estimated effects and confidence intervals are centred around the median transition-type specific transition times for the log baseline hazard rate function estimates and around the values at age 50 for effects of age.

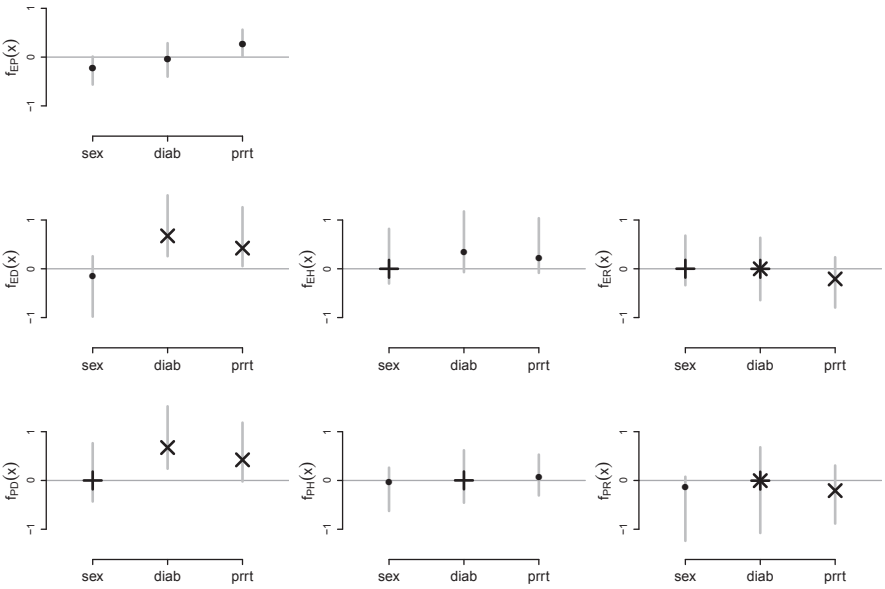

Figure 6: Estimated transition-type specific effects of sex, diabetes, and PRRT for the piece-wise exponential modelling in the peritoneal dialysis program data application with the minimum AIC penalty parameter combination $(\lambda_1, \lambda_2) = (9, 5.5)$. Black plotting symbols represent the resulting effect by the structured fusion Lasso penalised multi-state model, grey vertical lines illustrate 95% confidence intervals of the benchmark model using the software BayesX. Black bullet points ($\bullet$) represent non-fused effects different to 0, plus signs ($+$) represent effects equal to 0, crosses ($\times$) represent fused effects. Combinations of a plus sign and a cross appear as stars ($*$) and signify fused effects equal to 0.

and PRRT). Furthermore, the diabetes effect for the transition-types ER, PH, and PH are set to 0. Note here that any small coefficient differences preventing setting an effect to 0 or fusing two effects appear also for penalty parameter combinations $(9, 6)$, $(9.5, 5.5)$, and $(9.5, 6)$. Hence they are not attributable to a too coarse grid of penalty parameter combination candidates.

4 Discussion

With the structured fusion Lasso penalised multi-state modelling approach introduced by this article, we establish a data-driven way to perform a structured analysis of multi-state models with potential cross-transition-type effect and variable selection. We presented the approach for partial likelihood and piece-wise constant baseline hazard rate models and proposed an algorithm that is applicable to a broad class of multi-state models. This is achieved by the use of a penalised iterative reweighed least squares algorithm that is close to estimation algorithms known from generalized linear models. We are able to use this algorithm by the use of a local quadratic approximation of the L_1-norm. The best combination of Lasso and fusion penalty parameters is selected using a grid search for the minimal Akaike information criterion value.

The application to peritoneal dialysis data showed that we are able to work out interesting insights into the structural relationships between transition-types, a problem that has been addressed several times in multi-state modelling literature, but has never been entrusted with a suitable data-driven estimation concept.

Acknowledgements

We like to thank Laetitia Teixeira, Anabela Rodrigues, and Denisa Mendonça from the University of Porto, Portugal, and Carmen Cadarso-Suárez from the University of Santiago de Compostela, Spain, for kindly providing us the data used for the peritoneal dialysis program data application. The authors are supported by the German Research Foundation (DFG) research training group 1644 *Scaling Problems in Statistics*.

Appendix A

A design matrix may be build up in transition-type specific blocks of 16 transition-type specific covariates, e.g. by the following R (R Development Core Team, 2014) command using variables stored in data-set `d`:

```
X.EP <- as.matrix(d[, c("trans.EP", "bhr.1.EP", "bhr.2.EP", "age.3.EP",
            "bhr.4.EP", "bhr.5.EP", "bhr.6.EP", "age.1.EP", "age.2.EP",
            "age.3.EP", "age.4.EP", "age.5.EP", "age.6.EP", "sex.EP",
            "diab.EP", "prrt.EP")])
X.ED <- as.matrix(d[, c("trans.ED", "bhr.1.ED", ...)])
...
X <- cbind(1, X.EP[, -1], X.ED, X.EH, X.ER, X.PD, X.PH, X.PR)
```

Here, `X` used for piece-wise exponential modelling introduces a global, or in other words across all transition-types, constant baseline hazard rate in the first column and defines the transition-type EP as the reference transition-type with respect to the constant baseline hazard rate function – which is technically implemented by `cbind(1, X.EP[, -1], ...)`. The resulting design matrix `X` finally contains 112 columns. Note that the specification of a reference transition-type, as well as taking into account any baseline hazard rate function columns, is unneeded for partial likelihood modelling.

The build-up of the penalty structure matrix is conveniently separated into the Lasso part which is caught by matrix `PSM1`, and the fusion part which is caught by matrix `PSM2`. The Lasso part penalty structure matrix `PSM1` is composed by an identity matrix with the dimension matching the number of columns of `X`:

```
PSM1 <- diag(ncol(X))
```

Constant baseline hazard rate components equal to 0 seem to be a too restrictive null model for an unbalanced number of transition-type observations and we therefore leave the constant baseline hazard rates unpenalised:

```
PSM1[1, 1] <- PSM1[17, 17] <- PSM1[33, 33] <- PSM1[49, 49] <-
              PSM1[65, 65] <- PSM1[81, 81] <- PSM1[97, 97] <- 0
```

The penalty structure matrix `PSM2` for the fusion part consists of an equal number of columns as the design matrix `X` and 45 rows, since we want to penalise 15 covariate effects (the constant baseline hazard rates stay again unpenalised) for each of three transition-type pairs sharing an equal exit state-type (`ED` and `PD`; `EH` and `PH`; `ER`, `PR`):

```
PSM2 <- matrix(ncol = ncol(X), nrow = 45, 0)
colnames(PSM2) <- colnames(X)
PSM2[ 1, which(colnames(PSM2) %in% c("bhr.1.ED", "bhr.1.PD"))] <- c(-1, 1)
```

```
...
PSM2[ 6, which(colnames(PSM2) %in% c("bhr.6.ED", "bhr.6.PD"))] <- c(-1, 1)
PSM2[ 7, which(colnames(PSM2) %in% c("age.1.ED", "age.1.PD"))] <- c(-1, 1)
...
PSM2[12, which(colnames(PSM2) %in% c("age.6.ED", "age.6.PD"))] <- c(-1, 1)
PSM2[13, which(colnames(PSM2) %in% c("diab.ED", "age.PD"))] <- c(-1, 1)
PSM2[14, which(colnames(PSM2) %in% c("sex.ED", "age.PD"))] <- c(-1, 1)
PSM2[15, which(colnames(PSM2) %in% c("prrt.ED", "age.PD"))] <- c(-1, 1)
...
PSM2[45, which(colnames(PSM2) %in% c("prrt.ER", "prrt.PR"))] <- c(-1, 1)
```

References

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2012). BayesX: Software for Bayesian Inference in Structured Additive Regression Models. *Software published online on www.BayesX.org*. Version 2.1.

Carstensen, B. and Center, S. D. (2005). Demography and epidemiology: Practical use of the lexis diagram in the computer age. In *Annual meeting of Finnish Statistical Society*, volume 23, page 24.

Carstensen, B. and Plummer, M. (2011). Using Lexis Objects for Multi-State Models in R. *Journal of Statistical Software*, 38(6):1–18.

Carstensen, B., Plummer, M., Laara, E., and Hills, M. (2014). `Epi`: A Package for Statistical Analysis in Epidemiology. *R add-on package published online on the Comprehensive R Archive Network*. R package version 1.1.67.

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival Analysis Part IV: Further concepts and methods in survival analysis. *British Journal of Cancer*, 89:781–786.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, 4(4):2150–2180.

Goeman, J. J. (2010). L_1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal*, 52(1):70–84.

Goeman, J. J. (2012). `penalized`: L_1 (lasso and fused lasso) and L_2 (ridge) penalized estimation in GLMs and in the Cox model. *R add-on package published online on the Comprehensive R Archive Network*. R package version 0.9-42.

Gray, R. J. (1992). Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.

Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004.

Johansen, S. (1983). An Extension of Cox's Regression Model. *International Statistical Review / Revue Internationale de Statistique*, 51(2):pp. 165–174.

Kneib, T. and Hennerfeind, A. (2008). Bayesian semi parametric multi-state models. *Statistical Modelling*, 8:169–198.

Lambert, P. C., Smith, L. K., Jones, D. R., and Botha, J. L. (2005). Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine*, 24(24):3871–3885.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer journal*, 7(4):308–313.

Oelker, M.-R. and Tutz, G. (2013). A General Family of Penalties for Combining Differing Types of Penalties in Generalized Structured Models. *LMU Munich, Department of Statistics: Technical Reports, Nr. 139*.

Parmar, M. S. (2002). Chronic renal disease. *BMJ*, 325(7355):85–90.

Petry, S., Flexeder, C., and Tutz, G. (2011). Pairwise Fused Lasso. *LMU Munich, Department of Statistics: Technical Reports, Nr. 102*.

Puig, A. T., Wiesel, A., Fleury, G., and Hero, A. O. (2011). Multidimensional Shrinkage-Thresholding Operator and Group LASSO Penalties. *Signal Processing Letters, IEEE*, 18(6):363–366.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

R Development Core Team (2014). R: A Language and Environment for Statistical Computing. *Software published online on the Comprehensive R Archive Network*.

Reulen, H. (2015). `penMSM`: Estimating Regularized Multi-state Models Using L_1 Penalties. *R add-on package published online on the Comprehensive R Archive Network*. R package version 0.99.

Rodríguez-Girondo, M., Kneib, T., Cadarso-Suárez, C., and Abu-Assi, E. (2013). Model building in nonproportional hazard regression. *Statistics in Medicine*, 32(30):5301–5314.

Teixeira, L., Cadarso-Suárez, C., Rodrigues, A., and Mendonça, D. (2015). Assessing the discrimination ability of semiparametric multi-state models in the presence of competing risks. analysis of a peritoneal dialysis program. *in remission*.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. et al. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.