Regularization in Discrete Survival Models

Stephanie Möst

München 2014

Regularization in Discrete Survival Models

Stephanie Möst

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Stephanie Möst (geb. Rubenbauer)
aus München

München, den 29.04.2014

Erster Berichterstatter: Prof. Dr. Gerhard Tutz
Zweiter Berichterstatter: Prof. Dr. Harald Binder
Tag der Disputation: 09.07.2014

Zusammenfassung

Die Überlebenszeitanalyse beschreibt verschiedene statistische Methoden, die die Zeit bis zu einem bestimmten Ereignis analysieren. Bei der Analyse dieser Beobachtungsdauer wird die Zeit bis zum Eintritt eines Ereignisses oftmals als kontinuierlich angenommen. Allerdings ist in vielen Studien lediglich bekannt, dass das Ereignis zwischen zwei aufeinander folgenden Beobachtungszeitpunkten aufgetreten ist oder die Beobachtungsdauer tatsächlich diskret ist. Daher kann eine beachtliche Anzahl von Bindungen in den Daten vorkommen, was zu Problemen bei der Benutzung von Likelihood-Methoden führt. Um auftretende Bindungen zu berücksichtigen, werden in dieser Arbeit Überlebenszeitmodelle für diskrete Zeit betrachtet. Nach einer Umstrukturierung der üblichen Form von Time-To-Event-Daten, können diskrete Überlebenszeitmodelle als generalisierte lineare Modelle aufgefasst werden. Diese Umstrukturierung der Daten führt dazu, dass einige Beobachtungen nur sehr selten vorkommen, insbesondere wenn viele Zeitpunkte vorhanden sind. Die vorliegende Datensituation wird noch diffiziler, wenn zeit-variierende Koeffizienten in das Modell aufgenommen werden. Es kommt nicht selten vor, dass Maximum-Likelihood-Methoden (ML) in diesen Situationen nicht angewendet werden können, da die ML-Schätzungen entartet sind oder gar nicht existieren. Um stabile und zuverlässige Schätzungen zu erhalten, eignet sich die Anwendung von Regularisierungstechniken. In diesem Zusammenhang konzentriert sich die Arbeit auf Penalisierungsmethoden.
Da Time-To-Event-Daten über die Zeit hinweg gemessen werden, liegt es nahe, zeitvariierende Koeffizienten in das Modell einzubinden. Nachfolgend werden speziell auf diese Situation abgestimmte Penalisierungsterme vorgestellt. Diese Penalisierungsterme ermöglichen beispielsweise glatte zeit-variierende Koeffizienten oder eine Variablenselektion. Letzteres bedeutet, dass Kovariablen komplett aus dem Modell entfernt werden können. Die Stärke der Penalisierung wird durch Tuningparameter gesteuert, deren Wahl im Folgenden systematisch untersucht wird.
Die zugrunde liegenden Daten aus Überlebenszeitmodellen resultieren aus wiederholten Messungen, was zu unbeobachteter Heterogenität in den Daten führt. Um unbeobachtete Heterogenität zu berücksichtigen, werden Frailty-Modelle, beziehungsweise Modelle mit zufälligen Effekten, betrachtet. Dabei werden die linearen Prädiktoren der eben erwähnten Penalisierungsmethoden für diskrete Zeit um zufällige Effekte erweitert.
In zahlreichen Anwendungen in der Überlebenszeitanalyse, ist die Analyse von mehr als zwei möglichen Ereignissen von Interesse. Dabei kann für jedes Individuum eines von k ($k \geq 2$) möglichen Ereignissen (Competing Risks) auftreten. Diese Modelklasse wird für diskrete Beobachtungsdauer vorgestellt. Diskrete Competing Risk Modelle können in die Theorie von multinomialen Regressionsmodellen eingebettet werden. Um eine Variablenselektion durchführen zu können und um die Anzahl der dabei auftretenden Parametern zu bewältigen, empfiehlt es sich geeignete Penalisierungstechniken einzusetzen, welche für Competing Risk Modelle vorgestellt werden.

Summary

Survival analysis describes a collection of statistical procedures that explore the time until an event occurs. In the framework of analyzing failure time or time to event data, time is often considered to be continuously observed. However, in many studies it is only known that the event occurs between a pair of consecutive follow-ups or time is truly discrete. Hence, there are many ties in the data. This causes problems when likelihood methods for continuous-time models are used. Hence, to account for the issue of tied observations, in this thesis discrete-time survival models are considered. After a restructuring of the typical form of time-to-event data, survival models for discrete duration time can be understood as generalized linear models. This complex data restructuring process results in a large number of observations that are only rarely observed, especially when there are many time periods. This data situation becomes even more difficult when time-varying covariate effects are incorporated. Ordinary Maximum-Likelihood (ML) methods cannot be applied as the ML-estimates are deteriorated or even do not exist. To obtain stable and reliable estimates, the application of regularization methods is necessary. Thereby, this thesis focuses on penalization methods.

Survival data of an individual are usually collected over time. Hence, it is of great interest to incorporate time-varying covariate effects in the model. To regularize discrete survival regression models, different penalty terms that cope with this special case are proposed. For example, these penalty terms allow for smooth time-varying coefficients or provide a variable selection. The latter means that covariates can be completely removed from the model. The strength of penalization is steered by tuning parameters. In this thesis, it is systematically investigated, how these tuning parameters have to be chosen.

The underlying data in survival models deal with repeated measurements leading to certain heterogeneity in the data. To control for unobserved heterogeneity, frailty models are considered and a corresponding penalty term is introduced. That is, the described penalization methods are combined with the incorporation of frailty effects.

In many applications concerning survival analysis, the investigation of more than one terminating event is of interest. Hence, for each object one of k ($k \geq 2$) causes may occur, called competing risks. This model class is introduced in the context of discrete survival times. Discrete competing risks models can be embedded into the framework of multinomial regression models. To provide variable selection and to account for the large amount of parameters that arise with the use of this model type, a penalization technique for discrete-time competing risks models is introduced.

Vorwort und Danksagung

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftliche Mitarbeiterin am Institut für Statistik der Ludwig-Maximilians-Universität München. In erster Linie möchte ich mich bei meinem Doktorvater Herrn Prof. Dr. Gerhard Tutz bedanken, der mir diese Stelle angeboten hat und mich über all die Jahre hinweg hervorragend betreute. Mein Dank gilt weiterhin Herrn Prof. Dr. Harald Binder, der freundlicherweise die Aufgabe des Zweitgutachters übernommen hat.
Ebenso möchte ich meinen aktuellen sowie früheren Kollegen am Institut für Statistik danken, die mich mit Problemlösungen, Hilfestellungen und Ratschlägen unterstützten. Ein besonderer Dank geht dabei an Margret und Lisa, mit denen ich mich sowohl fachlich gut austauschen konnte, als auch viele Pausen genießen konnte. Vor allem aber auch möchte ich meinen Kollegen am Seminar für Stochastik danken und zwar insbersondere für das Korrektur lesen dieser Arbeit und die offenen Türen die ich immer vorgefunden habe. Hervorheben möchte ich dabei die besonders angenehme Arbeitsatmosphäre und die gute Unterhaltung abseits des Arbeitsalltags.
Abschließend danke ich meinem Mann Matthias, der mich mit viel Geduld während dem gesamten Entstehen dieser Arbeit unterstützt hat. Schließlich gilt mein aufrichtiger Dank meiner ganzen Familie, die mich stets unterstützen und immer hinter mir stehen. Auch all meinen Freunden möchte ich danken, für all die offenen Ohren und die wohltuende Ablenkung.

München, im April 2014 *Stephanie Möst*

Contents

1. Introduction

Survival Analysis

Survival analysis describes a collection of statistical procedures that explore the time until an event occurs. Thereby, *time* means any kind of time unit from the beginning of a follow-up of an object until the occurrence of an event. An *event* defines any designated experience of interest that may happen to an object, for example death, disease incidence or recovery. Thus, a number of mutually exclusive states can be considered at each point in time. The patterns for each object are described by the time spent within each state, and the dates of each transition made. Hence, each change of state is equivalent to an event.

In a *single spell* analysis, the survival time of an object is observed from the beginning of follow-up up to one absorbing event. When an object can experience one of several different types of absorbing events, the statistical problem is characterized as a *competing risk* problem. Competing risk theory has an intriguing history going back to a memoir read in 1760 by Daniel Bernoulli and published in 1765 (Klein and Moeschberger, 2003). It is about the merits of smallpox inoculation and asks "What would be the effect on mortality if the occurrence of one or more causes of death were changed?".
Furthermore, *recurrent events* that determine a multiple spell analysis have to be distinguished. They consider a single event that may occur more than once over the follow-up time for a given object. Thus, in this case the event is not absorbing. Aalen (1988) provided theoretical and practical motivation for such models. Of course, a mixture of recurrent events and competing risks, predominantly named multistate models, are possible as well. An example for a multistate model is the stages of sleep with the recurrent events rapid eye movement (REM), non-rapid eye movement (NREM) and awake (Loomis et al., 1937).

Modeling survival data, also denoted as time-to-event data, has been well investigated and is used in many different areas of research like medicine, biology, social science, economics or demography. In biological organisms, time often refers to the survival of an object, whereas in mechanical systems, time usually depicts a failure. Generally, time-to-event data also are denoted as event-history-data (e.g. social science) or survival data (e.g. biology or medicine) according to the area of application. More general, time-to-event data are named duration data, transition data or failure time data as well.

Censoring and Truncation

A common feature of survival data is the incorporation of either censored or truncated data. Censored data arise when an object's survival time has not been fully observed. Thus, the survival time is not known exactly. For instance, this is the case if an object does not experience the event before the end of the study or the object is dropped out during the study period. Possible common censoring types are *right censoring*, *left censoring* and *interval censoring*.

Right Censoring

Right censoring is given if the beginning of a follow-up time is known, but the time when the event arises is not observed. At the time of observation, the relevant event (transition out of the current state) had not yet occurred. Thus, it is only known that the survival time is larger than the observed time. This can be the case if an object dies from another cause, independently of the cause of interest, the study ends while the subject survives or the subject is lost to the study, by dropping out or moving to a different area. If it is not stated otherwise right censoring is assumed in this thesis.

Left Censoring

An object is said to be left-censored if it is known that the event of interest occurred at some time before the observation date but it is not known exactly when. Consequently, the observed survival time is larger than the true survival time. It happens, for example, if the date of a medical exam, revealing a disease, is specified, but it is not observed when the patient has been infected.

Interval Censoring

An object is defined as interval censored if the event occurs in a time interval but it is not known exactly when in the interval. It can appear, if an object is regularly checked and one time the event is experienced. So, the only information is that the incidence appears between two checks.

By contrast, for truncated survival time data, survival times are systematically excluded from a sample, and the sample selection effect depends on the survival time itself (Jenkins, 2005). An object whose event time is not in a certain observational interval is not observed. This is in contrast to censoring, where at least partial information on each object exists. It is to distinguish between *left* and *right truncation*.

Right truncation

Only objects with event times less than a specific threshold are included in the sample. So, relatively long survival times are systematically excluded from the study. A general example for right truncation, regardless of time, are stars that are too far away and not visible and are thus not incorporated in the estimation of the distribution of stars (Klein and Moeschberger, 2003). An example of a right truncated sample with respect to survival time is a mortality study based on death records.

Left truncation

Only those objects can be observed, whose event time exceeds some truncation threshold. Consequently, especially short survival times are systematically excluded from the observation sample. A common example of left truncation is the issue of estimating the distribution of the diameters of microscopic particles. On the basis of the resolution of the microscope, only particles big enough to be seen can be observed, whereas smaller particles do not come to the attention of the investigator (Klein and Moeschberger, 2003). Referring to survival time, the truncation event may be the occurrence of some intermediate event such as graft-versus-host disease after a bone marrow transplantation.

The issues of censoring and truncation are a main challenge in terms of analyzing survival data. Using counting process methodology has allowed for substantial advances in the statistical theory to account for censoring and truncation in survival experiments. Aalen (1975) first developed this approach by combining elements of stochastic integration, continuous time martingale theory and counting process theory. The resulting methodology easily concedes to the development of inference techniques for survival quantities based on censored and truncated data. Thereby, a relatively simple development of large sample properties of such statistics is possible (Klein and Moeschberger, 2003). A detailed description of this special field can be found in the books of Andersen et al. (1993) and Fleming and Harrington (2005).

Why does survival analysis need special statistical methods?

The methods of survival analysis are very specific and differ from commonly used methods of regression models. There are several reasons for these distinctions shortly mentioned in the following.

The goal of survival analysis is the modeling of survival time as response variable. All survival times are non-negative and distributions of survival times are typically skewed (Jenkins, 2005). These features of survival times have to be considered in modeling. A new modeling scheme is necessary since the survival times of some objects are usually not completely observed. Hence, the methods of survival analysis has to include the partial information provided by censored observations.

An ordinary regression model has only a single dependent variable. In survival analysis,

however, an object is observed for a period of time and multiple values of a time-varying covariate exist belonging to only one value of the response. Should one value of the time-varying covariate, that is the most representative, be chosen or how can time-varying covariates be handled?
In the context of survival analysis, the hazard rate is of special interest. It depicts the stochastic behavior of the survival time and is used as an alternative representation of it. It is a common way to choose a parametrization of the hazard rate to model time-to-event data, leading to special statistical methods.
The need of distinctive statistical models for survival analysis is described in more detail in Jenkins (2005). As survival analysis is widely used in several fields of application a considerable number of books are available like Klein and Moeschberger (2003), Kleinbaum and Klein (2013), Kalbfleisch and Prentice (2002), Therneau and Grambsch (2000) and Hosmer et al. (2011). Much more details on censoring and truncation can especially be found in Klein and Moeschberger (2003).

Regularization ideas

In general, modeling discrete survival data is done by using parametric regression models. Estimation procedures for discrete survival models require rearrangements of the data. This often leads to designs, especially when incorporating time-varying coefficients, where computational problems arise and estimates may become unstable. Therefore, the use of regularization techniques is recommended. In addition, depending on the technique, regularization techniques coincide with a reduction of the predictor space. This is a convenient effect when many predictors are available.
A typical regularization technique is *penalization*. Penalization means to add a penalty term to the log-likelihood yielding shrinkage of the estimates towards zero. Depending on the penalty, it is even possible to set particular estimates exactly to zero. One of the oldest penalization methods is the *ridge* method (Hoerl and Kennard, 1970), that uses a L_2-type penalization of the regression coefficients. However, no variable selection can be performed by using this penalty term. An alternative penalty term that has become very popular is the *lasso* penalty term, using a L_1-type penalty on the regression coefficients. In this case, variable selection can be carried out. As the lasso merely selects individual predictors, the penalty is unsatisfactory in the case of grouped data, for example, with categorical predictors. The *group lasso* proposed by Yuan and Lin (2006), can overcome these problems. To get consistent estimates of the parameters, Zou (2006) extended the lasso to the *adaptive lasso* by including weights on the penalized coefficients. Several further improvements for the lasso method have been designed in the last decades, for example *fused lasso* (Tibshirani et al., 2005), *SCAD* (Fan and Li, 2001), *elastic net* (Zou and Hastie, 2005), *Dantzig selector* (Candes and Tao, 2007) and *DASSO* (James et al., 2009).

Another regularization approach, developed in the machine learning community is based on *boosting methods*. Boosting is a powerful learning idea that was originally designed for

classification problems. However, it can also be applied to regression. The general idea of combining several weak learners to a final strong learner was introduced by Schapire (1990). A detailed overview on boosting algorithms can be found in Bühlmann and Hothorn (2007). The focus of this thesis is on penalization and to boosting techniques will not be gone into any further.

Guideline through the thesis

The main part of this thesis consists of four basic chapters, which show possibilities of penalization for discrete survival models. Chapters 3 and 4 deal with the penalization of discrete-time survival models with single spells and the corresponding choice of the tuning parameters. In Chapter 5, an additional frailty is incorporated in the model and finally Chapter 6 deals with penalized competing risks models for discrete duration time. Some background on survival analysis is given in Chapter 2. In order to keep individual chapters self-contained, some passages repeat themselves with only small modifications and adjustments due to the different frameworks.

Chapter 2: Discrete Survival Analysis

This chapter describes the basic aspects of univariate survival data and contains notation and important statistical methods with regard to discrete duration time models. Based on single spells, basic concepts of survival analysis like the hazard rate and the survival function are treated. Two commonly used regression models for discrete time survival analysis, are presented. Moreover, the incorporation of time-varying covariates and time-varying coefficients in discrete survival regression models is depicted.

Chapter 3: Lasso-Type Penalties

Using time-varying coefficients in a linear predictor of a discrete survival model results in a large number of parameters that have to be estimated. This might lead to unstable estimates or computational problems. To overcome these issues, in this chapter penalization methods are incorporated in discrete-time survival models. To get access to uncertainty measures, standard errors and confidence intervals for the parameters resulting from the penalized estimation are provided. Several simulation studies judge the performance of the presented method. The proposed method is applied to the Munich founder study and to a fertility study.

Chapter 4: Choice of Tuning Parameter

Building on Chapter 3, in this chapter, the choice of tuning parameters is systematically investigated. To this end, the conventional loss function, that is the predictive deviance, is substituted by different alternative loss functions and possible associated model improvements are investigated. Several measures of prediction accuracy are presented, whereby well-known measures used for continuous survival outcomes are adopted to discrete failure time analysis. The performance of the shown alternative loss functions is investigated by means of a simulation study.

Chapter 5: Penalization in Survival Models with Frailties

In general, survival data are based on repeated measurements leading to certain heterogeneity in the data. This chapter deals with the incorporation of random effects or frailties in survival models for discrete duration time. These frailties control for existing unobserved heterogeneity. The methodology and the estimation of discrete survival models with frailties are described and a penalty term that allows for variable selection is incorporated in the model. The performance of the proposed method is judged by means of a simulation study. To compare the results, the proposed method is applied to the same real data examples as in Chapter 3.

Chapter 6: Penalization in Competing Risks Models

In many applications concerning survival analysis, the investigation of more than one terminating event is of interest. Hence, for each object one of k ($k \geq 2$) causes may occur, referred to as competing risks. This model class is introduced in the context of discrete survival times. Discrete competing risks models can be embedded into the framework of multinomial regression models. Due to the large amount of parameters that arise with the use of this model type, a penalization technique for discrete-time competing risks models is introduced. The proposed method is applied to career paths of Congressmen in the United States and to characteristics of unemployment in Germany.

Software

All computations were carried out using the statistical program R (R Development Core Team, 2013) and related packages. The corresponding packages are indicated in the respective chapters and sections. The basis of the implementation of the approaches from Chapters 3 and 4 is the R add-on package `gvcm.cat` (Oelker and R Development Core Team, 2013). The functions of this package were adapted to survival models for discrete duration time. Moreover, for the approach of Chapter 5 the program code is enhanced to enable the computation of frailties for discrete survival models. That means, that the algorithms `pendsm` and `fpendsm` were completely implemented by means of the statistical program R. The used functions are available on request. For fitting penalized competing risks models, the R package `MLSP` was extended. The original package can be downloaded from `http://www.statistik.lmu.de/~poessnecker/software.html` and will be available as a proper R add-on package via CRAN (see `http://cran.r-project.org`) in the near future.

2. Discrete Survival Analysis

This chapter describes basic aspects of univariate survival data, contains notation and important statistical methods with regard to discrete duration time models. In Section 2.1, continuous time is generally opposed to discrete time in the context of survival analysis. Based on single spells, basic concepts of discrete-time survival analysis like the hazard rate and the survival function are treated in Section 2.2. Therein, the grouped proportional hazard model and the continuation ratio logit model, two commonly used regression models for discrete-time survival analysis, are presented as well. Moreover, Section 2.3 and Section 2.4 deal respectively with the incorporation of time-varying covariates and time-varying coefficients in discrete survival regression models.

2.1. Continuous time versus discrete time in Survival Analysis

The basic interest in survival analysis is the survival time. So far, it is implicitly assumed that the event of interest can arise at any particular time. In general, time is a continuum and the length of a spell can be measured by a non-negative real number. However, in fact, duration time is often grouped or banded into discrete intervals. For example, in some research fields time is measured in days or years. That is, time as a continuum is divided into an infinite sequence of continuous time periods. So, for survival times a set of positive integers can be used. Hence, the resulting transition process is more discrete than continuous. These kind of data are named *grouped data* and are commonly used in the context of survival data. The underlying process arises in continuous time, however, it is observed in a discrete manner. Biostatisticians typically refer to this situation as one of *interval censoring* (Jenkins, 2005). It is evident that in grouped survival data, some objects have an identical survival time, also named ties. The occurrence of ties may be an indicator of interval censoring. In statistical models for continuous time, it is usually assumed that no ties are present. When ties are present regarding continuous time models, it has to be clarified whether the ties are original or do they occur due to grouping at the observation or reporting stage.

On the other hand survival time can be *truly discrete*. In this case, Jenkins (2005) denoted the underlying transition process as intrinsically discrete. That means, that the measurements of survival time represent natural numbers. Consider, for example, the number of attempts at a puzzle before it is solved. A fraction of the number of attempts does not

make sense. Another similar example is the modeling of fertility, especially the time from puberty to first childbirth. It is more instinctual to model time in terms of numbers of menstrual cycles rather than the time up to pregnancy because the cycle length varies amongst women, and a woman ovulates only once per menstrual cycle (Kleinbaum and Klein, 2013).

In general, there is no difference in applying statistical methods when these two kinds of discrete data form the basis of modeling. Therefore, they are only referred to as discrete time models. The general distinction between discrete time data and continuous time data is the more important one.

2.2. Basic Concepts

In the following, survival data with regard to discrete time are assumed. That means, the time scale is truly discrete or continuous survival times are observed only in intervals. Let T represent a non-negative discrete random variable assuming values from $\{1, ..., q\}$.

In the case of intrinsically discrete time, the probabilities

$$f(t) = P(T = t),$$

where $t \in \{1, ..., q\}$ is a set of positive integers, are obtained. Thereby, t defines, for example, the numbers of menstrual cycles. Also definitions without referring to time are possible. That means, t might also denote the number of attempts to solve a puzzle, for example. For expositional purposes, t is generally considered in the context of time.

In contrast, in the case of grouped survival times, continuous time is divided into $q + 1$ intervals

$$[a_0, a_1), [a_1, a_2), ..., [a_{q-1}, a_q), [a_q, \infty), \tag{2.1}$$

where usually $a_0 = 0$ is assumed and a_q denotes the end of the observation period. Instead of observing continuous time the discrete time T is observed. Analogous to truly discrete survival time, in the grouped case, the random variable T takes values from a set of positive integers $\{1, ..., q\}$, where $T = t$ denotes an event within the interval $[a_{t-1}, a_t)$. Thereby, t denotes, for example, the month in which an event is occurred. In (2.1), it is presumed that the date marking the beginning of the interval is included. The interval ends at the instant before the date marking the end of the interval. This approach can also be defined vice versa, but the choice is largely irrelevant concerning the theory.

Besides the survival time, for each object i, $i = 1, ..., n$, a p-dimensional vector $\boldsymbol{x}_i$ of covariates is collected. For this vector $\boldsymbol{x}_i$, an impact on the survival time is assumed. Therefore, it is usually referred to conditional probability density functions $f(t|\boldsymbol{x}) = P(T = t|\boldsymbol{x})$ and conditional cumulative density functions $F(t|\boldsymbol{x})$, where $\boldsymbol{x}^T = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$. Initially, the covariates are supposed to be time-independent. The comprehension of time-dependent covariates is discussed in Section 2.3.

2.2.1. Hazard Rate and Survival Function

At the beginning of this chapter, single spells were assumed. As each object can experience the event of interest only once, event occurrence is inherently conditional. An object can experience the event at time t, $t = 1, .., q$, only if the event did not already occur at any earlier time. Similarly, once the event has occurred it cannot arise again after time t.

In survival analysis, the modeling of survival time T is general based on the hazard rate. It captures the intrinsically conditionality of the event occurrence and the dynamic feature of survival time. The discrete hazard function, given the covariates, is defined by

$$\lambda(t|\boldsymbol{x}) = P(T = t \mid T \geq t, \boldsymbol{x}), \quad t = 1, ..., q, \tag{2.2}$$

which is the conditional probability that an event occurs in interval $[a_{t-1}, a_t)$, given the interval is reached. In the same way, it can be interpreted for truly discrete survival times: $\lambda(t|\boldsymbol{x})$ is the conditional probability of an event at time t, given that no event has occurred prior to time t. The conditional probability (2.2) belongs to the most important parameters of the discrete-time survival process. A main issue of survival analysis is the estimation of the hazard function $\lambda(t|\boldsymbol{x})$ and the investigation of the influence of covariates on it (Singer and Willett, 1993). Since the discrete hazard rate (2.2) denotes a probability, it ranges between 0 and 1. Usually, the hazard rate is plotted subject to time t, visualizing the risk of the event occurring in each time period, respectively, on the condition that the event not having occurred at any earlier time. For further details on examples of hazard functions and the information that can be retrieved from them, compare Singer and Willett (1991).

In addition, the conditional probability for surviving interval $[a_{t-1}, a_t)$ or surviving time t, respectively, is given by

$$P(T > t|T \geq t, \boldsymbol{x}) = 1 - \lambda(t|\boldsymbol{x}).$$

The discrete survival function $S(t)$ gives the probability that an object survives longer than some specified discrete time t. $S(t)$ defines the probability that the random variable T exceeds the specified time t:

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}) = \prod_{j=1}^{t} (1 - \lambda(j|\boldsymbol{x})) = 1 - F(t|\boldsymbol{x}). \tag{2.3}$$

Thereby, the survival function (2.3) is directly linked to the discrete hazard function as well as the corresponding cumulative density function. In Singer and Willett (1993), it was explained why the hazard function and not the survival function forms the cornerstone of survival analysis.

Finally, the unconditional probability for an event in interval $[a_{t-1}, a_t)$ or at time t, respectively, is denoted by

$$P(T = t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}) \prod_{j=1}^{t-1} (1 - \lambda(j|\boldsymbol{x})). \tag{2.4}$$

2.2.2. Regression Models

A simple method to describe survival data for the total sample or for sub-populations of interest is by life table. The influence of covariates is disregarded in this case. It is a useful nonparametric estimation approach and can be applied to discrete survival times as well as grouped survival times. Life table estimates are shortly discussed in Section 4.2.2 and more detailed in Fahrmeir and Tutz (2001) or Hamerle and Tutz (1989).

The central focus of regression modeling of survival data is to obtain estimates of the hazard rate after adjusting for measured covariates, something that is not possible with life table estimators. As the hazard function $\lambda(t|\boldsymbol{x})$ is a probability, Cox (1972) proposed an parametrization by means of binary regression models. This is attained by modeling binary transitions. That means, does an event occur or not at time t, given the corresponding object reaches time t. To put it another way, the binary outcomes $\{T = t|T \geq t, \boldsymbol{x}\}$ and $\{T > t|T \geq t, \boldsymbol{x}\}$ are distinguished (Hamerle and Tutz, 1989). Thus, binary regression models with response variable $Y \in \{0, 1\}$ can be used. More details on the class of binary regression models can be found in Tutz (2012), Agresti (2013) and McCullagh and Nelder (1989). Using the representation of binary regression models, the discrete hazard rate has the form

$$\lambda(t|\boldsymbol{x}) = P(T = t \mid T \geq t, \boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = F(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}), \tag{2.5}$$

where F is an appropriate cumulative distribution function and $\boldsymbol{\gamma}$ contains the effects of the covariates. Moreover, γ_{0t} denotes a time-varying intercept and can be seen as a baseline effect disregarding any set of covariates. For the model (2.5) an appropriate distribution function has to be chosen. In this context, a directly associated function is the link function $g(\cdot)$ that "links" the linear predictor to the hazard function $\lambda(t|\boldsymbol{x})$. It is the inverse function of the cumulative distribution function F in Equation (2.5) and is given by

$$g(\lambda(t|\boldsymbol{x})) = \gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}.$$

The model (2.5) constitutes a sequential model and represents the common structure of discrete-time survival models. A sequential model can be seen as an extension of generalized linear models with a particular cumulative distribution function or link function. More details on that issue can be found in Fahrmeir and Tutz (2001).

To indicate if an observation i, i=1,...,n, is right-censored or not, a censoring indicator variable is introduced. In the case of grouped survival data, the censoring indicator is given by

$$\delta_i = \begin{cases} 1, & T_i \leq C_i, \text{that is failure in interval } [a_{t_i-1}, a_{t_i}) \\ 0, & T_i > C_i, \text{that is censoring in interval } [a_{t_i-1}, a_{t_i}) \,, \end{cases} \tag{2.6}$$

where $[a_{t_i-1}, a_{t_i})$ denotes the last observed time interval of object i with $t_i \leq q$, for all t_i. In definition (2.6), it is implicitly assumed that censoring occurs at the end of the interval $[a_{t_i-1}, a_{t_i})$. A censoring indicator for truly discrete survival time can be defined analogous.

To use binary regression models it is necessary to restructure the original data by setting up the binary transitions. Let R_t be a index set of objects that are respectively at risk in interval $[a_{t-1}, a_t)$ or at time t:

$$R_t = \{i : t \leq t_i\} \,.$$

Binary event indicators for $i \in R_t$ are defined as follows

$$y_{it} = \begin{cases} 1, & \text{for } t = t_i \text{ and } \delta_i = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

Consequently, in the resulting data set, denoted as long format, each observation of object i consists of t_i rows, where an additional variable defines the running time. In the case of a censored observation i ($\delta_i = 0$), this leads to

object	binary response	time	censoring	design
i	0	1	δ_i	$\boldsymbol{x}_i^T$
i	0	2	δ_i	$\boldsymbol{x}_i^T$
i	0	3	δ_i	$\boldsymbol{x}_i^T$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	0	t_i	δ_i	$\boldsymbol{x}_i^T$.

On the other hand, the data rows of a non-censored observation i with $\delta_i = 1$ are given by

object	binary response	time	censoring	design
i	0	1	δ_i	$\boldsymbol{x}_i^T$
i	0	2	δ_i	$\boldsymbol{x}_i^T$
i	0	3	δ_i	$\boldsymbol{x}_i^T$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	0	$t_i - 1$	δ_i	$\boldsymbol{x}_i^T$
i	1	t_i	δ_i	$\boldsymbol{x}_i^T$.

For example, the data representation of the long format of the following three observations $\{ (t_1 = 3, \delta_1 = 0, \boldsymbol{x}_1), (t_2 = 2, \delta_2 = 1, \boldsymbol{x}_2), (t_3 = 4, \delta_3 = 1, \boldsymbol{x}_3) \}$ is obtained by

object	binary response	time	censoring	design
1	0	1	0	$\boldsymbol{x}_1^T$
1	0	2	0	$\boldsymbol{x}_1^T$
1	0	3	0	$\boldsymbol{x}_1^T$
2	0	1	1	$\boldsymbol{x}_2^T$
2	1	2	1	$\boldsymbol{x}_2^T$
3	0	1	1	$\boldsymbol{x}_3^T$
3	0	2	1	$\boldsymbol{x}_3^T$
3	0	3	1	$\boldsymbol{x}_3^T$
3	1	4	1	$\boldsymbol{x}_3^T$.

The resulting structure of the data set enables the immediate application of software for binary regression models. In doing so, the running time, denoted as *time* in the previous examples, has to be included as a factor to the model, that means, with an appropriate coding with regard to a categorical variable.

Grouped Proportional Hazard Model

For continuous survival time the hazard function is given by

$$\lambda_{cont}(t|\boldsymbol{x}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

A possibility to model the hazard rate λ_{cont} subject to the covariates $\boldsymbol{x}$, is the model of Cox (Cox, 1972), that is one of the most popular regression models for continuous survival time. It is given by

$$\lambda_{cont}(t|\boldsymbol{x}) = \lambda_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\gamma}), \tag{2.8}$$

where the baseline hazard function $\lambda_0(t)$ is assumed to be the same for all observations and is independent of the covariates. Furthermore, no specific structure is assumed for the baseline hazard.

The Cox model is also denoted *proportional hazard model*, as the hazard ratio of two vectors of covariates $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ does not depend on time

$$\frac{\lambda_{cont}(t|\boldsymbol{x})}{\lambda_{cont}(t|\tilde{\boldsymbol{x}})} = \exp((\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \boldsymbol{\gamma}). \tag{2.9}$$

Let the underlying time be continuous but the observed survival times are interval-censored. Thus, let survival time T be a discrete random variable with $T = t$ denoting an event in

interval $[a_{t-1}, a_t)$. Then, the assumption of (2.8) yields the grouped proportional hazard model

$$\lambda(t|\boldsymbol{x}) = 1 - \exp(-\exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})), \tag{2.10}$$

deriving the parameters

$$\gamma_{0t} = \log(\exp(\theta_t) - \exp(\theta_{t-1})), \quad \text{with} \quad \theta_t = \log \int_0^{a_t} \lambda_0(u)du$$

from the baseline hazard function $\lambda_0(u)$ (Kalbfleisch and Prentice, 1973, 2002). In the model (2.10), the baseline hazard function exists no longer in an explicit form, rather the parameter γ_{0t} contains the information of the baseline hazard. Moreover, γ_{0t} is a discrete parametrization of $\lambda_0(t)$ of the underlying Cox model for continuous time. The model (2.10) is presumed as a Cox model for discrete survival times. The model (2.10) also holds, if truly discrete survival times instead of grouped survival times are assumed (Kalbfleisch and Prentice, 2002).
The parameter $\boldsymbol{\gamma}$ is identical in the discrete Cox model (2.10) as well as in the continuous Cox model (2.8). That means, the interpretation of the parameter $\boldsymbol{\gamma}$ is the same in both models.
As the discrete hazard rate is given by $\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x})$, the grouped proportional hazard model is a sequential model with a Gompertz distribution $F(x) = 1 - \exp(-\exp(x))$, that is an extreme-value-distribution. The corresponding link function is constituted by the complementary log-log link also known as clog-log link $g(\lambda(t|\boldsymbol{x})) = \log(-\log(1 - \lambda(t|\boldsymbol{x})))$. Because of this link function, the discrete Cox model is often referred to as clog-log model, and as stated by Kalbfleisch and Prentice (2002), this model "is the uniquely appropriate one for grouped data from the continuous proportional hazards model".

While the property (2.9) holds in the continuous Cox model, it no longer holds in the grouped proportional hazard model. Thus, the name grouped proportional hazard model sometimes is deceptive. A further feature of the continuous Cox model is the proportionality of logarithmic survival functions. Only the latter property can be transferred to the discrete Cox model. Compare Hamerle and Tutz (1989) for more details. An immediate generalization of the grouped proportional hazard was provided by Aranda-Ordaz (1983). Instead of considering a multiplicative model as in the Cox model (2.8), the proposed model of Aranda-Oraz is based on an additive form for continuous survival time, whereof a model for discrete survival time can be derived.

Finally, note that the clog-log model is not the only model that is consistent with a continuous time model and interval-censored survival time data. In Sueyoshi (1995), it was shown how, for example, a logistic hazard model (as considered in the following section) with interval specific intercepts, may be consistent with an underlying continuous time model in which the within-interval durations follow a log-logistic distribution (Jenkins, 2005).

Discrete Logistic Model

An alternative choice of the distribution function in Equation (2.5), is the logistic distribution function yielding the *discrete logistic model*. The resulting sequential model for the discrete hazard function is defined by

$$\lambda(t|\boldsymbol{x}) = \frac{\exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})}. \tag{2.11}$$

The model (2.11) was originally developed for truly discrete survival times but may also be applied to grouped survival times. The choice of the logistic distribution function is equivalent to the logit link function given by

$$g(\lambda(t|\boldsymbol{x})) = \frac{\lambda(t|\boldsymbol{x})}{1 - \lambda(t|\boldsymbol{x})}.$$

An alternative formulation of the model is obtained by

$$\log \frac{P(T = t|\boldsymbol{x})}{P(T > t|\boldsymbol{x})} = \gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}.$$

In this representation, the comparison of the response category t to the response categories $t + 1, ..., q$ is evident. The logits $P(T = t|\boldsymbol{x})/P(T > t|\boldsymbol{x})$ are also denoted *continuation ratio logits* (Tutz, 2012). This sometimes leads to the name continuation ratio logit model instead of discrete logistic model.

Originally, this model was proposed by Cox (1972). He extended the proportional hazard model to intrinsically discrete time by working with conditional odds of an event at time t, given survival up to that point. It is closely related to the Mantel-Haenszel approach (Mantel, 1966). The original model given by Cox is of the following form

$$\frac{\lambda(t|\boldsymbol{x})}{1 - \lambda(t|\boldsymbol{x})} = \frac{\lambda^0(t)}{1 - \lambda^0(t)} \exp(\boldsymbol{x}^T\boldsymbol{\gamma}),$$

where $\lambda(t|\boldsymbol{x})$ defines the conditional discrete-time hazard rate and $\lambda^0(t)$ defines the corresponding baseline hazard that arises when $\boldsymbol{x} = \mathbf{0}$. Taking the logarithms, the logit of the hazard rates, respectively the conditional probability of an event at t, given survival up to that time, is obtained by

$$\operatorname{logit} \lambda(t|\boldsymbol{x}) = \log \frac{\lambda(t|\boldsymbol{x})}{1 - \lambda(t|\boldsymbol{x})} = \alpha_t + \boldsymbol{x}^T\boldsymbol{\gamma}, \tag{2.12}$$

where $\alpha_t = \operatorname{logit} \lambda^0(t)$ is the logit of the baseline hazard. Thereby, time is treated as a discrete factor, as for each time t a parameter α_t is introduced. Interpretation of the

parameters $\boldsymbol{\gamma}$ is analogous to logistic regression. Moreover, an alternative formulation of the model (2.12) is given by

$$\lambda(t|\boldsymbol{x}) = \frac{1}{1 + \exp(-\alpha_t - \boldsymbol{x}^T\boldsymbol{\gamma})}. \tag{2.13}$$

Finally, the model (2.13) is equivalent to the model (2.11). A property concerning specifically the interpretation of the model can be derived from the proportion of odds of two populations $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$

$$\frac{P(T = t|\boldsymbol{x})/P(T > t|\boldsymbol{x})}{P(T = t|\tilde{\boldsymbol{x}})/P(T > t|\tilde{\boldsymbol{x}})} = \exp((\boldsymbol{x} - \tilde{\boldsymbol{x}})^T\boldsymbol{\gamma}.$$

It is seen that the proportion of odds does not depend on time. That means, it is the same for all time periods allowing for a simple interpretation of effects. Due to this feature the continuous ratio logit model is also known as a proportional odds model (Jenkins, 2005). Thompson (1977) dealed with the proportional odds model in detail, and showed that it leads back to Cox's proportional hazard model when the lengths of the grouping intervals approach zero (see also Fleming and Harrington, 2005). However, as stated in Prentice and Gloeckler (1978), the meaning of the regression coefficient in the discrete logistic model depends on the choice of grouping intervals.

Additional to the extreme-value distribution and the logistic distribution, for the choice of F in Equation (2.5), further cumulative distribution functions are possible. For example, possible distributions are the normal distribution yielding a probit model and the Gumbel distribution. More details can be found in Cox and Oakes (1984).

2.2.3. Discrete time versus continuous time models

In practice, it should be determined which kind of survival data are on hand: continuous survival times, grouped survival times (interval censoring) or intrinsically discrete survival times. According to the available data an appropriate model has to be chosen. However, in practice, it is common to use the widespread Cox model instead of prevalent models for discrete time.
Many approaches for continuous time assume that equal survival times of a sample have probability zero. This assumption is incorrect if many ties are present. As the Cox model is the most popular model for survival data, it is often preferred to models for discrete time. However, it can be inappropriate when analyzing event history data. The Cox model implies a continuous-time specification whereas the observations are often obtained by means of grouped data. As a result, the observation of ties is unavoidable. The partial likelihood approach of the Cox model requires, however, chronologically ordered duration times (Hess and Persson, 2012). Kalbfleisch and Prentice (2002) pointed out, that the incidence of ties causes asymptotic bias in both the estimation of the regression coefficients and in the estimation of the corresponding covariance matrix. There are several suggestions

how to deal with this issue. Breslow (1974) proposed one of the most popular approaches to handle ties. It is based on an approximation of the exact marginal likelihood but it becomes inaccurate in the case of many ties. More precisely, this leads to an increasing asymptotic bias of the parameter estimates (Prentice and Gloeckler, 1978; Hsieh, 1995). Another approximation of the exact marginal likelihood, being more appropriate, was proposed by Efron (1977), but it is still inaccurate in the presence of heavy ties. Scheike and Sun (2007) have investigated the performance of the methods of Breslow and Efron. They arrived at the conclusion that the impact of tied survival times depends on the number of ties in comparison to the spell size.

A further drawback of the Cox model is that it supposes the individual hazard functions to be proportional. When the assumption of proportional hazards is not met, the estimated covariate effects tend to be biased (Hess and Persson, 2012). The issue of constant covariates over survival time is well investigated in the literature and several tests exists to examine the proportionality assumption (McCall, 1994; Klein and Moeschberger, 2003). Two reasons exist why the proportional hazards assumption may fail to hold. First, the effect of covariates on the hazard may be intrinsically non-proportional. Second, if it is not accounted for unobserved individual heterogeneity, the influence of observed regressors depend on survival time, even if the underlying model satisfies the proportional hazard assumption (Lancaster and Nickell, 1980).

Consequently, the existing data should be well investigated before the analysis of survival time models. Statistical models for discrete survival time can overcome the handicaps just mentioned. Discrete-time models are also preferable for computational reasons. Conventional regression models for binary response data can be used to model hazard rates for discrete duration times. These models are computationally less demanding than the Cox model.

2.2.4. Estimation

As specified in the introduction, the objects' survival times often cannot be fully observed. Therefore, right censoring is included in the estimation approach. Estimates of the unknown parameters can be obtained by the maximum likelihood (ML) method. The considered model has the form

$$\lambda(t|\boldsymbol{x}) = F(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}),$$

as it is already defined in Equation (2.5). If censored data are existent in a sample, besides the survival time T the censoring time C (time until censoring of an object) is of interest. For each object i, the random variables survival time T_i and censoring time C_i are available. The observed time t_i for each object is the minimum of survival time and censoring time obtained by $t_i = \min(T_i, C_i)$, where $t_i \leq q$. It is generally assumed that the random variables T_i and C_i are independent and that the tupel (T_i, C_i), $i = 1, ..., n$, are measured independently (Hamerle and Tutz, 1989). This assumption is called *random censoring*.

For an object i, an observation is defined by the triple $(t_i, \delta_i, \boldsymbol{x}_i)$, $i = 1, ..., n$, where $\boldsymbol{x}_i$ contains the object's characteristics and δ_i is the censoring indicator variable of definition (2.6) with

$$\delta_i = \begin{cases} 1, & T_i \leq C_i, \text{ that means failure in interval } [a_{t_i-1}, a_{t_i}) \\ 0, & T_i > C_i, \text{ that means censoring in interval } [a_{t_i-1}, a_{t_i}) \,. \end{cases} \tag{2.14}$$

Thereby, $[a_{t_i-1}, a_{t_i})$ denotes the last observed time interval of object i with $t_i \leq q$. In definitions (2.6) and (2.14), it is implicitly assumed that censoring occurs at the end of the interval. This assumption is of special interest with respect to time-varying coefficients and holds during the whole thesis. For all details concerning censoring at the beginning of an interval, compare Fahrmeir and Tutz (2001).
By omitting covariates and using the assumption of random censoring, the probability of observing $(t_i, \delta_i = 1)$ is obtained by

$$P(T_i = t_i, \delta_i = 1) = P(T_i = t_i)P(C_i \geq t_i), \tag{2.15}$$

where censoring is supposed to occur at the end of the interval. On the other hand, the probability of observing $(t_i, \delta_i = 0)$ is given by

$$P(T_i = t_i, \delta_i = 0) = P(C_i = t_i)P(T_i > t_i). \tag{2.16}$$

The combination of probability (2.15) and probability (2.16) leads to the likelihood contribution of observation (t_i, δ_i)

$$L_i = P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i},$$

where, according to Allison (1982), a separate expression for censored objects and an expression for uncensored objects is derived. Let $c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ be the contribution of the censoring distributions. If it is presumed that c_i does not depend on the parameters determining the survival time, that means, the censoring mechanism is non-informative (see Kalbfleisch and Prentice, 2002), the reduced likelihood contribution is obtained by

$$L_i = c_i P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i}.$$

Inserting the definition of the discrete hazard function and including covariates it follows

$$L_i = c_i \lambda(t_i|\boldsymbol{x}_i)^{\delta_i} (1 - \lambda(t_i|\boldsymbol{x}_i))^{1-\delta_i} \prod_{j=1}^{t_i-1} (1 - \lambda(j|\boldsymbol{x}_i)).$$

By incorporating the binary event indicators, the likelihood can be written as

$$L_i = c_i \prod_{j=1}^{t_i} \lambda(j|\boldsymbol{x}_i)^{y_{ij}} (1 - \lambda(j|\boldsymbol{x}_i))^{1-y_{ij}}, \tag{2.17}$$

where

$$(y_{i1}, ..., y_{it_i}) = \begin{cases} (0, ..., 0, 0), & \text{if } \delta_i = 0 \\ (0, ..., 0, 1), & \text{if } \delta_i = 1. \end{cases} \tag{2.18}$$

The data structure (2.18) is already described in Section 2.2.2. The likelihood contribution (2.17) describes that of a binary regression model. Thus, the likelihood function of a discrete survival model is equivalent to the likelihood of a binary regression model (Laird and Olivier, 1981; Allison, 1982; Brown, 1975). Finally, the log-likelihood of a discrete-time survival model is given by

$$l = \sum_{i=1}^{n} \sum_{j=1}^{t_i} (y_{ij} \log(\lambda(j|\boldsymbol{x}_i)) + (1 - y_{ij}) \log(1 - \lambda(j|\boldsymbol{x}_i))).$$

The contribution of each object to the design matrix are t_i rows whereas the complete design matrix consists of $\sum_i t_i$ rows. It has to be noted, that the equivalence of the likelihood of discrete survival time models and binary regression models is only valid concerning the estimation approach. The asymptotic propositions of binary regression models, that is, the distributions of test statistics cannot be adopted. The presented estimation approach of models for discrete survival times was illustrated on the basis of grouped survival times. It can be carried out analogous to intrinsically discrete time.

2.3. Time-Varying Covariates

By means of the baseline effects, a time variation is implicitly incorporated in the models of the previous section. This time variation is ensured by including the running time as a factor variable into the model. The observed covariates of each object, though, are assumed to be constant over survival time. However, in many applications it might be of special interest that the covariates vary over the duration time.
For the i-th object, $\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{it}$, $i = 1, ..., n$, $t = 1, ..., t_i$, denote the time-dependent observations of covariates until time t_i. The vector $\boldsymbol{x}_{it}^T = (x_{it1}, ..., x_{itp})$, with p denoting the number of covariates, is observed at the end of interval $[a_{t-1}, a_t)$ or is determined at discrete time t. The resulting hazard rate incorporating time-dependent covariates is given by

$$\lambda(t|\boldsymbol{x}_i(t)) = P(T = t|T \geq t, \boldsymbol{x}_i(t)) = F(\boldsymbol{z}_{it}^T \boldsymbol{\beta}),$$

where the temporal sequence of covariates $\boldsymbol{x}_i^T(t) = (\boldsymbol{x}_{i1}^T, ..., \boldsymbol{x}_{it}^T)$ influence the hazard rate and $\boldsymbol{z}_{it}$ is composed from $\boldsymbol{x}_{it}$. A very simple way to specify $\boldsymbol{z}_{it}$ is given by

$$\boldsymbol{z}_{it}^T \boldsymbol{\beta} = \gamma_{0t} + \boldsymbol{x}_{it}^T \boldsymbol{\gamma},$$

where $\boldsymbol{z}_{it}^T = (0, ..., 1, ..., 0, \boldsymbol{x}_{it}^T)$, $\boldsymbol{\beta}^T = (\gamma_{01}, ..., \gamma_{0q}, \boldsymbol{\gamma}^T)$ and $\boldsymbol{\gamma}^T = (\gamma_1, ..., \gamma_p)$. Further specifications of $\boldsymbol{z}_{it}$, in terms of time-varying coefficients or time lags, can be found in Fahrmeir

and Tutz (2001). For modeling the hazard rate and incorporating time-varying covariates, the models of Section 2.2.2 can be modified by

Discrete Logistic Model

$$\lambda(t|\boldsymbol{z}_{it}) = \frac{\exp(\boldsymbol{z}_{it}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{z}_{it}^T\boldsymbol{\beta})}.$$

Grouped Proportional Hazard Model

$$\lambda(t|\boldsymbol{z}_{it}) = 1-\exp(-\exp(\boldsymbol{z}_{it}^T\boldsymbol{\beta})). \tag{2.19}$$

Due to the incorporation of time-dependent covariates, the model (2.19) cannot be derived as a grouped version of the continuous Cox model (Hamerle and Tutz, 1989).

Moreover, Kalbfleisch and Prentice (2002) distinguished between *external* and *internal* covariates that are outcomes of a stochastic process. With external covariates, the sequence of a covariate vector is not influenced by the duration time of an object. Conversely, it can thoroughly impact the duration time. A possible example for external covariates is environmental factors. Let the sequence of observations of covariates $\boldsymbol{x}_{i1},...,\boldsymbol{x}_{it}$ be an output of a stochastic process. This output may be considered external if the condition

$$P(\boldsymbol{x}_{i,t+1},...,\boldsymbol{x}_{iq}|\boldsymbol{x}_i(t),\boldsymbol{y}_i(t)) = P(\boldsymbol{x}_{i,t+1},...,\boldsymbol{x}_{iq}|\boldsymbol{x}_i(t)), \quad t=1,...,q, \tag{2.20}$$

holds, where $\boldsymbol{y}_i(t) = (y_{i1},...,y_{it})$. The condition (2.20) implies that failure does not influence the path of the covariate process.

In contrast, an internal time-dependent covariate depends on the individual survival, its path may carry information about the time of failure occurrence. It is related to the behavior of the individual and can be observed only as long the object is in the study and alive. Thus, the hazard rate only incorporates the path until time t

$$\lambda(t|\boldsymbol{x}_i(t)) = P(y_{it}=1|\boldsymbol{x}_i(t), y_{i1}=0,...,y_{i,t-1}=0).$$

As the covariates $x_{i,t+1},...,x_{iq}$ no longer have any meaning if an event occurs in interval $[a_{t-1},a_t)$ the condition (2.20) cannot be presumed to hold. This statement can be adapted to intrinsically discrete survival times, as well.

In addition, Hamerle and Tutz (1989) have shown that for external time-dependent covariates, an analogous formulation of the survival function (2.3) can be derived

$$S(t|\tilde{\boldsymbol{x}}_t) = \prod_{s=1}^{t}(1-\lambda(s|\tilde{\boldsymbol{x}}(s))),$$

where $\tilde{\boldsymbol{x}}(t) = (\boldsymbol{x}(1),...,\boldsymbol{x}(t))$. The previous equation leads to the usability of the same likelihood function that is used for the models with time-independent covariates (Fahrmeir and Tutz, 2001). For the type of internal covariates, however, the simple connection between

survival function and hazard function, given in Section 2.2.1, no longer hold. With internal covariates the likelihood of time-independent covariates cannot be used, especially if internal covariates are strict informative. A strictly informative internal covariate has to fulfill the condition $P(T \geq t|\boldsymbol{x}(1), ..., \boldsymbol{x}_t) = 1$. In Hamerle and Tutz (1989), a likelihood function was introduced that is appropriate with respect to internal covariates.

2.4. Time-Varying Coefficients

So far, the regression parameters of the models in Section 2.2.2 are assumed to be time-constant. A further extension is to allow the regression parameters to vary over time. Time-varying coefficients were systematically introduced by Hastie and Tibshirani (1993). Varying coefficient models arise from various statistical contexts. For example, Hoover et al. (1998) considered a time-varying coefficient model for continuous longitudinal data using smoothing splines and local polynomial estimators. Huang et al. (2002) proposed a basis function approximation method to estimate the time-varying coefficients, whereas Tutz and Kauermann (2003) used generalized local likelihood estimation.
The discrete-time hazard model with time-varying coefficients has the form

$$\lambda(t|\boldsymbol{x}) = F(\boldsymbol{\eta}_t) = F(\beta_0(t) + \boldsymbol{x}_1\beta_1(t) + ... + \boldsymbol{x}_p\beta_p(t)) = F(\beta_{0t} + \boldsymbol{x}_1\beta_{1t} + ... + \boldsymbol{x}_p\beta_{pt}), \quad (2.21)$$

where $\beta_1(t), ..., \beta_p(t)$, $t = 1, ..., q$ are unspecified functions to be estimated and $\boldsymbol{\eta}_t$ is the linear predictor depending on time. If covariates are also time varying, the model (2.21) can be extended to

$$\lambda(t|\boldsymbol{x}_{it}) = F(\eta_{it}) = F(\beta_{0t} + x_{it1}\beta_{1t} + ... + x_{itp}\beta_{pt}). \quad (2.22)$$

The models (2.21) and (2.22) have the form of generalized linear models (Fahrmeir and Tutz, 2001). In time-varying coefficient models, time t is considered as an effect modifier as it modifies the effect on the predictors. In the case of discrete-time survival models, time can supposed to be a categorical effect modifier (Gertheiss and Tutz, 2012). The variation of the parameters across the effect modifier time t, may be seen as an interaction of time and covariate j, $j = 1, ..., p$.
Although the models with time-varying coefficients are given as a linear combination, the functions $\beta_1(t), ..., \beta_p(t)$ are not parametrically specified and have to be estimated by smoothing techniques. To this end, regression splines are used for the estimation of the functions $\beta_j(t)$, $j = 1, .., p$. In the last decades, regression splines have been widely used for the estimation of additive structures, see, for example, Marx and Eilers (1998) and Wood (2006). In regression spline approaches the unspecified functions $\beta_j(t)$ are approximated by basis functions of the form

$$\beta_j(t) = \beta_{jt} = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t),$$

where $B_{jm}(t)$ are basis functions such as B-splines or truncated power series and m_j, $j = 1, ..., p$, denotes the number of basis functions for each time-varying parameter. The choice of basis functions and their number influence the flexibility of the basis function approach. B-splines are very famous and widely used because the basis functions are strictly local defined entailing a big numerical advantage compared to other basis function approaches. Using a B-spline basis of degree d yields

$$\beta_j(t) = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t; d) = \boldsymbol{\alpha}_j^T \boldsymbol{b}_{jt},$$

where $\boldsymbol{\alpha}_j^T = (\alpha_{j1}, ..., \alpha_{jm_j})$ denotes the unknown parameter vector of the j-th smooth function, $\boldsymbol{b}_j(t)^T = (B_{j1}(t; d), ..., B_{jm_j}(t; d))$ represents the vector-valued evaluations of the m_j basis functions and p denotes the number of time-varying parameters. For simplicity reasons, the degree d is often omitted. Generally, cubic B-splines ($d = 3$) and equidistant knots are assumed (Eilers and Marx, 1996). For B-splines of degree $d = 0$ and knots at each point in time, that means, $m_j = q$, it results $\boldsymbol{\beta}_j = \boldsymbol{\alpha}_j$, with $\boldsymbol{\beta}_j^T = (\beta_{j1}, ..., \beta_{jq})$.

By the use of B-splines for the time-varying coefficients, the resulting linear predictor can be written as

$$\begin{aligned}\eta_{it} &= \boldsymbol{\alpha}_0^T \boldsymbol{b}_0(t) + x_{it1} \boldsymbol{\alpha}_1^T \boldsymbol{b}_1(t) + \cdots + x_{itp} \boldsymbol{\alpha}_p^T \boldsymbol{b}_p(t) \\ &= \boldsymbol{b}_0(t)^T \boldsymbol{\alpha}_0 + (x_{it1} \boldsymbol{b}_1(t))^T \boldsymbol{\alpha}_1 + \cdots + (x_{itp} \boldsymbol{b}_p(t))^T \boldsymbol{\alpha}_p.\end{aligned}$$

Moreover, not necessarily all of the p covariates have to vary over time. It is also possible, that some of the functions $\beta_j(t)$ are assumed to be constant, that means $\beta_j(t) = \gamma_j$, for all t, resulting in terms with simple linear effects. Altogether this leads to the linear predictor

$$\eta_{it} = \boldsymbol{b}_0(t)^T \boldsymbol{\alpha}_0 + (z_{it1} \boldsymbol{b}_1(t))^T \boldsymbol{\alpha}_1 + \cdots + (z_{itr} \boldsymbol{b}_r(t))^T \boldsymbol{\alpha}_r + \boldsymbol{x}_{it}^T \boldsymbol{\gamma}, \tag{2.23}$$

with parameter vector $\boldsymbol{\gamma}^T = (\gamma_1, ..., \gamma_s)$ and $\boldsymbol{x}_{it}^T = (x_{it1}, ..., x_{its})$, $i = 1, ..., n$, $t = 1, ..., t_i$. Thereby, z_{itj}, $i = 1, ..., n$, $t = 1, ..., t_i$, $j = 1, ..., r$, denotes the observations of the covariates that are allowed to exhibit time-varying effects, and $\boldsymbol{x}_{it}$ define the covariates that are restricted to have constant effects. That is, the linear predictor (2.23) consists of $p = r + s$ covariates with r time-dependent covariates and s time-constant covariates. By collecting observations over time, that is $\boldsymbol{X}_i^T = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{it_i})$ for the i-th object, a simpler form of the model is given by

$$\boldsymbol{\eta}_i = \tilde{\boldsymbol{Z}}_{i0} \boldsymbol{\alpha}_0 + \tilde{\boldsymbol{Z}}_{i1} \boldsymbol{\alpha}_1 + ... + \tilde{\boldsymbol{Z}}_{ip} \boldsymbol{\alpha}_r + \boldsymbol{X}_i \boldsymbol{\gamma},$$

where $\tilde{\boldsymbol{Z}}_{i0}^T = (\boldsymbol{b}_0(1)^T, ..., \boldsymbol{b}_0(t_i)^T)$ denotes the transposed B-spline design matrix of the time-varying intercept for the i-th object and $\tilde{\boldsymbol{Z}}_{ij}^T = ((z_{i1j} \boldsymbol{b}_j(1))^T, ..., (z_{it_ij} \boldsymbol{b}_j(t_i))^T)$, $j = 1, ..., r$, represents the transposed B-spline design matrix for the i-th object and the j-th time-

dependent covariate effect.
With $\boldsymbol{X}^T = \left[\boldsymbol{X}_1^T, ..., \boldsymbol{X}_n^T\right]$, the model in matrix form is given by

$$\boldsymbol{\eta} = \tilde{\boldsymbol{Z}}_0\boldsymbol{\alpha}_0 + \tilde{\boldsymbol{Z}}_1\boldsymbol{\alpha}_1 + ... + \tilde{\boldsymbol{Z}}_p\boldsymbol{\alpha}_r + \boldsymbol{X}_i\boldsymbol{\gamma},$$

where $\tilde{\boldsymbol{Z}}_j^T = \left[\tilde{\boldsymbol{Z}}_{1j}^T, ..., \tilde{\boldsymbol{Z}}_{nj}^T\right]$ defines the transposed B-spline design matrix of the j-th smooth function. By collecting $\tilde{\boldsymbol{Z}} = \left[\tilde{\boldsymbol{Z}}_0, \tilde{\boldsymbol{Z}}_1, ..., \tilde{\boldsymbol{Z}}_r\right]$ and $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, ..., \boldsymbol{\alpha}_r^T)$ the linear predictor can be further reduced to

$$\boldsymbol{\eta} = \tilde{\boldsymbol{Z}}\alpha + \boldsymbol{X}\gamma.$$

3. Lasso-Type Penalties

Using time-varying coefficients in a linear predictor of a model, results in a large number of parameters that have to be estimated. Especially in the context of discrete survival models this might lead to wiggly time-varying covariate effects, unstable estimates or computational problems (Section 3.1). To overcome these issues, regularization techniques in terms of penalization methods are incorporated in discrete-time survival models (Section 3.2). To get access to uncertainty measures, Section 3.3 provides standard errors and confidence intervals for the parameters resulting from the penalized estimation. In addition, in Section 3.4, several simulation studies are conducted to judge the performance of the presented method. Finally, two real data examples are discussed in Section 3.5. Section 3.6 contains concluding remarks. Some parts of the following chapter are based on Hess et al. (2014), the result of a close cooperation. In the following, only the notation and explanations with respect to grouped survival times are considered, but they can easily be modified to truly discrete survival times.

3.1. Introduction

High-dimensional time-varying coefficients, for example, were treated by Xue and Qu (2012), Wei et al. (2011) or Wang and Xia (2009), but they do not consider survival analysis. The issue of discrete duration time models with time-varying coefficients has been investigated, for example, by Tutz and Binder (2004) using penalized B-splines for estimating time-varying coefficients. Fahrmeir (1994) introduced a dynamic modeling approach by means of penalized likelihood techniques. A fully Bayesian approach using Markov Chain Monte Carlo techniques was proposed by Fahrmeir and Knorr-Held (1997). However, the latter publications disregard regularization techniques.
Survival analysis with discrete duration times is not well investigated. In contrast, a rich literature dealing with continuous-time survival models is available. Unfortunately, in practice, this encourages the use of continuous survival models, also in situations where discrete-time survival models are more appropriate leading to biased estimates.
The data referring to survival analysis usually consist of repeated measurements over time. As discrete-time survival data often cover rather long time periods, the question arises whether the effects of the explanatory variables vary over time. However, in many studies, it is simply assumed that regression coefficients are time-constant. The incorporation of time-varying coefficients often leads to a large number of coefficients to be estimated and consequently to instability in the estimation process, especially, if the data become sparse for larger values of time t. Hence, alternative tools such as penalization methods

are needed. In addition, to obtain coefficient estimates that vary flexibly over time, interactions of covariates and time have to be incorporated in the regression model. Using this approach, the problem of overfitting of the model may arise. That means, it leads to overly wiggly and hard-to-interpret covariate effects. The issue of overfitting can be avoided by using penalized regression methods. The basic idea of this chapter is to incorporate smooth flexible interactions of covariates and time into the regression model. By using that time intervals are naturally ordered, the difference between coefficients of adjacent time periods are penalized by the use of regularization techniques.
This approach offers several benefits. First, the flexible interactions of regressors and time allow for covariate effects that vary across time without any restrictions on parametric assumptions. Second, the use of penalized differences between coefficients of adjacent time periods solves the problem of overfitting. Third, the type of penalization of the proposed method is rather flexible. Different kinds of penalty terms can be used. The type of penalty enables different interpretation possibilities that allows for a large flexibility. Depending on the particular application, for each predictor can be chosen which one is the most empirically reasonable penalty. Finally, the penalty term must not only affect the differences between coefficients of adjacent time periods. Due to the fact that several penalty types can shrink coefficients to be exactly zero, the proposed approach can also be used for model selection. To reduce the inhibitions regarding discrete survival models, in the following an approach for a discrete survival model that accounts for time-varying coefficients and that incorporates penalization methods is proposed.

3.2. Lasso-Type Penalties

Let y_{it} denote the binary outcome of an object i, $i = 1, ..., n$, in period t, $t = 1, ..., t_i$, and let $(\boldsymbol{z}_{it}, \boldsymbol{x}_{it})^T = (z_{it1}, ..., z_{itr}, x_{it1}, ..., x_{its})$ with $p = r + s$ be a vector of realizations of explanatory variables that may vary over time. Thereby, the binary outcome y_{it} denotes an event in $[a_{t-1}, a_t)$ if $y_{it} = 1$ and no event or censoring if $y_{it} = 0$ leading to the data situation 2.18. Thus, the existing data structure is of the same form as described in Section 2.2.2, regarding binary regression models in the context of discrete survival analysis. The discrete hazard function $\lambda(t|\boldsymbol{z}_{it}, \boldsymbol{x}_{it})$ is fitted by the binary regression model (2.5)

$$\lambda(t|\boldsymbol{z}_{it}, \boldsymbol{x}_{it}) = P(y_{it} = 1|\boldsymbol{z}_{it}, \boldsymbol{x}_{it}) = F(\eta_{it}). \tag{3.1}$$

By incorporating time-varying as well as time-constant regression coefficients the linear predictor for object i, $i = 1, ..., n$, and time period t, $t = 1, ..., q$, is given by

$$\eta_{it} = \beta_{0t} + \sum_{j=1}^{r} z_{itj}\beta_{jt} + \sum_{l=1}^{s} x_{itl}\gamma_l, \tag{3.2}$$

where the parameters β_{0t} represent the baseline effects that are the same for all individuals. Moreover, let $\boldsymbol{z}_{\bullet\bullet 1}, ..., \boldsymbol{z}_{\bullet\bullet r}$ with $\boldsymbol{z}_{\bullet\bullet j}^T = (z_{11j}, ..., z_{1t_ij}, ..., z_{n1j}, ..., z_{nt_ij})$ denote the observations of the covariates that are allowed to exhibit time-varying effects, and $\boldsymbol{x}_{\bullet\bullet 1}, ..., \boldsymbol{x}_{\bullet\bullet s}$ with

$\boldsymbol{x}_{\bullet\bullet l}^T = (x_{11l}, ..., x_{1t_il}, ..., x_{n1l}, ..., x_{nt_il})$ define the observations of covariates that are restricted to have time-constant effects. In other words, the model implies time-varying coefficients $\beta_{0t}, \beta_{1t}, ..., \beta_{rt}$ including a time-varying intercept β_{0t}, whereas $\boldsymbol{\gamma}^T = (\gamma_1, ..., \gamma_s)$ is assumed to be time-constant. When the covariates do not depend on time, the time index is omitted; that is $z_{itj} = z_{ij}$ and $x_{itl} = x_{il}$.

As the time-varying effects in model (3.1) are estimated for each time period t, the number of parameters can be very large and estimating them using conventional maximum likelihood (ML) techniques may lead to unstable results. In extreme cases, the maximum likelihood estimates might not exist. Furthermore, the resulting wiggly coefficient vectors are hard to interpret in a meaningful manner. A simple solution to reduce the amount of parameters is the application of B-splines basis functions to the time-varying covariate effects (see also Section 2.4). The time-varying coefficients then are expanded in equally spaced B-splines given by $\beta_{jt} = \beta_j(t) = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t)$, $j = 0, ..., r$. By using this B-spline representation to model the time-varying effects β_{jt}, $j = 0, 1, ..., r$, $t = 1, ..., q$, the linear predictor (3.2) can be rewritten as

$$\eta_{it} = \tilde{\boldsymbol{z}}_0^T \alpha_0 + \sum_{j=1}^{r} \sum_{m=1}^{m_j} \tilde{z}_{itj} \alpha_{jm} + \sum_{l=1}^{s} x_{itl} \gamma_l. \tag{3.3}$$

Note, that $\tilde{z}_{itj}$ in the linear predictor (3.3) differs from z_{itj} in the linear predictor (3.2). The $\tilde{z}_{itj}$ contain the design of the interaction of the according covariate and the evaluations of the appropriate basis functions at time t. Moreover, α_{jm} denotes the unknown parameter values of the j-th smooth function. By using B-splines for the basis expansion of the time-varying coefficients the resulting model is a generalized additive model, but the linear predictor (3.3) can still be embedded into the framework of generalized linear models.
In general, additive models exhibit an identification problem so that it is necessary to fix the level of each flexible function. This is usually done by centering each flexible function around zero (Fahrmeir et al., 2009). However, in the context of time-varying coefficients centering of the time-varying functions $\beta_j(t)$ depends on the scale of the j-th interaction variable belonging to the observations $\boldsymbol{z}_{\bullet\bullet j}$, $j = 1, ..., r$. In the case of a metric interaction variable, the smooth terms does not have to be centered and it is not necessary to incorporate main effects of the j-th covariate belonging to observations $\boldsymbol{z}_{\bullet\bullet j}$ in the model. In contrast, if the interaction variable is categorical, centering constraints are applied to the smooths, which usually means that the variable itself should be included as a parametric term as well (Fahrmeir et al., 2009; Wood, 2006).
To yield smooth functions $\beta_j(t)$ that are not too wiggly, the differences between adjacent parameters of the smooths $\alpha_{jm} - \alpha_{j,m-1}$, $j = 1, ..., r$, $m = 2, ..., m_j$, are penalized. Some coefficients may not have an impact on the discrete hazard function and should be omitted from the model. This implies a need for variable selection, that is to determine whether $\beta_{jt} = 0$, for all t. In order to accomplish smooth time-varying coefficients as well as variable selection, penalized estimation techniques are proposed in the following.

3.2.1. Penalized Estimation

As stated in the previous sections, the modeling approach of discrete-time survival data utilizes the framework of binary regression models. Thus, the estimation of the parameters is performed by maximum likelihood estimation. In empirical applications it is often eligible to suppose that covariate effects vary rather smoothly over time. This implies that adjacent coefficients α_{jm} and $\alpha_{j,m-1}$ in the linear predictor (3.3) should be expected to be similar, or in other words, that differences $\zeta_{jm} = \alpha_{jm} - \alpha_{j,m-1}$, $j = 1, ..., r$, $m = 2, ..., m_j$, are small. Moreover, parameters that are not relevant should be removed from the model. These goals can be reached by penalization, leading to the maximization of the penalized log-likelihood given by

$$l_\xi(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = l(\boldsymbol{\alpha}, \boldsymbol{\gamma}) - J_\xi(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = l(\boldsymbol{\alpha}, \boldsymbol{\gamma}) - \xi J(\boldsymbol{\alpha}, \boldsymbol{\gamma}), \tag{3.4}$$

where $l(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ denotes the ordinary log-likelihood and $J(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ stands for a penalty term that depends on a scalar tuning parameter $\xi \geq 0$. The tuning parameter ξ controls the strength of penalization. Without a penalty, that is, with $\xi = 0$, ordinary ML-estimation is obtained.

As the choice of $J(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ determines the properties of the penalized estimator, the main issue is to choose an adequate penalty. Typical penalties are the *ridge* penalty (Hoerl and Kennard, 1970) that shrinks coefficients towards zero but do not perform variable selection, the *lasso* method (Tibshirani, 1996) that combines shrinkage and selection of single coefficients or the *fused lasso* (Tibshirani et al., 2005) applying the Lasso penalty to differences between adjacent parameters. In the latter case, parameters are shrunken towards each other and possibly are fused. A further penalty used in this thesis, is the *group lasso* (Yuan and Lin, 2006) that shrinks whole groups of parameters simultaneously towards zero until their selection. Hence, true variable selection can be obtained instead of simple parameter selection as, for example, in case of the lasso. For a detailed account of recently developed regularization approaches, including the lasso and versions of it, compare Bühlmann and van de Geer (2011).

Selection and Smoothing of Time-Varying Coefficients

A simple penalty that effects smoothness of the time-varying coefficients is given by

$$J_\xi(\boldsymbol{\alpha}) = \xi \sum_{j=1}^{r} \sum_{m=d}^{m_j} \left(\Delta^d \alpha_{jm}\right)^2 = \xi \sum_{j=1}^{r} \boldsymbol{\alpha}_j^T \boldsymbol{K}_{d,j} \boldsymbol{\alpha}_j, \tag{3.5}$$

where Δ defines a difference operator, operating on adjacent coefficients $\Delta\alpha_{jm} = \alpha_{jm} - \alpha_{j,m-1}$, $\Delta^2\alpha_{jm} = \Delta(\alpha_{jm} - \alpha_{j,m-1}) = \alpha_{jm} - 2\alpha_{j,m-1} + \alpha_{j,m-2}$ etc. The penalty does not depend on the time-constant parameters $\boldsymbol{\gamma}$ as currently only time-varying coefficients are considered. This penalty term was originally presented by Eilers and Marx (1996) in the context of penalized regression splines. The matrix $\boldsymbol{K}_d$ follows from representing the differences in matrix form and has a banded structure. By means of difference matrices, the

requested difference can be obtained recursively by $\boldsymbol{D}_d = \boldsymbol{D}_1\boldsymbol{D}_{d-1}$. For example, for $d = 2$ the $(d-2 \times d)$ difference matrix is given by

$$\begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

By means of the difference matrix $\boldsymbol{D}_d$, the sum of squared differences is simply given by $\boldsymbol{\alpha}_j^T \boldsymbol{K}_{d,j} \boldsymbol{\alpha}_j$ with $\boldsymbol{K}_d = \boldsymbol{D}_d^T \boldsymbol{D}_d$. Through penalization of squared differences between the adjacent parameters $\alpha_{jm} - \alpha_{j,m-1}$, $m = 2, ..., m_j$, large shifts in parameter values are avoided (see Gertheiss and Tutz, 2009). That is, for $\xi > 0$ large parameter differences have a negative impact on the penalized log-likelihood $l_\xi(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ that has to be maximized. Thus, estimated parameter differences will be smaller than they would have been in unpenalized models. Finally, the tuning parameter ξ is responsible for the roughness of the time-varying effects β_{jt}. The stronger the penalization, the smoother is the resulting curve of the covariate effect. However, the penalization term (3.5) disregards selection. To achieve a selection of all differences $\zeta_{jm} = \alpha_{jm} - \alpha_{j,m-1}$, $m = 2, ..., m_j$, belonging to the j-th time-varying coefficient, the following penalty can be applied

$$J_\xi(\boldsymbol{\alpha}) = \xi \sum_{j=1}^{r} \sqrt{\sum_{m=2}^{m_j} (\underbrace{\alpha_{jm} - \alpha_{j,m-1}}_{\zeta_{jm}})^2} = \xi \sum_{j=1}^{r} \sqrt{\boldsymbol{\alpha}_j^T \boldsymbol{K}_{1,j} \boldsymbol{\alpha}_j}. \tag{3.6}$$

The penalty (3.6) is defined as the group lasso of Yuan and Lin (2006), extended to generalized linear models by Meier et al. (2008), applied to the differences $\zeta_{j2}, ..., \zeta_{jm_j}$ between adjacent parameters of the smooths regarding the time-varying coefficients. In order to maximize the penalized likelihood (3.4), the group lasso shrinks parameter differences ζ_{jm}, thereby generating smooth time-varying coefficients. For a value of ξ large enough, the group lasso simultaneously forces the whole group of parameter differences $\zeta_{j2}, ..., \zeta_{jm_j}$ to be zero, implying that the effect of covariate j is constant over time. For estimation of the parameters by using the group lasso penalty, Meier et al. (2008) proposed an algorithm that is implemented in the R add-on package `grplasso` (Meier, 2013). Hence, the penalty term (3.6) can be simply applied by using the freely available standard software R. The only requirement for application of standard software is the reparametrization of the linear predictor (3.3) using parameters ζ_{jm} instead of α_{jm}, that is accomplished by using split coding of the covariates. This approach is described in more detail in (Gertheiss et al., 2011).

However, the penalty (3.6) is still not satisfactory in the context of discrete survival modeling: First, a penalty with regard to the time-varying intercept has to be incorporated. It is essential that the time-varying intercept is not excluded from the model as it represents a baseline hazard. Hence, a further tuning parameter is needed only responsible for the

strength of penalization of the intercept. Second, by means of the penalty term (3.6) only the differences ζ_{jm} can be selected out of the model but no variable selection is performed. Moreover, if penalization of the $\boldsymbol{\gamma}$-parameters is intended an additional penalty term has to be included. Incorporating these three issues, leads to the extended penalty

$$\begin{aligned} J_{\xi_0,\xi}(\boldsymbol{\alpha},\boldsymbol{\gamma}) =& \xi_0 \sum_{m=2}^{m_0} \left(\alpha_{0m} - \alpha_{0,m-1}\right)^2 \\ &+ \xi\left(\phi \sum_{j=1}^{r} \psi_j \left\|\boldsymbol{\zeta}_j\right\|_2\right) + \xi\left((1-\phi)\sum_{j=1}^{r} \varphi_j \left\|\boldsymbol{\alpha}_j\right\|_2\right) + \xi J(\boldsymbol{\gamma}), \end{aligned} \tag{3.7}$$

where $\|\cdot\|_2$ denotes the L_2-norm, $\boldsymbol{\zeta}_j^T = (\zeta_{j2}, ..., \zeta_{jm_j})$ and $\boldsymbol{\alpha}_j^T = (\alpha_{j1}, ..., \alpha_{jm_j})$. The first term enforces shrinkage of the differences between adjacent B-spline coefficients of the baseline hazard with the objective of a smooth function over time. This is related to a ridge penalty of the parameters ζ_{0m}, $m = 2, ..., m_j$. This part of penalization is predominantly incorporated due to stability reasons. Hence, the tuning parameter ξ_0 should be chosen rather small, for example $\xi_0 = 0.001$. By using a group lasso penalty with respect to the differences $\zeta_{j2}, ..., \zeta_{jm_j}$, $j = 1, ..., r$, the second term steers the smoothness of the time-varying covariate effects, but for a value of ξ large enough, all differences $\zeta_{j2}, ..., \zeta_{jm_j}$ are removed from the model resulting in a constant covariate effect. The third term steers the selection of covariates and corresponds to a group lasso penalty with regard to the parameters α_{jm}, $m = 1, ..., m_j$, belonging to the j-th covariate. If the tuning parameter ξ exceeds a certain value, $\alpha_{j1}, ..., \alpha_{jm_j}$ are set to zero and the j-th covariate is removed from the model. That means, the penalty term may distinguish if a covariate effect is incorporated smooth or constant in the model or if it is completely removed from the model. The last term has the intention to penalize the time-constant parameters $\gamma_1, ..., \gamma_s$. If only shrinkage of γ_l is desired a ridge penalty can be used, whereas for exclusion of the model, that means shrinking the coefficient γ_l to zero, the lasso penalty has to be applied. In addition, many further penalty types are available. For more details compare Oelker and Tutz (2013). The penalties of time-constant parameters γ_l are summarized in $J(\boldsymbol{\gamma})$. Except the penalty regarding the time-varying intercept, the remaining penalty terms have a shared tuning parameter ξ. Weighting of the selection part and the smoothing part of a covariate is obtained by the parameter ϕ, that is a further tuning parameter. The terms $\psi_j = \sqrt{m_j - 1}$ and $\varphi_j = \sqrt{m_j}$ are weights that assign different amounts of penalization to different parameter groups.

To sum up, the penalty (3.7) performs smooth time-varying or time-constant coefficients as well as their selection out of the model. Additionally, the baseline effects are slightly penalized and optional penalty terms regarding the time-constant parameters $\boldsymbol{\gamma}_l$ may be added.

Selection and Fusion of Time-Varying Coefficients

In some situations it may be reasonable that the time-varying effect of a covariate is assumed to be piecewise constant over time. That means, the effect of covariate j only changes for some distinct time periods $t_1 < t_2 < \ldots$. Some external events may be the cause of relevant changes. The objective is to identify these breakpoints or jumps when estimating the regression coefficients. For this issue, the linear predictor (3.2) is considered. Hence, for time-varying coefficients β_{jt}, each time period determines a separate coefficient. This corresponds to a B-spline representation with degree $d = 0$ and knots at the observed interval limits and jumps really refer to time periods instead of the chosen knots. For the simultaneous estimation of regression coefficients and the identification of jumps, fused lasso-type penalties can be employed (Tibshirani et al., 2005). As time t can be seen as an ordinal effect modifier, a regularization method for categorical effect modifiers using the fused lasso in the context of generalized linear models can be applied (Oelker et al., 2014). Then, the penalty term of the penalized likelihood (3.4) has the form

$$J_\xi(\boldsymbol{\beta}) = \xi \left(\sum_{j=1}^{r} \left(\sum_{t=2}^{q} |\beta_{jt} - \beta_{j,t-1}| + \kappa_j \sum_{t=1}^{q} |\beta_{jt}| \right) \right), \tag{3.8}$$

where κ_j is an indicator that activates the second term if needed. Thus, using the L_1-norm, the penalty encourages sparsity of the coefficients (second term of the penalty) on the one hand and on the other hand sparsity of their differences (first part of the penalty), that means, time constancy of the coefficient profile. In other words, variable selection is obtained by a penalization, strong enough to set some of the β_{jt} to zero. Moreover, a distinction of time-varying or time-constant coefficients is obtained by the selection of relevant jumps. That is, separate differences $\beta_{jt} - \beta_{j,t-1}$ can be fitted as exactly zero. In some situations it might be reasonable to introduce a weighting on the selection and the fusion part to emphasize either of them (see Tibshirani et al., 2005). More concrete, this results in

$$J_{\xi,\phi}(\boldsymbol{\beta}) = \xi \left(\sum_{j=1}^{r} \left(\phi \sum_{t=2}^{q} |\beta_{jt} - \beta_{j,t-1}| + (1-\phi)\kappa_j \sum_{t=1}^{q} |\beta_{jt}| \right) \right), \tag{3.9}$$

where the tuning parameter ϕ is restricted to $[0, 1]$ in order to separate it strictly from tuning parameter ξ. However, a further tuning parameter is introduced. A drawback of the lasso-type penalty (3.8) is the occurrence of much more parameters compared to the B-spline approach using $\beta_{jt} = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t)$, $j = 0, 1, ..., r$, where only m_j parameters per covariate are used and usually $m_j < q$ is assumed.

To get a time-varying intercept with piecewise time-constant coefficients the corresponding penalty can be incorporated in penalties (3.8) or (3.9) as well. However, the intercept should rather vary smoothly over time. Hence, for penalizing the intercept, the corresponding part of penalty term (3.7) can be used. Furthermore, it is possible to combine the penalty terms (3.7) and (3.9) depending on whether a covariate effect should vary smoothly over time, be

piecewise time-constant or, in the case of time-constant coefficients, is regularized by any lasso-type penalty.

By using a penalization approach, the importance of variables can be assessed. That is, if the contribution of covariates or their time interactions is relatively little to the maximization of the likelihood, they may be removed from the model.

Adaptive Penalties

Variable selection procedures aim to identify the right subset model. Let A denote the *active set* of parameters of a corresponding model, that means, all non-zero coefficients are collected in A. A selection procedure is consistent if asymptotically the right subset model is found, that is $\lim_n P(A_n = A) = 1$, where A_n is the active set for n observations. It has been shown by Zou (2006) that for the ordinary lasso the induced variable selection can be inconsistent in certain scenarios. To overcome this selection inconsistency, Zou (2006) proposed an adaptive version of the lasso. This adaptive approach was extended to the group lasso by Wang and Leng (2008). Further applications of adaptive penalties, can be found, for example in Zhang and Lu (2007), Meier et al. (2009) or Gertheiss and Tutz (2012). The decisive modification is to weight the penalty terms by the inverse of the respective unpenalized parameter estimates. For example, given penalty (3.7), the adaptive version is obtained by replacing the weights ψ_j and φ_j by

$$\psi_j^a = \frac{\sqrt{m_j - 1}}{||\hat{\boldsymbol{\zeta}}_j^{\mathrm{ML}}||}, \quad \varphi_j^a = \frac{\sqrt{m_j}}{||\hat{\boldsymbol{\alpha}}_j^{\mathrm{ML}}||}, \tag{3.10}$$

where $\hat{\boldsymbol{\zeta}}_j^{\mathrm{ML}}$ and $\hat{\boldsymbol{\alpha}}_j^{\mathrm{ML}}$ denote the according ML-estimates. The intuition behind this weighting procedure is rather straightforward. With very large data sets, unpenalized point estimates can be expected to be rather accurate. Thus, the norm of ML-estimates of parameter groups belonging to relevant predictors is rather large. Consequently, the corresponding penalization should be small. In contrast, a strong penalization goes along with parameter groups belonging to irrelevant predictors and, hence, leading to a small norm of the ML-estimates. Moreover, the ML-estimates employed in the adaptive weights (3.10), can be replaced by any $\sqrt{n}$-consistent estimates. For example, a ridge penalty can be used in situations where the ML-estimates do not exist.

3.2.2. Computational Issues

In the following, some details regarding the penalized estimation of the parameters are described. Some details on the estimation are outlined and tuning parameter selection for discrete penalized survival models is discussed. As the proposed method deals with **PEN**alization of **D**iscrete **S**urvival **M**odels, it is denoted by `pendsm`.

Estimation

Using the penalty term (3.6), the parameters can be estimated by the algorithm proposed in Meier et al. (2008). By using the penalty term (3.7) or (3.9), this is no longer possible as a combination of different types of penalties is used. The application of several penalties turns out as a challenge since various potentially non-differentiable terms can arise. To cope with these different penalties that employ different norms, local quadratic approximations in a penalized iteratively re-weighted least squares (PIRLS) algorithm is used. This idea is based on Fan and Li (2001), who approximate the non-convex SCAD penalty quadratically. An adoption to lasso-type penalties was proposed by Ulbricht (2010). Oelker and Tutz (2013) showed how penalties that are norms of scalar linear transformations of the coefficient vector, can be approximated quadratically in generalized linear models. Their approach is based on the ideas of Fan and Li (2001) and Ulbricht (2010) and is defined such that a variety of penalty types, like for example the lasso, group lasso, fused lasso, ridge or elastic net are embedded. The approximation allows to combine all these penalties in one model. In common generalized linear models, the unpenalized optimization problem is given by $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} -l(\boldsymbol{\theta})$. Thereby, let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$ be defined with respect to the penalized likelihood (3.4). This equation is solved iteratively by

$$\hat{\boldsymbol{\theta}}_{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} - H(\hat{\boldsymbol{\theta}}^{(k)})^{-1} s(\hat{\boldsymbol{\theta}}^{(k)}), \tag{3.11}$$

where $s(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta})/\partial(\boldsymbol{\theta})$ denotes the score function and $H(\boldsymbol{\theta}) = \partial^2 l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ the Hessian matrix. Equation (3.11) can be transformed to a Fisher scoring algorithm or an iteratively re-weighted least squares algorithm. Oelker and Tutz (2013) proposed penalized versions of the score function $s(\boldsymbol{\theta}^{(k)})$ and the Hessian matrix $H(\boldsymbol{\theta}^{(k)})$ or the Fisher matrix, respectively. They are needed to use a conventional penalized iteratively re-weighted least squares (PIRLS) algorithm for the penalized optimization problem (3.4). By using the penalized score function, the same optimization problem as for usual generalized linear models is obtained by solving $s_{pen}(\boldsymbol{\theta}) = 0$. For more details see Oelker and Tutz (2013).
An implementation of the PIRLS algorithm is provided by the R add-on package `gvcm.cat` (Oelker and R Development Core Team, 2013). As the discrete survival model with time-varying coefficients can be estimated by means of a binary regression model, `pendsm` can be embedded into `gvcm.cat`. However, some profound implementation modifications have to be executed to enable the application to `pendsm`.

Tuning parameter selection

In practice, a remaining task is the selection of the tuning parameters ξ and ϕ. For this purpose, a common approach for selecting tuning parameters, is K- fold cross-validation, whereby the data are (randomly) split into K (roughly) equal-sized parts. To include the whole information of an object, the splitting refers to objects instead of individual data points. Then, for each part $k = 1, ..., K$, the model is fitted to the remaining $K - 1$ parts of the data. The data set, on which the estimation is based, constitutes the learning

sample. Afterwards, the prediction error of the fitted model is calculated when predicting the outcome variables y_{it} of the k-th sub-sample denoting the test sample. Lastly, the K estimates of the prediction error are combined, and the resulting measure of prediction error is minimized as a function of the tuning parameter of interest (see Hastie et al., 2009).

A possible approach of assessing the predictive performance of a model is the predictive deviance. For a new observation $(t_i^{pred}, \delta_i^{pred}, \boldsymbol{z}_i^{pred}, \boldsymbol{x}_i^{pred})$, with $\boldsymbol{z}_i^{pred} = (\boldsymbol{z}_{i1}^{pred}, ..., \boldsymbol{z}_{it_i}^{pred})^T$ and $\boldsymbol{x}_i^{pred} = (\boldsymbol{x}_{i1}^{pred}, ..., \boldsymbol{x}_{it_i}^{pred})^T$ the predictive deviance is defined by

$$\mathrm{D}_i = -2 \sum_{t=1}^{t_i} \left\{ y_{it}^{pred} \log(\hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})) + (1 - y_{it}^{pred}) \log(1 - \hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})) \right\},$$

where $\lambda(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred}) = P(T_i = t|T_i \geq t, \boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})$ and $(y_{i1}^{pred}, ..., y_{it_i}^{pred})$ denotes the transitions over periods of object i. To choose simultaneously multiple tuning parameters by cross-validation, a two- or more-dimensional grid of the tuning-parameters is used on which the cross-validation is performed. Hence, for choosing the tuning parameters ξ and ϕ a two-dimensional grid of possible parameters is used, on which the optimal parameter combination is chosen.

3.3. Standard Errors and Confidence Intervals

A drawback of regularization methods is the missing standard error of the unknown parameters. Especially, in practical applications the user demands for a measure of uncertainty. To solve this problem, Efron (1979) proposed *bootstrapping*. When no information about the distribution of the standard errors exists, a nonparametric bootstrap approach is conducted: Let F denote the underlying true distribution function of a random variable and $\hat{F}$ its empirical probability distribution referring to the observations $x_1, ..., x_n$. By drawing B random samples of size n with replacement, the bootstrap samples $x_b = x_1^*, ..., x_n^*$, $b = 1, ..., B$, are obtained. As in survival analysis multiple measures per object are given, the bootstrap sampling will be based on samples with replacement from objects instead of individual data points. This cluster sampling takes into account any correlation structure that might exist within subjects. For each bootstrap replication b, the model of interest is fitted leading to B estimates of the parameter. For a single parameter γ_l, this results in $\hat{\gamma}_l^{(b)}$ leading to the following estimation of the standard error

$$\hat{se}(\hat{\gamma}_l) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\gamma}_l^{(b)} - \bar{\hat{\gamma}}_l^{(b)} \right)^2}.$$

Thereby, the tuning parameters are set to the chosen values and hence, are hold constant for the fitting procedure of all bootstrap replications.
In Efron and Tibshirani (1993), it was shown that 50 to 100 bootstrap replications are

generally sufficient for standard error estimation. As the B random samples are drawn referring to individuals instead of single data points, it has been shown that substantially more bootstrap replications are necessary to yield stable standard error estimates. Hence, in the following the number of bootstrap replications is constituted to 1000. The resulting bootstrap estimates can also be used for deriving bootstrap confidence intervals, applying the percentile bootstrap. For a 95% confidence interval, this is achieved by the 0.025- and the 0.975-percentile of the empirical distribution of the bootstrapped parameters $\hat{\gamma}_l^{(b)}$.
In this thesis, also time-varying coefficients β_{jt} are considered. Hence, bootstrap standard error bands are of particular interest. In this context, Hoover et al. (1998) used the simple case of ± 2 pointwise bootstrap standard error bands. However, in the following it is preferred to use the more exact pointwise percentile bootstrap approach for β_{jt}. Since smoothing bias has not been taken into account, these bootstrap bands are not actually confidence intervals in the usual sense (Hoover et al., 1998). For estimating confidence intervals based on bootstrap percentiles, Efron and Tibshirani (1993) suggested a number of bootstrap replications of about 500 to 1000.

3.4. Simulation Study

In this section, the performance of `pendsm`, that means, combining different penalty terms in a discrete-time survival model, is evaluated. Moreover, it is compared to a method using conventional generalized additive models (GAM) implemented in the function `gam` of the R add-on package `mgcv` (Wood, 2006). The study aims to investigate in which data situations `pendsm` can outperform `gam`. As well as in `pendsm`, in `gam` it is possible to penalize the parameters representing the baseline hazard separately. However, the selection part is conducted differently. In `pendsm`, it can be chosen which covariate effects might be set to zero and `pendsm` allows to distinguish whether an effect is time-varying or time-constant. The `mgcv` package enables a model selection removing complete smoothing terms from the model by adding an extra penalty. This is achieved by setting the option `select=TRUE`. Though, this selection affects all smoothing terms. An individual choice which smoothing terms are allowed to be removed from the model and which are not, is not possible. Furthermore, any penalties applied to the parametric model terms can be specified in the option `paraPen`. As `gam` only works with quadratic penalty approaches, variable selection is not feasible for parametric terms. Moreover, the use of a penalty resulting in piecewise time-constant coefficients is not provided. Hence, `gam` is not as flexible as the `pendsm` method in the context of discrete survival models with time-varying covariate effects.

The simulation approach is performed as follows: The simulated true survival time T_i is obtained by inversion sampling. Thereby, it is aimed to sample from $f(x)$, with $F(x) = u$ and $F^{-1}(u) = x$. As $F(x)$ is a cumulative distribution function, it is bounded to $(0, 1)$. Thus, sample a random value u from $\mathcal{U}(0, 1)$ and compute F^{-1} to obtain x, where x is drawn from $f(x)$. To simulate discrete time survival times, $F(\cdot)$ is denoted by the cumulative distribution function of a binary regression model. This regression model includes the

simulated values of the covariates and the true coefficients. Moreover, for all settings, that are described in the following section, the complementary log-log link is used. This results in the employed cumulative distribution function $F(\boldsymbol{\eta}) = 1 - \exp(1 - \exp(\boldsymbol{\eta}))$, where $\boldsymbol{\eta}$ is the linear predictor of the corresponding model. For more details on inversion sampling, see Kolonko (2006). The true individual censoring times C_i are drawn from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$. Hence, the vector $\boldsymbol{p}_c$ determines the censoring rate, that means, the relative frequency of censoring for all observations and is specified for each setting in the following section. The minimum of the survival time and the censoring time defines the observed survival time $t_i = \min(T_i, C_i)$. The censoring indicator δ_i then follows from definition (2.6). Afterwards, the data has to be restructured as proposed in Section 2.2.2 to yield a binary regression model. Furthermore, the covariates of the whole simulation study are assumed to be time-constant throughout.
Additionally, in the simulation study, the predictive accuracy of the models is investigated. Statistically significant covariates do not guarantee a high prognostic value of a statistical model incorporating these covariates (Korn and Simon, 1990). By using regularization techniques, no significances are obtained for the estimated parameters. However, the feature of model selection ensures a resulting model which only includes the covariates having a large influence on the dependent variable. So, the investigation of prediction accuracy might also be useful in the context of regularization methods. To assess the predictive performance, for each simulation setting n_p additional observations are independently sampled.
The components of the simulation settings required for computing the underlying true linear predictor

$$\eta_{it}^{true} = \beta_{0t} + \sum_{j=1}^{r} z_{ij}\beta_{jt} + \sum_{l=1}^{s} x_{il}\gamma_l$$

are given in the following.

3.4.1. Settings

The simulation study assessing the performance of `pendsm` compared to `gam` consists of four simulation settings. In each setting, both time-constant and time-varying coefficients are included in the model. Several penalties performing variable selection are investigated. Moreover, the number of time periods and the number of covariates is modified and one setting considers correlated covariates.

Setting 1

For the first scenario, $n = 100$ realizations of six covariates are simulated according to $Z_{i1}, X_{i1}, ..., X_{i5} \overset{iid}{\sim} \mathcal{U}(0, 1)$. Only three covariates have an effect on the survival time, whereas

the remaining tree covariates are noise variables. The realizations of covariates $z_{i1}, x_{i1}, ..., x_{i5}$ are used to simulate survival times according to the linear predictor

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4 + x_{i5}\gamma_5,$$

where the time-varying coefficient effects β_{jt}, $j = 0, 1$, are given by

$$\begin{aligned}\beta_{0t} &= (\ -1, -0.5, -1.25, -1.5, -1.75, -1.5, -1.9),\\ \beta_{1t} &= (\ -3, \quad -2, \quad -1, -0.5, \quad 1, \quad 1.5, \quad 2).\end{aligned}$$

Finally, the time-constant coefficient effects γ_l are defined by $\gamma_1 = 0.5$, $\gamma_2 = -1$, $\gamma_3 = \gamma_4 = \gamma_5 = 0$. For the time-varying coefficients β_{jt}, a cubic B-spline approach is used, where the number of equidistant inner knots is set to four. The censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$, where $\boldsymbol{p}_c$ is defined by $\boldsymbol{p}_c^T = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.4)$. This leads to a censoring rate of approximately 35%. The simulation scheme for Setting 1 is replicated 100 times. To evaluate the predictive accuracy, $n_p = 200$ further independent observations are sampled.

Setting 2

In simulation Setting 2, the number of time periods is increased to $q = 15$. The model consists of 12 covariates, whereof six are noise variables. For the study $n = 400$ realizations of the covariates are simulated according to

$$\begin{aligned} Z_{i1}, X_{i1}, X_{i4}, X_{i5} &\overset{iid}{\sim} \mathcal{U}(0, 2),\\ X_{i6}, X_{i7} &\overset{iid}{\sim} \mathcal{U}(0, 1),\\ Z_{i2}, X_{i2} &\overset{iid}{\sim} \mathcal{N}(2, 1),\\ Z_{i3}, X_{i3}, X_{i8}, X_{i9} &\overset{iid}{\sim} \mathcal{B}(0.5).\end{aligned}$$

Hence, the model contains six uniformly distributed, two normal distributed and four binary distributed covariates. The survival times are sampled by means of the realizations of the covariates with the linear predictor given by

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + z_{i2}\beta_{2t} + z_{i3}\beta_{3t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + ... + x_{i9}\gamma_9.$$

Thereby, the time-varying effects β_{jt}, $j = 0, 1, 2, 3$, are defined by

$$\begin{aligned}\beta_{0t} &= (-1.50, -0.75, -0.52, -0.46, -0.50, -0.58, -0.68, -0.78, -0.88, -0.97, -1.06, -1.13, -1.19, -1.24, -1.29),\\ \beta_{1t} &= (-0.49, -0.93, -1.28, -1.47, -1.48, -1.28, -0.93, -0.50, -0.07, \ 0.28, \ 0.48, \ 0.47, \ 0.28, -0.07, -0.51),\\ \beta_{2t} &= (\ 0.18, \ 0.19, \ 0.21, \ 0.23, \ 0.26, \ 0.29, \ 0.33, \ 0.38, \ 0.43, \ 0.50, \ 0.58, \ 0.67, \ 0.79, \ 0.93, \ 1.09),\\ \beta_{3t} &= (\ -0.5, \ -0.6, \ -0.7, \ -0.8, \ \ 0.9, \ \ \ 1, \ \ 1.1, \ \ 0.9, \ \ 0.7, \ \ 0.8, \ \ 0.9, \ \ 1.2, \ \ 1.3, \ \ 1.5, \ \ 1.7),\end{aligned}$$

and the time-constant coefficients γ_l, $l = 1, ..., 9$, are given by $\gamma_1 = -0.5$, $\gamma_2 = -1.7$, $\gamma_3 = 1$, $\gamma_4 = ... = \gamma_9 = 0$. Analogous to Setting 1, the time-varying coefficients β_{jt} are modeled in terms of cubic B-splines, but the number of equidistant inner knots is set to eight. In this setting, the censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$ with $\boldsymbol{p}_c^T = (0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.1, 0.1, 0.1, 0.1, 0.1)$. This results in a censoring rate of approximately 65%. The number of replications is 100. The data set for judging prediction accuracy consists of $n_p = 800$ additionally independently sampled observations.

Setting 3

In the third setting, the impact of correlated covariates is investigated. To this end, $n = 250$ realizations of eight correlated covariates following a normal distribution are simulated with $Z_{i1}, Z_{i2}, X_{i1} \sim \mathcal{N}(1, 1)$ and $Z_{i3}, Z_{i4}, X_{i2}, X_{i3}, X_{i4} \sim \mathcal{N}(0, 1)$. The corresponding correlation structure is specified by the correlation matrix $\boldsymbol{R}$:

$$\boldsymbol{R} = \begin{pmatrix} 1.0 & 0.0 & 0.5 & -0.3 & 0.2 & -0.1 & 0.3 & 0.0 \\ 0.0 & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.1 & -0.3 \\ 0.5 & 0.1 & 1.0 & 0.3 & 0.7 & 0.2 & 0.0 & 0.25 \\ -0.3 & 0.0 & 0.3 & 1.0 & 0.4 & -0.4 & -0.2 & 0.1 \\ 0.2 & 0.0 & 0.7 & 0.4 & 1.0 & 0.0 & 0.1 & 0.0 \\ -0.1 & 0.0 & 0.2 & -0.4 & 0.0 & 1.0 & -0.1 & 0.0 \\ 0.3 & 0.1 & 0.0 & -0.2 & 0.1 & -0.1 & 1.0 & 0.2 \\ 0.0 & -0.3 & 0.25 & 0.1 & 0.0 & 0.0 & 0.2 & 1.0 \end{pmatrix}. \tag{3.12}$$

By using the realizations of these covariates, the survival times are sampled with the linear predictor defined by

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + z_{i2}\beta_{2t} + z_{i3}\beta_{3t} + z_{i4}\beta_{4t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4,$$

where the time-varying coefficients β_{jt}, $j = 0, 1, 2, 3, 4$, are given by

$$\begin{aligned}
\beta_{0t} &= (\ -2, -0.8, -0.5, -1.25, -1.125, -1.5, -2, -2, -1.5, -1), \\
\beta_{1t} &= (\ -4, -3, -2, -1, 0, 0.5, 1.25, 1.5, 1.7, 2), \\
\beta_{2t} &= (\ -2, -1.9, -1.7, -1.5, -1.25, -1, -0.75, -0.5, -1.5, 1), \\
\beta_{3t} &= (\ 0, 0.1, 0.5, 0.8, 0.9, 1.1, 1.5, 1.6, 1.8, 2.5), \\
\beta_{4t} &= (\ -0.8, -0.7, -0.6, -0.5, 0.5, -0.6, -0.7, -0.8, -0.9, -1).
\end{aligned}$$

Time-constant covariate effects are defined by $\gamma_1 = -0.5$, $\gamma_2 = 1$ and $\gamma_3 = \gamma_4 = 0$. Again, the time-varying coefficients β_{jt} are expanded in cubic B-splines with four equidistant inner knots. Moreover, the censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$ with probability vector $\boldsymbol{p}_c^T = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, leading to a censoring rate of approximately 65%. For this simulation setting, 100 simulation runs are

executed. To evaluate the predictive accuracy, $n_p = 500$ further independent observations with correlated covariates are sampled.

Setting 4

For the last scenario, $n = 200$ realizations of six covariates are simulated according to $Z_{i1}, X_{i1}, ..., X_{i5} \overset{iid}{\sim} \mathcal{U}(0,1)$. Thereby, three covariates have an effect on the survival times and the remaining three covariates are noise variables. The covariate outcomes are the basis of simulating the survival times with the linear predictor given by

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4 + x_{i5}\gamma_5.$$

The linear predictor includes the time-varying coefficients β_{jt}, $j = 0, 1$ with

$$\begin{aligned}\beta_{0t} &= (\ -1.25, \quad -1, -0.75, \quad -2, -1.85, -1.75, -1.9),\\ \beta_{1t} &= (\quad -3, -2.5, \quad -1, -0.5, \quad 1.5, \quad 2, \quad 3),\end{aligned}$$

as well as the time-constant coefficients $\gamma_1 = -0.5$, $\gamma_2 = -1$, $\gamma_3 = \gamma_4 = \gamma_5 = 0$. For simulation Setting 4, a penalty resulting in piecewise time-constant coefficients is used. Hence, the coefficients β_{jt} have to be estimated for all time periods, that is $t = 1, ..., q$. The censoring times are drawn from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$, with $\boldsymbol{p}_c^T = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.4)$ leading to a censoring rate of approximately 50%. The simulation scheme for Setting 4 is replicated 100 times. Further $n_p = 400$ independent observations are sampled that are used to evaluate the predictive accuracy.

To allow for maximal flexibility in modeling, for all coefficients time-varying effects are assumed. For Settings 1-3, the time-varying covariate effects are expanded in B-spline basis functions resulting in the linear predictor

$$\eta_{it}^{model} = \tilde{\boldsymbol{z}}_0^T \alpha_0 + \sum_{j=1}^{r} \sum_{m=1}^{m_j} \tilde{z}_{itj}\alpha_{jm},$$

where $\tilde{z}_{(\cdot)}$ represents the interaction of the corresponding covariate and the evaluation of the B-splines at time t (see Section 3.2). In contrast, the linear predictor for Setting 4 is given by

$$\eta_{it}^{model} = \beta_{0t} + \sum_{j=1}^{p} z_{ij}\beta_{jt},$$

where for each time period a single parameter is estimated. Moreover, the number of time-varying coefficients depends on the simulation setting and $r = p$ holds.
Two different types of penalties, referring to the estimation of the time-varying covariate effects, are used. For Settings 1-3, the penalty term

$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{m_0} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi\left(\phi \sum_{j=1}^{r} \psi_j \left\|\boldsymbol{\zeta}_j\right\|_2 + (1-\phi) \sum_{j=1}^{r} \varphi_j \left\|\boldsymbol{\alpha}_j\right\|_2\right),$$

setting	n	time intervals	covariate effects varying	constant	noise	penalty	correlation
1	100	7	1	2	3	smoothing & selection	-
2	400	15	3	3	6	smoothing & selection	-
3	250	10	4	2	2	smoothing & selection	✓
4	200	7	1	2	3	fusion & selection	-

Table 3.1. Overview simulations settings of Chapter 3.

is used. It allows for stable baseline effects and steers smoothing, constant effects and selection of the time-varying coefficients, whereas fusion and selection of the time-varying coefficients is performed by the penalty term

$$J_{\xi_0,\xi}(\boldsymbol{\beta}) = \xi_0 \sum_{m=2}^{m_0} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi \left(\sum_{j=1}^{r} \sum_{t=2}^{q} |\beta_{jt} - \beta_{j,t-1}| + \sum_{j=1}^{r} \sum_{t=1}^{q} |\beta_{jt}| \right),$$

used for Setting 4. In the context of survival analysis, it is proposed that a time-varying intercept β_{0t} remains in the model. Thus, the penalization of β_{0t} is predominantly executed due to stability reasons. It is defined by $\xi_0 = 0.001$ in all simulation settings corresponding to a ridge penalty on differences between adjacent weights of the B-spline basis functions. Moreover, adaptive versions of the penalties are used for the estimation.

Finally, the simulation settings are summarized in Table 3.1. Therein, n denotes the number of observations of each setting. The number of time-varying and time-constant covariate effects as well as the number of noise variables are shown in the columns *varying*, *constant* and *noise*, respectively. Moreover, *penalty* describes which penalties are used and *correlation* declares if a correlation of the covariates is incorporated.

3.4.2. Results

Before estimation, the simulated data sets has to be adapted to the appropriate binary regression design. In other words, the data are transformed to the long format described in Section 2.2.2. To be on comparable scales, all covariates are standardized to have equal variance in order to avoid that coefficient values are scale dependent. The tuning parameters ξ and ϕ are chosen by 5-fold cross-validation.
The results of `pendsm` are compared to the results obtained by the function `gam` of the R add-on package `mgcv` (Wood, 2006), by fitting analogous models. That means, for the time-varying intercept a slight ridge penalty with tuning parameter 0.001 is used, whereas the time-varying covariate effects are estimated by cubic B-splines with penalized first differences between adjacent parameters of the smooth functions. Moreover, the option `select` is set to `TRUE`, adding a penalty to each smooth, to allow it to be penalized out of the model.

The assessment of parameter estimations is evaluated in general, and separately for truly time-varying and truly time-constant parameters. For each simulation run, the according mean squared errors are computed by

$$\begin{aligned} \text{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) &= \frac{1}{r}\sum_{j=1}^{r}\left(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\right)^2 + \frac{1}{s}\sum_{l=1}^{s}\left(\tilde{\boldsymbol{\beta}}_l - \hat{\boldsymbol{\beta}}_l\right)^2, \\ \text{MSE}_{vary}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) &= \frac{1}{r}\sum_{j=1}^{r}\left(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\right)^2, \qquad \text{MSE}_{const}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \frac{1}{s}\sum_{l=1}^{s}\left(\tilde{\boldsymbol{\beta}}_l - \hat{\boldsymbol{\beta}}_l\right)^2, \end{aligned} \tag{3.13}$$

where $\tilde{\boldsymbol{\beta}}_l = (\gamma_l, ..., \gamma_l)$ and $p = r + s$. Hence, $\boldsymbol{\beta}_j$ and $\tilde{\boldsymbol{\beta}}_l$ denote the true parameter values, whereas $\hat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{\beta}}_l$ define the estimates. That means, as all components are estimated time-varying, $\boldsymbol{\gamma}$ is compared to $\hat{\boldsymbol{\beta}}$ as well. For reference, the ordinary maximum likelihood (ML) estimation is used. However, for many simulation runs, the ML-method does not converge. To stabilize the estimates, a slight ridge penalty of 0.001 is applied to all parameters (denoted by ML_{ridge}). Hence, the ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ can be interpreted in a meaningful manner. In the following plots or outcomes, these ratios are marked by a star *. The predictive accuracy is judged by considering the predictive deviance using the n_p additional independently sampled observations. This results in the covariates $(\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})$ and the corresponding predictive deviance is given by

$$\text{D}_{pred} = -2\sum_{i=1}^{n}\sum_{t=1}^{t_i}\left\{y_{it}^{pred}\log(\hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})) + (1 - y_{it}^{pred})\log(1 - \hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred}))\right\},$$

where $\lambda(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred}) = P(T_i = t|T_i \geq t, \boldsymbol{z}_{it}^{pred}, x_{it}^{pred})$ and $(y_{i1}^{pred}, ..., y_{it_i}^{pred})$ denotes the transitions over periods of object i.

After estimation of the coefficients $\boldsymbol{\beta}_{jt}$, $j = 1, ..., r$, $t = 1, ..., q$, and $\boldsymbol{\gamma}_l$, $l = 1, ..., s$, the results are compared to the true parameters. Analogous to Oelker et al. (2014), for the evaluation of the selection performance, false positive rates (FPR) and false negative rates (FNR) are considered for each simulation run. Thereby, false positive means that a single parameter value that is truly zero is set to non-zero. In contrast, false negative means that a single non-zero parameter value is set to zero. The corresponding rates are defined by

$$\text{FPR} = \frac{\#(\text{truly zero set to non-zero})}{\#(\text{truly zero})} \qquad FNR = \frac{\#(\text{truly non-zero set to zero})}{\#(\text{truly non-zero})}.$$

The results of the simulation settings are initially summarized in tables. Therein, the results of the ordinary ML-estimates with a slight ridge penalty of 0.001 can be found in the column ML_{ridge}. The outcomes of `pendsm` and of `gam` are shown in the corresponding columns. The first three rows of the table contain the absolute values of the mean squared errors for all covariates (MSE) as well as for truly time-varying (MSE_{vary}) and truly time-constant covariates (MSE_{const}). A detailed definition of these mean squared er-

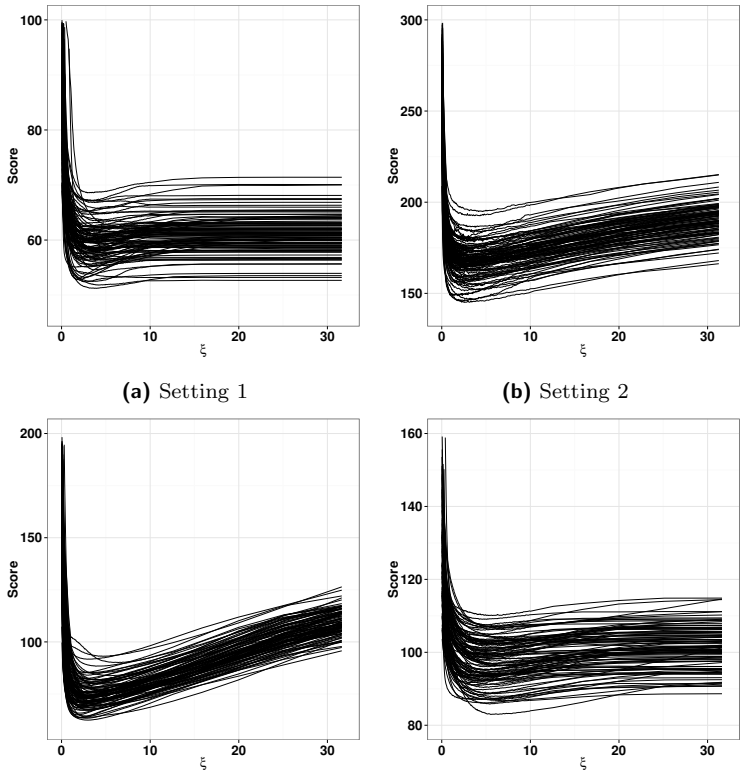


Figure 3.1. Cross-validation scores of the 100 simulation runs subject to penalty parameter ξ for Setting 1-4. Tuning parameter ϕ is set to 0.5.

rors is given in Equation (3.13). The mean squared errors marked with * correspond to the ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$, a logarithmic relation to ML_{ridge}. Therefore, the first column has no values in this case. In the same way, the predictive deviances D_{pred} and D^*_{pred}=$\log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$ are tabulated. Finally, the false positive rate FPR and the false negative rate FNR are shown. All presented values correspond to the mean values over all simulation runs. Moreover, the results are illustrated in boxplots regarding the ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ and $\log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$. For the sake of interpretability, outliers are omitted in single cases.

For all settings, the cross-validation scores referred to the predictive deviance subject to the tuning parameter ξ is illustrated in Figure 3.1. Therein, all simulation runs are incorporated. Note, that the tuning parameter ϕ is set to 0.5 for all plots. All cross-validation curves imply that penalization techniques clearly outperform the ML-procedure that is nearly obtained for $\xi = 0$ as ξ_0 is set to the very small value of 0.001. This is indicated by the strong decrease of the curves for small values of ξ.

	ML_{ridge}	pendsm	gam
MSE	7.31 (5.87)	0.47 (0.27)	0.66 (0.31)
MSE_{vary}	7.77 (8.93)	1.08 (0.80)	1.36 (0.68)
MSE_{const}	7.13 (6.03)	0.22 (0.19)	0.38 (0.34)
MSE^*	-	-2.66	-2.27
MSE^*_{vary}	-	-1.87	-1.46
MSE^*_{const}	-	-3.48	-3.05
D_{pred}	896.04 (268.74)	586.34 (28.78)	594.58 (36.72)
D^*_{pred}	-	-0.39	-0.38
FPR	1	0.53	0.42
FNR	0	0.13	0.12

Table 3.2. Results for Setting 1 for the estimated mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE^*, MSE^*_{vary}, MSE^*_{const}), the predictive deviances D_{pred}, D^*_{pred} referring to test data and the false positive rate (FPR) as well as the false negative rate (FNR) for the ML_{ridge}, `pendsm` and `gam`. The displayed values represent the means over all simulation runs. Estimated standard errors are given in parentheses.

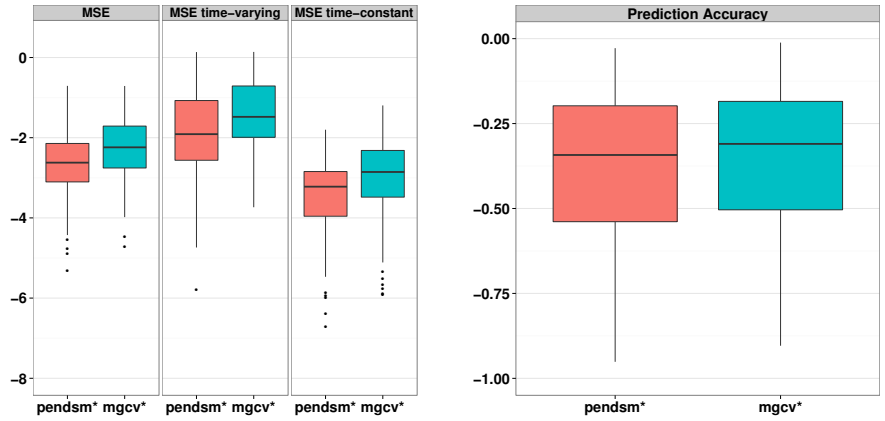
Setting 1

The MSE of an ordinary ML-model with regard to Setting 1 is $3.03 \cdot 10^{30}$ (sd=$1.57 \cdot 10^{31}$). This value is extremely large and it is not meaningful as the ML-algorithm does not converge. The corresponding simulation results are summarized in Table 3.2. It can be seen, that a slight ridge penalty improves the MSE value enormously. Penalization methods achieve even better values of the MSE, as the estimation is more stable and variable selection is performed. Furthermore, for all MSE values `pendsm` outperforms `gam`.
In addition, the predictive deviance D_{pred} yields a bad value for the ordinary maximum likelihood compared to `pendsm` and `gam`. By using penalization approaches, it can be considerably improved. Thereby, `pendsm` produces lower values regarding the prediction accuracy than `gam`. The competence of the algorithm with respect to variable selection is performed by FPR and FNR. For ML-methods, FPR is always equal to one and FNR always equal to zero, since the ML-method cannot set coefficient values to zero. FNR is quite similar for `pendsm` and `gam`. However, the FPR of `gam` is lower than the FPR of `pendsm`.

The boxplots of `pendsm`* and `gam`* of the log ratios $\log(\text{MSE}(\cdot)/\text{MSE}(ML_{ridge}))$ show that `pendsm` outperform `gam` (Figure 3.2a). As expected, MSE^*_{const} has the lowest boxes, whereas MSE^*_{vary} has the highest. Finally, in Figure 3.2b the boxplots referring to the log ratios $\log(D_{pred}(\cdot)/D_{pred}(ML_{ridge}))$ are illustrated. Therein, it is seen that the predictive accuracy is slightly better for `pendsm` as for `gam`.

Setting 2

The second setting is the most complex one. Different covariate distributions are incorporated in the model and the number of time periods is set to 15. This complex setting leads to a MSE of the ML-model of $1.22 \cdot 10^{30}$ ($1.71 \cdot 10^{30}$). Moreover, the ML-algorithm does not converge. The summary of the simulation results are shown in Table 3.3. Like in the



(a) Boxplots of the mean squared errors for `pendsm*` and `gam*` regarding all coefficients, truly time-varying and truly time-constant coefficients.

(b) Boxplots of the predictive deviance for `pendsm*` and `gam*`.

Figure 3.2. Boxplots of the mean squared errors and the predictive deviance for Setting 1

previous setting, a small ridge penalty improves the MSE values. By using the penalization methods `pendsm` and `gam` much more reasonable MSE values are obtained. Again, `pendsm` performs better than `gam` for all MSE values. Even the values of the predictive deviances D_{pred} and D^*_{pred} are slightly smaller than the predictive deviances of `gam`. However, the false positive rate FPR has a smaller value for `gam`.

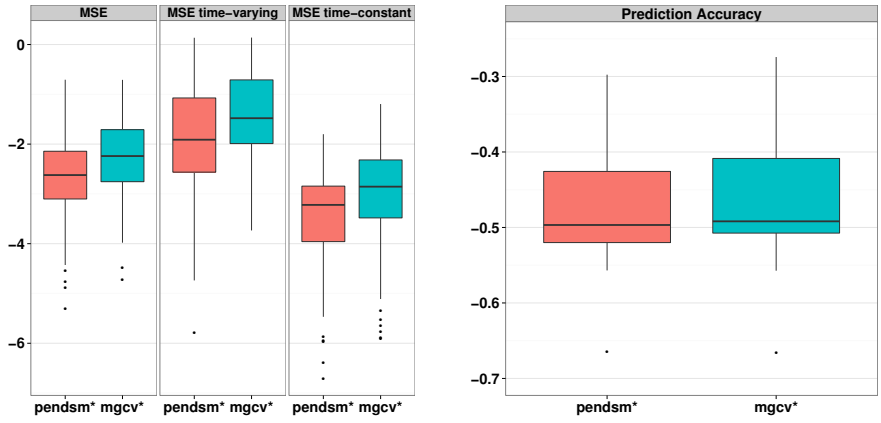
Regarding the boxplots that stem from the empirical distribution of the log ratios with $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$, it is obvious that `pendsm` outperforms `gam` for all boxplots (Figure 3.3a). The predictive performance of `pendsm` yields slightly better results than that of `gam` shown in the boxplots of the log ratios $\log(D_{pred}(\cdot)/D_{pred}(\text{ML}_{ridge}))$ (Figure 3.3b).

Setting 3

In simulation Setting 3, correlated covariates according to the correlation matrix (3.12) are incorporated in the model. In this setting it should be investigated whether this correlation affects the estimation results. The MSE of the ordinary ML-method constitutes a value of $9.70 \cdot 10^{33}$ $(2.39 \cdot 10^{34})$, that is a higher value compared to the other settings. The results for Setting 3 are summarized in Table 3.4. On closer examination of the MSE values, it can be seen that `pendsm` outperforms `gam` clearly. In addition, both predictive deviance values D_{pred} and D^*_{pred} perform best for `pendsm`. Only for the FPR, `gam` yields better results than `pendsm`. However, the FPR values of `pendsm` and `gam` are higher compared to the other settings. This leads to the conclusion that variable selection performs much worse in the case of correlated variables.

	ML_{ridge}	pendsm	gam
MSE	28.20 (27.37)	0.15 (0.07)	0.22 (0.11)
MSE_{vary}	37.31 (38.47)	0.42 (0.23)	0.57 (0.27)
MSE_{const}	24.15 (27.56)	0.04 (0.02)	0.07 (0.06)
MSE^*	-	-4.86	-4.49
MSE^*_{vary}	-	-4.11	-3.76
MSE^*_{const}	-	-6.04	-5.65
D_{pred}	3192.03 (701.13)	1640.77 (80.54)	1647.16 (87.01)
D^*_{pred}	-	-0.64	-0.64
FPR	1	0.65	0.56
FNR	0	0.00	0.00

Table 3.3. Results for Setting 2 for the estimated mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE^*, MSE^*_{vary}, MSE^*_{const}), the predictive deviances D_{pred}, D^*_{pred} referring to test data and the false positive rate (FPR) as well as the false negative rate (FNR) for the ML_{ridge}, `pendsm` and `gam`. The displayed values represent the means over all simulation runs. Estimated standard errors are given in parentheses.



(a) Boxplots of the mean squared errors for `pendsm*` and `gam*` regarding all coefficients, truly time-varying and truly time-constant coefficients.

(b) Boxplots of the predictive deviance for `pendsm*` and `gam*`.

Figure 3.3. Boxplots of the mean squared errors and the predictive deviance for Setting 2

	ML_{ridge}	pendsm	gam
MSE	5.87 (3.06)	0.34 (0.13)	0.80 (1.14)
MSE_{vary}	8.16 (5.11)	0.56 (0.21)	1.22 (1.78)
MSE_{const}	3.01 (2.59)	0.06 (0.05)	0.29 (0.49)
MSE^*	-	-2.74	-2.07
MSE^*_{vary}	-	-2.54	-1.55
MSE^*_{const}	-	-4.12	-2.76
D_{pred}	1546.15 (405.13)	730.40 (51.82)	862.70 (194.54)
D^*_{pred}	-	-0.72	-0.57
FPR	1	0.75	0.62
FNR	0	0.03	0.05

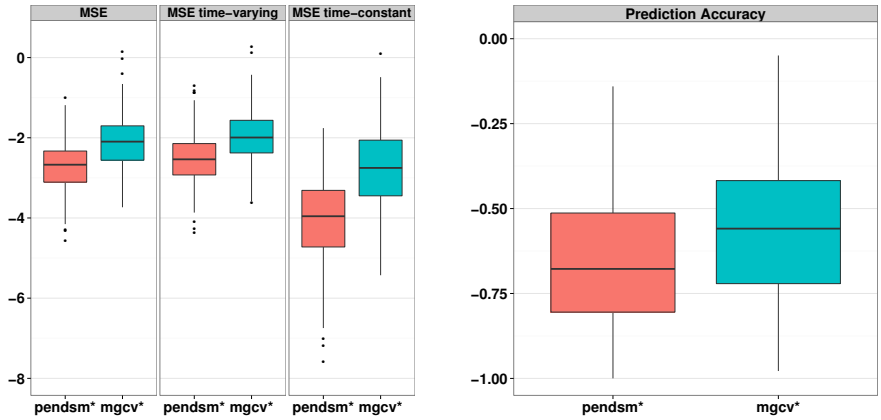
Table 3.4. Results for Setting 3 for the estimated mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE^*, MSE^*_{vary}, MSE^*_{const}), the predictive deviances D_{pred}, D^*_{pred} referring to test data and the false positive rate (FPR) as well as the false negative rate (FNR) for the ML_{ridge}, `pendsm` and `gam`. The displayed values represent the means over all simulation runs. Estimated standard errors are given in parentheses.

In addition, the boxplots of the log ratios $\log(MSE(\cdot)/MSE(ML_{ridge}))$ related to `pendsm` outperforms the boxplots related to `gam` (Figure 3.4a). Especially, for the mean squared errors resulting from the true time-constant parameters `pendsm` yield extremely better results than `gam`. Additionally, the boxplots assessing the predictive performance by the log ratios $\log(D_{pred}(\cdot)/D_{pred}(ML_{ridge}))$ show that the boxplot of `pendsm` outperforms the boxplot of `gam` (Figure 3.4b). Summarizing the outcomes of this setting, `pendsm` performs well in the case of correlated covariates for the mean squared errors and the predictive deviance. However, for both methods `pendsm` and `gam`, the performance assessing the variable selection yields bad results.

Setting 4

The employed penalty of the last setting indicates piecewise time-constant coefficients by employing a $L1$-type penalty. However, the penalty regarding the time-varying intercept constitutes a ridge penalty of the differences between coefficients of adjacent B-spline parameters. The mean squared error of the ML-procedure is $3.13{\cdot}10^{27}$ $(3.13 \cdot 10^{28})$ and the $L1$-type penalty achieves a clear improvement. This can be seen in the summary of the results (Table 3.5). As the `gam` procedure only deals with quadratic penalties, this setting cannot exactly be implemented using the `gam` function of `mgcv`. To obtain a comparative method, the features of `gam` are conformed to the `pendsm` preferably equivalent, except the penalty itself. The illustrated results serve only as exemplification. The MSE values exhibit that `pendsm` with $L1$-type penalty performs better than `gam`, whereas, as expected, both `pendsm` with $L1$-type penalty and `gam` outperform ML_{ridge}.

Also the boxplots of the log ratios $\log(MSE(\cdot)/MSE(ML_{ridge}))$ confirm that the $L1$-penalty of `pendsm` is more appropriate than the $L2$-penalty regarded in `gam` (Figure 3.5a). The illustration of the predictive performance referring to the ratios $\log(D_{pred}(\cdot)/D_{pred}(ML_{ridge}))$ shows that `pendsm` yields slightly better results than `gam` (Figure 3.5b).



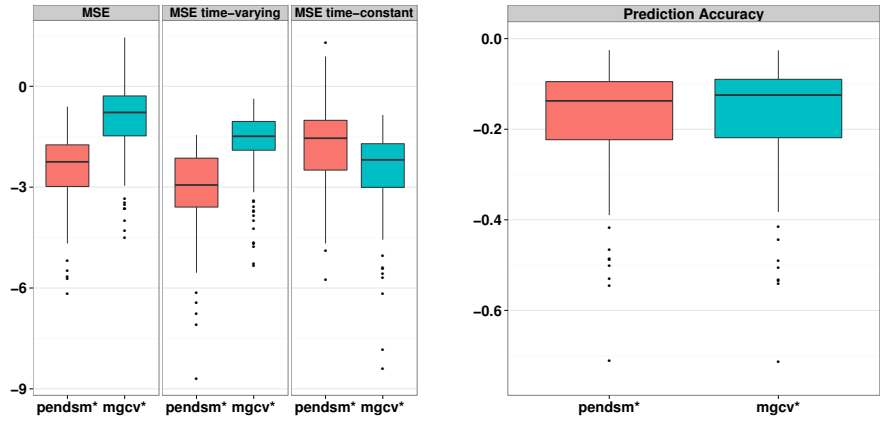
(a) Boxplots of the mean squared errors for `pendsm*` and `gam*` regarding all coefficients, truly time-varying and truly time-constant coefficients.

(b) Boxplots of the predictive deviance for `pendsm*` and `gam*`.

Figure 3.4. Boxplots of the mean squared errors and the predictive deviance for Setting 3

	ML_{ridge}	`pendsm`	`gam`
MSE	9.28 (20.66)	0.38 (0.21)	0.69 (0.28)
MSE_{vary}	11.08 (24.48)	0.85 (0.59)	1.53 (0.45)
MSE_{const}	8.56 (21.51)	0.19 (0.14)	0.35 (0.33)
MSE^*	-	-2.42	-1.74
MSE^*_{vary}	-	-1.79	-1.04
MSE^*_{const}	-	-3.16	-2.56
D_{pred}	1194.40 (246.54)	966.37 (41.69)	973.35 (46.25)
D^*_{pred}	-	-0.20	-0.19
FPR	1	0.49	0.55
FNR	0	0.12	0.11

Table 3.5. Results for Setting 4 for the estimated mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE^*, MSE^*_{vary}, MSE^*_{const}), the predictive deviances D_{pred}, D^*_{pred} referring to test data and the false positive rate (FPR) as well as the false negative rate (FNR) for the ML_{ridge}, `pendsm` and `gam`. The displayed values represent the means over all simulation runs. Estimated standard errors are given in parentheses.



(a) Boxplots of the mean squared errors for `pendsm*` and `gam*` regarding all coefficients, truly time-varying and truly time-constant coefficients.

(b) Boxplots of the predictive deviance for `pendsm*` and `gam*`.

Figure 3.5. Boxplots of the mean squared errors and the predictive deviance for Setting 4

3.5. Applications

In this section, `pendsm` is applied to two real data examples. The first data set depicts the Munich founder study, whereas the second describes a fertility study. For comparison, the `gam` function is applied to the data examples as well.
By incorporating time-varying covariate effects, these two data examples require a large number of parameters. These parameters can be reduced by using basis expansions for the time-varying coefficients. However, to use a penalty term that allows for piecewise time-constant time-varying coefficients, like penalty term 3.9, for each time period and each covariate effect a single parameter has to be estimated. Due to this large number of parameters, heavily computational problems occur when fitting these two data examples. Therefore, the application of a penalty allowing for piecewise time-constant time-varying coefficients is not possible in this section.

3.5.1. The Munich Founder Study

The practical use of `pendsm` is illustrated by considering the Munich founder study. A detailed description of the data can be found in Brüderl et al. (1992). Therein, the survival of newly founded firms in the area of Munich and Upper Bavaria is investigated. In the local Chamber of Commerce, 28646 business registrations were listed in 1985-1986. From this total number, a stratified random sample of about 6000 companies was drawn. In 1990, 1849 business founders were interviewed. For the analysis, only the complete cases that means, observations with no missing values for any covariate are used, resulting in 1224 observations.
The dependent variable defines the transition process of a newly founded company up to

Variable	Description
Time	Time (in quarters) until insolvency of a company (effect modifier)
Sector	Economic sector 0: industry, manufacturing and building sector, 1: commerce, 2: service industry
Legal	Legal form 0: Small, 1: Partnership
Seed	Seed capital 0: ≤ 25000 , 1: > 25000
Equity	Equity capital 0: no, 1: yes
Debt	Debt capital 0: no, 1: yes
Market	Target market 0: local, 1: national
Clientele	Clientele 0: wide spread, 1: small amount of important customer, one important customer
Degree	Education degree 0: no A-levels, 1: A-Levels
Gender	Gender 0: female, 1: male
Experience	Professional experience 0: <10 years,1: $\geq$ 10 years
Employees	Number of employees excluding the company founders 0: 0 or 1, 1: >2
Age	Age of the founder at formation of the company, centered around 43 (sample mean: 43.22)

Table 3.6. Description of the variables used for the Munich founder study.

insolvency, denoting the event. Thereby, the duration time until insolvency is measured in quarters, where a maximum of 22 quarters can be reached. A company that was still "alive" at the time of the registration of the interview is treated as right-censored. Based on the results of the analysis of (Brüderl et al., 1992), only a part of the covariates of the Munich founder study are incorporated in the model. Moreover, to reduce the number of categories of some variables, several categories were combined. In Table 3.6, an overview of the employed variables is given.

The data were reorganized according to Section 2.2.2, to conduct a binary regression model with complementary log-log link corresponding to a discrete-time survival model. The long format of the data consists of $\sum_{i=1}^{n} t_i = 17736$ rows. To be on comparable scales, all covariates are standardized to have equal variance in order to avoid that coefficient values are scale dependent. For company i and measurement at quarter t, the considered model has the form

$$\begin{aligned}\eta_{it} = \beta_{0t} &+ \text{Sector}_{it}^{(1)}\beta_{1t} + \text{Sector}_{it}^{(2)}\beta_{2t} + \text{Legal}_{it}\beta_{3t} + \text{Seed}_{it}\beta_{4t} + \text{Equity}_{it}\beta_{5t} + \text{Debt}_{it}\beta_{6t} \\ &+ \text{Market}_{it}\beta_{7t} + \text{Clientele}_{it}\beta_{8t} + \text{Degree}_{it}\beta_{9t} + \text{Gender}_{it}\beta_{10t} \\ &+ \text{Experience}_{it}\beta_{11,t} + \text{Employees}_{it}\beta_{12,t} + \text{Age}_{it}\beta_{13,t}.\end{aligned}$$

In Table 3.7, the frequencies of the dependent variable versus the quarters are illustrated. It can be seen that in the first three years after establishing of the companies no censoring had occurred. Moreover, with increasing number of quarters the frequency of failure decreases.

y \ quarter	1	2	3	4	5	6	7	8	9	10	11
0	0	0	0	0	0	0	0	0	0	0	0
1	17	28	30	45	26	22	28	29	21	15	24

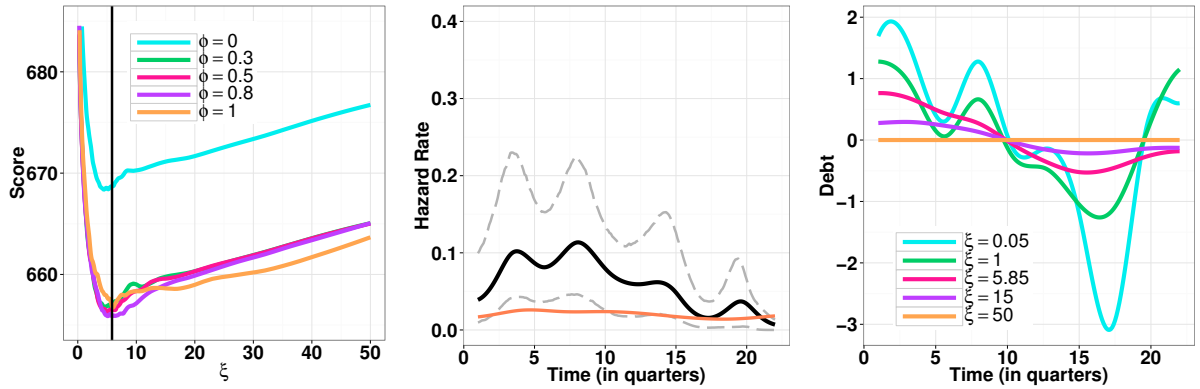
y \ quarter	12	13	14	15	16	17	18	19	20	21	22
0	0	19	109	110	99	119	119	77	101	96	16
1	15	11	21	7	7	4	0	6	2	1	0

Table 3.7. Frequencies for the dependent variable y and the quarters.

For all covariate effects cubic B-splines are used, that means $\beta_{jt} = \sum_m \alpha_{jm} B_{jm}(t)$, $j = 0, 1, ..., 13$, with 10 equidistant inner knots resulting in 12 basis functions. The used penalty is given by

$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{12} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi \left(\phi \sum_{j=1}^{13} \psi_j \left\| \boldsymbol{\zeta}_j \right\|_2 + (1-\phi) \sum_{j=1}^{13} \varphi_j \left\| \boldsymbol{\alpha}_j \right\|_2 \right),$$

where adaptive weights (3.10) for ψ_j and φ_j are employed. The penalty allows for smooth or constant covariate effects or their selection from the model. For estimation, the tuning parameter regarding the penalization of the time-varying intercept ξ_0 is set to 0.001. For selection of the tuning parameters ξ and ϕ, a 5-fold cross-validation, based on the predictive deviance, is conducted. In Figure 3.6a, the corresponding scores are illustrated. The vertical black line determines the chosen tuning parameters with ξ set to 5.85 and ϕ set to 0.8. The run of all curves indicates that penalization clearly improves ordinary ML-estimation nearly obtained for $\xi = 0$ as ξ_0 is set to the very small value of 0.001.
For a comparison of the results, a generalized additive model using the function `gam` of R add-on package `mgcv` (Wood, 2006) is applied to the Munich founder study. For this purpose, for the time-varying intercept a slight ridge penalty with tuning parameter 0.001 is used, whereas the time-varying covariate effects are estimated by cubic B-splines with penalized first differences between the parameters of the smooth functions and 10 equidistant inner knots. Moreover, the option `select` is set to `TRUE`, adding a penalty to each smooth, to allow it to be penalized out of the model. However, as time-varying categorical covariates are incorporated as smooth effects in modeling, in addition to the interaction of time and the categorical covariate, `gam` enforces the inclusion of the corresponding main effect (see Section 3.2 and Wood, 2006). This is due to internal centering constraints of `gam`. For example, the interaction *Legal:Time* is incorporated as a smooth time-varying effect in the model. As *Legal* is categorical, the main effect of *Legal* has to be incorporated as well. Thus, in the case of time-varying categorical covariates, `gam` can only remove the smooth terms from the model. The constant terms still remain in the model as `gam` cannot perform variable selection with respect to parametric terms. For metric covariates like *Age*, no main effect has to be incorporated and the smooth time-varying effect of *Age* can be excluded from the model by `gam`. Hence, in the context of time-varying coefficients, `gam` can



(a) Cross-validation scores subject to penalty parameter ξ

(b) Estimated hazard rate for all predictors at reference. Dashed lines represent the pointwise 95% bootstrap confidence interval. The red line mark the `gam` estimate.

(c) Time-varying effect of the covariate *Debt* subject to different tuning parameters ξ and fixed $\phi = 0.8$.

Figure 3.6. Plots corresponding to the Munich founder study

only perform true variable selection for metric covariates. The parametrization of `pendsm` allows for variable selection of any measurement scale of the covariates.

Figure 3.6b shows the resulting hazard rate of the fitted model when all covariate characteristics are set at reference. That is, a female founder at the age of 43 whose firm has the following characteristics: industry, manufacturing and building sector, small legal form, seed capital $\leq$ 25000, with equity capital, with debt capital, local target markets, wide spread clientele, no A-levels, professional experience $<$ 10 years and 0 or 1 employees. The probability of becoming insolvent tends to increase in the first 8 quarters, but from quarter 9 on the risk of failure has a decreasing tendency over time. However, this hazard rate shows some variation. Maybe, this variation stems from seasonal fluctuation. This issue could be investigated in a further time series analysis. The values of the corresponding hazard function resulting from the `gam` function (red line) are almost completely below the hazard rate resulting from `pendsm` and shows no large variation. The dashed lines specify the pointwise 95% confidence interval based on 1000 bootstrap replications. See Section 3.3 for more details.

In the context of penalization methods, it is common to plot coefficient paths subject to the tuning parameter ξ. For time-varying coefficients, the illustration of such coefficients paths would be too complex and is not meaningful. To illustrate the impact of the tuning parameter ξ on the smoothness of a time-varying coefficient, exemplary the effect of the variable *Debt* is plotted in Figure 3.6c with five different values of ξ. Thereby, tuning parameter ϕ is fixed to 0.8, the resulting value of the cross-validation. With only a slight penalization of $\xi = 0.05$ the effect of *Debt* varies rather erratically over time. Moreover, the impression of a rather decreasing effect of *Debt* is provided. By introducing a penalty of $\xi = 1$ the changes in the effect of *Debt* are shrunk, making the function rather smooth. For a penalty of $\xi = 5.85$, that is equivalent to the cross-validated tuning parameter ξ,

the curve is smoother. Increasing the penalization to ξ=15 let the effect of *Debt* almost vanishing. Finally, a penalty of $\xi = 50$ is large enough to set the whole effect to zero. That means, the effect of *Debt* is removed from the model.

In Figure 3.7, the estimates $\hat{\beta}_{jt}$ resulting from the model with $\xi = 5.85$ and $\phi = 0.8$ are summarized. The solid black lines denote the parameter estimates, whereas the dashed lines specify the corresponding 95% confidence intervals. The intervals are based on 1000 bootstrap replications and are computed pointwise. Moreover, the solid red lines define the estimates resulting from the `gam` function. Both procedures come to the conclusion that *Legal*, *Market*, *Clientele*, *Degree* and *Experience* have a linear effect in the predictor, whereas *Age* is removed from the model. By using `pendsm`, it is suggested that the predictor referring to the sector commerce has a time-varying effect, and the effect of the variable *Employees* is estimated time-constant. In contrast, for `gam`, the situation is vice versa. In general, the course of the curves of the time-varying coefficients follows a similar pattern.

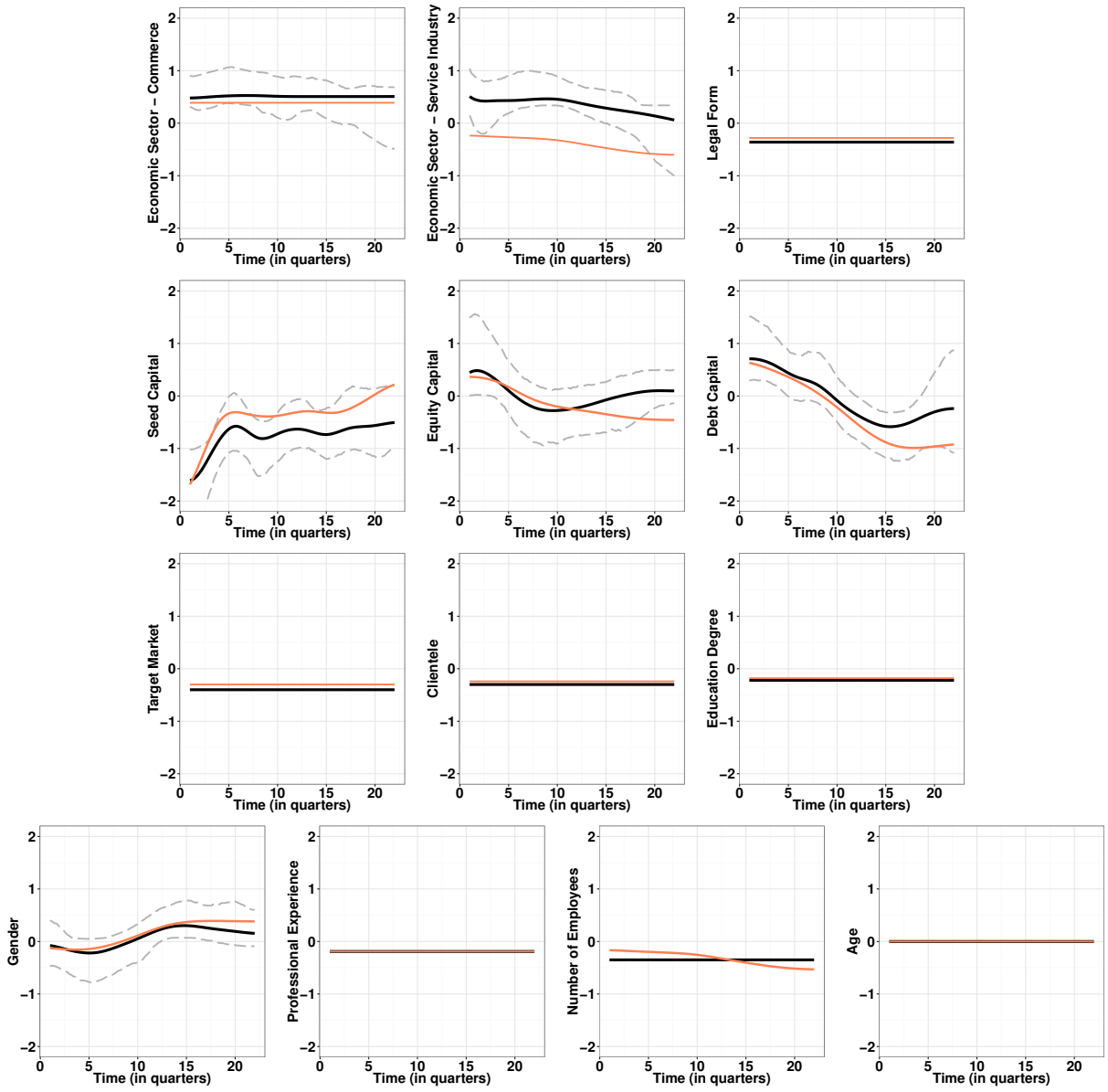


Figure 3.7. Estimates of `pendsm` for the Munich founder study using cubic B-splines. Black lines mark the estimates resulting from `pendsm`, whereas red lines stand for the estimates from `gam`. Dashed lines represent pointwise 95% bootstrap confidence intervals.

Variable	Description
Time	Time (in years) until pregnancy (effect modifier)
Job	Labour Status 0: unemployed/seeking work/housewife, 1: full-time/self-employed, 2: marginal/part-time employed, 3: school, 4: no info
Education	Educational attainment 0: school leaver/(general) certificate of secondary education (low-level), 1: A-levels/apprenticeship (A-levels), 2: polytechnic degree/university degree/PhD (university), 3: no info
Relationship	Relationship Status 0: single, 1: cohabit, 2: married
Siblings	Number of siblings 0: no, 1: yes
ClassParents	Parents educational attainment 0: school leaver/(general) certificate of secondary education (low-level), 1: A-levels/apprenticeship (A-levels), 2: polytechnic degree/university degree/PhD (university), 3: no info
Cohort	Birth cohort 0: 1971-1973, 1: 1981-1983, 2: 1991-1993

Table 3.8. Description of the variables used for the fertility study. Employed labels are indicated in brackets.

3.5.2. Fertility Study

The fertility study investigates whether labor force participation of women influences the transition to motherhood. The underlying data are used in Abedieh (2013). Therein, it is aimed to validate the study of Schröder and Brüderl (2008). In Abedieh (2013), the data were extracted from *pairfam* (Nauck et al., 2012), a multi-disciplinary longitudinal study analyzing cooperative and life forms of families in Germany.
The panel has started in 2008 and is based on 12000 randomly chosen observations of the birth cohorts 1971-73, 1981-83 and 1991-1993. Three survey waves (wave 1: 2008/2009, wave 2: 2009/2010 and wave 3: 2010/2011) are incorporated in the data. In the meantime, a fourth wave is available. The dependent variable is the transition to pregnancy with duration time (in years) until pregnancy. For modeling the duration time, the start of the observation process is set to 14 years. The maximum value is 27 years until pregnancy. Furthermore, several covariates are included in the model. An essential variable is the time-varying covariate *Job*. It describes the employment status of the women in the study and originally consists of 23 parameter values, but it is summarized to four categories. The variable *Education* contains the highest degree of a woman during the study. The current relationship status is defined in *Relationship*. In *Siblings*, it is indicated if a woman has siblings or not. By means of the covariate *ClassParents*, the educational attainment of the parents is captured. Thereby, the value is based on the higher degree of one parent. Finally, *Cohort* describes the birth cohort. To sum up, in this application, the covariates *Job* and *Relationship* are time-varying, that is, the covariate values per object may vary over the observed period. For more details to the underlying data see Abedieh (2013). An overview is shown in Table 3.8.

The used data set consists of 2468 observations. Restructuring of the data was executed according to Section 2.2.2, to conduct a binary regression model with complementary log-

log link corresponding to a model for discrete duration time. The long format of the data contain $\sum_{i=1}^{n} t_i = 34601$ rows. To be on comparable scales, all covariates are standardized to have equal variance. For duration year t of individual i, the considered model has the form

$$\begin{aligned}\eta_{it} = &\beta_{0t} + \text{Job}_{it}^{(1)}\beta_{1t} + \text{Job}_{it}^{(2)}\beta_{2t} + \text{Job}_{it}^{(3)}\beta_{3t} + \text{Job}_{it}^{(4)}\beta_{4t} + \text{Education}_{it}^{(1)}\beta_{5t} \\ &+\text{Education}_{it}^{(2)}\beta_{6t} + \text{Relationship}_{it}^{(1)}\beta_{7t} + \text{Relationship}_{it}^{(2)}\beta_{8t} + \text{Siblings}_{it}\beta_{9t} \\ &+\text{ClassParents}_{it}^{(1)}\beta_{10t} + \text{ClassParents}_{it}^{(2)}\beta_{11t} + \text{ClassParents}_{it}^{(3)}\beta_{12t} \\ &+\text{Cohort}_{it}^{(1)}\gamma_{11} + \text{Cohort}_{it}^{(2)}\gamma_{12},\end{aligned}$$

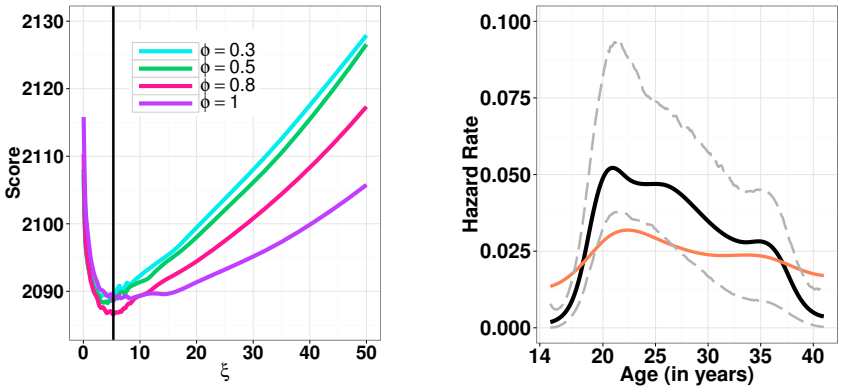
where all covariates, except the covariate *cohort*, are incorporated as time-varying effects using cubic B-splines. That means $\beta_{jt} = \sum_m \alpha_{jm} B_{jm}(t)$, $j = 0, 1, ..., 12$, with 8 equidistant inner knots resulting in 10 basis functions. The employed penalty is given by

$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{10} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi \left(\phi \sum_{j=1}^{12} \psi_j \left\| \boldsymbol{\zeta}_j \right\|_2 + (1-\phi) \sum_{j=1}^{12} \varphi_j \left\| \boldsymbol{\alpha}_j \right\|_2 + \sqrt{\gamma_{11}^2 + \gamma_{12}^2} \right)$$

and it allows for smooth or constant time-varying covariate effects or their selection from the model, whereas for the covariate *cohort* a group lasso penalty is used. This means that the coefficients of the covariate *cohort* can be shrunk simultaneously until the variable, including all categories, is removed from the model. Moreover, adaptive weights (3.10) are employed for estimation. Tuning parameter ξ_0 is set to 0.001 and tuning parameters ξ and ϕ are chosen by 5-fold cross-validation with the predictive deviance as loss criterion. They are set to 5.25 and 0.8, respectively. The corresponding cross-validation scores are shown in Figure 3.8a, where the vertical black line marks the chosen tuning parameters ξ and ϕ. Thereby, the score referring to $\phi = 0$, meaning that the weight of the penalty was completely assigned to the selection part, is omitted. This was done due to a heavy erratically run of the score curve with extreme peaks. The run of the presented curves indicates that penalization clearly improves ordinary ML-estimates nearly obtained for $\xi = 0$ as ξ_0 is set to the very small value of 0.001.

For a comparison of the results, a generalized additive model using the function `gam` of R add-on package `mgcv` (Wood, 2006) is applied as well. For this purpose, for the time-varying intercept a slight ridge penalty with tuning parameter 0.001 is used, whereas the time-varying covariate effects are estimated by cubic B-splines with penalized first differences between the parameters of the smooth functions and 10 equidistant inner knots. Moreover, the option `select` is set to `TRUE`, adding a penalty to each smooth, to allow it to be penalized out of the model. The covariate *cohort* is included unpenalized. As stated in Section 3.5.1, in the context of smooth time-varying coefficients, `gam` can only perform complete variable selection for metric covariates and not for categorical covariates.

Resulting plots of the estimated coefficients can be found in Figures 3.8b and 3.9 and Table 3.9. In the figures, the labeling of the abscissa is adapted to the real age of the women,



(a) Cross-validation score subject to penalty parameter ξ

(b) Estimated hazard rate for all predictors at reference. Dashed lines represent the pointwise 95% bootstrap confidence interval. The red line mark the `gam` estimate.

Figure 3.8. Plots corresponding to the fertility study.

where the observation period starts at the age of 14. Solid lines denote the parameter estimates, whereas the dashed lines specify the 95% confidence intervals. The intervals are based on 1000 bootstrap replications and are computed pointwise (see Section 3.3). Moreover the solid red lines mark the estimates resulting from the `gam` function. As only categorical covariates are incorporated in the model, `gam` cannot set any predictor to zero. Figure 3.8b shows the resulting hazard rate of the fitted model when all covariate characteristics are set at reference. That is, an unemployed single woman born between 1971-1973 with low-level education, no siblings and low-level education of the parents. The effect of the hazard rate is as expected: Initially, the curve ascends strongly but then descends as it approaches the early twenties and, finally, steeps from the age of about 36. The course resulting from `gam` is similar but is clearly flatter.

The plots for the remaining time-varying coefficients are depicted in Figure 3.9. `pendsm` set the coefficients belonging to the part-time labour status and *Siblings* to zero. For *Siblings*, `gam` suggests a very small constant effect of 0.006 as it cannot be removed from the model. Several covariate effects are considered to be time-constant by `pendsm` and `gam`. It can be seen, that the course of the curves is quite different for single effects. For example, the coefficient for *Relationship - Married* is thoroughly positive for `pendsm` and thoroughly negative for `gam`. However, being married should rather raise the probability of getting pregnant. Hence, the outcome of `pendsm` is more meaningful than that of `gam`. The same behavior can be recognized for further predictors.

	pendsm				gam	
			95%-CI			
	estimate	standard error	lower	upper	estimate	standard error
γ_{11} (cohort 1981-83)	-0.061	0.045	-0.164	0.000	-0.301	0.167
γ_{12} (cohort 1991-93)	-0.327	0.327	-1.169	0.000	-0.612	0.283

Table 3.9. Parameter estimates of *cohort* of the fertility study.

The predictor *cohort* was incorporated as a time-constant effect in the model and the obtained estimates as well as the standard errors and the 95% confidence intervals are shown in Table 3.9. The impact on the fertility of cohort 1981-83 is hardly smaller than for cohort 1971-73, whereas the effect of cohort 1991-93, given by $\gamma_{12} = -0.33$, is considerably stronger. Compared to the estimates of `gam`, the estimates obtained by `pendsm` are smaller due to the incorporated penalization technique. The results of *cohort* agree with the statement that birth rates have been decreased over the last decades in Germany.

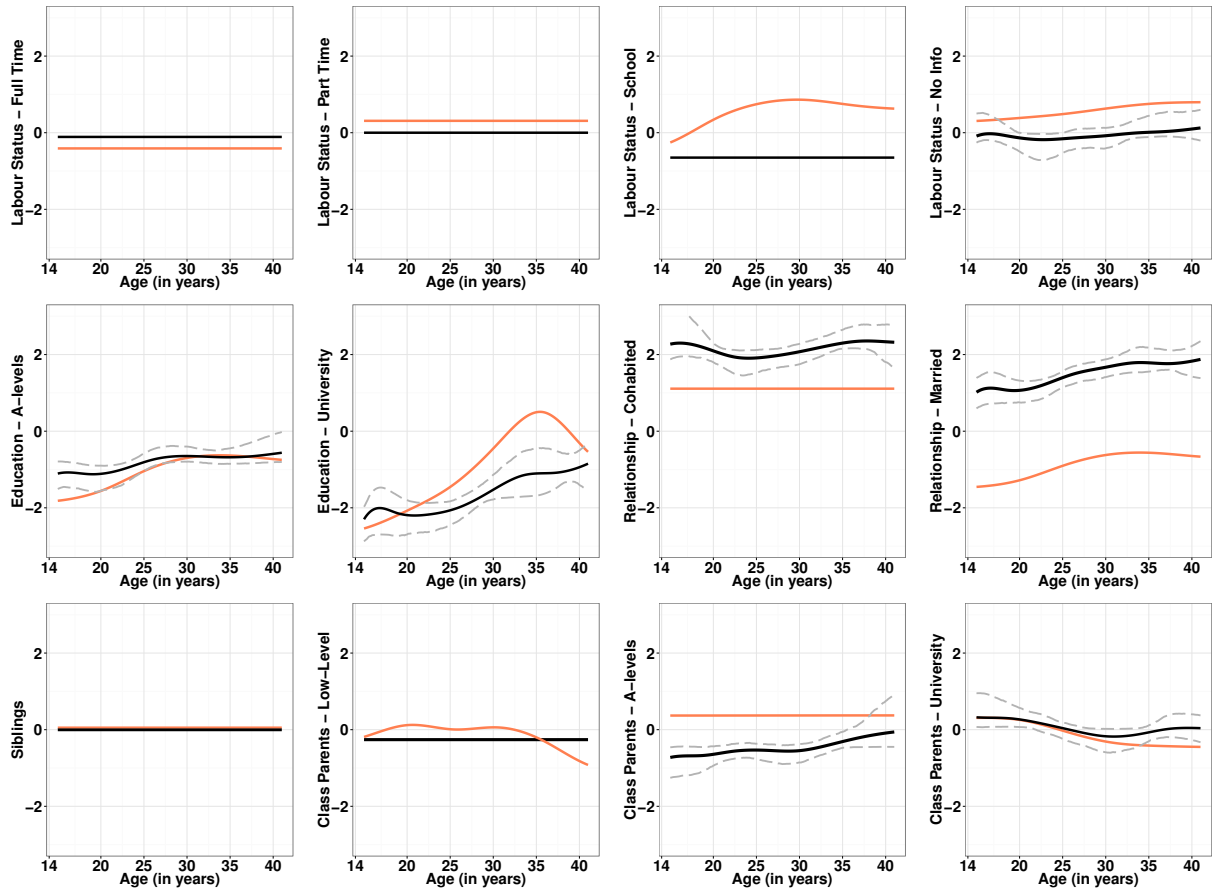


Figure 3.9. Estimates of `pendsm` for the time-varying coefficients of the fertility study using cubic B-splines. Black lines mark the estimates resulting from `pendsm`, whereas red lines stand for the estimates from `gam`. Dashed lines represent pointwise 95% bootstrap confidence intervals.

3.6. Concluding Remarks

In this chapter, discrete-time survival models with time-varying coefficients are investigated. Due to the fact, that the incorporation of time-varying coefficients leads to many parameters and possibly overfitting, a regularization method is provided. This penalization method `pendsm` allows for different types of penalties and even a combination of them can be employed. One of the main benefits of this approach is the predominantly smooth temporal variation of time-varying covariate effects. Furthermore, the proposed method performs variable selection leading to interpretable and parsimonious models. Hence, the resulting procedure is considerably flexible and can be applied to a variety of applications. The challenge of estimation, in the case where the model incorporates different types of penalties, is solved by a local quadratic approximation for the penalties. This issue is based on ideas of Fan and Li (2001), Ulbricht (2010) and Oelker and Tutz (2013). The approximation is updated in a PIRLS algorithm. The existing `gvcm.cat` R add-on package is modified and extended to apply the proposed method to discrete-time survival data. The computation of standard errors and confidence intervals are conducted using bootstrap methods.
A simulation study was conducted to assess the performance of the proposed method `pendsm`. Therein, `pendsm` outperforms the existing `gam` function of the R add-on package `mgcv` (Wood, 2006). However, due to poor results with regard to variable selection for all methods, caution is recommended in the case of correlated variables.
Moreover, in the context of smooth time-varying coefficients, only `pendsm` can perform variable selection for metric and for categorical covariates, whereas `gam` can only remove metric predictors from the model.

4. Choice of Tuning Parameter

In the previous chapter, the predictive deviance is the employed loss criterion for the choice of tuning parameters. It will be substituted by an alternative loss function in this chapter and possible associated model improvements are investigated. Some background on prediction measures is given in Section 4.1. In Section 4.2, several measures of prediction accuracy are presented, whereby well-known measures used for continuous survival outcomes were adopted to discrete failure time analysis. These are useful measures for the choice of tuning parameter ξ. The performance of the shown alternative loss functions is investigated by means of a simulation study whose settings and results are provided in Section 4.3. Finally, Section 4.4 sums up the results. In the following, only the notation and explanations with respect to grouped survival times are considered, but they can easily be modified to truly discrete survival times.

4.1. Introduction

The tuning parameter ξ in Chapter 3, steers the strength of penalization and is chosen by cross-validation based on the predictive deviance. In this chapter, only tuning parameter ξ is regarded, whereas tuning parameter ϕ is ignored or is set to 0.5, respectively. The procedure to be undertaken is cross-validation and the predictive deviance constitutes the loss function (Wald, 1950).
In discrete survival analysis, the present data consists of repeated measurements per object for several time periods. The typical long format data structure of discrete time survival models particularly reflects these repeated measurements. For each object i, the predictive accuracy is computed independently for all $t = 1, ..., t_i$ using the predictive deviance. Hence, this loss function ignores the dependence structure of the repeated measurements of the t_i observations belonging to one object. In the context of discrete survival times, alternative measures might be more appropriate for the choice of tuning parameter ξ. Therefore, the performance of further loss functions, that assess the predictive performance and incorporate the observations' dependency, will be investigated.

Duration time models are widely used for predicting the survival of objects. Especially in biomedical research, where the forecast of patient's survival is of particular interest. To assess prognostic performance, statistical measures are needed for evaluating the prediction accuracy of survival models. These measures are intended to form a rational basis for further decisions. Due to the occurrence of censoring, the evaluation of prognostic models is not trivial. Moreover, results of predictive accuracy might be biased due to miss-specification

of the model. Consequently, universally valid evaluation criteria that are independent of the underlying model have to be used.

The topic of evaluating the prognostic performance of prediction for continuous survival outcomes is frequently discussed in the literature of recent methodology, wherein it is aimed to overcome the problem of biased predictions in the presence of censored observations (Schemper and Stare, 1996). For this purpose, various new approaches have been suggested that can be classified into three groups (Schmid et al., 2011):

- likelihood-based approaches (Kent and O'Quigley, 1988; Nagelkerke, 1991; Xu and O'Quigley, 1999; O'Quigley et al., 2005)
- ROC-based approaches (Heagerty et al., 2000; Heagerty and Zheng, 2005; Cai et al., 2006; Uno et al., 2007; Pepe et al., 2008)
- distance-based approaches (Korn and Simon, 1990; Graf et al., 1999; Schemper and Henderson, 2000; Gerds and Schumacher, 2006, 2007; Schoop et al., 2008).

The likelihood-based approaches often set the log-likelihood of a prediction model in relation to the corresponding log-likelihood obtained from a null model where no covariate information is incorporated. When utilizing ROC-based approaches the dependent variable is considered as binary, as for each time t the outcomes "event" or "no event" may occur. This entails the adaption of established concepts for the evaluation of binary classification rules to time-to-event data. In distance-based approaches, the prediction error is measured in terms of the distances between predicted and observed survival functions of observations in a sample.

For discrete-time survival analysis such approaches, with respective comparisons, are not well investigated, so far. Hence, in the following, some approaches defined for continuous outcomes are adapted to discrete-time. Furthermore, these measures of prediction accuracy can substitute the predictive deviance in the cross-validation procedure and provide suitable loss functions to determine the tuning parameter ξ.

4.2. Choice of Tuning Parameter ξ

Before the choice of the tuning parameter ξ is investigated, several adequate measures of prediction accuracy are presented. Thereby, likelihood-based, distance-based as well as ROC-based measures are considered. Usually, measures of prediction accuracy are computed based on an i.i.d. test sample using models which were fitted based on an i.i.d. learning sample. In order to avoid any misunderstandings, the superscript $\mathcal{L}$ is used for expressions related to the learning data and $\mathcal{T}$ marks expressions related to the training data.

To remind some notation, some basic concepts of discrete survival analysis are given in the following. Let T denote a non-negative discrete random variable taking values from

$\{1, ..., q\}$. The survival time is represented by discrete time T, where $T = t$ defines an event in interval $[a_{t-1}, a_t)$, $t = 1, ..., q$. Additionally, a p-dimensional time-constant vector $\boldsymbol{x}$ of covariates is given. The conditional survival function given the covariates, is denoted by

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}).$$

For the i-th subject let C_i denote the individual censoring time independent of survival time T_i, that is, random censoring is assumed. Consequently, the observed survival time is defined by $t_i = \min(T_i, C_i)$. The censoring indicator $\delta_i = I(T_i \leq C_i)$ determines whether observation i is right censored ($\delta_i = 0$) or an event occurs ($\delta_i = 1$) with respect to the interval $[a_{t-1}, a_t)$. Finally, the conditional discrete-time hazard rate is given by

$$\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x}).$$

Let y_{it} denote the binary outcome of an object i, $i = 1, ..., n$, in period t, $t = 1, ..., t_i$. Thereby, the binary outcome y_{it} denotes an event in interval $[a_{t-1}, a_t)$ if $y_{it} = 1$ and no event or censoring if $y_{it} = 0$. Thus, the existing data structure is of the form in Section 2.2.2 regarding binary regression models leading to the discrete survival model

$$\lambda(t|\boldsymbol{x}_i) = P(y_{it} = 1|\boldsymbol{x}_i) = F(\beta_{0t} + \boldsymbol{x}_i^T\boldsymbol{\gamma}),$$

where F denotes an appropriate cumulative distribution function and $\boldsymbol{x}_i$ collects time-independent characteristics of the i-th object. In general, it holds $F(\cdot) = 1 - \exp(-\exp(\cdot))$, defining the complementary log-log link.

4.2.1. Measures of Prediction Accuracy

Predictive Deviance

A likelihood-based measure of prediction accuracy is the predictive deviance. It is evaluated on the test data and for discrete survival times it is given by

$$\begin{aligned} \mathrm{D} &= -2\sum_{i=1}^{n^{\mathcal{T}}} \left(\delta_i^{\mathcal{T}} \log\left(\hat{P}^{\mathcal{L}}(T_i^{\mathcal{T}} = t_i^{\mathcal{T}})\right) + (1 - \delta_i^{\mathcal{T}}) \log\left(\hat{P}^{\mathcal{L}}(T_i^{\mathcal{T}} > t_i^{\mathcal{T}})\right)\right) \\ &= -2\sum_{i=1}^{n^{\mathcal{T}}} \sum_{t=1}^{t_i^{\mathcal{T}}} \left(y_{it}^{\mathcal{T}} \log(\hat{\lambda}^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}})) + (1 - y_{it}^{\mathcal{T}}) \log(1 - \hat{\lambda}^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}))\right), \end{aligned}$$

where $y_{i1}, ..., y_{it_i}$ denote the binary transitions over time periods for object i. The expression $\hat{\lambda}^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}))$ indicates that the model for fitting the hazard rate is based on the learning sample but the evaluation is based on the test sample. The predictive deviance is equivalent to the negative log-likelihood of a binomial regression model evaluated on the test data. Therefore, prediction accuracy is large if the given deviance is small. The predictive deviance is the employed loss criterion in Chapter 3. The question that arises in the context

of repeated measurements is whether it makes a difference to base the construction of the cross-validation on the whole observation of an object or on individual data points.

Another likelihood-based prediction measure making use of the predictive deviance is the R^2 coefficient. To facilitate interpretation, it is often convenient to use such a bounded measure that is given by

$$R^2 = \frac{1 - \exp(1/n^{\mathcal{T}}(D - D_0))}{1 - \exp(1/n^{\mathcal{T}} D_0)},$$

where D_0 is the deviance obtained from a "null model", that is the model without any covariates. If the covariates have no impact on the prediction accuracy R^2 is equal to zero and a R^2 value equal to one indicates perfect prediction (Nagelkerke, 1991). In the following, only the predictive deviance among the likelihood-based measures is considered.

Discrete Ranked Probability Score

As the predictive deviance only takes single observations at particular time periods t into account, it seems more compelling to define scoring rules directly in terms of predictive cumulative distribution functions. The estimated survival function for an observation i, $i = 1, ..., n^{\mathcal{T}}$, is given by $\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) = \prod_{j=1}^{t}(1 - \hat{\lambda}^{\mathcal{L}}(j|\boldsymbol{x}_i^{\mathcal{T}})$. A degenerate survival function for an observation $(t_i^{\mathcal{T}}, \delta_i^{\mathcal{T}} = 1, \boldsymbol{x}_i^{\mathcal{T}})$, $i = 1, ..., n^{\mathcal{T}}$, is obtained by the following simple step function

$$S_i^{\mathcal{T}}(t) = \begin{cases} 1, & \text{if } t < t_i^{\mathcal{T}} \text{ and } \delta^{\mathcal{T}} = 1 \\ 0, & \text{if } t \geq t_i^{\mathcal{T}} \text{ and } \delta^{\mathcal{T}} = 1. \end{cases} \tag{4.1}$$

On the other hand, a right-censored observation $(t_i^{\mathcal{T}}, \delta_i^{\mathcal{T}} = 0, \boldsymbol{x}_i^{\mathcal{T}})$ leads to a truncated survival function given by

$$S_i^{\mathcal{T}}(t) = 1, \text{ if } t \leq t_i^{\mathcal{T}} \text{ and } \delta^{\mathcal{T}} = 0. \tag{4.2}$$

A score that measures the discrepancy between the estimated survival function $\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}})$ and the observed survival function $S_i^{\mathcal{T}}(t)$ with respect to discrete duration time, is defined by

$$\widehat{\mathrm{PR}} = \sum_{i=1}^{n^{\mathcal{T}}} \left(\delta_i^{\mathcal{T}} \sum_{t=1}^{q} (\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) - S_i^{\mathcal{T}}(t))^2 + (1 - \delta_i^{\mathcal{T}}) \sum_{t=1}^{t_i} (\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) - S_i^{\mathcal{T}}(t))^2 \right).$$

The discrete ranked probability score measures the quadratic difference between the observed and fitted survival functions. This measure is not based on single observations, but rather on the whole distribution of an object i. Moreover, the discrete ranked probability score is strongly related to the continuous ranked probability score (Gneiting and Raftery, 2007). The continuous ranked probability score, which lately has attracted much attention, enjoys appealing properties and serves as a standard score for evaluating probabilistic and distributional forecasts of real-valued variables. As PR constitutes a distance-based

approach, the lower the PR the better is the prediction performance. A simple requirement regarding distance-based prediction measures is the existence of estimates of the survival function $S(t|\boldsymbol{x}_i)$ for each time period t, $t = 1, ..., q$.

Brier Score

A very traditional measure to assess prediction accuracy is the Brier score (Brier, 1950), originally developed for judging the inaccuracy of probabilistic weather forecasts. It is applicable to tasks in which predictions must assign probabilities to a set of mutually exclusive discrete outcomes. In case of discrete-time survival analysis, the set of possible outcomes is binary. The Brier score defines the mean squared deviation of a binary outcome y_{it}, $i = 1, .., n^{\mathcal{T}}$, and the corresponding fitted probability for a given time period t, $t = 1, .., q$. Adapted to discrete survival times the Brier score is given by

$$\begin{aligned}\widehat{\text{BR}} = \sum_{i=1}^{n^{\mathcal{T}}} \Bigg(& \delta_i^{\mathcal{T}} ((1 - \hat{P}^{\mathcal{L}}(T_i^{\mathcal{T}} = t_i^{\mathcal{T}} | \boldsymbol{x}_i^{\mathcal{T}}))^2 + \sum_{t=1}^{t_i^{\mathcal{T}}-1} \hat{P}^{\mathcal{L}}(T_i^{\mathcal{T}} = t | \boldsymbol{x}_i^{\mathcal{T}})^2) \\ & + (1 - \delta_i^{\mathcal{T}}) \sum_{t=1}^{t_i} \hat{P}^{\mathcal{L}}(T_i^{\mathcal{T}} = t | \boldsymbol{x}_i^{\mathcal{T}})^2 \Bigg).\end{aligned} \tag{4.3}$$

The Brier score can have values between zero and one and the lower the Brier score is for a set of predictions, the better the predictions are calibrated. Nevertheless, the Brier score (4.3) does not take into account the whole distribution of an observation i, $i = 1, ..., n^{\mathcal{T}}$, but only the single values of observation i at time t, $t = 1, ..., t_i^{\mathcal{T}}$. Hence, a modification of the Brier score based on inverse probability of censoring weights (IPCW) is suggested.

Modified Brier Score based on IPCW

The formula of the original Brier Score can be applied to the distance between survival functions. This approach takes the whole distribution into account by incorporating the squared distance between the predicted survival functions $\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) = \prod_{j=1}^{t}(1 - \hat{\lambda}^{\mathcal{L}}(j|\boldsymbol{x}_i^{\mathcal{T}})$, $i = 1, ..., n^{\mathcal{T}}$, and the corresponding observed survival functions $S_i^{\mathcal{T}}(t))$, defined in (4.1) and (4.2). However, the resulting score is not robust against model misspecification, why particular weights are included in the formula. Regarding discrete survival times, this leads to the modified Brier score

$$\begin{aligned}\widehat{\text{BS}}(t) &= \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} w_i^{\mathcal{T}} \left(\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) - S_i^{\mathcal{T}}(t) \right)^2 \\ &= \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \left[\frac{\delta_i^{\mathcal{T}}(1 - S_i^{\mathcal{T}}(t))}{\hat{G}^{\mathcal{L}}(t_{i-1}^{\mathcal{T}}|\boldsymbol{x}_i^{\mathcal{T}})} + \frac{S_i^{\mathcal{T}}(t)}{\hat{G}^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}})} \right] \left(\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) - S_i^{\mathcal{T}}(t) \right)^2,\end{aligned}$$

where $\hat{G}^{\mathcal{L}}(\cdot|\boldsymbol{x}_i^{\mathcal{T}})$ denotes the survival function of the censoring process estimated from the learning data. Section 4.2.2 deals with the choice of the underlying model determining the

survival function G. The weights $w_i^{\mathcal{T}}$, $i = 1, ..., n^{\mathcal{T}}$, account for the inverse probability that an observation in the test data is censored at t and thus ensure the consistency of $\text{BS}(t)$ (Gerds and Schumacher, 2006). The underlying assumption to achieve consistency of the estimator is random censoring.
Corresponding to the definition, the Brier score becomes small if the predicted survival functions agree closely with the observed survival functions. It can further be shown that the Brier score reaches its minimum if $\hat{S}$ is equal to the true survival function (Gneiting and Raftery, 2007). A corresponding time-independent coefficient is given by the integrated Brier score $\widehat{\text{BS}} = \sum_t \widehat{\text{BS}}(t) \cdot \hat{P}^{\mathcal{L}}(T = t)$. Generally, the integrated Brier score of well-predicting models should be smaller than 0.25 which is the integrated Brier Score obtained from the non-informative model with $\hat{S}(t) = 0.5$, for all t.

Modified Schemper-Henderson Estimator based on IPCW

The Schemper-Henderson estimator originally was introduced by Schemper and Henderson (2000). In contrast to the Brier score, the Schemper-Henderson estimator is based on the absolute deviations between survival functions instead of quadratic deviations. However, as the estimator of Schemper-Henderson is distance-based it is not robust against model misspecification. In analogy to the modified Brier score, this problem is solved by incorporating weights. This leads to the modified Schemper-Henderson estimator, proposed by Schmid et al. (2011). For discrete survival times, it is defined via the absolute distance between the predicted survival functions $\hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) = \prod_{j=1}^{t}(1 - \hat{\lambda}^{\mathcal{L}}(j|\boldsymbol{x}_i^{\mathcal{T}})$, $i = 1, ..., n^{\mathcal{T}}$, and the corresponding observed survival functions $S_i^{\mathcal{T}}(t)$, defined in (4.1) and (4.2),

$$\begin{aligned}\widehat{\text{SH}}(t) &= \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} w_i^{\mathcal{T}} \left| \hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) - S_i^{\mathcal{T}}(t) \right| \\ &= \frac{1}{n^{\mathcal{T}}} \sum_{i=1}^{n^{\mathcal{T}}} \left[\frac{\delta_i^{\mathcal{T}}(1 - S_i^{\mathcal{T}}(t))}{\hat{G}^{\mathcal{L}}(t_{i-1}^{\mathcal{T}}|\boldsymbol{x}_i^{\mathcal{T}})} + \frac{S_i^{\mathcal{T}}(t)}{\hat{G}^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}})} \right] \left| \hat{S}_i^{\mathcal{L}}(t|\boldsymbol{x}_i^{\mathcal{T}}) - S_i^{\mathcal{T}}(t) \right|.\end{aligned}$$

Thereby, $\hat{G}^{\mathcal{L}}(\cdot|\boldsymbol{x}_i^{\mathcal{T}})$, again, denotes the survival function of the censoring time C estimated from the learning data and $w_i^{\mathcal{T}}$ denote the inverse probability of censoring weights. Schmid et al. (2011) have shown that the modified Schemper-Henderson estimator is robust against model misspecification under the random censoring assumption. Possible approaches of estimating $\hat{G}^{\mathcal{L}}$ are discussed in Section 4.2.2. A corresponding time-independent measure is given by the integrated Schemper-Henderson estimator $\widehat{\text{SH}} = \sum_t \widehat{\text{SH}}(t) \cdot \hat{P}^{\mathcal{L}}(T = t)$.

Concordance Index

An alternative approach that characterizes the predictive performance is based on discrimination measures. This type of measures makes use of the linear predictor η_{it}, $i = 1, ..., n$, $t = 1, ..., t_i$. Discrimination measures consider survival outcomes as time-dependent binary

outcomes with categories "event at t" (denoted as *cases*) and "event after t" (denoted as *controls*). Consequently, η_{it} has a high prediction accuracy if it has a high discriminative power, that means to distinguish between cases and controls in the test data. This results in an interpretation of binary outcomes (cases versus controls) leading to the fact that established concepts for the evaluation of binary classification rules can be adapted to the analysis of time-to-event data. One of the most commonly used discrimination measures is the concordance index, which has its roots in receiver operating characteristics (ROC) methodology.
When outcomes are binary, the discriminative power can be summarized through the *time-dependent sensitivity* and the *time-dependent specificity* (Heagerty and Zheng, 2005), which are given by

$$\text{sensitivity}(c,t) = P(\eta_{it} > c | T = t) \tag{4.4}$$
$$\text{specificity}(c,t) = P(\eta_{it} \leq c | T > t). \tag{4.5}$$

Thereby, c is a threshold of the linear predictor η_{it}. Summarizing sensitivity (4.4) and specificity (4.5) yield a time-dependent ROC curve, which is defined by

$$\text{ROC}(c,t) = \{1 - \text{specificity}(c,t), \text{sensitivity}(c,t)\},$$

with $c \in \mathbb{R}$. Another measure applied to binary outcomes is the *area under the curve (AUC)* , that means, the area under the time-dependent ROC curve for each time t. This results in the time-dependent AUC curve, denoted by AUC(t). By definition, the time-dependent AUC curve quantifies the discriminative ability of a linear predictor at each time under consideration. As a value of 0.5 corresponds to the AUC value obtained by a model without covariate information, only AUC values larger than 0.5 are meaningful.
A time-independent measure of discriminative power is given by the consideration of the area under the time-dependent AUC curve. Heagerty and Zheng (2005) suggested the index

$$C^* = \int_t \text{AUC}(t) w(t) dt,$$

with weights $w(t) = P(T = t)P(T > t) / \sum_t P(T = t)P(T > t)$. According to Schmid et al. (2014) and Uno et al. (2007), sensitivity and specificity can be estimated by

$$\widehat{\text{sensitivity}}(c,t) = \frac{\sum_i \delta_i^{\mathcal{T}} I(\hat{\eta}_{it}^{\mathcal{T}} > c \cap t_i^{\mathcal{T}} = t) / \hat{G}^{\mathcal{L}}(t_i^{\mathcal{T}} - 1 | \boldsymbol{x}_i^{\mathcal{T}})}{\sum_i \delta_i^{\mathcal{T}} I(t_i^{\mathcal{T}} = t) / \hat{G}^{\mathcal{L}}(t_i^{\mathcal{T}} - 1 | \boldsymbol{x}_i^{\mathcal{T}})}$$
$$\widehat{\text{specificity}}(c,t) = \frac{\sum_i I(\hat{\eta}_{it}^{\mathcal{T}} \leq c \cap t_i^{\mathcal{T}} > t)}{\sum_i I(t_i^{\mathcal{T}} > t)},$$

where $\hat{\eta}_{it}^{\mathcal{T}}$, $i = 1, ..., n^{\mathcal{T}}$, denotes the estimated linear predictor using observations from the test data but the estimated parameters were obtained from the learning sample. Similar to the Brier score and the Schemper-Henderson estimator, the weights $1/\hat{G}^{\mathcal{L}}(t_i^{\mathcal{T}} - 1 | \boldsymbol{x}_i^{\mathcal{T}})$ ensure the consistency under the random censoring assumption. Esti-

mates of AUC(t) can be obtained by using numerical integration of the estimated ROC curve $\left\{1 - \widehat{\text{specificity}}(c,t), \widehat{\text{sensitivity}}(c,t)\right\}$. By means of the estimated area under the curve $\widehat{\text{AUC}}(t)$, the concordance index C^* can be estimated by

$$C^* = \sum_t \frac{\hat{P}^{\mathcal{L}}(T=t) \cdot \hat{P}^{\mathcal{L}}(T>t)}{\sum_t \hat{P}^{\mathcal{L}}(T=t) \cdot \hat{P}^{\mathcal{L}}(T>t)} \widehat{AUC}(t).$$

While prediction rules based on random guessing yield $C^* = 0.5$, a perfectly discriminating linear predictor leads to $C^* = 1$. In contrast to the likelihood measures and the distance-based measures described above, the concordance index has to be maximized.

4.2.2. Choice of the Censoring Distribution

For estimating the modified Brier score and the modified Schemper-Henderson estimator, the underlying model to determine $\hat{G}^{\mathcal{L}}(\cdot|\boldsymbol{x}_i^{\mathcal{T}})$ has to be chosen. In general, the conditional survival function of the censoring process is denoted by $G(t|\boldsymbol{x}_i) = P(C > t|\boldsymbol{x}_i)$.

A possible approach to determine $G(t|\boldsymbol{x}_i)$ is a *marginal censoring model*. This model ignores any covariates resulting in $G(t|\boldsymbol{x}_i) = G(t)$. For example, Graf et al. (1999) used a marginal censoring model for estimating $G(t|\boldsymbol{x}_i)$ with regard to continuous survival outcomes. Referring to discrete-time survival analysis, life table estimates can be used with respect to marginal censoring models. Under the assumption of random censoring, the life table estimator is consistent (Breslow and Crowley, 1974). The life table estimator, assuming censoring at the end of an interval $[a_{t-1}, a_t)$, is given by

$$\lambda(t) = \frac{d_t}{n_t}, \tag{4.6}$$

where d_t denotes the number of observed events (deaths) in the interval $[a_{t-1}, a_t)$ and $n_t = n_{t-1} - d_{t-1}$ denotes the number of observations at risk in the interval $[a_{t-1}, a_t)$.

In contrast, $G(t|\boldsymbol{x}_i)$ can be modeled by incorporating covariates. A method to model the dependence of the censoring survival function on covariates is given by generalized additive models. For the following simulation study, the linear predictor of the model consists of a time-varying intercept, whereas the covariates are incorporated as linear effects. The corresponding hazard rate of the censoring process is given by

$$\lambda_G(t|\boldsymbol{x}_i) = 1 - \exp(\exp(\beta_{0t} + \boldsymbol{x}_i^T \boldsymbol{\gamma})). \tag{4.7}$$

Thereby, $\beta_{0t} = \sum_{m=1}^{m_0} \alpha_{0m} B_{0m}(t)$ indicates a time-varying intercept expanded in B-splines. The estimate of the corresponding survival function $G(t|\boldsymbol{x}_i)$ is derived from the estimates of model (4.7) by

$$\hat{G}(t|\boldsymbol{x}_i) = \prod_{j=1}^{t} (1 - \hat{\lambda}_G(j|\boldsymbol{x}_i)).$$

To ensure robustness of the estimated model parameters a small ridge penalty is applied to model (4.7) resulting in the following penalized likelihood

$$l_{\xi_0,\xi}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}) = l(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}) - \left(\xi_0 \left(\sum_{m=2}^{m_0} (\alpha_{0m} - \alpha_{0,m-1})^2 \right) + \xi \left(\sum_{l=1}^{p} \gamma_l^2 \right) \right), \tag{4.8}$$

where $l(\boldsymbol{\alpha}_0, \boldsymbol{\gamma})$ denotes the ordinary log-likelihood. The first penalty term in (4.8) yield stable estimates of the time-varying intercepts, whereas the second term slightly shrinks the parameters referring to the linear effects to yield stable results.

4.3. Simulation Study

In this section, the choice of the tuning parameter ξ is investigated in terms of a simulation study. Generally, the simulation procedure is equivalent to that of Chapter 3. Thereby, in Chapter 3 the tuning parameter ξ was chosen by cross-validation based on the predictive deviance. In the following simulation study, for the choice of ξ a cross-validation procedure is used as well, but as loss criterion the measures presented in Section 4.2.1 are used. It is of special interest, if another loss function performs better than the predictive deviance. Furthermore, it is investigated how the construction of the cross-validation parts affects the predictive performance since the cross-validation parts can be based on the whole information of an object or on individual data points. This issue is analyzed by means of the predictive deviance.
The components of the simulation settings required for computing the underlying true linear predictor

$$\eta_t^{true} = \beta_{0t} + \sum_{j=1}^{r} z_{ij} \beta_{jt} + \sum_{l=1}^{s} x_{il} \gamma_l$$

are given in the following. Thereby, $(z_{i1}, ..., z_{ir}, x_{i1}, ..., x_{is})$, with $p = r + s$, is a vector of realizations of explanatory variables that do not vary over time, and $z_{i1}, ..., z_{ir}$ denote the observations of the covariates that are allowed to exhibit time-varying effects and $x_{i1}, ..., x_{is}$ define the observations of the covariates that are restricted to have constant effects. In other words, the model implies time-varying coefficients $\beta_{jt}^T = (\beta_{0t}, \beta_{rt}, ..., \beta_{rt})$ including an time-varying intercept β_{0t}, whereas $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_s)$ is assumed to be time-constant. The time-varying coefficients are expanded in equally spaced B-splines given by $\beta_{jt} = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t)$, $j = 0, ..., r$.

Finally, the prediction accuracy of the resulting models based on different loss functions is assessed. To ensure the comparability of the predictive accuracy for all measures, it is judged by considering the predictive deviance irrespective of the underlying loss function. To this end, n_p additional independent observation are sampled resulting in the vector of covariates $(\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})$ and the corresponding predictive deviance is given by

$$\mathrm{D}_{pred} = -2 \sum_{i=1}^{n} \sum_{t=1}^{t_i} \left\{ y_{it}^{pred} \log(\hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})) + (1 - y_{it}^{pred}) \log(1 - \hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})) \right\},$$

where $\lambda(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred}) = P(T_i = t|T_i \geq t, \boldsymbol{z}_{it}^{pred}, x_{it}^{pred})$ and $(y_{i1}^{pred}, ..., y_{it_i}^{pred})$ denotes the transitions over periods of object i.

4.3.1. Settings

The simulation study consists of five simulation settings. In each setting, both time-constant and time-varying coefficients are incorporated in the model and several penalties are investigated. Moreover, the number of time periods and covariates is modified and in one setting correlated covariates are incorporated. In a further setting, censoring depends on the covariates. The covariates of the whole simulation study are time-constant throughout.

Setting 1

For the first setting, $n = 120$ realizations of six covariates are simulated according to $Z_{i1}, X_{i1}, ..., X_{i5} \overset{iid}{\sim} \mathcal{U}(0, 1)$. Only three covariates have an effect on the survival time and the remaining three covariates are noise variables. The covariate realizations $z_{i1}, x_{i1}, ..., x_{i5}$ are used to simulate survival times according to the linear predictor

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4 + x_{i5}\gamma_5,$$

where the time-varying coefficient effects β_{jt}, $j = 0, 1$ are given by

$$\begin{aligned} \beta_{0t} &= (\ -1.25, \quad -1, -0.75, -1.5, -1.85, -1.75, -1.9) \\ \beta_{1t} &= (\quad -3, -2.5, \quad -1, -0.5, \quad 1.5, \quad 2, \quad 3). \end{aligned}$$

Furthermore, the time-constant coefficient effects are constituted by $\gamma_1 = 0.5$, $\gamma_2 = -1$, $\gamma_3 = \gamma_4 = \gamma_5 = 0$. The time-varying coefficients β_{jt} are expanded in cubic B-splines, where the number of equidistant inner knots is set to four. The censoring times are drawn from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$, with $\boldsymbol{p}_c^T = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.4)$. This leads to a censoring rate of approximately 60%. The simulation scheme for Setting 1 is replicated 100 times. To evaluate the predictive accuracy, $n_p = 240$ further independent observations are sampled.

Setting 2

In simulation Setting 2, the number of time periods is increased to $q = 10$. The considered model contains four time-varying covariate effects and one time-constant covariate effect. In addition, three noise variables are incorporated. For the covariates, $n = 200$ realizations are simulated according to

$$\begin{aligned} Z_{i1}, Z_{i2}, X_{i3}, X_{i4} &\overset{iid}{\sim} \mathcal{U}(0,1), \\ Z_{i3}, Z_{i4} &\overset{iid}{\sim} \mathcal{N}(-1,1), \\ X_{i1}, X_{i2} &\overset{iid}{\sim} \mathcal{B}(0.5). \end{aligned}$$

Hence, the model consists of four uniformly distributed, two normal distributed and two binary distributed covariates. The survival times are sampled by means of the realizations of the covariates using the linear predictor

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + z_{i2}\beta_{2t} + z_{i3}\beta_{3t} + z_{i4}\beta_{4t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4.$$

Thereby, the time-varying effects β_{jt}, $j = 0, ..., 4$, are defined by

$$\begin{aligned} \beta_{0t} &= (\ -2, -1.5, -1.3, -1, -0.8, -0.5, -1.25, -1.3, -2, 0.01), \\ \beta_{1t} &= (-2.5, -2, -1, -0.5, 0.01, 0.5, 0.75, 1.5, 1.8, 1.3), \\ \beta_{2t} &= (\ -2, -1.9, -1.8, -1.7, -1.6, -1.5, -1.25, -0.75, -0.5, 0.01), \\ \beta_{3t} &= (\ 1, 0.25, 0.25, 0.01, -0.25, 0.25, 0.5, 0.5, 0.75, 0.8), \\ \beta_{4t} &= (-1.5, -1, -1, 0.01, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7), \end{aligned}$$

and the time-constant coefficients are set to $\gamma_1 = -1.5$, $\gamma_2 = \gamma_3 = \gamma_4 = 0$. Analogous to Setting 1, for the time-varying coefficients β_{jt}, B-spline basis functions are used, but the number of equidistant inner knots is set to seven. In this setting, the censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$, with $\boldsymbol{p}_c^T = (0.05, 0.05, 0.05, 0.05, 0.1, 0.1, 0.1, 0.1, 0.1, 0.3)$, resulting in a censoring rate of approximately 60%. In this simulation scheme, the number of replications is 100. The data set for investigating prediction accuracy consists of $n_p = 400$ independently sampled observations.

Setting 3

In the third setting the impact of correlated covariates is investigated. To this end, $n = 300$ realizations of eight correlated covariates following a normal distribution are simulated with $Z_{i1}, Z_{i2}, X_{i1}, X_{i3} \sim \mathcal{N}(1,1)$ and $Z_{i3}, Z_{i4}, X_{i2}, X_{i4} \sim \mathcal{N}(0,1)$. The sampling is carried out in consideration of the corresponding correlation matrix R:

$$\boldsymbol{R} = \begin{pmatrix} 1.0 & 0.0 & 0.5 & -0.3 & 0.2 & -0.1 & 0.1 & 0.2 \\ 0.0 & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 & 0.2 & -0.2 \\ 0.5 & 0.1 & 1.0 & 0.3 & 0.7 & 0.2 & 0.1 & 0.1 \\ -0.3 & 0.0 & 0.3 & 1.0 & 0.4 & -0.4 & 0.05 & 0.0 \\ 0.2 & 0.0 & 0.7 & 0.4 & 1.0 & 0.0 & 0.3 & 0.4 \\ -0.1 & 0.0 & 0.2 & -0.4 & 0.0 & 1.0 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.1 & 0.05 & 0.3 & 0.01 & 1.0 & -0.3 \\ 0.2 & -0.2 & 0.1 & 0.0 & 0.4 & 0.1 & -0.3 & 1.0 \end{pmatrix}. \tag{4.9}$$

The corresponding survival times are sampled using the realizations of the covariates and the linear predictor defined by

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + z_{i2}\beta_{2t} + z_{i3}\beta_{3t} + z_{i4}\beta_{4t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4.$$

Thereby, the time-varying coefficients β_{jt}, $j = 0, ..., 4$, are given by

$$\begin{aligned}
\beta_{0t} &= (\ -2, -0.8, -0.5, -1.25, -1.125, -1.5, -2, -2, -1.5, -1), \\
\beta_{1t} &= (\ -4, -3, -2, -1, 0.01, 0.5, 1.25, 1.5, 1.7, 2), \\
\beta_{2t} &= (\ -2, -1.9, -1.7, -1.5, -1.25, -1, -0.75, -0.5, -1.5, 1), \\
\beta_{3t} &= (\ 0.01, 0.1, 0.5, 0.8, 0.9, 1.1, 1.5, 1.6, 1.8, 2.5), \\
\beta_{4t} &= (\ -0.8, -0.7, -0.6, -0.5, 0.5, -0.6, -0.7, -0.8, -0.9, -1),
\end{aligned}$$

and the time-constant coefficients are given by $\gamma_1 = -0.5$, $\gamma_2 = 1$ and $\gamma_3 = \gamma_4 = 0$. Again, the time-varying coefficients β_{jt} are modeled in terms of cubic B-splines with seven equidistant inner knots. Moreover, the censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$ with probability vector $\boldsymbol{p}_c^T = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$, leading to a censoring rate of approximately 65%. For this simulation setting, 100 simulation runs are executed. To evaluate the predictive accuracy, $n_p = 600$ further independent observations are sampled.

Setting 4

In simulation Setting 4, the number of time periods is equal to $q = 10$. The considered model contains four time-varying covariate effects and two time-constant covariate effects. In addition, two noise variables are incorporated. For the study $n = 150$, realizations of the covariates were simulated according to

$$\begin{aligned}
Z_{i1}, Z_{i2}, X_{i3}, X_{i4} &\overset{iid}{\sim} \mathcal{U}(0, 1), \\
Z_{i3}, Z_{i4} &\overset{iid}{\sim} \mathcal{N}(-1, 1), \\
X_{i1}, X_{i2} &\overset{iid}{\sim} \mathcal{B}(0.5),
\end{aligned}$$

that is, the model contains four uniformly distributed, two normal distributed and two binary distributed covariates. The survival times are sampled depending on the realizations of these covariates with the linear predictor

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + z_{i2}\beta_{2t} + z_{i3}\beta_{3t} + z_{i4}\beta_{4t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4.$$

Thereby, the time-varying effects β_{jt}, $j = 0, ..., 4$, are defined by

$$\begin{aligned}
\beta_{0t} &= (\ \ -2, -0.8, -0.5, -1.25,\ \ -1.3, -1.7, 0.01),\\
\beta_{1t} &= (-2.5,\ \ -2,\ \ -1,\ \ 0.01,\ \ 0.75,\ \ 1.5,\ \ 1.3),\\
\beta_{2t} &= (\ \ -2, -1.7, -1.5, -1.25, -0.75, -0.5, 0.01),\\
\beta_{3t} &= (\ \ \ 1,\ 0.25,\ 0.01, -0.25,\ \ 0.25,\ \ 0.5, 0.75),\\
\beta_{4t} &= (-1.5,\ \ -1,\ 0.01,\ \ \ 0.3,\ \ \ 0.4,\ \ 0.5,\ \ 0.6),
\end{aligned}$$

and the time-constant coefficients γ_l are $\gamma_1 = -1.5$, $\gamma_2 = 0.5$ and $\gamma_3 = \gamma_4 = 0$. Analogous to the previous settings, the time-varying coefficients β_{jt} are expanded in cubic B-splines, where the number of equally spaced inner knots is set to four. In this setting, the vector of probabilities for simulating the censoring times from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$ depends on the combination of the covariates X_{i5} and X_{i6}. As X_{i5} and X_{i6} are drawn from a binomial distribution, there exist four different vectors of censoring probabilities:

$$\begin{aligned}
X_{i5} = 0 \wedge X_{i6} = 0: &\quad \boldsymbol{p}_{c1}^T = (0.05, 0.05, 0.05, 0.05, 0.1, 0.1, 0.6),\\
X_{i5} = 0 \wedge X_{i6} = 1: &\quad \boldsymbol{p}_{c2}^T = (0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 0.2),\\
X_{i5} = 1 \wedge X_{i6} = 0: &\quad \boldsymbol{p}_{c3}^T = (0.1, 0.1, 0.25, 0.25, 0.05, 0.05, 0.2),\\
X_{i5} = 1 \wedge X_{i6} = 1: &\quad \boldsymbol{p}_{c4}^T = (1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7),
\end{aligned}$$

This leads to an overall censoring rate of approximately 70%. In this simulation scheme the number of replications is 100. The data set for investigating prediction accuracy consists of $n_p = 300$ independently sampled observations.

Setting 5

For the last setting, $n = 120$ realizations of six covariates are simulated according to $Z_{i1}, X_{i1}, ..., X_{i5} \overset{iid}{\sim} \mathcal{U}(0, 1)$. Thereby, three covariates have an effect on the survival time and three covariates are noise variables. The covariate outcomes are the basis for simulating the survival time with the linear predictor given by

$$\eta_{it} = \beta_{0t} + z_{i1}\beta_{1t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4 + x_{i5}\gamma_5.$$

The linear predictor includes the time-varying coefficients β_{jt}, $j = 0, 1$ given by

$$\begin{aligned}
\beta_{0t} &= (\ -1.25,\ \ -1, -0.75,\ \ -2, -1.85, -1.75, -1.9),\\
\beta_{1t} &= (\ \ \ -3, -2.5,\ \ \ -1, -0.5,\ \ \ 1.5,\ \ \ \ 2,\ \ \ \ 3),
\end{aligned}$$

and the time-constant coefficients $\gamma_1 = 0.5$, $\gamma_2 = 1$, $\gamma_3 = \gamma_4 = \gamma_5 = 0$. Consequently, simulation scheme 5 is equivalent to Setting 1, but in contrast to Setting 1 a penalty resulting in piecewise time-constant coefficients is used. Hence, the time-varying coefficients have to be estimated for all time periods $t = 1, ..., q$. The censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$, with $\boldsymbol{p}_c^T = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.4)$ leading to a censoring rate of approximately 55%. The simulation scheme for Setting 5 is replicated 100 times. Additionally, $n_p = 240$ independent observations are sampled to evaluate the predictive accuracy.

To allow for maximal flexibility in modeling, for all coefficients, time-varying effects are assumed. This leads to the linear predictor

$$\eta_{it}^{model} = \beta_{0t} + \sum_{j=1}^{p} z_{ij}\beta_{jt}.$$

Thereby, the number of time-varying coefficients depends on the simulation setting and $r = p$ holds. For Setting 1-4, the time-varying coefficients are expanded in B-spline basis functions with

$$\beta_{0t} = \sum_{m=1}^{m_0} \alpha_{0m} B_{0m}(t) \quad \text{and} \quad \beta_{jt} = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t).$$

For the simulation study, two different types of penalties, with regard to the estimation of the time-varying covariate effects, are employed. For Settings 1-4, the penalty term

$$J_{\xi_0,\xi}(\boldsymbol{\zeta}, \boldsymbol{\alpha}) = \xi_0 \sum_{m=2}^{m_0} \left(\alpha_{0m} - \alpha_{0,m-1}\right)^2 + \xi \sum_{j=1}^{r} \psi_j \left\|\boldsymbol{\zeta}_j\right\|_2,$$

where $\boldsymbol{\zeta}_j^T = (\zeta_{j2}, ..., \zeta_{jm_j})$, $\zeta_{jm} = \alpha_{jm} - \alpha_{j,m-1}$, $m = 2, ..., m_j$, is used. It allows for stable baseline effects and steers the smoothness of the time-varying covariate effects. Moreover, the differences $\boldsymbol{\zeta}_j^T = (\zeta_{j2}, ..., \zeta_{jm_j})$ can be simultaneously shrunk to zero until their selection from the model leading to constant effects. On the other hand fusion and selection of the time-varying coefficients is performed by the following penalty term

$$J_{\xi_0,\xi}(\boldsymbol{\beta}) = \xi_0 \sum_{m=2}^{m_0} \left(\alpha_{0m} - \alpha_{0,m-1}\right)^2 + \xi \left(\sum_{j=1}^{r} \sum_{t=2}^{q} |\beta_{jt} - \beta_{j,t-1}| + \sum_{j=1}^{r} \sum_{t=1}^{q} |\beta_{jt}| \right),$$

used for Setting 5. In the context of survival analysis, it is proposed that a time-varying intercept β_{0t} remains in the model. Thus, the penalization of β_{0t} is only executed due to stability reasons. It is defined by $\xi_0 = 0.001$ in all simulation settings. This corresponds to a ridge penalty on differences between adjacent parameters of the B-spline basis functions. For more details on the penalty, compare Chapter 3.

Finally, the simulation settings are summarized in Table 4.1. Therein, n denotes the number of observations of each setting. The number of time-varying and time-constant covariate

	n	time intervals	covariate effects varying	constant	noise	penalty	correlation	dependent censoring
1	120	7	1	2	3	smoothing	-	-
2	400	10	4	1	3	smoothing	-	-
3	250	10	4	2	2	smoothing	✓	-
4	200	7	4	2	2	smoothing	-	✓
5	200	7	1	2	3	fusion & selection	-	-

Table 4.1. Overview simulations settings of Chapter 4.

effects as well as the number of noise variables are shown in the columns *varying*, *constant* and *noise*, respectively. Moreover, *penalty* describes which penalties are used and *correlation* declares if a correlation of the covariates is incorporated. Finally, *dependent censoring* marks if the censoring depends on the covariates.

4.3.2. Results

The execution of the simulation study and the corresponding estimation procedure is equivalent to the proceeding described in Section 3.4, but the loss function of the cross-validation procedure is varied. The simulated data sets are reorganized to the long format before estimation and all covariates are standardized to have equal variance. The tuning parameter ξ is chosen by 5-fold cross-validation incorporating the loss functions presented in Section 4.2.1. The model is estimated by a binary regression model with complementary log-log link. Estimates are obtained by maximizing the corresponding penalized likelihood (see Section 3.2.1).

In Gerds and Schumacher (2006), the authors investigated if the censoring model to determine the survivor function G should depend on covariates. Gerds and Schumacher (2006) concluded that a censoring model depending on covariates should be preferred to a marginal censoring model. Hence, for each setting the estimation of the censoring model is carried out two times. First, a marginal censoring model by means of the life table estimator (4.6), and second, a regression censoring model as defined in Section 4.2.2 is used. That means, the censoring regression model of the simulation study is given by

$$\eta_{it}^G = \beta_{0t} + \sum_{l=1}^{p} x_{li}\gamma_l, \qquad \beta_{0t} = \sum_{m=1}^{m_0} \alpha_{0m} B_{0m}(t),$$

with penalty term

$$J_\xi^G(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}) = \xi_0 \left(\sum_{m=2}^{m_0} \left(\alpha_{0m} - \alpha_{0,m-1}\right)^2 \right) + \xi \left(\sum_{l=1}^{p} \gamma_l^2 \right).$$

For the time-varying intercept a cubic B-spline approach is used and all covariates are incorporated as linear effects. The penalty allows for stable baseline effects, whereas the coefficients γ_l are subject to a slight ridge penalty, where $\xi_0 = \xi = 0.001$. The number of equidistant inner knots for estimating β_{0t} in the censoring model is equivalent to the number of equidistant inner knots used in the corresponding model for fitting the survival model.

The assessment of parameter estimates is evaluated as a whole, and separately for truly time-varying and truly time-constant parameters. For each simulation run, the according mean squared errors are computed according to definition (3.13).
Thereby, $\boldsymbol{\beta}_j$ and $\tilde{\boldsymbol{\beta}}_l$ denote the true parameter values, whereas $\hat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{\beta}}_l$ define the estimates. That means, as all components are estimated time-varying, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are compared as well. The ordinary maximum likelihood (ML) estimation is used as reference. Since many algorithms of the simulation settings did not converge using the ML-method, a slight ridge penalty of 0.001 is installed (denoted by ML_{ridge}). Hence, the ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ can be interpreted in a meaningful manner.

The results of all simulation settings are initially summarized in tables. Therein, the ordinary ML-estimates with a slight ridge penalty of 0.001 can be found in column ML_{ridge}. The columns denoted with D and D.a contain the results using the predictive deviance as loss function. In the case of D, the parts of the cross-validation are selected by object, hence, all observations of an object are selected at a time. In contrast, the splitting refers to individual data points for D.a. Consequently, not all observations of an object are forced to be in the same part of the cross-validation. The remaining abbreviations relate to the measures given in Section 4.2.1. Thereby, the subscript $marg$ defines a marginal censoring model used for estimating the survival function G of the censoring process. Additionally, reg refers to a generalized regression censoring model for G used in the IPCW estimators (see Section 4.3.2). The tables contain the mean squared errors, obtained from the definition (3.13). Thereby, MSE_{vary} refers to the truly time-varying coefficients and MSE_{const} to the truly time-constant coefficients. The mean squared errors marked with * correspond to the ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$, a logarithmic relation to ML_{ridge}. In D_{pred} and $\text{D}^*_{pred} = \log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$ the predictive deviances based on the n_p additionally independently sampled observations are computed. D_{pred} and D^*_{pred} are computed for all loss functions. All presented values correspond to the mean values over all simulation runs. Moreover, the results are illustrated in boxplots where the ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ and $\log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$ are depicted. For the sake of interpretability, outliers are omitted in single cases.

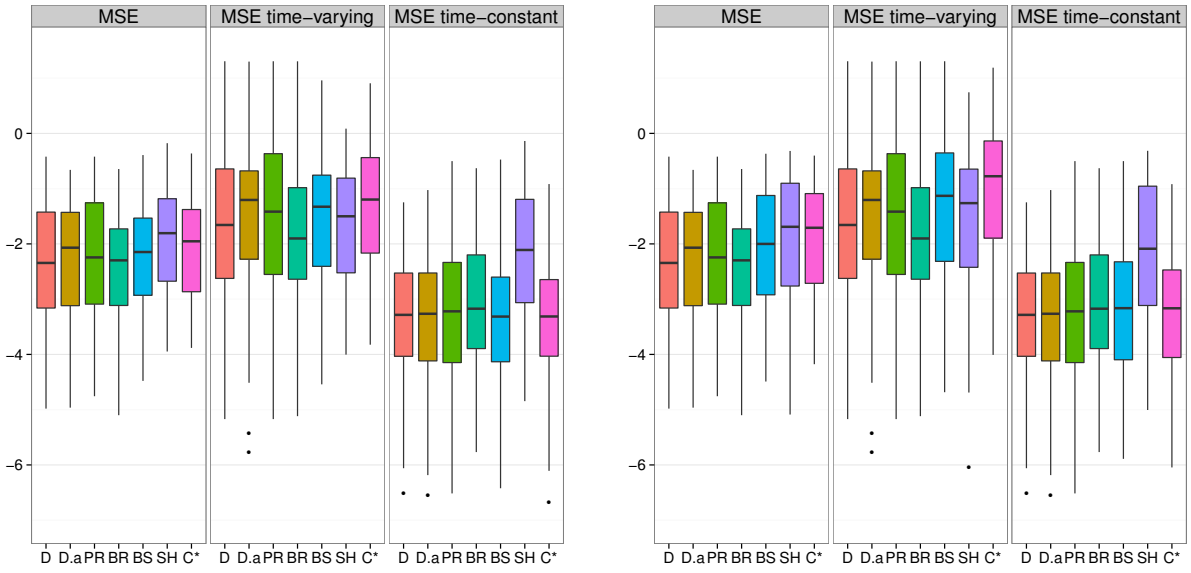
	ML_{ridge}	D	D.a	PR	BR	BS_{marg}	BS_{reg}	SH_{marg}	SH_{reg}	C^*_{marg}	C^*_{reg}
MSE	11.80	**0.69**	0.80	0.84	0.74	0.94	0.95	1.49	1.39	1.01	1.00
MSE_{vary}	14.94	1.72	2.03	2.15	**1.50**	2.43	2.35	2.29	2.13	2.81	2.84
MSE_{const}	10.54	0.28	0.31	0.31	0.43	0.35	0.39	1.17	1.09	0.29	**0.26**
MSE*	-	-2.44	-2.27	-2.24	**-2.46**	-2.22	-2.09	-1.94	-1.86	-2.10	-1.95
MSE^*_{vary}	-	-1.73	-1.47	-1.46	**-1.89**	-1.47	-1.29	-1.73	-1.57	-1.27	-1.02
MSE^*_{const}	-	-3.32	-3.37	-3.31	-3.11	**-3.38**	-3.17	-2.16	-2.16	-3.37	-3.28
D_{pred}	888.90	**610.71**	612.63	615.01	612.90	628.34	620.96	662.41	646.18	626.80	617.56
D^*_{pred}	-	-0.34	-0.34	-0.33	-0.34	**-0.36**	-0.33	-0.31	-0.29	-0.36	-0.33

Table 4.2. Results for simulation Setting 1 for the mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE^*, MSE^*_{vary}, MSE^*_{const}), the predictive deviance referring to the additional sampled test data (D_{pred}, D^*_{pred}) for ML_{ridge} and the loss functions D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C^* (concordance index). The subscripts *marg* and *reg* denote a marginal censoring model and a regression censoring model, respectively. The displayed values represent the means over all 100 simulation runs. Bold values indicate the best value in each case.

Setting 1

The simulation results for Setting 1 are summarized in Table 4.2. To get an impression of the magnitude, the corresponding MSE of the ordinary ML-estimation denotes $1.88 \cdot 10^{29}$. A slight penalty already improves the MSE substantially which can be seen in column ML_{ridge}. The penalization approaches outperform ML_{ridge} in all cases. Moreover, the predictive deviance D yields better results than D.a for all MSE values (except MSE^*_{const}). Apart from that, the predictive deviance D outperforms for the MSE and D_{pred}. In comparison to the other measures, the Brier score (BR) performs best for MSE_{vary}, MSE^*_{vary} and MSE^*. The modified Brier score using a marginal censoring model provides the best values for MSE^*_{const} and D^*_{pred}, whereas C^*_{reg} outperforms the other measures for MSE_{const}. No clear preference for the marginal or the regression censoring model can be given for the IPCW estimates.

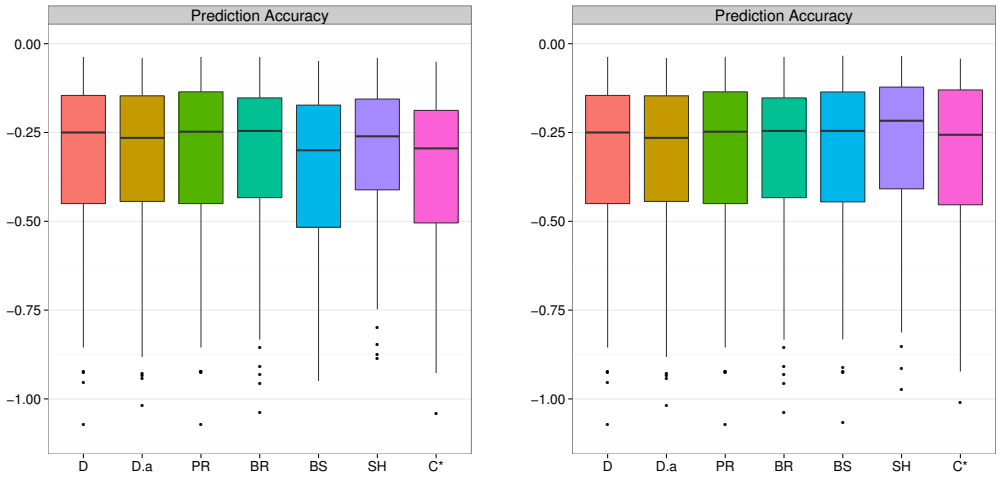
The corresponding boxplots of the ratios $\log(MSE(\cdot)/MSE(ML_{ridge}))$ are shown in Figure 4.1, where Figure 4.1a contains the boxplots using a marginal censoring model for the IPCW estimators and Figure 4.1b contains the boxplots using a regression censoring model. It has to be mentioned that the boxplots for D, D.a, PR and BR are identical in both cases as no censoring distributions are utilized for this measures. Comparing the boxplots for the marginal and the regression censoring model, only a small shift is observable for the mean squared error of the truly time-varying coefficients using C^* as loss function. Again, the boxplots of the log ratios $\log(D_{pred}(\cdot)/D_{pred}(ML_{ridge}))$ assessing the predictive performance are identical for D, D.a, PR and BR (Figure 4.2). The predictive performance seems to be somewhat better in the case of the marginal censoring distribution for the IPCW estimators. For the other measures the prediction accuracy is on a similar level.



(a) Marginal censoring model

(b) Regression censoring model

Figure 4.1. Boxplots of the mean squared errors for Setting 1 for D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index)



(a) Marginal censoring model

(b) Regression censoring model

Figure 4.2. Boxplots of the predictive accuracy for Setting 1 for D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index)

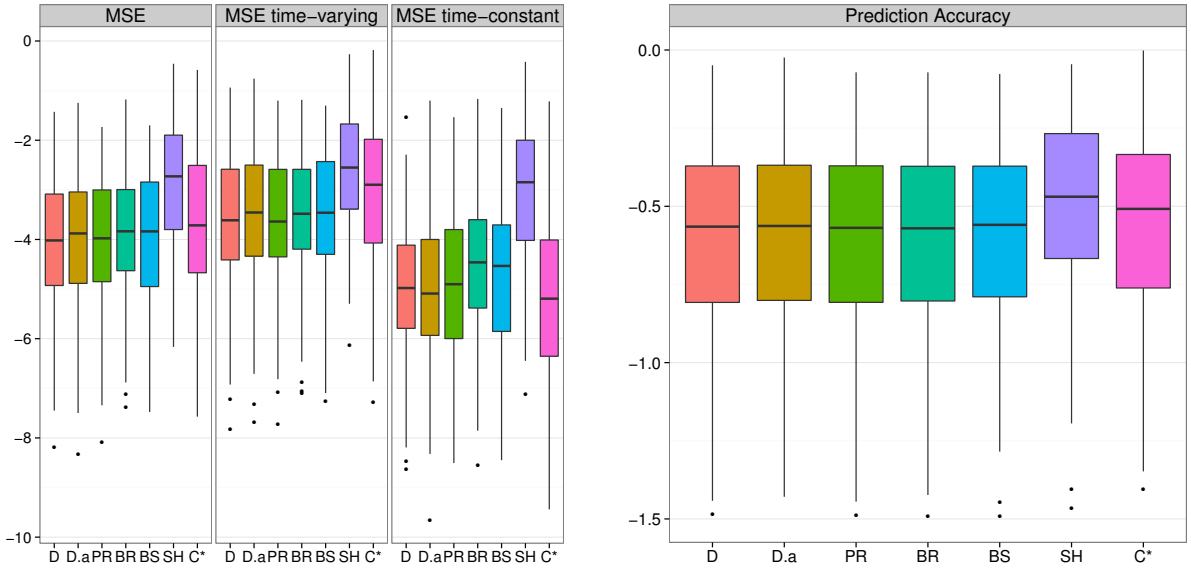
Setting 2

The summary of the simulations results for Setting 2 are given in Table 4.3. The MSE of the corresponding ML-estimation is $1.10 \cdot 10^{30}$. As expected, the penalization approaches outperform ML_{ridge} in all cases. Except for MSE^*_{const}, the predictive deviance D outperforms D.a for all values. Furthermore, the predictive deviance D yields best results for MSE, MSE_{const}, MSE*, D_{pred} and D^*_{pred}. In addition, the discrete ranked probability score (PR) performs well and shows the best results for MSE, MSE_{vary}, MSE^*_{vary}, D_{pred} and D^*_{pred}. C^*_{marg} and C^*_{reg} outperform the other measures for MSE^*_{const}. The values of the IPCW estimators are very similar, that means, that the estimates using the marginal censoring model and the estimates using a regression censoring model do not outperform each other. Finally, the predictive deviance D and the discrete ranked probability score PR turn out to perform best in this simulation scheme.

There are no visible differences between the boxplots of the log-ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ and $\log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$ using a marginal censoring model for the IPCW estimators (Figure 4.3) and the respective boxplots using a censoring regression model (Figure A.1). Comparing the mean squared errors (Figure 4.3a) and the predictive accuracy (Figure 4.3b) for the different loss functions, the boxplots of the Schemper-Henderson estimator perform worst, whereas the boxplots of D and PR perform quite well. Except for the predictive accuracy of SH and C*, the boxplots are on a similar level.

	ML_{ridge}	D	D.a	PR	BR	BS_{marg}	BS_{reg}	SH_{marg}	SH_{reg}	C^*_{marg}	C^*_{reg}
MSE	46.21	**0.36**	0.40	**0.36**	0.44	0.49	0.51	1.73	1.80	0.69	0.69
MSE_{vary}	43.83	0.50	0.55	**0.48**	0.56	0.66	0.67	1.95	1.99	1.07	1.07
MSE_{const}	49.19	**0.17**	0.20	0.20	0.28	0.27	0.31	1.45	1.57	0.22	0.22
MSE*	-	**-4.08**	-3.99	-4.07	-3.92	-3.96	-3.92	-2.85	-2.83	-3.65	-3.65
MSE^*_{vary}	-	-3.62	-3.52	**-3.64**	-3.56	-3.54	-3.53	-2.66	-2.65	-3.05	-3.05
MSE^*_{const}	-	-4.96	-5.05	-4.86	-4.48	-4.71	-4.62	-3.06	-3.05	**-5.27**	**-5.27**
D_{pred}	1986.5	**1035.6**	1043.1	**1035.6**	1039.3	1044.8	1046.1	1160.5	1163.0	1098.1	1098.2
D^*_{pred}	-	**-0.61**	-0.60	**-0.61**	-0.60	-0.60	-0.60	-0.50	-0.50	-0.55	-0.55

Table 4.3. Results for simulation Setting 2 for the mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE*, MSE^*_{vary}, MSE^*_{const}), the predictive deviance referring to the additional sampled test data (D_{pred}, D^*_{pred}) for ML_{ridge} and the loss functions D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index). The subscripts *marg* and *reg* denote a marginal censoring model and a regression censoring model, respectively. The displayed values represent the means over all 100 simulation runs. Bold values indicate the best value in each case.



(a) Mean squared errors

(b) Predictive accuracy

Figure 4.3. Boxplots of the mean squared errors and predictive accuracy for Setting 2 using a marginal regression model for IPCW estimators for D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index).

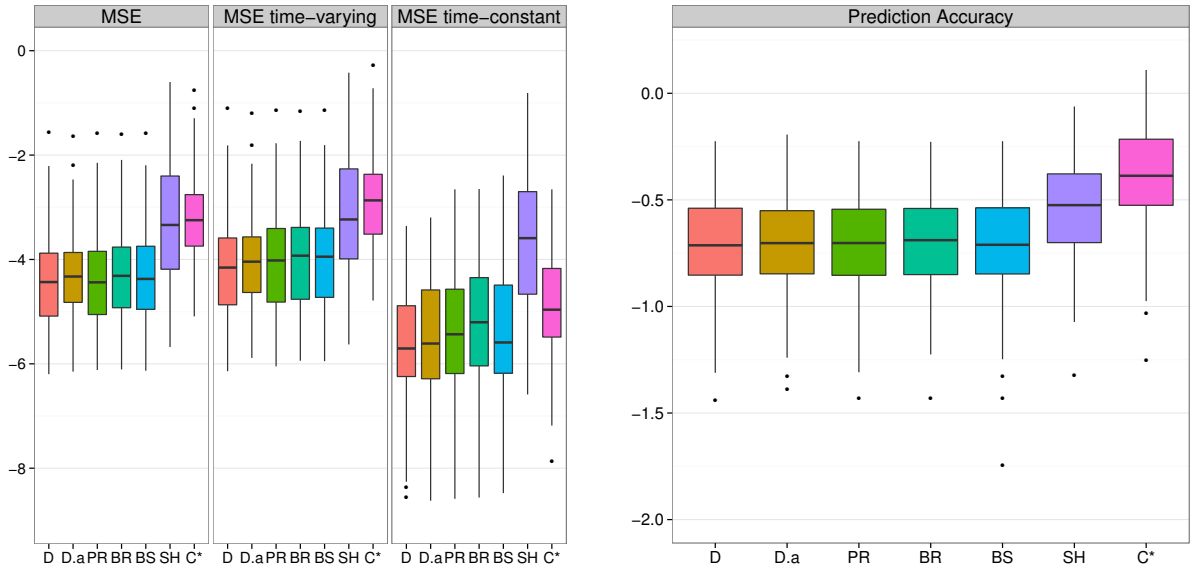
Setting 3

In simulation Setting 3 correlated covariates according to the correlation matrix (4.9) are incorporated in the model. The dimension for Setting 3 is the same as in Setting 2. The MSE of the underlying ML-estimation denotes $4.55 \cdot 10^{33}$. A summary of the results for Setting 3 can be found in Table 4.4. Compared to the other measures the predictive deviance D achieves predominately best results. In particular, the predictive deviance D yields better or similar results than D.a. Apart from that, the results of the modified Brier score BS_{reg} outperform all other measures for MSE*, MSE^*_{vary} and D^*_{pred}. Again, no evidence can be found if the use of marginal or regression censoring models is more advisable.

The boxplots dealing with marginal regression models (Figure 4.4) and the boxplots dealing with censoring regression models for the IPCW estimators (Figure A.2) are in high accordance. The Schemper Henderson estimator SH and the concordance index C* perform worst in terms of the log ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ and $\log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$ using a marginal censoring model for the IPCW estimators. The boxplots of the remaining measures are similar. The conclusion for the predictive accuracy is equivalent (Figure 4.4b). In general, correlated covariates seem to have no negative impact on the estimation approach as the absolute magnitudes of the results are not considerably different compared to the previous simulation Settings.

	ML_{ridge}	D	D.a	PR	BR	BS_{marg}	BS_{reg}	SH_{marg}	SH_{reg}	C^*_{marg}	C^*_{reg}
MSE	31.87	**0.32**	0.34	0.34	0.36	0.35	0.40	1.40	1.57	1.06	1.05
MSE_{vary}	41.28	**0.53**	0.56	0.55	0.56	0.56	0.62	1.98	2.24	1.81	1.80
MSE_{const}	20.11	**0.07**	**0.07**	0.08	0.11	0.08	0.11	0.67	0.73	0.12	0.12
MSE*	-	-4.29	-4.24	-4.25	-4.21	-4.23	**-4.30**	-3.21	-3.26	-3.05	-3.23
MSE^*_{vary}	-	-3.99	-3.93	-3.96	-3.94	-3.94	**-4.02**	-3.06	-3.09	-2.70	-2.88
MSE^*_{const}	-	**-5.56**	**-5.56**	-5.48	-5.24	-5.46	-5.42	-3.58	-3.68	-4.71	-4.88
D_{pred}	1816.6	**871.2**	877.9	874.4	886.5	876.1	888.0	1041.4	1082.7	1209.4	1211.9
D^*_{pred}	-	-0.71	-0.70	-0.70	-0.69	-0.70	**-0.76**	-0.54	-0.58	-0.38	-0.45

Table 4.4. Results for simulation Setting 3 for the mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE*, MSE^*_{vary}, MSE^*_{const}), the predictive deviance referring to the additional sampled test data (D_{pred}, D^*_{pred}) for ML_{ridge} and the loss functions D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index). The subscripts *marg* and *reg* denote a marginal censoring model and a regression censoring model, respectively. The displayed values represent the means over all 100 simulation runs. Bold values indicate the best value in each case.



(a) Mean squared errors

(b) Predictive accuracy

Figure 4.4. Boxplots of the mean squared errors and predictive accuracy for Setting 3 using a marginal regression model for IPCW estimators D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index)

	ML_{ridge}	D	D.a	PR	BR	BS_{marg}	BS_{reg}	SH_{marg}	SH_{reg}	C^*_{marg}	C^*_{reg}
MSE	35.66	**0.60**	0.64	0.65	0.78	0.85	0.85	2.45	2.02	1.86	1.98
MSE_{vary}	30.16	**0.83**	0.91	0.87	0.97	1.08	1.07	2.45	2.24	2.09	2.21
MSE_{const}	42.54	**0.31**	**0.31**	0.38	0.55	0.56	0.58	2.46	1.75	1.57	1.68
MSE*	-	**-3.78**	-3.69	-3.72	-3.65	-3.63	-3.61	-2.81	-2.94	-3.08	-3.05
MSE^*_{vary}	-	**-3.17**	-3.05	-3.13	-3.12	-3.06	-3.07	-2.48	-2.56	-2.54	-2.51
MSE^*_{const}	-	**-4.95**	-4.89	-4.82	-4.52	-4.66	-4.49	-3.14	-3.35	-4.02	-3.98
D_{pred}	1480.87	**569.10**	578.88	572.05	586.29	582.95	580.63	708.25	674.47	666.38	672.09
D^*_{pred}	-	**-0.89**	-0.87	-0.88	-0.86	-0.87	-0.87	-0.71	-0.75	-0.76	-0.75

Table 4.5. Results for simulation Setting 4 for the mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE^*, MSE^*_{vary}, MSE^*_{const}), the predictive deviance referring to the additional sampled test data (D_{pred}, D^*_{pred}) for ML_{ridge} and the loss functions D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C^* (concordance index). The subscripts *marg* and *reg* denote a marginal censoring model and a regression censoring model, respectively. The displayed values represent the means over all 100 simulation runs. Bold values indicate the best value in each case.

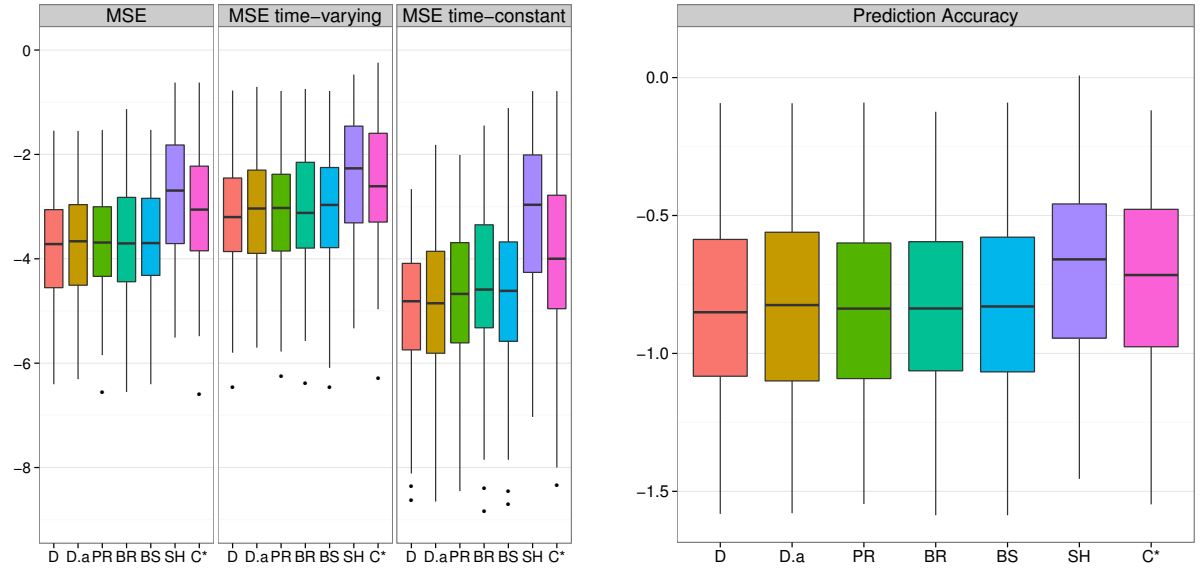
Setting 4

In simulation scheme 4 the censoring process depends on covariates. This is challenging, because all IPCW estimators of Section 4.2.1 regarding discrete survival times assume random censoring. Here, the MSE of the ML-method denotes $1.26 \cdot 10^{30}$. The results for Setting 4 are summarized in Table 4.5. Therein, the results of the predictive deviance D outperforms the results of D.a in all cases. Apart from that, the predictive deviance yields best results for all MSE values, D_{pred} and D^*_{pred}. The Schemper-Henderson estimator using a regression censoring model SH_{reg} performs slightly better compared to the estimator SH_{marg} using a marginal censoring model. On the other hand, for the concordance index the use of a marginal censoring model C^*_{marg} tends to perform better than the use of a regression censoring model (C^*_{reg}).

Again, there are no considerable differences between the boxplots of the log ratios $\log(MSE(\cdot)/MSE(ML_{ridge}))$ and $\log(D_{pred}(\cdot)/D_{pred}(ML_{ridge}))$ for the marginal censoring model (Figure 4.5) and the regression censoring model for the IPCW estimators (Figure A.3). Analogous to Setting 3, the Schemper Henderson estimator SH and the concordance index C^* yield worst results, whereas all other measures perform similarly. This is equivalent for the MSE and the prediction accuracy (Figure 4.5a and Figure 4.5b). However, in case of violating the random censoring assumption the overall level of the MSE as well as the prediction accuracy are somewhat worse compared to the other settings.

Setting 5

The results of simulation Setting 5 are summarized in Table 4.6. In comparison to the previous settings, a penalty achieving piecewise time-constant coefficient estimates is used in Setting 5. The corresponding MSE of the underlying ML-model denotes 155.23. It is not obvious if the predictive deviance D or D.a yields better results for all cases. The



(a) Mean squared errors

(b) Predictive accuracy

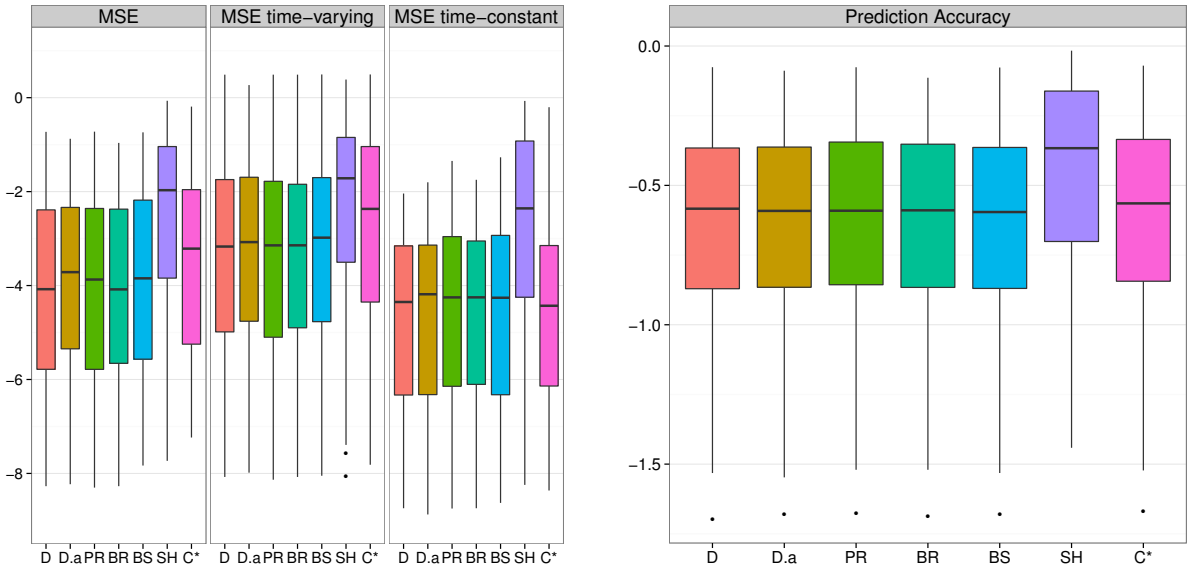
Figure 4.5. Boxplots of the mean squared errors and predictive accuracy for Setting 4 using a marginal regression model for IPCW estimators for D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index)

MSE is minimal for the predictive deviance D, the discrete ranked probability score PR and the Brier score BR. MSE_{vary} yields minimal results for PR. Except for the Schemper-Henderson estimator SH the values for MSE_{const} are similar for all measures. There is no stability concerning the performance of the different loss functions, since D performs best for MSE*, D.a performs best for MSE^*_{vary} and C^*_{marg} performs best for MSE^*_{const}. However, for the predictive deviances D_{pred} and D^*_{pred} the results are very similar over all measures, except for SH yielding worse results. In general, no measure outperforms the others.

The boxplots of the log ratios $\log(\text{MSE}(\cdot)/\text{MSE}(\text{ML}_{ridge}))$ and $\log(\text{D}_{pred}(\cdot)/\text{D}_{pred}(\text{ML}_{ridge}))$ dealing with marginal regression models (Figure 4.6) and the boxplots dealing with censoring regression models for the IPCW estimators (Figure A.4) are in high accordance. It can be seen, that the predictive deviance D possesses a lower median MSE value than for D.a. Again, the Schemper Henderson estimator SH and the concordance index C* yield the worst results.

	ML_{ridge}	D	D.a	PR	BR	BS_{marg}	BS_{reg}	SH_{marg}	SH_{reg}	C^*_{marg}	C^*_{reg}
MSE	33.66	**0.79**	0.82	**0.79**	**0.79**	0.93	0.93	2.44	2.44	1.06	1.05
MSE_{vary}	34.35	1.87	2.08	**1.82**	1.72	2.48	2.44	3.50	3.50	3.10	3.08
MSE_{const}	21.54	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	**0.25**	1.86	1.76	**0.25**	**0.25**
MSE*	-	**-3.78**	-3.55	-3.73	-3.89	-3.72	-3.62	-1.89	-1.87	-3.17	-3.17
MSE^*_{vary}	-	-3.02	**-3.06**	-3.00	-3.04	-2.96	-2.66	-1.68	-1.64	-2.34	-2.33
MSE^*_{const}	-	-4.31	-4.19	-4.14	-4.13	-4.26	-4.30	-2.33	-2.32	**-4.43**	-4.23
D_{pred}	1119.57	**617.39**	618.57	619.24	617.47	626.26	626.26	688.99	687.49	630.41	630.41
D^*_{pred}	-	-0.58	-0.59	-0.59	-0.59	**-0.60**	-0.59	-0.37	-0.35	-0.56	-0.56

Table 4.6. Results for simulation Setting 5 for the mean squared errors (MSE, MSE_{vary}, MSE_{const}, MSE*, MSE^*_{vary}, MSE^*_{const}), the predictive deviance referring to the additional sampled test data (D_{pred}, D^*_{pred}) for ML_{ridge} and the loss functions D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index). The subscripts *marg* and *reg* denote a marginal censoring model and a regression censoring model, respectively. The displayed values represent the means over all 100 simulation runs. Bold values indicate the best value in each case.



(a) Mean squared errors

(b) Predictive accuracy

Figure 4.6. Boxplots of the mean squared errors and predictive accuracy for Setting 5 using a marginal regression model for IPCW estimators for D (deviance based on objects), D.a (deviance based on individual data points), PR (ranked probability score), BR (Brier score), BS (modified Brier score), SH (Schemper-Henderson estimator), C* (concordance index)

4.4. Summary and Conclusion

This chapter is based on the modeling approach in Chapter 3. That means, in both chapters, a penalized regression model using a binary regression model with complementary log-log link was used for fitting discrete survival models. In Chapter 3, the tuning parameter of this penalization approach was chosen by cross-validation with the predictive deviance as loss function. In this chapter, it was investigated if the performance of such penalized regression models can be improved by modifying the loss criterion. Therefore, several prediction measures used in the framework of continuous survival outcomes are adapted to discrete time survival outcomes. The following measures were used: the predictive deviance, the discrete ranked probability score, the Brier score, the modified Brier score, the Schemper-Henderson estimator and the concordance index. The performance of these alternative loss functions is investigated by means of a simulation study. All five settings of the simulation study are conducted twice, one time with a marginal censoring model and one time with a regression censoring model for the survival function of the censoring process used in the IPCW estimates. Unlike the statement of Gerds and Schumacher (2006), the simulation results do not provide general evidence that either a marginal censoring model or a regression censoring model performs better.
Furthermore, the predictive deviance D (regarding complete cases per object with respect to the cross-validation splits) was compared to the predictive deviance D.a (using individual data points with respect to the cross-validation splits). In conclusion, with regard to all simulation settings, it can be said that D should be preferred to D.a. In most of the simulation cases, D outperforms D.a considerably or at least yield similar results.
Generally, the predictive deviance D is recommended for the choice of the loss function. In many cases D outperforms the other measures or yields only slightly worse results. This can be explained by the predictive deviance being a likelihood-based measure. Hence, the optimization criteria maximizing the penalized likelihood as well as the predictive deviance referring to the loss function, are based on the likelihood.

Finally, some remarks to the concordance index have to be made. In several cases it outperforms the other measures. However, throughout all simulation settings the concordance index has chosen very high values for the tuning parameter ξ. This leads to a large amount of parameters that are set to almost zero. Apart from that, the related curves of the cross-validation scores were sometimes degenerated. Although the concordance index is derived by time-dependent measures, as the time-dependent sensitivity and specificity and the time-dependent AUC curve, it seems to be inappropriate in the context of discrete-time survival analysis with respect to time-varying coefficients.

5. Penalization in Survival Models with Frailties

In this chapter the incorporation of random effects or frailties for survival models for discrete duration time is considered. These frailties control for unobserved heterogeneity. After a short introduction (Section 5.1), a small simulation study is presented that illustrates the issue of unobserved heterogeneity. The methodology and the estimation of discrete survival models with frailties are described in Section 5.2. In analogy to the previous chapters, a penalty term is incorporated that also allows for variable selection (Section 5.3) and the performance of the proposed method is judged by means of a simulation study (Section 5.4). The resulting method is applied to the same real data examples as in Chapter 3 illustrated in Section 5.5. Section 5.6 contains concluding remarks. In the following, only the notation and explanations with respect to grouped survival times are considered, but they can easily be modified to truly discrete survival times.

5.1. Introduction

In the previous chapters, it has been implicitly assumed that the considered population is homogeneous. However, the underlying data in survival models deal with repeated measurements that cause certain heterogeneity in the data. That means, the hazard rates after comprehension of all relevant covariates may differ for several objects. In this case, *unobserved heterogeneity* is existent. If unobserved heterogeneity is disregarded the hazard rates tend to a bias towards negative duration dependence (Heckman and Singer, 1984a). In addition, ignoring unobserved heterogeneity may lead to considerable bias for the effects of the observed covariates (e.g. Lancaster, 1990; Hougaard et al., 1994; van den Berg, 2001). A possible approach to deal respectively with unobserved heterogeneity or repeated measurements, is the incorporation of unobserved latent variables. These *random effects* or *frailty* components control for the variation between the objects. Random effects are shared by the measurements of an object and introduce a correlation between the measurements. The way of incorporating frailties to control for unobserved heterogeneity was first introduced for continuous duration models (Lancaster, 1990; Heckman and Singer, 1984a; van den Berg, 2001). In the area of biostatistics and demographics, models with incorporated latent variables are denoted as frailty models and were firstly discussed in Vaupel et al. (1979). Tuma et al. (1979) introduced these models in the context of event history analysis in social sciences. In econometrics, unobserved heterogeneity was initially covered

by Heckman and Singer (1982, 1984a,b), Flinn and Heckman (1982) and Elbers and Ridder (1982). Unobserved heterogeneity in the context of discrete survival analysis was discussed, for example, by Scheike and Jensen (1997), Jenkins (1995) or Muthén and Masyn (2005).

Random effects models are also known as *mixed models* and were introduced by Fisher (1919). Mixed models contain two kinds of effects: population-specific fixed effects and individual-specific random effects. They focus on the conditional distribution of each response value conditional on the corresponding random effect. An alternative method to control for the dependence structure of repeated measurements is the *generalized estimation equation* approach proposed by Liang and Zeger (1986). In contrast to mixed models, the response values are modeled marginally by using only population-specific effects.
The incorporation of frailties in the context of discrete survival analysis leads to generalized linear mixed models. This model class is widely used to model correlated and clustered responses. The computational issues in generalized mixed models require special tools leading to a great deal of research, accelerating in the 1990s. This results in a wide range of estimation methods. For example, Breslow and Clayton (1993), Schall (1991) and Wolfinger and O'connell (1993) proposed a joint maximization approach. That means, maximizing the joint likelihood of the observed data and the random effects simultaneously. Additionally, numerical computational techniques (e.g. Booth and Hobert, 1999; McCulloch, 1997) as well as fully Bayesian approaches (e.g. Zeger and Karim, 1991; Clayton, 1996) were introduced.

A survival model for discrete duration time is considered by means of a binary regression model (see Chapter 2). By incorporating time-varying covariate effects using B-spline basis functions for the flexible functions, this model can be perceived as a generalized linear model (see Chapter 3). Furthermore, the inclusion of random effects yield generalized linear/additive mixed models. Hence, for generalized linear/additive mixed models, the computation procedure tremendously influences the estimation approach. Especially in the case when many covariates are incorporated in the model, leading to a large number of parameters, the estimation of the parameters become unstable or even may be nonexistent. This problem becomes even more apparent if covariate effects are also assumed to be time-varying. This usually leads to an incorporation of only few covariates in the model. Hence, to avoid such problems, methods selecting relevant predictors are of particular importance.

Simple conventional variable selection methods are represented by *Forward-* and *Backward-Stepwise Selection* (e.g. Hastie et al., 2009). However, these methods exhibit stability problems that are based on the inherent discreteness of the method (Breiman, 1996).
An alternative model selection approach that is more up-to-date is based on regularization techniques. Thereby, penalization is an approved regularization approach. Adding a penalty term to the log-likelihood yields shrinkage of the estimates towards zero. Depending on the penalty, it is even possible to set particular estimates exactly to zero. One of the oldest penalization methods is the *ridge* method (Hoerl and Kennard, 1970) that

uses a L_2-type penalty on the regression coefficients. However, no variable selection can be performed by using this penalty term. An alternative penalty term that has become very popular is the *lasso* penalty term using a L_1-type penalty on the regression coefficients. In this case, variable selection can be carried out. As the lasso merely selects individual predictors, the penalty is unsatisfactory in the case of grouped data, for example with categorical predictors. The *group lasso* proposed by Yuan and Lin (2006), can overcome these problems. To get consistent estimates of the parameters, Zou (2006) extended the lasso to the *adaptive lasso* by including weights on the penalized coefficients.
Several further improvements for the lasso method have been designed in the last decade, for example the *fused lasso* (Tibshirani et al., 2005), *SCAD* (Fan and Li, 2001), *elastic net* (Zou and Hastie, 2005), *Dantzig selector* (Candes and Tao, 2007) and *DASSO* (James et al., 2009). Using penalization techniques for the selection of variables in mixed models is still in the beginning. For example, Groll and Tutz (2014) introduced a L_1-penalization for generalized linear models. A coordinate descent algorithm for generalized linear mixed models including the elastic net was proposed by Friedman et al. (2010). To utilize the features of penalization approaches and simultaneously control for unobserved heterogeneity, a penalization approach for discrete survival models with frailties is proposed in the following. For this purpose, the models and the likelihood of Chapter 3 are respectively extended by a random effect. To estimate the resulting penalized likelihood, the PIRLS algorithm of Chapter 3 is modified.

Simulation: The effect of unobserved heterogeneity

In this section a short simulation study to evaluate the consequences of ignoring unobserved heterogeneity is conducted. This issue already has been investigated by Nicoletti and Rondinelli (2010) but they do not consider time-varying coefficients. Baker and Melino (2000) and Gaure et al. (2007) investigated the omitting of unobserved heterogeneity as well, but they assumed a non-parametric distribution for the unobserved heterogeneity.

The underlying data needed for the simulation to model the discrete hazard rate $\lambda(t|\boldsymbol{z}) = P(T = t|T \geq t, \boldsymbol{z}) = F(\boldsymbol{\eta})$, where $F(\cdot)$ denotes an appropriate cumulative distribution function, are generated along the lines of Section 3.4. The only difference is the incorporation of a random intercept in the linear predictor. Apart from that, the survival time T_i is obtained by inversion sampling and the censoring times C_i are sampled from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$ with $\boldsymbol{p}_c = 1/\sum_t(\exp(t/10))\exp(t/10)$, $t = 1, ..., 30$. The minimum of survival time and censoring time defines the observed survival time $t_i = \min(T_i, C_i)$ and the censoring indicator δ_i, indicating right censoring, then follows from definition (2.6). Afterwards, the data has to be restructured as proposed in Section 2.2.2. The true linear predictor is generally defined by

$$\eta_{it}^{true} = b_i + \beta_{0t} + \sum_{j=1}^{r} z_{ij}\beta_{jt},$$

where the random effects b_i are specified by $b_i \sim \mathcal{N}(0, \sigma_b^2)$ with the scenarios $\sigma_b = (0, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2)$. The remaining components are specified in more detail in the corresponding sections of the simulation settings. Moreover, the complementary log-log link is used resulting in a binary generalized additive mixed model. For estimation, the `gam` function supplied by the R add-on package `mgcv` (Wood, 2006) is used. Thereby, the smooth components are estimated by cubic regression splines. The estimation of mixed models in the `gam` function is carried out by penalized maximum likelihood techniques. For both settings, 100 simulation runs are executed.
Two different types of models are estimated for each setting. The fixed effects are estimated in both models identically but the first model incorporates random intercepts, whereas the second (marked with tilde) ignores them. With $\boldsymbol{\beta}^T = (\beta_{01}, ..., \beta_{0q}, ..., \beta_{r1}, ..., \beta_{rq})$ and $\tilde{\boldsymbol{\beta}}^T = (\tilde{\beta}_{01}, ..., \tilde{\beta}_{0q}, ..., \tilde{\beta}_{r1}, ..., \tilde{\beta}_{rq})$, the performance of the estimates is evaluated separately for the structural components and the variance by the following mean squared errors

$$MSE_\beta = ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\,||^2, \qquad MSE_{\tilde{\beta}} = ||\boldsymbol{\beta} - \hat{\tilde{\boldsymbol{\beta}}}\,||^2, \qquad MSE_{\sigma_b} = \|\sigma_b - \hat{\sigma}_b\|^2 .$$

Setting 1

At the beginning, the focus lies on a simple intercept model given by

$$\eta_{it}^{true} = b_i + \beta_{0t},$$

where $i = 1, ..., n$, $n = 500$ and $t = 1, ..., 30$. The time-varying intercept is defined by

$$\beta_{0t^*} = 1.5 \cdot \Gamma(\nu, \alpha) - 3,\ \nu = 2,\ \alpha = 1 \text{ and } t^* = (t-1)/6,$$

where $\Gamma(\nu, \alpha)$ denotes the density of a gamma distribution with shape parameter ν and scale parameter α. By incorporating $\sigma_b = (0, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2)$, eight different scenarios are provided. The two models used for fitting are given by

$$\lambda^{model}(t|\boldsymbol{z}) = F(b_i + \beta_{0t}) \qquad \tilde{\lambda}^{model}(t|\boldsymbol{z}) = F(\tilde{\beta}_{0t}),$$

where $F(\cdot) = 1 - \exp(-\exp(\boldsymbol{\eta}))$ defines the complementary log-log link. The estimates of β_{0t} results from a simple intercept model that rightly incorporates random effects as in the true model and the estimates of $\tilde{\beta}_{0t}$ from a simple intercept model that does not consider random effects. Hence, estimates of σ_b are only provided by the first model.

Results for Setting 1 are shown in Table 5.1 and Figures 5.1, 5.2. The best mean squared errors referred to the variance are obtained for $\sigma_b = 0.25$ (Table 5.1). For $\sigma_b \geq 0.25$ all MSE values increase continuously and for $\sigma_b > 0.1$, MSE_β yields better MSE values than $\text{MSE}_{\tilde{\beta}}$.

In Figure 5.1, the performance of the MSE values of σ_b, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ is illustrated. In particular, for higher values of σ_b all MSE values increase considerably. For small values of σ_b the MSE

σ_b	MSE_σ	MSE_β	$\text{MSE}_{\tilde{\beta}}$
0.00	0.052	0.039	0.032
0.10	0.019	0.039	0.037
0.25	0.007	0.035	0.045
0.50	0.110	0.063	0.079
0.75	0.431	0.136	0.165
1.00	0.850	0.252	0.366
1.50	1.401	0.560	0.873
2.00	2.192	0.866	1.359

Table 5.1. Mean squared errors for σ_b, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ for Setting 1. Thereby, $\boldsymbol{\beta}$ indicates the estimates resulting from a model with incorporated random effects and $\tilde{\boldsymbol{\beta}}$ the estimates resulting from a model ignoring the random effects. The displayed values represent the means over all simulation runs.

values of the structural components are quite similar but for higher values of σ_b the MSE values belonging to $\boldsymbol{\beta}$ are considerably lower than those belonging to $\tilde{\boldsymbol{\beta}}$. When the values of σ_b exceed 0.75, the corresponding MSE values become larger. This is due to the fact, that in this cases the value of the random intercept b_i has a high impact on the linear predictor compared to the remaining fixed effects leading to a more unstable data basis.

Exemplarily for the case $\sigma_b = 0.5$, the estimates of the baseline effects β_{0t} are shown in Figure 5.2. It is obvious that the model ignoring the random effects $\tilde{\lambda}(t|\boldsymbol{z})$ underestimates the baseline effects. The increasing deviation of the estimated baseline effects at the end of the time periods is due to the data situation. That means, due to censoring and the occurrence of events the number of objects that are at risk diminish with increasing time.

Setting 2

In this setting, the realizations of a time-independent normal distributed covariate $Z_i \overset{iid}{\sim} \mathcal{N}(1, 0.5)$ are included in the model. This leads to the true linear predictor

$$\eta_{it}^{true} = b_i + \beta_{0t} + z_i\beta_{1t},$$

where $i = 1, ..., n$, $n = 550$ and $t = 1, ..., 30$. The time-varying coefficients β_{0t} and β_{1t} are defined by

$$\begin{aligned}
&\beta_{0t^*} = 1.5 \cdot \Gamma(\nu, \alpha) - 3,\ \nu = 2,\ \alpha = 1 \text{ and } t^* = (t-1)/6,\\
&\beta_{1t^*} = Exp(1) - 1,\ t^* = t/5,
\end{aligned}$$

where $\Gamma(\nu, \alpha)$ denotes the density of a gamma distribution with shape parameter ν and scale parameter α. Moreover, $Exp(\cdot)$ denotes the density of an exponential distribution.

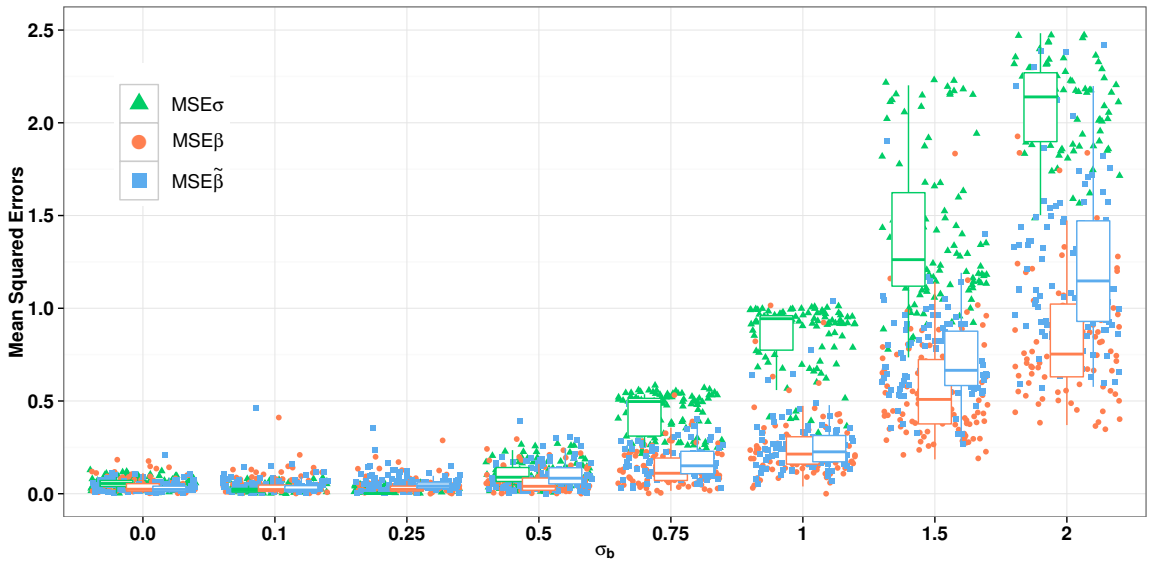


Figure 5.1. Mean squared errors and corresponding boxplots for σ_b, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ for Setting 1. Thereby, $\boldsymbol{\beta}$ indicates the estimates resulting from a model with incorporated random effects and $\tilde{\boldsymbol{\beta}}$ the estimates resulting from a model ignoring the random effects.

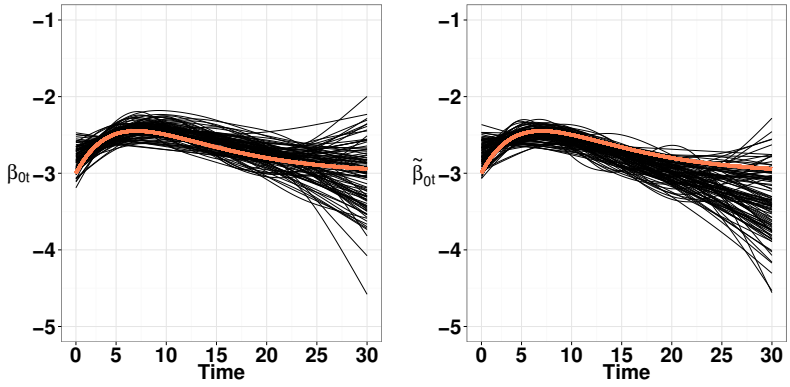


Figure 5.2. Baseline hazard functions for Setting 1 for $\sigma_b = 0.5$. Thereby, $\boldsymbol{\beta}_{0t}$ indicates the estimates resulting from a model with incorporated random effects and $\tilde{\boldsymbol{\beta}}_{0t}$ the estimates resulting from a model ignoring the random effects. The red line indicates the true baseline hazard function.

σ_b	MSE_σ	MSE_β	$\text{MSE}_{\tilde{\beta}}$
0.00	0.002	0.085	0.080
0.10	0.007	0.078	0.096
0.25	0.044	0.069	0.083
0.50	0.212	0.060	0.099
0.75	0.409	0.066	0.117
1.00	0.602	0.108	0.168
1.50	1.175	0.257	0.424
2.00	2.323	0.365	0.750

Table 5.2. Mean squared errors for σ_b, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ for Setting 2. Thereby, $\boldsymbol{\beta}$ indicates the estimates resulting from a model with incorporated random effects and $\tilde{\boldsymbol{\beta}}$ the estimates resulting from a model ignoring the random effects. The displayed values represent the means over all simulation runs.

By incorporating $\sigma_b = (0, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2)$, eight different scenarios are provided. The two models used for fitting are given by

$$\lambda^{model}(t|\boldsymbol{z}) = F(b_i + \beta_{0t} + \boldsymbol{z}\beta_{1t}) \qquad \tilde{\lambda}^{model}(t|\boldsymbol{z}) = F(b_i + \tilde{\beta}_{0t} + \boldsymbol{z}\tilde{\beta}_{1t}),$$

where $F(\cdot) = 1 - \exp(-\exp(\boldsymbol{\eta}))$ defines the complementary log-log link. The first fitted model rightly incorporates random intercepts as in the true model, whereas the second (denoted by tilde) ignores random intercepts.

The MSE_σ values increase continuously for higher values of σ_b (Table 5.2). Except for the setting with $\sigma_b = 0$, the values of MSE_β outperform that of $\text{MSE}_{\tilde{\beta}}$. In particular, for $\sigma_b \geq 0.1$ MSE_β is considerably smaller than $\text{MSE}_{\tilde{\beta}}$. The MSE values in Figure 5.3, where the MSE values of all simulation runs including the corresponding boxplots are illustrated, arrive to the same conclusion. The corresponding baseline effects are shown exemplary for $\sigma_b = 0.5$ (Figure 5.4). Compared to the red line marking the true baseline hazard the curves of $\tilde{\beta}_{0t}$ cannot cope with the underlying data structure. In contrast, β_{0t} yield good results especially for early observation times.

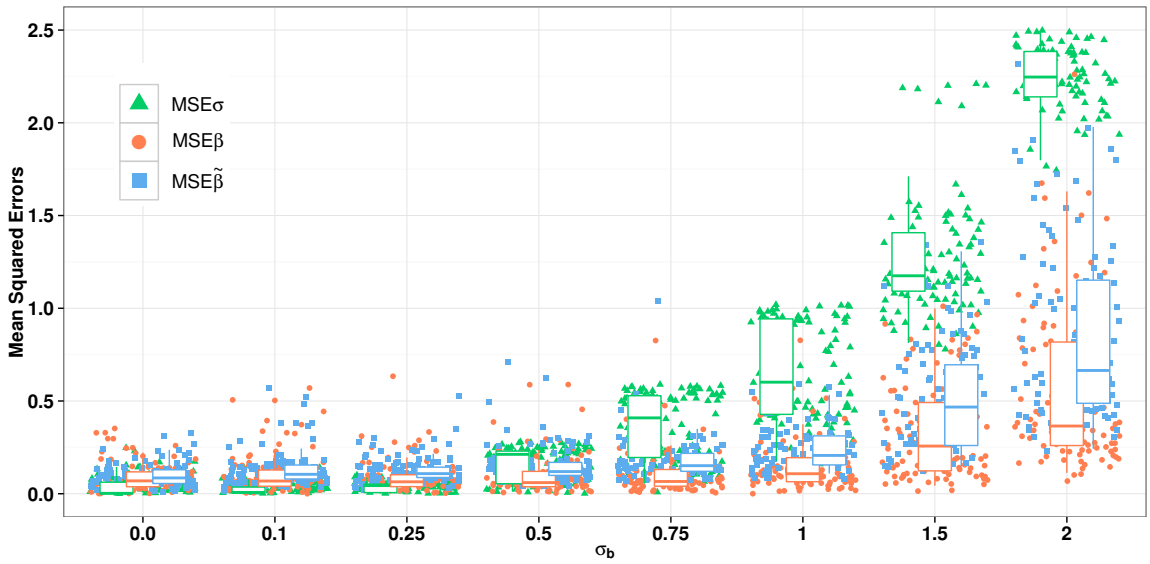


Figure 5.3. Mean squared errors and corresponding boxplots for σ_b, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ for Setting 2. Thereby, $\boldsymbol{\beta}$ indicates the estimates resulting from a model with incorporated random effects and $\tilde{\boldsymbol{\beta}}$ the estimates resulting from a model ignoring the random effects.

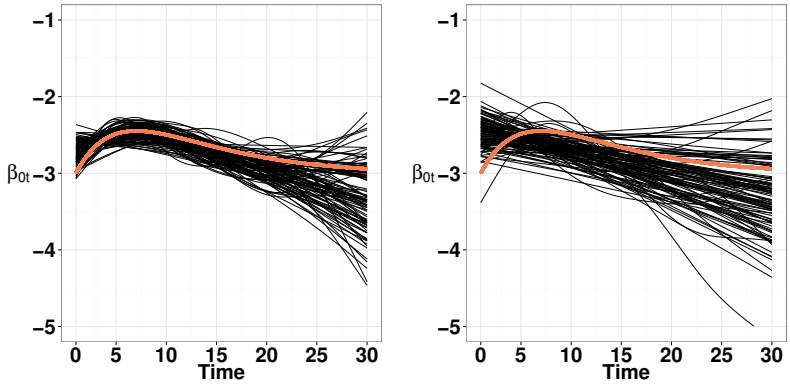


Figure 5.4. Baseline hazard functions for Setting 2 for $\sigma_b = 0.5$. Thereby, $\boldsymbol{\beta}_{0t}$ indicates the estimates resulting from a model with incorporated random effects and $\tilde{\boldsymbol{\beta}}_{0t}$ the estimates resulting from a model ignoring the random effects. The red line indicates the true baseline hazard function.

5.2. Discrete Survival Models with Frailties

The previous simulations have shown that ignoring unobserved heterogeneity may lead to problems with the estimation results. In many applications with regard to discrete-time survival analysis, a considerable variation in the measurements of an object might occur. Therefore, in this section, random intercepts are incorporated in discrete survival models to control for the unobserved heterogeneity.

5.2.1. Methodology

In analogy to the previous chapters, time T is considered as a non-negative random variable taking values from $\{1, ..., q\}$. Let time T be divided into $q + 1$ intervals $[a_0, a_1), [a_1, a_2), ..., [a_{q-1}, a_q), [a_q, \infty)$, where $T = t$ denotes an event within interval $[a_{t-1}, a_t)$. The main issue in survival analysis is the modeling of the discrete hazard function given by

$$\lambda(t|\boldsymbol{z}_{it}, \boldsymbol{x}_{it}) = P(T = t|T \geq t, \boldsymbol{z}_{it}, \boldsymbol{x}_{it}) = F(\eta_{it}),\ i = 1, ..., n,\ t = 1, ..., q, \tag{5.1}$$

where $F(\cdot)$ is an appropriate cumulative distribution function (see Chapter 2). In the following, $F(\cdot) = 1 - \exp(-\exp(\cdot))$ is assumed. A common feature of survival data is censoring that have to be incorporated in the modeling approach. For the i-th subject let C_i denote the individual censoring time. Moreover, random censoring is assumed, that is, the censoring time C is independent of the survival time T. Consequently, the observed survival time is defined by $t_i = \min(T_i, C_i)$. The censoring indicator $\delta_i = I(T_i \leq C_i)$ determines whether observation i is right censored ($\delta_i = 0$) or an event occurs ($\delta_i = 1$) with regard to the interval $[a_{t_i-1}, a_{t_i})$. By defining binary event indicators $y_{it} = 1$, for $t = t_i$ and $\delta_i = 1$ and $y_{it} = 0$, otherwise, the model (5.1) can be considered as a binary regression model (see Chapter 2). Hence, let y_{it} denote the binary outcome of an object i, $i = 1, ..., n$, in period t, $t = 1, ..., t_i$, and let $(\boldsymbol{z}_{it}, \boldsymbol{x}_{it})^T = (z_{it1}, ..., z_{itr}, x_{it1}, ..., x_{its})$ with $p = r + s$ be a vector of realizations of explanatory variables that may vary over time.

The linear predictor for object i, $i = 1, ..., n$, and time period t, $t = 1, ..., q$, is defined by

$$\eta_{it} = \beta_{0t} + \sum_{j=1}^{r} z_{itj}\beta_{jt} + \sum_{l=1}^{s} x_{itl}\gamma_l, \tag{5.2}$$

where the parameter β_{0t} represents the baseline hazard function that is the same for all individuals, that means unobserved heterogeneity is ignored. Moreover, $\boldsymbol{z}_{\bullet\bullet 1}, ..., \boldsymbol{z}_{\bullet\bullet r}$ with $\boldsymbol{z}_{\bullet\bullet j}^T = (z_{11j}, ..., z_{1t_ij}, ..., z_{n1j}, ..., z_{nt_ij})$ denote the observations of the covariates that are allowed to exhibit time-varying effects and $\boldsymbol{x}_{\bullet\bullet 1}, ..., \boldsymbol{x}_{\bullet\bullet s}$ with $\boldsymbol{x}_{\bullet\bullet l}^T = (x_{11l}, ..., x_{1t_il}, ..., x_{n1l}, ..., x_{nt_il})$ define the observations of the covariates that are restricted to have constant effects. In other words, the model implies time-varying coefficients $\beta_{jt}^T = (\beta_{0t}, \beta_{rt}, ..., \beta_{rt})$ including a time-varying intercept β_{0t}, whereas $\boldsymbol{\gamma}^T = (\gamma_1, ..., \gamma_s)$ is

assumed to be time-constant. When the covariates do not depend on time, the time index is omitted, that is, $z_{itj} = z_{ij}$ and $x_{itl} = x_{il}$.

To control for unobserved heterogeneity, the linear predictor (5.2) can be extended by a random intercept resulting in

$$\eta_{it} = b_i + \beta_{0t} + \sum_{j=1}^{r} z_{itj}\beta_{jt} + \sum_{l=1}^{s} x_{itl}\gamma_l,$$

where $\boldsymbol{\beta}_{jt}$, $j = 0, ..., r$, and $\boldsymbol{\gamma}_l$, $l = 1, .., s$, $t = 1, ..., q$, determine the fixed effects. Moreover, b_i is considered to be an individual-specific random effect. With explanatory variables $\boldsymbol{z}_{it}^T = (\boldsymbol{z}_{it1}, ..., \boldsymbol{z}_{itr})$ and $\boldsymbol{x}_{it}^T = (\boldsymbol{x}_{it1}, ..., \boldsymbol{x}_{its})$ the conditional means $\lambda(t|\boldsymbol{z}_{it}, \boldsymbol{x}_{it}, b_i) = E(y_{it}|\boldsymbol{z}_{it}, \boldsymbol{x}_{it}, b_i)$ are considered. The individual-specific parameters b_i are assumed to be independent with $E(b_i) = 0$ and have density $p(b_i)$. The mean of the b_i is set to zero because the population mean is already represented by the fixed baseline hazard parameters β_{0t}. The individual-specific random effects are assumed to be normally distributed with

$$\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}),$$

where $\boldsymbol{Q} = diag(\sigma_b^2, ..., \sigma_b^2)$ and $\boldsymbol{b}^T = (b_1, ..., b_n)$. Including individual-specific random effects b_i, responses y_{it} given b_i are conditionally independent. Furthermore, the random effects b_i are assumed to be independent from the covariates.

Along the lines of Chapter 2, under random censoring the probability of observing (t_i, δ_i) can be written as

$$P(t_i, \delta_i|\boldsymbol{z}_i, \boldsymbol{x}_i, b_i) = P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i},$$

where the vectors of covariates $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ are assumed to be time-independent. This probability is defined given covariates and random effect b_i, which are suppressed on the right hand side of the equation. Under the assumption that censoring contributions do not depend on the parameters that determine the survival time (non-informative in the sense of Kalbfleisch and Prentice, 2002), the factor $c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ can be omitted and a simpler form is given by

$$P(t_i, \delta_i|\boldsymbol{z}_i, \boldsymbol{x}_i, b_i) = \prod_{t=1}^{t_i} \lambda(t|\boldsymbol{z}_i, \boldsymbol{x}_i, b_i)^{y_{it}} (1 - \lambda(t|\boldsymbol{z}_i, \boldsymbol{x}_i, b_i))^{1-y_{it}}. \tag{5.3}$$

Moreover, analogous to Chapter 2, binary event indicators representing the binary sequences, are incorporated in Equation (5.3). They are defined by

$$y_{it} = \begin{cases} 1, & \text{for } t = t_i \text{ and } \delta_i = 1 \\ 0, & \text{otherwise.} \end{cases}$$

As the probability $P(t_i, \delta_i | \boldsymbol{z}_i, \boldsymbol{x}_i, b_i)$ is defined given the random effect b_i, the unconditional probability can be obtained by

$$\begin{aligned} P(t_i, \delta_i | \boldsymbol{z}_i, \boldsymbol{x}_i) &= \int P(t_i, \delta_i | \boldsymbol{z}_i, \boldsymbol{x}_i, b_i) p(b_i) db_i \\ &= \int \prod_{t=1}^{t_i} \lambda(t | \boldsymbol{z}_i, \boldsymbol{x}_i, b_i)^{y_{it}} (1 - \lambda(t | \boldsymbol{z}_i, \boldsymbol{x}_i, b_i))^{1-y_{it}} p(b_i) db_i. \end{aligned}$$

5.2.2. Estimation

The fixed parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and σ_b^2 are estimated by maximizing the marginal log-likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_b^2) = \sum_{i=1}^{n} \log \left(\int \prod_{t=1}^{t_i} \lambda(t | \boldsymbol{z}_i, \boldsymbol{x}_i, b_i)^{y_{it}} (1 - \lambda(t | \boldsymbol{z}_i, \boldsymbol{x}_i, b_i))^{1-y_{it}} p(b_i) db_i \right). \tag{5.4}$$

The intractable integral in the marginal log-likelihood (5.4) is the main impediment to apply mixed models for discrete survival models. There are two general types of solutions of the integral. The first is to approximate the integral numerically, so that the marginal likelihood can be computed and optimized. The second is the approximation of the integrand leading to a closed form of the integral of the approximation. An overview regarding the statistical inference of generalized linear mixed models can be found in Tuerlinckx et al. (2006) and Fahrmeir and Tutz (2001).
An approximation of the marginal log-likelihood (5.4) can be obtained, for example, via (adaptive) Gauss-Hermite quadrature (e.g. Bock and Aitkin, 1981; Stroud and Secrest, 1966; Pinheiro and Bates, 1995) or Monte Carlo integration (e.g. Tuerlinckx et al., 2006; Fahrmeir and Tutz, 2001). Two approaches for maximizing the resulting approximated likelihood may be used. Firstly, a direct maximization approach using fitting techniques for generalized linear models (e.g. Fahrmeir and Tutz, 2001) and secondly, an indirect maximization approach based on the expectation-maximization algorithm (e.g. Hinde, 1982; Brillinger and Preisler, 1983; Booth and Hobert, 1999).

Instead of numerically approximating the integral, the integrand itself may be approximated. The goal is then to find an approximation leading to a tractable integral, so that the closed-form expression that follows from it can be maximized. A possible method is the penalized quasi-likelihood (PQL) approach, which has been suggested by Breslow and Clayton (1993), Schall (1991) and Stiratelli et al. (1984). Thereby, the estimation of the variance σ_b^2 is separated from the estimation of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and b_i, that are collected in $\boldsymbol{\omega}^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, b_1, ..., b_n)$. As a discrete survival model can be represented by a binary regression model whose log-likelihood has a considerable simpler form, the likeli-

hood of a binary regression model is used in the following. This results in the marginal log-likelihood

$$l(\boldsymbol{\omega}, \sigma_b^2) = \sum_{i=1}^{n} \log \left(\int f(\boldsymbol{y}_i | \boldsymbol{\omega}, \sigma_b^2) p(b_i) db_i \right), \tag{5.5}$$

where $f(\boldsymbol{y}_i | \boldsymbol{\omega}, \sigma_b^2) = \prod_{t=1}^{t_i} \lambda(t|\boldsymbol{z}_i, \boldsymbol{x}_i, b_i)^{y_{it}} (1 - \lambda(t|\boldsymbol{z}_i, \boldsymbol{x}_i, b_i))^{1-y_{it}}$ with $\boldsymbol{y}_i^T = (y_{i1}, ..., y_{it_i})$ defining the transitions of object i and $p(b_i)$ denotes the density function of the random effects. The approximation along the lines of Breslow and Clayton (1993) is based on a Laplace approximation yielding the penalized likelihood

$$l^{app}(\boldsymbol{\omega}, \sigma_b^2) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i | \boldsymbol{\omega}, \sigma_b^2) - \frac{\sigma_b^2}{2} \sum_{i=1}^{n} b_i^2, \tag{5.6}$$

that can be maximized in place of the log-likelihood (5.5). The penalty term $\frac{\sigma_b^2}{2} \sum_{i=1}^{n} b_i^2$ stems from the approximation based on the Laplace method (see Appendix A.2). The PQL approach uses the concept of joint maximization of the penalized likelihood with respect to the parameters and the random effects appended by the estimation of the variance of the random effects. That means, given an estimate $\hat{\sigma}_b^2$ the profile likelihood $l^{app}(\boldsymbol{\omega}, \hat{\sigma}_b^2)$ is maximized and separately the random effects parameter σ_b^2 is estimated.

Additionally, the penalized log-likelihood (5.6) can also be motivated as a posterior mode estimation (see Fahrmeir and Tutz, 2001; Tutz, 2012; Fahrmeir et al., 2013). The PQL method is implemented in the functions `glmmPQL` and `gamm` of the R add-on packages `MASS` (Venables and Ripley, 2002) and `mgcv` (Wood, 2014, 2006), respectively. However, for modeling discrete frailty survival times with time-varying coefficients the R-function `glmmPQL` yields no meaningful results. This could possibly be due to the fact that `glmmPQL` is based on fitting ordinary ML-estimates that often do not converge. Hence, in this framework the R function `glmmPQL` cannot be recommended.

5.3. Penalization

To include regularization techniques in discrete frailty survival models, the penalized log-likelihood (5.6) has to be extended by a further penalty term that determines the properties of the estimated fixed effects. Hence, a penalty term following Chapter 3 can be incorporated yielding the penalized log-likelihood

$$l^{pen}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma_b^2) = l^{pen}(\boldsymbol{\omega}, \sigma_b^2) = l^{app}(\boldsymbol{\omega}, \sigma_b^2) - J_{\xi_0, \xi}(\boldsymbol{\beta}, \boldsymbol{\gamma}).$$

The estimation of the random effects b_i and the corresponding variance σ_b^2 is already very extensive. To yield more parsimonious models, the time-varying coefficients are expanded

in equally spaced B-splines with $\beta_{jt} = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t)$, $j = 0, ..., r$. The corresponding penalized log-likelihood is obtained by

$$l^{pen}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma_b^2) = l^{app}(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{b}, \sigma_b^2) - J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}). \tag{5.7}$$

A penalty that enforces variable selection and smoothness of the baseline effects is given by

$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{m_0} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi \left(\phi \sum_{j=1}^{r} \psi_j \left\| \boldsymbol{\zeta}_j \right\|_2 \right) + \xi \left((1-\phi) \sum_{j=1}^{r} \varphi_j \left\| \boldsymbol{\alpha}_j \right\|_2 \right), \tag{5.8}$$

where $\|\cdot\|_2$ denotes the L_2-norm and $\boldsymbol{\zeta}_j^T = (\zeta_{j2}, ..., \zeta_{jm_j})$, $\zeta_{jm} = \alpha_{jm} - \alpha_{j,m-1}$, $m = 2, ..., m_j$, $\boldsymbol{\alpha}_j^T = (\alpha_{j1}, ..., \alpha_{jm_j})$. The first term of $J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ enforces shrinkage of the differences between adjacent B-spline coefficients of the baseline hazard with the objective of a smooth function over time. This part of penalization is predominantly incorporated due to stability reasons. Hence, the tuning parameter ξ_0 should be chosen rather small, for example $\xi_0 = 0.001$. By using a group lasso penalty with respect to the differences $\zeta_{j2}, ..., \zeta_{jm_j}$, the second term steers the smoothness of the time-varying covariate effects, but for a value of ξ large enough, all differences $\zeta_{j2}, ..., \zeta_{jm_j}$ are removed from the model resulting in a constant covariate effect. Finally, the third term steers the selection of covariates (see Chapter 3). The latter term corresponds to a group lasso penalty with regard to the parameters α_{jm}, $m = 1, ..., m_j$, belonging to the j-th covariate. If the tuning parameter ξ exceeds a certain value, the values of the parameters $\alpha_{j1}, ..., \alpha_{jm_j}$ are set to zero and the covariate j is removed from the model. That means, the penalty term may distinguish if a covariate effect is incorporated smooth or constant in the model or if it is removed from the model. Weighting of the second part and the selection part is obtained by parameter ϕ, that is a further tuning parameter. Both tuning parameters ξ and ϕ have to be chosen by an appropriate technique. The terms $\psi_j = \sqrt{m_j - 1}$ and $\varphi_j = \sqrt{m_j}$ are weights that assign different amounts of penalization to different parameter groups, relative to the respective group size. In analogy to Chapter 3, the penalty term might be extended by a penalty regarding the time-constant parameters γ_l.

For given $\hat{\sigma}_b^2$ the optimization problem reduces to

$$\hat{\boldsymbol{\omega}}_\alpha = \arg\max_{\boldsymbol{\omega}_\alpha} l^{pen}(\boldsymbol{\omega}_\alpha, \hat{\sigma}_b^2) = \arg\max_{\boldsymbol{\omega}_\alpha} \left(l^{app}(\boldsymbol{\omega}_\alpha, \hat{\sigma}_b^2) - J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \right),$$

where the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{b}$ are collected in $\boldsymbol{\omega}_\alpha^T = (\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{b}^T)$. By considering the whole parameter vector $\boldsymbol{\omega}_\alpha^T = (\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{b}^T)$, the penalized log-likelihood (5.7) can be seen as a partially penalized approach.

5.3.1. Numerical Computation

Maximizing the penalized likelihood $l^{pen}(\boldsymbol{\omega}_\alpha, \hat{\sigma}_b^2)$ is obtained by solving

$$s_p(\boldsymbol{\omega}_\alpha, \sigma_b^2) = (\partial l^{pen}(\boldsymbol{\omega}_\alpha, \sigma_b^2)/\partial \boldsymbol{\alpha}^T, \partial l^{pen}(\boldsymbol{\omega}_\alpha, \sigma_b^2)/\partial \boldsymbol{\gamma}^T, \partial l^{pen}(\boldsymbol{\omega}_\alpha, \sigma_b^2)/\partial \boldsymbol{b}^T) = \mathbf{0},$$

where $s_p(\boldsymbol{\omega}_\alpha, \sigma_b^2)$ denotes the penalized score function. The closed form of the score function is given by

$$s_p(\boldsymbol{\omega}_\alpha, \sigma_b^2) = \tilde{\boldsymbol{X}}\boldsymbol{D}(\boldsymbol{\omega}_\alpha)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega}_\alpha)(\boldsymbol{y} - \lambda(t|\boldsymbol{z}, \boldsymbol{x})) - \boldsymbol{K}\boldsymbol{\omega}_\alpha,$$

where $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{Z}}|\boldsymbol{X}|\mathbf{1}]$, with $\boldsymbol{X}^T = \left[(\boldsymbol{x}_{11}, ..., \boldsymbol{x}_{1t_1})^T, ..., (\boldsymbol{x}_{n1}, ..., \boldsymbol{x}_{nt_n})^T\right]$, $\tilde{\boldsymbol{Z}}^T = \left[(\tilde{\boldsymbol{z}}_{11}, ..., \tilde{\boldsymbol{z}}_{1t_1})^T, ..., (\tilde{\boldsymbol{z}}_{n1}, ..., \tilde{\boldsymbol{z}}_{nt_n})^T\right]$ and $\tilde{\boldsymbol{z}}_{it}$ contains the design of the interactions of the according covariates and the evaluations of the appropriate B-spline basis functions (see Equation (3.3) and Chapter 2). Moreover, $\boldsymbol{D}(\boldsymbol{\omega}_\alpha) = \partial h(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$, $h(\cdot) = 1 - \exp(-\exp(\cdot))$, $\lambda(t|\boldsymbol{z}, \boldsymbol{x}) = h(\boldsymbol{\eta})$, $\boldsymbol{\Sigma}(\boldsymbol{\omega}_\alpha) = cov(\boldsymbol{y}|\boldsymbol{\omega}_\alpha)$ and

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}^{-1} \end{pmatrix}.$$

Thereby, the matrix $\boldsymbol{A}$ contains the penalization components belonging to the local quadratic approximations of the penalty term $J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ (see Chapter 3 and Oelker and Tutz (2013)) and $\boldsymbol{Q} = diag(\sigma_b^2, ..., \sigma_b^2)$ contains the variance components of the random effects. The corresponding penalized Fisher matrix is defined by

$$\boldsymbol{F}_p(\boldsymbol{\omega}_\alpha, \sigma_b^2) = \tilde{\boldsymbol{X}}^T\boldsymbol{W}(\boldsymbol{\omega}_\alpha)\tilde{\boldsymbol{X}}^T + \boldsymbol{K} \quad \text{with} \quad \boldsymbol{W}(\boldsymbol{\omega}_\alpha) = \boldsymbol{D}(\boldsymbol{\omega}_\alpha)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega}_\alpha)\boldsymbol{D}(\boldsymbol{\omega}_\alpha)^T.$$

By means of the penalized versions of the score function and the Fisher matrix, a penalized iteratively re-weighted least squares (PIRLS) algorithm, that is a pseudo Fisher scoring in this case, can be executed and is given in the following. It is named `fpendsm` for **F**railty **PEN**alized **D**iscrete **S**urvival **M**odels.

Algorithm `fpendsm`

Initialization
Compute starting values $\hat{\boldsymbol{\alpha}}^{(0)}$, $\hat{\boldsymbol{\gamma}}^{(0)}$, $\hat{\boldsymbol{b}}^{(0)}$, $(\hat{\sigma}_b^2)^{(0)}$ and the linear predictor $\hat{\boldsymbol{\eta}}^{(0)} = \boldsymbol{b}^{(0)} + \tilde{\boldsymbol{Z}}\hat{\boldsymbol{\alpha}}^{(0)} + \boldsymbol{X}\hat{\boldsymbol{\gamma}}^{(0)}$. Set $k = 0$.

Iteration

1. Determine $\hat{\boldsymbol{\alpha}}^{(k+1)}$, $\hat{\boldsymbol{\gamma}}^{(k+1)}$, $\hat{\boldsymbol{b}}^{(k+1)}$ by

$$\hat{\boldsymbol{\omega}}_\alpha^{(k+1)} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^{(k+1)} \\ \hat{\boldsymbol{\gamma}}^{(k+1)} \\ \hat{\boldsymbol{b}}^{(k+1)} \end{pmatrix} = (\tilde{\boldsymbol{X}}^T\boldsymbol{W}(\hat{\boldsymbol{\omega}}_\alpha^{(k)})\tilde{\boldsymbol{X}} + \boldsymbol{K})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}(\hat{\boldsymbol{\omega}}_\alpha^{(k)})\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\omega}}_\alpha^{(k)}),$$

with pseudo-observations $\tilde{\boldsymbol{\eta}}(\boldsymbol{\omega}_\alpha) = \tilde{\boldsymbol{X}}\boldsymbol{\omega}_\alpha + \boldsymbol{D}^{-1}(\boldsymbol{\omega}_\alpha)(\boldsymbol{y} - \lambda(t|\boldsymbol{z}, \boldsymbol{x}))$. Thereby, the current estimate of $\boldsymbol{Q} = diag(\sigma_b^2, ..., \sigma_b^2)$ is used. More details on the inversion of the pseudo-Fisher matrix $\boldsymbol{F}_p(\boldsymbol{\omega}_\alpha, \sigma_b^2)$ are given in Appendix A.3.

2. Computation of the variance components
   Estimates $(\hat{\sigma}_b^2)^{(k)}$ are typically obtained by an approximate EM-type algorithm or by an approximate REML-type algorithm (see Section 5.3.2). Thereby, the current estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, $\boldsymbol{b}$ are used.

Iterate between one step of Fisher scoring (1.) yielding an estimate for $\hat{\boldsymbol{\omega}}_\alpha$ and one step of updating $\hat{\sigma}_b^2$ (2.) until convergence.

5.3.2. Computational Details

In the following some details on the algorithm are described. Details on the initialization of the starting values are outlined and two estimation techniques for the variance components are described. Furthermore the tuning parameter selection for discrete frailty survival models is explained. Finally, adaptive penalties are presented.

Starting Values

For the initialization of the starting values $\hat{\boldsymbol{\alpha}}^{(0)}$, $\hat{\boldsymbol{\gamma}}^{(0)}$, $\hat{\boldsymbol{b}}^{(0)}$, $(\hat{\sigma}_b^2)^{(0)}$ from step 1 of the algorithm `fpendsm`, a corresponding generalized linear mixed model, using a slight ridge penalty of 0.001 for the fixed effects, can be fitted.

Variance-Covariance Components

The penalized log-likelihood approach of the previous section yields estimates of $\boldsymbol{\omega}_\alpha$ based on the assumption that σ_b^2 is known. Variance estimates can be obtained by an approximate EM-algorithm, using the posterior mode estimates $\hat{\boldsymbol{\omega}}_\alpha^{(k)}$ and the posterior curvatures $\hat{V}_{ii}^{(k)}$. With $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$, the corresponding formula for computing $(\hat{\sigma}_b^2)^{(k)}$ is given by

$$(\hat{\sigma}_b^2)^{(k)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{V}_{ii}^{(k)} + (\hat{b}_i^{(k)})^2),$$

where V_{ii} can be derived by

$$V_{ii} = F_{ii}^{-1} + F_{ii}^{-1}F_{i\tilde{\boldsymbol{\beta}}_{act}}(\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}_{act}\tilde{\boldsymbol{\beta}}_{act}} - \sum_{i=1}^{n} F_{\tilde{\boldsymbol{\beta}}_{act}i}F_{ii}^{-1}F_{i\tilde{\boldsymbol{\beta}}_{act}})^{-1}F_{\tilde{\boldsymbol{\beta}}_{act}i}F_{ii}^{-1}.$$

In analogy to Groll (2011), $\tilde{\boldsymbol{\beta}}_{act}$ is the set of "active" covariates, corresponding to the non-zero coefficients and $\tilde{\boldsymbol{X}}_{act}$ is the corresponding design matrix that consists only of the

columns belonging to non-zero coefficients. $\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}_{act}\tilde{\boldsymbol{\beta}}_{act}}$, $F_{i\tilde{\boldsymbol{\beta}}_{act}}$, $F_{\tilde{\boldsymbol{\beta}}_{act}i}$ and F_{ii} are the elements of the partitioned Fisher matrix, for details see Appendix A.3 with $\tilde{\boldsymbol{\beta}}_{act}$ and $\tilde{\boldsymbol{X}}_{act}$ in place of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{X}}$.

An alternative estimation of variances was proposed in Breslow and Clayton (1993), wherein the authors suggested to maximize the profile likelihood that is associated with the normal theory model. With $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\alpha}, \boldsymbol{\gamma})$ and replacing $\tilde{\boldsymbol{\beta}}$ by $\hat{\tilde{\boldsymbol{\beta}}}$, the following term

$$\begin{aligned} l(\sigma_b^2) = &- \frac{1}{2}\log(|\boldsymbol{V}(\hat{\boldsymbol{\omega}}_\alpha)|) - \frac{1}{2}\log(\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\omega}}_\alpha)\boldsymbol{X}\right|) \\ &- \frac{1}{2}(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\omega}}_\alpha) - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}})^T\boldsymbol{V}^{-1}(\hat{\boldsymbol{\omega}}_\alpha)(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\omega}}_\alpha) - \boldsymbol{X}\hat{\tilde{\boldsymbol{\beta}}}) \end{aligned}$$

has to be maximized with respect to σ_b^2, with pseudo-observations $\tilde{\boldsymbol{\eta}}(\boldsymbol{\omega}_\alpha) = \tilde{\boldsymbol{X}}\boldsymbol{\omega}_\alpha + \boldsymbol{D}^{-1}(\boldsymbol{\omega}_\alpha)(\boldsymbol{y} - \lambda(t|\boldsymbol{z}, \boldsymbol{x}))$ and $\boldsymbol{V}(\boldsymbol{\omega}_\alpha) = \boldsymbol{W}^{-1}(\boldsymbol{\omega}_\alpha)\mathbf{1}\boldsymbol{Q}\mathbf{1}^T$, $\boldsymbol{Q} = diag(\sigma_b^2, ..., \sigma_b^2)$, $\boldsymbol{W}(\boldsymbol{\omega}_\alpha) = \boldsymbol{D}(\boldsymbol{\omega}_\alpha)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega}_\alpha)\boldsymbol{D}(\boldsymbol{\omega}_\alpha)^T$ and $\boldsymbol{\Sigma}(\boldsymbol{\omega}_\alpha) = cov(\boldsymbol{y}|\boldsymbol{\omega}_\alpha)$. Having calculated $\hat{\boldsymbol{\omega}}_\alpha^{(k)}$, the obtained estimate $(\hat{\sigma}_b^2)^{(k)}$ is an approximate REML-type estimate for σ_b^2.

Tuning Parameter Selection

The tuning parameters ξ and ϕ are chosen by K-fold cross-validation (see Section 3.2.2). However, there is a special characteristic of the cross-validation with respect to discrete frailty survival models that has to be accounted for. The hazard rate for discrete survival models $\lambda(t|\boldsymbol{z}, \boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{z}, \boldsymbol{x})$ describes the conditional probability for the risk of failure in interval $[a_{t-1}, a_t)$, given the interval is reached. Therefore, it is not an adequate cross-validation approach to separate single observations from the entire measurement of an object. This fact was confirmed by the results of Chapter 4. Consequently, all observations of an object are located either in the learning sample or in the test sample. If the measurements of an object are located in the test sample, there are no corresponding estimates of the random effects that are available for prediction.
In this case, the random effects are assumed to be zero that is their most likely value as the random effects are normally distributed with $b_i \sim \mathcal{N}(0, \sigma_b^2)$. Effectively, for prediction, only the "fixed part" of the model is used.
In analogy to Chapter 3, the predictive performance in the cross-validation approach is assessed by the predictive deviance. For a new observation $(t_i^{pred}, \delta_i^{pred}, \boldsymbol{z}_i^{pred}, \boldsymbol{x}_i^{pred})$, with $\boldsymbol{z}_i^{pred} = (\boldsymbol{z}_{i1}^{pred}, ..., \boldsymbol{z}_{it_i}^{pred})^T$ and $\boldsymbol{x}_i^{pred} = (\boldsymbol{x}_{i1}^{pred}, ..., \boldsymbol{x}_{it_i}^{pred})^T$ the predictive deviance is defined by

$$\mathrm{D}_i = -2\sum_{t=1}^{t_i}\left\{y_{it}^{pred}\log(\hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})) + (1 - y_{it}^{pred})\log(1 - \hat{\lambda}(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred}))\right\},$$

where $\lambda(t|\boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred}) = P(T_i = t|T_i \geq t, \boldsymbol{z}_{it}^{pred}, \boldsymbol{x}_{it}^{pred})$ and $(y_{i1}^{pred}, ..., y_{it_i}^{pred})$ denotes the transitions over periods of object i. For choosing the tuning parameters ξ and ϕ a two-

dimensional grid of possible parameters is used, on which the optimal parameter combination is chosen.

Adaptive Penalties

Similar to Section 3.2.1, the idea of adaptive penalties can be used by weighting the penalty terms by the inverse of the respective unpenalized parameter estimates. This modification is applied due to the inconsistency of simple lasso or group lasso penalties (Zou, 2006; Wang and Leng, 2008). Given the penalty (5.8), the adaptive version is obtained by replacing the weights ψ_j and φ_j by

$$\psi_j^a = \frac{\sqrt{m_j - 1}}{||\hat{\boldsymbol{\zeta}}_j^{\mathrm{ML}}||_2}, \quad \varphi_j^a = \frac{\sqrt{m_j}}{||\hat{\boldsymbol{\alpha}}_j^{\mathrm{ML}}||_2}, \tag{5.9}$$

where $\hat{\boldsymbol{\zeta}}_j^{\mathrm{ML}}$ and $\hat{\boldsymbol{\alpha}}_j^{\mathrm{ML}}$ denote the according ML-estimates. The intuition behind this weighting procedure is rather straightforward. With very large data sets, unpenalized point estimates can be expected to be rather accurate. Thus, the norm of ML-estimates of parameter groups belonging to relevant predictors is rather large. Consequently, the corresponding penalization should be small. By contrast, a strong penalization goes along with parameter groups belonging to irrelevant predictors and, hence, leading to a small norm of ML-estimates. Moreover, the ML-estimates employed in the adaptive weights (5.9), can be replaced by any $\sqrt{n}$-consistent estimates. For example, a ridge penalty can be used in situations where the ML-estimates do not exist.

5.4. Simulation

In this section, the performance of `fpendsm`, that means, combining different penalty terms in discrete-time survival frailty models, is evaluated. Moreover, it is compared to methods using conventional generalized additive mixed models implemented in the functions `gam` and `gamm` of the R add-on package `mgcv` (Wood, 2014, 2006). For `gam`, the fitting approach of the smooth terms is carried out by a conversion to penalized regression terms. In `gamm`, the smooths are partly specified as fixed effects, but the wiggly components of the smooths are treated as random effects providing a penalized quasi-likelihood approach for generalized additive models. As in `fpendsm`, in `gam` it is possible to penalize the parameters representing the baseline hazard separately. However, the selection part is conducted differently. In `fpendsm`, it can be chosen which covariate effects might be set to zero and `fpendsm` allows to distinguish whether an effect is time-varying or time-constant. The `mgcv` package provides a model selection removing complete smoothing terms from the model by adding an extra penalty. This is achieved by setting the option `select=TRUE`. Though, this selection affects all smoothing terms. Moreover, a variable selection with regard to the parametric terms in the linear predictor is not available. In contrast, the `gamm` function cannot penalize the baseline effects separately and cannot perform variable selection.

The underlying data needed for the simulation are generated along the lines of Section 3.4. That means, the survival time T_i is obtained by inversion sampling and the censoring times C_i are sampled from a multinomial distribution. The minimum of survival time and censoring time defines the observed survival time $t_i = \min(T_i, C_i)$ and the censoring indicator δ_i then follows from definition (2.6). Afterwards, the data have to be restructured as proposed in Section 2.2.2 to yield a binary regression model. The difference to Section 3.4 lies in the computation of the linear predictor since a random intercept is incorporated. The components of the simulation settings required for computing the underlying true linear predictor

$$\eta_{it}^{true} = b_i + \beta_{0t} + \sum_{j=1}^{r} z_{ij}\beta_{jt} + \sum_{l=1}^{s} x_{il}\gamma_l$$

are given in the following. Thereby, the random effects are specified by $b_i \sim \mathcal{N}(0, \sigma_b^2)$ with the scenarios $\sigma_b = (0.1, 0.5, 0.7, 1, 2)$. For all settings, according to discrete duration times, the complementary log-log link is used. To be on comparable scales, all covariates are standardized. The tuning parameters ξ and ϕ are chosen by 5-fold cross-validation, where for the splits it is referred to whole observations of an object. Thereby, the predictive deviance is chosen as loss criterion.

Setting 1

The first scenario consists of $n = 300$ realizations of five covariates $X_{i1}, ..., X_{i5}$ independently drawn from a normal distribution $\mathcal{N}(1, 0.5)$. Only two covariates have an effect on the survival time, whereas the remaining three covariates are noise variables. There are $q = 10$ time periods considered. The covariate realizations $x_{i1}, ..., x_{i5}$ are used to simulate survival times according to the linear predictor

$$\eta_{it} = b_i + \beta_{0t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3 + x_{i4}\gamma_4 + x_{i5}\gamma_5.$$

Thereby, the time-varying intercept β_{0t} is given by

$$\beta_{0t^*} = 1.5 \cdot \Gamma(\nu, \alpha) - 1,$$

where $\Gamma(\nu, \alpha)$ denotes the density of a gamma distribution with shape parameter $\nu = 2.1$, scale parameter $\alpha = 1$ and $t^* = (t-1)/2.5$, $t = 1, ..., 10$. Furthermore, the time-constant coefficient effects γ_l are defined by $\gamma_1 = 0.5$, $\gamma_2 = -1$, $\gamma_3 = \gamma_4 = \gamma_5 = 0$. For the time-varying intercept β_{0t}, a cubic B-splines approach is used, where the number of equally spaced inner knots is set to six. The censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$, where $\boldsymbol{p}_c$ is defined by $\boldsymbol{p}_c^T = (0.02, 0.02, 0.02, 0.02, 0.02, 0.05, 0.05, 0.1, 0.15, 0.55)$. The simulation scheme for Setting 1 is replicated 50 times.

Setting 2

In simulation Setting 2, the number of time periods is set to $q = 10$ as well. The model consists of 5 covariates, whereof two are noise variables. For the study, $n = 300$ realizations of the covariates are simulated according to $Z_{i1}, Z_{i2}, X_{i1}, X_{i2}, X_{i3} \overset{iid}{\sim} \mathcal{N}(1, 0.5)$. The survival times are sampled by means of the realizations of the covariates with the linear predictor given by

$$\eta_{it} = b_i + \beta_{0t} + z_{i1}\beta_{1t} + z_{i2}\beta_{2t} + x_{i1}\gamma_1 + x_{i2}\gamma_2 + x_{i3}\gamma_3.$$

Thereby, the time-varying effects β_{jt}, $j = 0, 1, 2$, $t = 1, ..., 10$, are defined by

$$\begin{aligned}
\beta_{0t^*} &= 1.5 \cdot \Gamma(\nu, \alpha) - 1,\ \nu = 2.1,\ \alpha = 1 \text{ and } t^* = (t-1)/2.5, \\
\beta_{1t^*} &= Exp(1) - 1,\ t^* = t/3, \\
\beta_{2t} &= 0.5 \cdot \cos((t+4)/1.2),
\end{aligned}$$

where $\Gamma(\nu, \alpha)$ denotes the density of a gamma distribution with shape parameter ν and scale parameter α. Moreover, $Exp(\cdot)$ denotes the density of an exponential distribution. The time-constant coefficients γ_l, $l = 1, ..., 9$, are given by $\gamma_1 = -0.5$, $\gamma_2 = \gamma_3 = 0$. The time-varying coefficients β_{jt} are expanded in cubic B-splines with six equidistant inner knots. In this setting, the censoring times are simulated from a multinomial distribution $\mathcal{M}(n, \boldsymbol{p}_c)$ with $\boldsymbol{p}_c^T = (0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.1, 0.1, 0.1, 0.1, 0.1)$. The number of replications is 50.

To allow for maximal flexibility in modeling, for all coefficients, time-varying effects expanded in B-spline basis functions are assumed. This results in the linear predictor

$$\eta_{it}^{model} = \beta_{0t} + \sum_{j=1}^{5} z_{ij}\beta_{jt},$$

with

$$\beta_{0t} = \sum_{m=1}^{m_0} \alpha_{0m} B_{0m}(t) \quad \text{and} \quad \beta_{jt} = \sum_{m=1}^{m_j} \alpha_{jm} B_{jm}(t).$$

For the simulation study the following penalty term

$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{m_0} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi \left(\phi \sum_{j=1}^{r} \psi_j \left\| \boldsymbol{\zeta}_j \right\|_2 + (1-\phi) \sum_{j=1}^{r} \varphi_j \left\| \boldsymbol{\alpha}_j \right\|_2 \right)$$

where $\boldsymbol{\zeta}_j^T = (\zeta_{j2}, ..., \zeta_{jm_j})$, $\zeta_{jm} = \alpha_{jm} - \alpha_{j,m-1}$, $m = 2, ..., m_j$, is used. Therein, ψ_j and φ_j are replaced by adaptive weights (5.9). The penalty allows for stable baseline effects and steers smoothing, constant effects and selection of the time-varying coefficients. The penalization of β_{0t} is predominantly executed due to stability reasons. It is defined by $\xi_0 = 0.001$ in all simulation settings. For more details on the penalty term, see Section 5.3.

Results

The results of `fpendsm` are compared to the results obtained by the R functions `gam` and `gamm` of the R add-on package `mgcv`, by fitting analogous models. That means, using the `gam` function, for the time-varying intercept a slight ridge penalty with tuning parameter 0.001 is used. In contrast, in the `gamm` function, the time-varying intercept is penalized by a ridge penalty, where the tuning parameter is chosen internally. For both functions, `gam` and `gamm`, the time-varying covariate effects are estimated by cubic B-splines with penalized first differences between the parameters of the smooth functions. Moreover, the option `select` is set to `TRUE` for `gam`, adding a penalty to each smooth, to allow it to be removed from the model. Such an option is not provided by `gamm`.

The performance of the estimators is evaluated separately for the structural components and the variance. The assessment of parameter estimations is evaluated in general, and separately for truly time-varying and truly time-constant parameters. For each simulation run the corresponding mean squared errors are computed by

$$\begin{aligned} \text{MSE}(\sigma_b, \hat{\sigma}_b) &= (\sigma_b - \hat{\sigma}_b)^2, \\ \text{MSE}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) &= \frac{1}{r}\sum_{j=1}^{r}\left(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\right)^2 + \frac{1}{s}\sum_{l=1}^{s}\left(\tilde{\boldsymbol{\beta}}_l - \hat{\boldsymbol{\beta}}_l\right)^2, \\ \text{MSE}_{vary}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) &= \frac{1}{r}\sum_{j=1}^{r}\left(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\right)^2, \qquad \text{MSE}_{const}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \frac{1}{s}\sum_{l=1}^{s}\left(\tilde{\boldsymbol{\beta}}_l - \hat{\boldsymbol{\beta}}_l\right)^2, \end{aligned} \tag{5.10}$$

where $\tilde{\boldsymbol{\beta}}_l = (\gamma_l, ..., \gamma_l)$ and $p = r+s$. Hence, σ_b, $\boldsymbol{\beta}_j$ and $\tilde{\boldsymbol{\beta}}_l$ denote the true parameter values, whereas $\hat{\sigma}_b$, $\hat{\boldsymbol{\beta}}_j$ and $\hat{\tilde{\boldsymbol{\beta}}}_l$ define the corresponding estimates. That means, as all components are estimated time-varying, $\boldsymbol{\gamma}$ is compared to $\hat{\boldsymbol{\beta}}$ as well. Additional information on the stability of the algorithms is collected in n.c. (not converged), which defines the sum over the data sets, where numerical problems occurred during estimation. This issue solely affects the `gamm` function.

In analogy to Chapter 3, after estimation of the coefficients $\boldsymbol{\beta}_{jt}$, $j = 1, ..., r$, $t = 1, ..., q$, and $\boldsymbol{\gamma}_l$, $l = 1, ..., s$, the results are compared to the true parameters. For the evaluation of the selection performance, false positive rates (FPR) and false negative rates (FNR) are considered for each simulation run. Thereby, false positive means that a single parameter value that is truly zero is set to non-zero. In contrast, false negative means that a single non-zero parameter value is set to zero. The corresponding rates are defined by

$$\text{FPR} = \frac{\#(\text{truly zero set to non-zero})}{\#(\text{truly zero})} \qquad FNR = \frac{\#(\text{truly non-zero set to zero})}{\#(\text{truly non-zero})}.$$

Initially, the results of the simulation settings are summarized in tables. The outcomes of `fpendsm`, `gam` and `gamm` are shown in the corresponding columns. The first three rows of the tables contain the absolute values of the mean squared errors for all covariates (MSE)

as well as for truly time-varying (MSE_{vary}) and truly time-constant covariates (MSE_{const}). A detailed definition of these mean squared errors is given in Equation (5.10). Finally, the false positive rate FPR and the false negative rate FNR are shown. Moreover, for each setting a table containing the estimates of σ_b is provided, where the random effects variance components are obtained by an EM-type algorithm (see Section 5.3.1). All presented values correspond to the mean values over all simulation runs. Moreover, the results are illustrated in boxplots, where for the sake of interpretability, outliers are omitted in single cases.

Setting 1

For settings 1, only a time-varying intercept is chosen, that means that all covariates originally have a constant effect. The summary of the mean squared errors referring to $\boldsymbol{\beta}$ as well as FPR and FNR are shown in Table 5.3. For $\sigma_b = 0.1$ and $\sigma_b = 0.5$, `gamm` outperforms `fpendsm` and `gam` but poor results are yielded for the remaining values of σ_b. In general, `fpendsm` attains slightly worse results than `gam` but outperforms it for single cases. The FPR values of `gam` outperform those of `fpendsm`, whereas for FNR the results are quite similar. For `gamm`, FPR is always equal to one and FNR always equal to zero since no variable selection is provided.

The estimates of the random effects variance components derived by `fpendsm` outperforms those of gam, except for $\sigma_b = 2.0$ (Table 5.4). The corresponding results of `gamm` are the best for $\sigma_b = 0.7$ and $\sigma_b = 1.0$ but the worst for $\sigma_b = 0.1$ and $\sigma_b = 0.5$. For the case $\sigma_b = 2.0$, the estimate of σ_b is disproportionately high for `gamm`, but it is based on only one replication. Moreover, for at least 25 out of 50 simulation runs, the `gamm` method does not converge (see column n.c. in Table 5.4). Hence, `gamm` only converges for less than 50% of the replications and it can be expected that these cases are more simple with regard to the estimation. Thus, the results obtained by `gamm` cannot be considered as stable or reliable and `gamm` cannot be recommended.

In addition, the mean squared errors are illustrated in Figure 5.5, which shows the corresponding boxplots exemplarily for $\sigma_b = 0.5$. In this case, `fpendsm` performs quite well, in particular for the mean squared errors referring to σ_b.

Setting 2

In the second setting, additional to the time-varying intercept, two covariate effects are originally time-varying. The mean squared errors referring to $\boldsymbol{\beta}$ as well as FPR and FNR are shown in Table 5.5. Again, for some cases `gamm` outperforms `fpendsm` and `gam` but on the other hand also very poor results occur for `gamm`. The results of `fpendsm` are better than those of `gam` for single cases. The FNR values of `fpendsm` outperform those of `gam`, but for FPR it is the opposite case.

For the estimates of the random effects variance components, `fpendsm` outperforms `gam` for all cases (Table 5.6). The corresponding results of `gamm` are the best for $\sigma_b = 0.7$ and

	σ_b	fpendsm	gam	gamm	σ_b	fpendsm	gam	gamm
MSE	0.1	0.052	0.047	0.027	0.5	0.060	0.057	0.035
MSE_{vary}	0.1	0.194	0.157	0.065	0.5	0.183	0.161	0.094
MSE_{const}	0.1	0.024	0.026	0.019	0.5	0.035	0.037	0.024
FPR	0.1	0.410	0.261	1	0.5	0.503	0.355	1
FNR	0.1	0.005	0	0	0.5	0.013	0.007	0

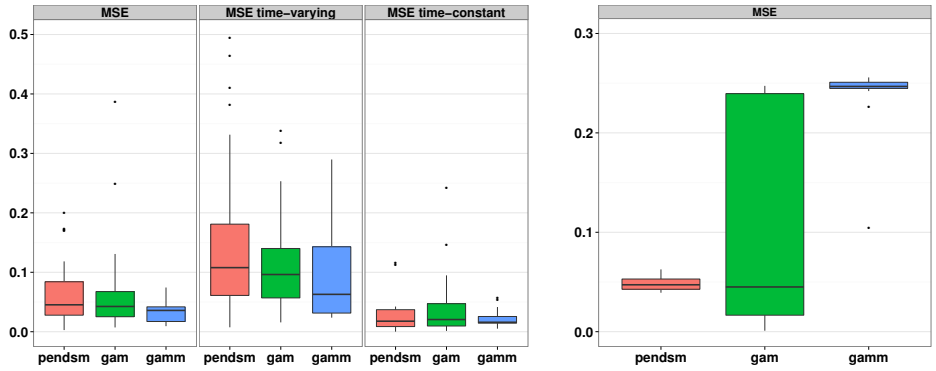
	σ_b	fpendsm	gam	gamm	σ_b	fpendsm	gam	gamm
MSE	0.7	0.091	0.070	0.345	1.0	0.135	0.116	0.155
MSE_{vary}	0.7	0.308	0.188	1.222	1.0	0.493	0.365	0.197
MSE_{const}	0.7	0.047	0.047	0.169	1.0	0.063	0.067	0.146
FPR	0.7	0.521	0.340	1	1.0	0.573	0.365	1
FNR	0.7	0.011	0.007	0	1.0	0.027	0.021	0

	σ_b	fpendsm	gam	gamm
MSE	2.0	0.446	0.338	$1.368 \cdot 10^{30}$
MSE_{vary}	2.0	1.946	1.367	$3.556 \cdot 10^{30}$
MSE_{const}	2.0	0.146	0.132	$9.305 \cdot 10^{29}$
FPR	2.0	0.581	0.377	1
FNR	2.0	0.058	0.095	0

Table 5.3. Results for Setting 1 for the estimated mean squared errors (MSE, MSE_{vary}, MSE_{const}) referring to β and the false positive rate (FPR) as well as the false negative rate (FNR) for `fpendsm`, `gam` and `gamm`.

σ_b	fpendsm MSE_σ	gam MSE_σ	gamm MSE_σ	gamm n.c.
0.1	0.028	0.032	0.808	25
0.5	0.048	0.095	0.236	37
0.7	0.168	0.169	0.069	42
1.0	0.480	0.490	0.057	48
2.0	2.897	2.389	$1.01 \cdot 10^{15}$	49

Table 5.4. Results for Setting 1 for the mean squared errors referring to σ_b for `fpendsm`, `gam`, and `gamm`.



(a) Mean squared errors for $\boldsymbol{\beta}$ **(b)** Mean squared errors for σ_b

Figure 5.5. Mean squared errors for Setting 1 for `fpendsm`, `gam` and `gamm` for $\sigma_b = 0.5$.

	σ_b	fpendsm	gam	gamm	σ_b	fpendsm	gam	gamm
MSE	0.1	0.163	0.108	0.096	0.5	0.220	0.159	0.099
MSE_{vary}	0.1	0.296	0.193	0.160	0.5	0.404	0.286	0.152
MSE_{const}	0.1	0.029	0.024	0.032	0.5	0.036	0.032	0.045
FPR	0.1	0.568	0.328	1	0.5	0.583	0.308	1
FNR	0.1	0.041	0.086	0	0.5	0.036	0.101	0

	σ_b	fpendsm	gam	gamm	σ_b	fpendsm	gam	gamm
MSE	0.7	0.218	0.178	0.068	1.0	0.302	0.266	0.147
MSE_{vary}	0.7	0.401	0.320	0.105	1.0	0.560	0.488	0.265
MSE_{const}	0.7	0.035	0.037	0.030	1.0	0.044	0.044	0.028
FPR	0.7	0.551	0.350	1	1.0	0.550	0.397	1
FNR	0.7	0.051	0.126	0	1.0	0.072	0.147	0

	σ_b	fpendsm	gam	gamm
MSE	2.0	0.761	0.629	$1.951 \cdot 10^{29}$
MSE_{vary}	2.0	1.454	1.181	$3.260 \cdot 10^{29}$
MSE_{const}	2.0	0.067	0.077	$6.415 \cdot 10^{28}$
FPR	2.0	0.497	0.358	1
FNR	2.0	0.263	0.311	0

Table 5.5. Results for Setting 2 for the estimated mean squared errors (MSE, MSE_{vary}, MSE_{const}) referring to β and the false positive rate (FPR) as well as the false negative rate (FNR) for `fpendsm`, `gam` and `gamm`. The displayed values represent the means over all simulation runs.

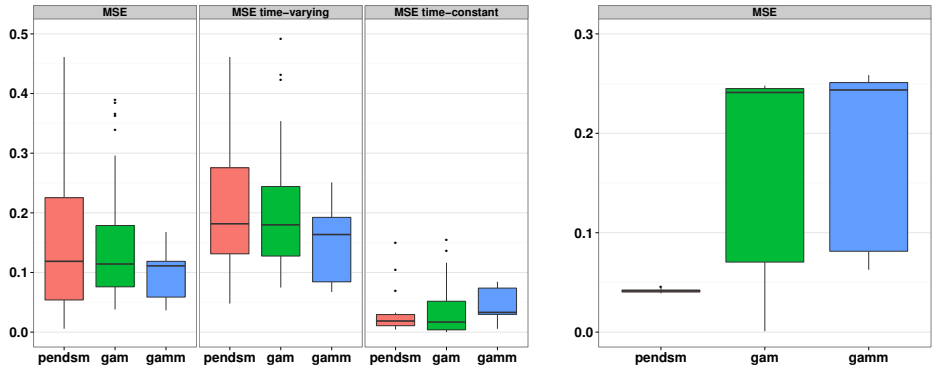
σ_b	fpendsm MSE_σ	gam MSE_σ	gamm MSE_σ	gamm n.c.
0.1	0.023	0.024	0.775	37
0.5	0.042	0.171	0.179	45
0.7	0.163	0.342	0.049	46
1.0	0.492	0.749	0.072	47
2.0	2.889	2.822	$6.25 \cdot 10^{14}$	42

Table 5.6. Results for Setting 2 for the mean squared errors referring to σ_b for `fpendsm`, `gam`, and `gamm`. The displayed values represent the means over all simulation runs.

$\sigma_b = 1.0$ but the worst for $\sigma_b = 0.1$, $\sigma_b = 0.5$ and $\sigma_b = 2.0$. For the latter case, the estimates of σ_b are disproportionately high for `gamm`. Moreover, for at least 37 out of 50 simulation runs, the `gamm` method does not converge (Table 5.6) leading to unstable results that cannot be reliably interpreted.

Exemplarily for $\sigma_b = 0.5$, the corresponding boxplots of the mean squared errors are illustrated in Figure 5.6. `fpendsm` yields good results, in particular for the mean squared errors referring to σ_b.

To conclude, the `gamm` method has a large variation within the outcomes and evidently convergence problems. Hence, it is not recommended to use `gamm`. `fpendsm` and `gam` have no convergence problems and the outcomes are comparable and seem to be stable. However, `fpendsm` is much more flexible as for example much more types of penalties are available.



(a) Mean squared errors for $\boldsymbol{\beta}$ **(b)** Mean squared errors for σ_b

Figure 5.6. Mean squared errors for Setting 2 for `fpendsm`, `gam` and `gamm` for $\sigma_b = 0.5$.

5.5. Applications

In this section, `fpendsm` is applied to two real data problems. To compare the results to the examples from Chapter 3, where `pendsm` is used, the Munich founder study and the fertility study are analyzed. However, no reference methods, for example, by applying the functions `gam` or `gamm` of the R add-on package `mgcv` (Wood, 2006) to the data, are available for discrete survival models with frailties. The `gamm` function is not able to compute either one of the examples as due to the incorporated time-varying coefficients the data are to complex and too many random intercepts have to be estimated. In contrast, the `gam` function provides results but after an excessive amount of time. However, the results are not meaningful. That means, for example, the variance parameter was estimated by an absolute value greater than 500. Consequently, the functions `gam` and `gamm` cannot cope with such complex data.

5.5.1. The Munich Founder Study

In this section, `fpendsm` is applied to the Munich founder study. Therein, the survival of newly founded firms in the area of Munich and Upper Bavaria is investigated. The dependent variable defines the transition process of a newly founded company up to insolvency, denoting the event. The duration time until the event of insolvency is measured in quarters, where a maximum of 22 quarters can be reached. A company that was still alive at the time of the registration of the interview is treated as right-censored. See Section 3.5.1 for more details on the data. The data were reorganized according to Section 2.2.2 and standardized, to conduct a binary regression model with complementary log-log link corresponding to a discrete-time survival model. The long format of the data consists of $\sum_{i=1}^{n} t_i = 17736$ rows. In contrast to Section 3.5.1, a frailty is incorporated to control for

the unobserved heterogeneity that is the basic difference between the observed firms. For company i and measurement at quarter t, the considered model has the form

$$\begin{aligned}\eta_{it} = \beta_{0t} &+ \text{Sector}_{it}^{(1)}\beta_{1t} + \text{Sector}_{it}^{(2)}\beta_{2t} + \text{Legal}_{it}\beta_{3t} + \text{Seed}_{it}\beta_{4t} + \text{Equity}_{it}\beta_{5t} + \text{Debt}_{it}\beta_{6t} \\ &+ \text{Market}_{it}\beta_{7t} + \text{Clientele}_{it}\beta_{8t} + \text{Degree}_{it}\beta_{9t} + \text{Gender}_{it}\beta_{10t} \\ &+ \text{Experience}_{it}\beta_{11,t} + \text{Employees}_{it}\beta_{12,t} + \text{Age}_{it}\beta_{13,t} + b_i,\end{aligned}$$

with individual-specific random intercepts $b_i \sim \mathcal{N}(0, \sigma_b^2)$. For all covariate effects cubic B-splines are used, that means $\beta_{jt} = \sum_m \alpha_{jm} B_{jm}(t)$, $j = 0, 1, ..., 13$, with 10 equidistant inner knots resulting in 12 basis functions. Hence, the used penalty is given by

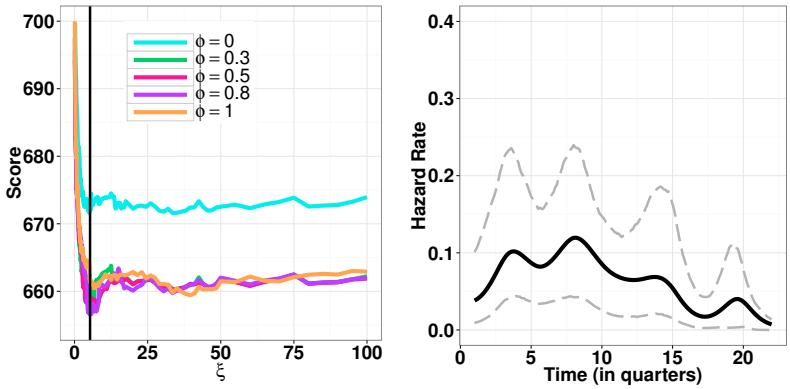
$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{12} \left(\alpha_{0m} - \alpha_{0,m-1}\right)^2 + \xi\left(\phi \sum_{j=1}^{13} \psi_j \left\|\boldsymbol{\zeta}_j\right\|_2 + (1-\phi) \sum_{j=1}^{13} \varphi_j \left\|\boldsymbol{\alpha}_j\right\|_2\right),$$

where adaptive weights (5.9) are incorporated in the penalty term. The penalty allows for smooth time-varying or time-constant covariate effects or their elimination from the model. The tuning parameter ξ_0 is set to 0.001 and for selection of the tuning parameters ξ and ϕ a 5-fold cross-validation, based on the predictive deviance, is conducted. In Figure 5.7a the corresponding scores are illustrated. The vertical black line determines the chosen tuning parameters with ξ set to 5.3 and ϕ set to 0.8. The chosen value of ξ is slightly smaller than that in Chapter 3.5.1. The run of all curves indicates that penalization clearly improves ordinary ML-estimation that is nearly obtained for $\xi = 0$ as ξ_0 is set to the very small value of 0.001. Moreover, the curves are wigglier than in Chapter 3.5.1.

In analogy to Chapter 3.5.1, Figure 5.7b shows the resulting hazard function of the fitted model when all covariates are set at reference. That is, a female founder at the age of 43 whose firm has the following characteristics: industry, manufacturing and building sector, small legal form, seed capital ≤ 25000, with equity capital, with debt capital, local target markets, wide spread clientele, no A-levels, professional experience < 10 years and 0 or 1 employees. The resulting hazard rate is quite similar to that in Chapter 3.5.1. The corresponding confidence interval is based on a nonparametric bootstrap method with 1000 bootstrap replications (see Section 3.3 for more details). They are somewhat larger than those in Chapter 3.5.1.

The estimation of σ_b^2 is derived by an approximate EM-algorithm (see Section 5.3.1) and the resulting estimates of the variance parameter for `fpendsm` are shown in Table 5.7, where the standard error and the confidence interval are based on 1000 bootstrap replications. To illustrate the failure of the `gam` function, the corresponding estimate of σ_b is set to 8900, and absolutely not meaningful.

In Figure 5.8 the estimates $\hat{\beta}_{jt}$ of the coefficients of the model with $\xi = 5.3$ and $\phi = 0.8$ are summarized. The solid black lines denote the parameter estimates, whereas the dashed lines specify the corresponding 95% confidence intervals. The intervals are based on 1000 bootstrap replications and are computed pointwise. By using `fpendsm`, it is suggested that



(a) Cross-validation scores subject to penalty parameter ξ

(b) Estimated hazard rate for all predictors at reference. Dashed lines represent the 95% confidence interval.

Figure 5.7. Plots corresponding to the Munich founder study

			95%-CI	
	estimate	standard error	lower	upper
$\hat{\sigma}_b$	0.309	0.002	0.306	0.312

Table 5.7. Estimates resulting from `fpendsm` for the standard deviation of the random effects, the corresponding standard error and the 95% bootstrap confidence interval for the Munich founder study.

Legal,*Market* and *Experience* have a linear effect in the predictor, whereas *Age* is removed from the model. These results are analogous to those of Section 3.5.1. In contrast to Section 3.5.1, for `fpendsm` the estimates of *Clientele*, *Degree* and *Employees* result in flexible time-varying functions. In general, the course of the curves of the time-varying coefficients follows a similar pattern compared to the corresponding curves in Section 3.5.1.

5.5.2. Fertility Study

Finally, `fpendsm` is applied to the fertility study. Therein, the question is investigated if labor force participation of women influences the transition to motherhood. The data are extracted from *pairfam* (Nauck et al., 2012), a multi-disciplinary longitudinal study analyzing cooperative and life forms of families in Germany, and are described in more detail in Abedieh (2013). The dependent variable is the transition to pregnancy with duration time (in years) until pregnancy. For modeling the duration time the start of the observation process is set to 14 years. The maximum value is 27 years until gravidity. See Section 3.5.2 for more details on the data. The data were reorganized according to Section 2.2.2 and standardized, to conduct a binary regression model with complementary log-log link corresponding to a discrete-time survival model. The long format of the data consists of $\sum_{i=1}^{n} t_i = 34601$ rows. In contrast to Section 3.5.2, a frailty is incorporated to control for

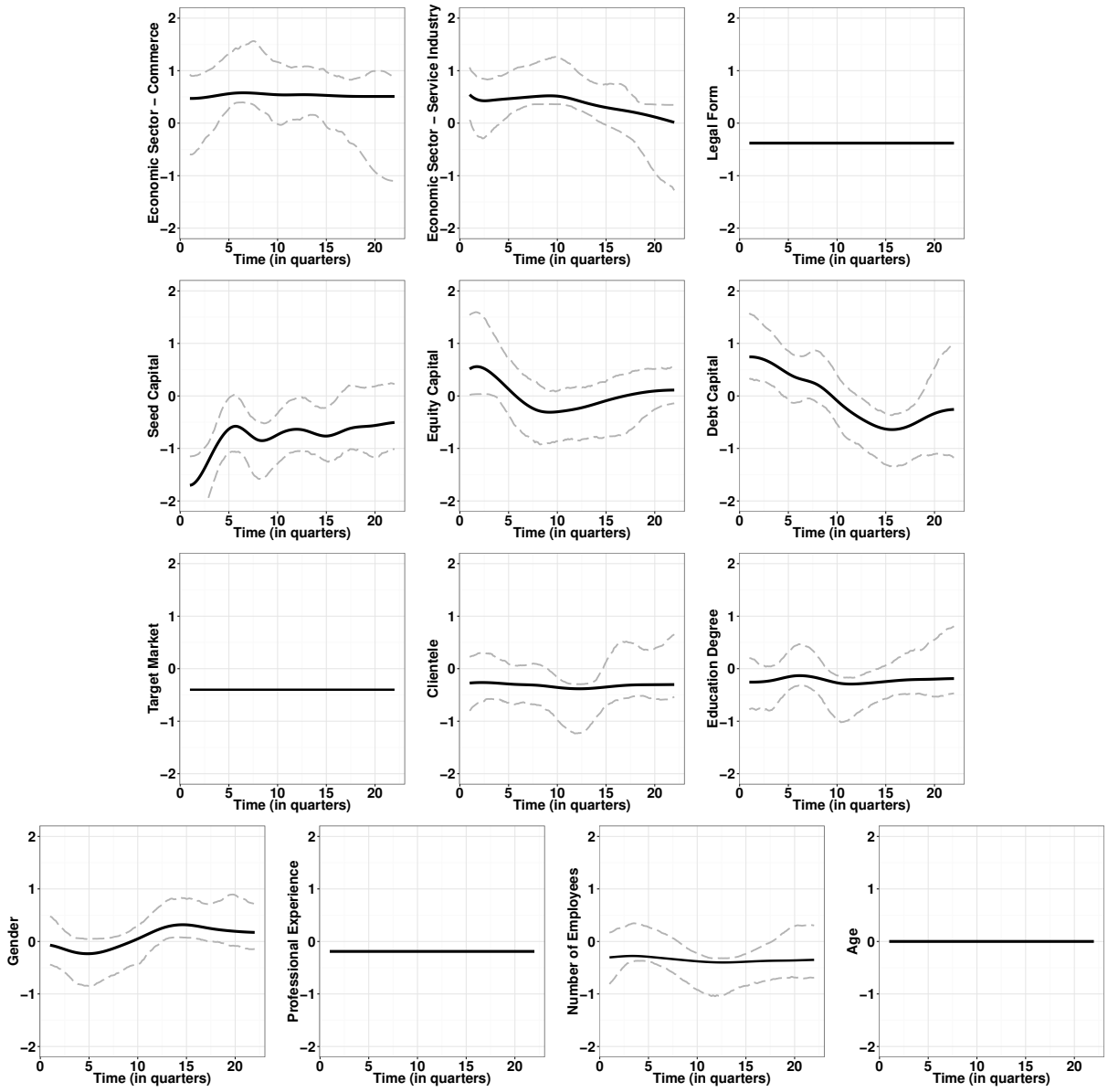


Figure 5.8. Estimates of `fpendsm` for the Munich founder study using cubic B-splines. Dashed lines represent pointwise 95% bootstrap confidence intervals.

the unobserved heterogeneity, that is, the basic differences between the observed women. For duration year t of individual i the considered model has the form

$$\begin{aligned}\eta_{it} = &\beta_{0t} + \text{Job}_{it}^{(1)}\beta_{1t} + \text{Job}_{it}^{(2)}\beta_{2t} + \text{Job}_{it}^{(3)}\beta_{3t} + \text{Job}_{it}^{(4)}\beta_{4t} + \text{Education}_{it}^{(1)}\beta_{5t} \\ &+\text{Education}_{it}^{(2)}\beta_{6t} + \text{Relationship}_{it}^{(1)}\beta_{7t} + \text{Relationship}_{it}^{(2)}\beta_{8t} + \text{Siblings}_{it}\beta_{9t} \\ &+\text{ClassParents}_{it}^{(1)}\beta_{10t} + \text{ClassParents}_{it}^{(2)}\beta_{11t} + \text{ClassParents}_{it}^{(3)}\beta_{12t} \\ &+\text{Cohort}_{it}^{(1)}\gamma_{12} + \text{Cohort}_{it}^{(2)}\gamma_{12} + b_i,\end{aligned} \tag{5.11}$$

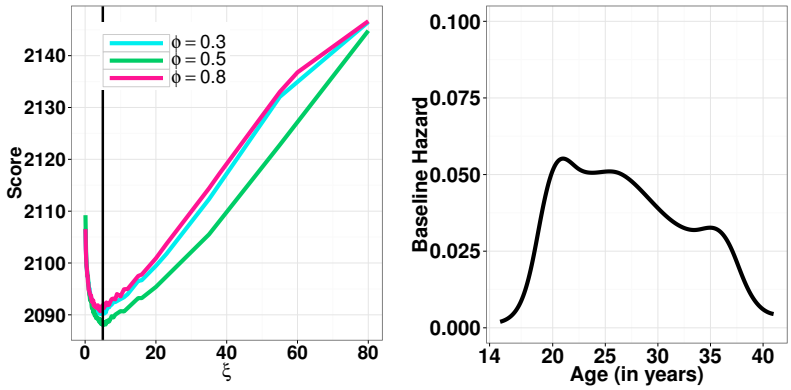
with individual-specific random intercepts $b_i \sim \mathcal{N}(0, \sigma_b^2)$. All covariates, except the covariate *cohort*, are incorporated as time-varying effects using cubic B-splines, that means $\beta_{jt} = \sum_m \alpha_{jm} B_{jm}(t)$, $j = 0, 1, ..., 12$, with 8 equidistant inner knots resulting in 10 basis functions. The used penalty is given by

$$J_{\xi_0,\xi}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_0 \sum_{m=2}^{10} (\alpha_{0m} - \alpha_{0,m-1})^2 + \xi \left(\phi \sum_{j=1}^{12} \psi_j \left\| \boldsymbol{\zeta}_j \right\|_2 + (1-\phi) \sum_{j=1}^{12} \varphi_j \left\| \boldsymbol{\alpha}_j \right\|_2 + \sqrt{\gamma_{11}^2 + \gamma_{12}^2} \right),$$

where adaptive weights (5.9) are used for estimation. The penalty term allows for smooth time-varying or time-constant covariate effects or their selection from the model, whereas for the covariate *cohort* a group lasso penalty is used. That means, that the coefficients of the covariate *cohort*, including all categories, can be shrunk simultaneously until the complete variable is removed from the model.

The tuning parameter ξ_0 is set to 0.001. Tuning parameters ξ and ϕ are chosen by 5-fold cross-validation with the predictive deviance as loss criterion. They are respectively set to 5.0 and 0.8. The corresponding cross-validation scores are shown in Figure 5.9a, where the vertical black line marks the chosen tuning parameters with $\xi = 5.0$ and $\phi = 0.8$, where the chosen ξ is slightly smaller than that in Section 3.5.2. Thereby, the score referred to $\phi = 0$ and $\phi = 1$, meaning that the weight of the penalty was completely assigned either to the selection part or the smoothing part, was omitted. This was done due to a heavy erratically run of the score curves attended by extreme peaks. The runs of the showed curves indicate that penalization clearly improves ordinary ML-estimation that is nearly obtained for $\xi = 0$ as ξ_0 is set to the very small value of 0.001.

Figure 5.9b shows the resulting hazard function of the fitted model when all covariates are set at reference. That is, an unemployed single women born between 1971-1973 with low-level education, no siblings and low-level education of the parents. The resulting hazard rate is quite similar to that in Chapter 3.5.2. For this very complex data example with the linear predictor (5.11), that is, a data set with 34601 observations, time-varying covariate effects and included frailty effects, the computation of bootstrap standard errors and confidence intervals is not feasible. As the computation of a corresponding model using `gam` is not meaningful, only the estimates of `fpendsm` are shown. Resulting plots of the estimated time-varying coefficients can be found in Figure 5.10. Therein, the labeling of the abscissa is adapted to the real age of the women, where the observation period starts at the age of



(a) Cross-validation scores subject to penalty parameter ξ

(b) Estimated hazard rate for all predictors at reference.

Figure 5.9. Plots corresponding to the fertility study.

14. In general, the course of the curves of the time-varying coefficients follows a similar pattern compared to the corresponding curves in Section 3.5.2. The estimation of σ_b^2 is derived by an approximate EM-algorithm (see Section 5.3.1) and is given by $\hat{\sigma}_b = 0.295$. The estimates for *cohort* are $\gamma_{11} = -0.061$ and $\gamma_{12} = -0.327$, that are identical to the corresponding estimates in Section 3.5.2.
In spite of the large amount of parameters and the complex estimation approach due to frailties, `fpendsm` yields stable and meaningful results.

5.6. Concluding Remarks

This Chapter extends the method proposed in Chapter 3, where lasso-type penalties are treated, by an incorporation of frailty effects. Thereby, it is controlled for unobserved heterogeneity since ignoring unobserved heterogeneity may lead to biased estimates. Hence, complex penalty terms are combined with random effects for survival models for discrete duration time. Moreover, time-varying coefficients are regarded in the linear predictor. The incorporation of these different issues may lead to very complex and difficult data situations. The proposed method is even able to yield stable estimates for this cases, whereas existing methods provided by the functions `gam` and `gamm` of the R add-on package `mgcv` (Wood, 2006) or the function `glmmPQL` of the R add-on package `MASS` (Venables and Ripley, 2002) typically fail. However, it has to be mentioned that the computation of `fpendsm` is very time-consuming. Hence, a further challenge is the optimization of the estimation algorithm with the aim to be more efficient.
It has to be noted that actually an improved version of the `gamm` function is available. This improved function is `gamm4` from the corresponding R add-on package `gamm4` (Wood and Scheipl, 2013) and is numerically more robust than `gamm` by using `lme4` (Bates et al., 2014) as the underlying fitting engine. However, since the release of 3.0.0 in April 2013 the functions `gamm4` and `lme4` are no longer compatible. Hence, unfortunately, the function `gamm4` cannot be used for the analysis of this chapter.

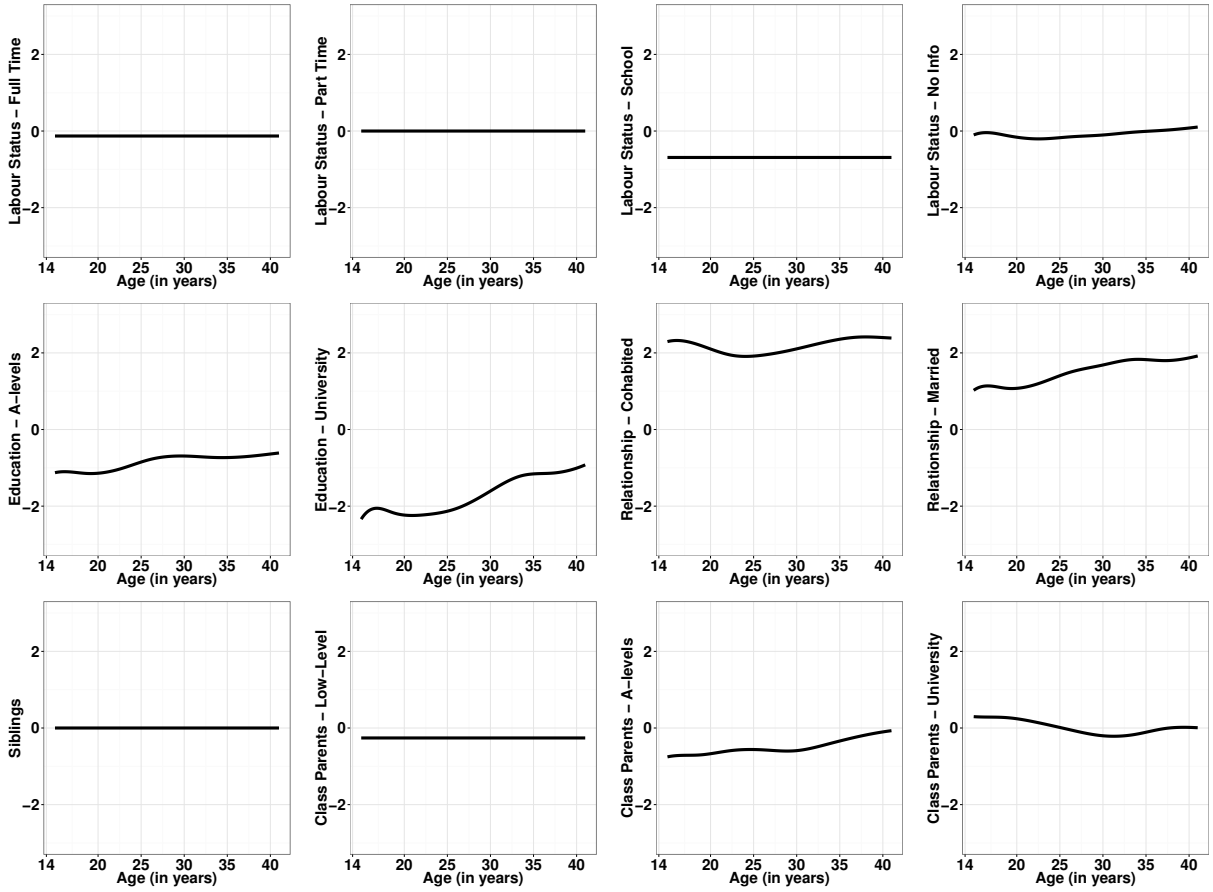


Figure 5.10. Estimates of `fpendsm` for the time-varying coefficients of the fertility study using cubic B-splines.

6. Penalization in Competing Risks Models

In many applications concerning survival analysis, the investigation of more than one terminating event is of interest. Hence, for each object one of k ($k \geq 2$) causes may occur, called competing risks. Background on competing risks models is given in Section 6.1. This model class is considered in Section 6.2 with respect to discrete duration time. In the context of discrete survival times, competing risks models can be embedded into the framework of multinomial logit models. However, a large number of parameters arises with the use of this model type. Therefore, in Section 6.3 a penalization technique for discrete-time competing risks models is introduced. Some details regarding the corresponding estimation approach are described in Section 6.4. In Section 6.5 the proposed method is applied to two applications. Finally, concluding remarks are summarized in Section 6.6. In the following, only the notation and explanations with respect to grouped survival times are considered, but they can easily be modified regarding truly discrete survival times.

6.1. Introduction

In the preceding sections, the duration time of an object until it reaches one absorbing event has been considered. However, in many applications it has to be distinguished between several distinct types of terminating events. That means, a subject may fail due to one of k ($k \geq 2$) multiple causes. In survival analysis, the events may stand for several causes of death, whereas, for example, in the case of government duration, a government may collapse in one of two ways: dissolution (new elections) or replacement (no new elections) (Diermeier and Stevenson, 1999). Since only the transition to one of multiple different states can be observed, models for this type of data are referred to as competing risks models. In some applications however, the expression competing chances would be more appropriate than competing risks. Most of the literature on competing risks considers the case of continuous time. Some examples treating competing risks for continuous time are Beyersmann et al. (2011), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003) and Kleinbaum and Klein (2013). If time is discretely observed, ties cause problems in the estimation procedure and the model might become inappropriate, especially for a low number of time periods (see Chapter 2). Hence, appropriate methods are required to model competing risks with discrete time. In this context, for example, Tutz (1995) proposed two fundamental approaches to the modeling of competing risks with nominal or ordinal categories of

response. Furthermore, Fahrmeir and Wagenpfeil (1996) introduced smooth estimation of hazard functions and time-varying effects in a flexible way. Frailty-based competing risks models for discrete time were exemplarily considered by Enberg et al. (1990) or Gorfine and Hsu (2011).

As the target events may be seen as unordered categorical responses, the multinomial logit model is one of the most widely used models for competing risks with discrete time. Moreover, if the causes are ordered more parameter economic parametrizations are available, for example, by using ordinal response models (McCullagh, 1980; Agresti, 2013). Their application to competing risks with discrete time can be found in Tutz (1995).

The multinomial logit model is associated with a large number of parameter estimates since it employs several coefficients for each explanatory variable. Hence, maximum likelihood estimates tend to deteriorate quickly and interpretability suffers as well, leading to the restriction to applications with few predictors. Therefore, applying regularization methods that induce variable selection leading to interpretable and reliable models is very advantageous.

Simple conventional variable selection methods are represented by *Forward-* and *Backward-Stepwise Selection* (e.g. Hastie et al., 2009). However, these methods exhibit stability problems and cannot be recommended. A more current alternative model selection approach is based on regularization techniques. Thereby, penalization is an approved regularization approach. Adding a penalty term to the log-likelihood yields shrinkage of the estimates towards zero. Depending on the penalty, it is even possible to set particular estimated parameters exactly to zero. One of the oldest penalization methods is the *ridge* method that uses a L_2-type penalty on the regression coefficients. However, no variable selection can be performed by using this penalty term. An alternative penalty term that has become very popular, is the *lasso* penalty using a L_1-type penalty on the regression coefficients. In this case, variable selection can be carried out. As the lasso merely selects individual predictors, the penalty is unsatisfactory in the case of grouped data, for example with categorical predictors. The *group lasso*, proposed by Yuan and Lin (2006), can overcome these problems. To obtain consistent estimates of the parameters, Zou (2006) extended the lasso to the *adaptive lasso* by including different weights on the penalty for different coefficients. Several further improvements for the lasso method have been designed in the last decade, for example the *fused lasso* (Tibshirani et al., 2005), *SCAD* (Fan and Li, 2001), *elastic net* (Zou and Hastie, 2005), *Dantzig selector* (Candes and Tao, 2007) and *DASSO* (James et al., 2009).

However, these methods are designed for models with univariate response. As the multinomial logit model is not a common univariate generalized linear model, these methods cannot be applied immediately. The effect of one predictor variable is represented by several parameters. Hence, there is a difference in providing variable selection and parameter selection, where variable selection is only obtained if all parameters belonging to one predictor are simultaneously set to zero. The available penalty techniques for multinomial logit models (Krishnapuram et al., 2005; Friedman et al., 2010) use L_1-type penalties that shrink all parameters separately. Thus, they pursue the goal of parameter selection and not the goal of variable selection as the lasso method does not enforce that all coefficients be-

longing to a covariate are shrunk to zero. This problem was overcome by Tutz et al. (2012) and Tutz (2012), where several penalization methods for the multinomial logit model are described in detail. In particular, the authors perform true variable selection by simultaneously removing all effects of one predictor from the model. In the following, on the basis on Tutz et al. (2012) variable selection in competing risks models for discrete time is executed by means of an appropriate penalization approach. This penalization approach also allows for smooth time-varying cause-specific baseline effects. For this purpose, the likelihood of a discrete competing risks model is extended by a penalty term.

6.2. Competing Risks Models for Discrete Time

In this section, a competing risks model for discrete duration time is considered. The model is defined and maximum likelihood estimation is embedded into the framework of multivariate generalized linear models.

6.2.1. Methodology

In the following, time T is considered as a non-negative random variable taking values from $\{1, ..., q\}$. The values of T may be original observations, that is, T is intrinsically discrete. Alternatively, discreteness may be due to interval censoring. Let time T be divided into $q + 1$ intervals

$$[a_0, a_1), [a_1, a_2), ..., [a_{q-1}, a_q), [a_q, \infty),$$

where usually $a_0 = 0$ is assumed and a_q denotes the final follow-up. Then, $T = t$ is observed if failure occurs within the interval $[a_{t-1}, a_t)$. Let the distinct terminating causes be denoted by $R \in \{1, ..., k\}$. The *cause-specific discrete hazard function* resulting from cause or risk r is given by the conditional probability

$$\lambda_r(t|\boldsymbol{x}) = P(T = t, R = r|T \geq t, \boldsymbol{x}),$$

where $\boldsymbol{x}$ is a vector of covariates and $r = 1, ..., k$, $t = 1, ..., q$. Summarizing the k hazard functions $\lambda_1(t|\boldsymbol{x}), ..., \lambda_k(t|\boldsymbol{x})$ to an overall hazard function, regardless of cause, yields

$$\lambda(t|\boldsymbol{x}) = \sum_{r=1}^{k} \lambda_r(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x}).$$

The survival function and the unconditional probability of an event in period t have the same form as in the simple case of one target event (compare Equations (2.3) and (2.4)) and are given by

$$S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{x}) = \prod_{j=1}^{t} (1 - \lambda(j|\boldsymbol{x})) = 1 - F(t|\boldsymbol{x})$$

and

$$P(T = t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}) \prod_{j=1}^{t-1} (1 - \lambda(j|\boldsymbol{x})) = \lambda(t|\boldsymbol{x}) S(t-1|\boldsymbol{x}).$$

If an individual reaches interval $[a_{t-1}, a_t)$, there are $k+1$ possible outcomes. That is, end of the duration by transition to one of the k target events or survival. The corresponding conditional response probabilities are given by

$$\lambda_1(t|\boldsymbol{x}), ..., \lambda_k(t|\boldsymbol{x}), 1 - \lambda(t|\boldsymbol{x}),$$

where $1 - \lambda(t|\boldsymbol{x})$ is the probability for survival. Therefore, given an individual reaches interval $[a_{t-1}, a_t)$, the $k+1$ possible events may be seen as unordered categorical responses and a natural parametric model for the hazards is the multinomial logit model given by

$$\lambda_r(t|\boldsymbol{x}) = \frac{\exp(\beta_{0tr} + \boldsymbol{x}^T \boldsymbol{\gamma}_r)}{1 + \sum_{i=1}^{k} \exp(\beta_{0ti} + \boldsymbol{x}^T \boldsymbol{\gamma}_i)}, \tag{6.1}$$

where $t = 1, ..., q$, and $r = 1, ..., k$. Then the parameters $\beta_{01r}, ..., \beta_{0qr}$ determine the cause-specific baseline hazard functions and $\boldsymbol{\gamma}_r$ contains the cause-specific effects of covariates. It suffices to specify the conditional probability of the target events $1, ..., k$ since conditional survival corresponds to the reference category in the multinomial logit model. Conditional probability of survival is implicitly determined by

$$P(T > t|T \geq t, \boldsymbol{x}) = 1 - \sum_{i=1}^{k} \lambda_r(t|\boldsymbol{x}) = \frac{1}{1 + \sum_{i=1}^{k} \exp(\beta_{0ti} + \boldsymbol{x}^T \boldsymbol{\gamma}_i)}.$$

With $R \in \{0, 1, ..., k\}$, where $R = 0$ denotes the conditional survival, the conditional probabilities are given by $\lambda_0(t|\boldsymbol{x}) = P(T > t|T \geq t, \boldsymbol{x}), \lambda_1(t|\boldsymbol{x}), ..., \lambda_k(t|\boldsymbol{x})$, which sum up to one. To simplify the cause-specific baseline effects, they can be expanded in basis functions, for example, in equally spaced B-splines resulting in

$$\beta_{0tr} = \sum_{m=1}^{m_r} \alpha_{0m}^{(r)} B_m(t).$$

The incorporation of B-splines also results in more parsimonious models when $m_r < q$ is chosen. Some more formal details referring to the incorporation of time-varying baseline effects can be found in Section 2.4.

6.2.2. Estimation

In this section, the derivation of the ML estimates for the multinomial logit model is shown. For subject i, the data are given by $(t_i, r_i, \delta_i, \boldsymbol{x}_i)$. Thereby, $t_i = \min(T_i, C_i)$ is the observed discrete duration time, where C is a random variable indicating the censoring

time and random censoring is assumed. Moreover, $r_i \in \{1, ..., k\}$ indicates the type of the terminating event, $\boldsymbol{x}_i$ a covariate vector and δ_i denotes the censoring indicator with

$$\delta_i = \begin{cases} 1, & T_i \leq C_i, \text{ that means failure in interval } [a_{t_i-1}, a_{t_i}) \\ 0, & T_i > C_i, \text{ that means censoring in interval } [a_{t_i-1}, a_{t_i}) . \end{cases}$$

This definition of the censoring indicator implicitly assumes that censoring occurs at the end of the interval. The corresponding likelihood contribution of the i-th observation for model (6.1) is given by

$$L_i = P(T_i = t_i, R_i = r_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i},$$

where for notational simplicity, the condition to the covariate vector $\boldsymbol{x}_i$ is omitted. Under the assumption that censoring does not depend on the parameters that determine the survival time (non-informative censoring, Kalbfleisch and Prentice, 2002), the factor $c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ can be omitted, resulting in the reduced likelihood

$$L_i = \lambda_{r_i}(t_i|\boldsymbol{x}_i)^{\delta_i}(1 - \lambda(t_i|\boldsymbol{x}_i))^{1-\delta_i} \prod_{t=1}^{t_i-1}(1 - \lambda(t|\boldsymbol{x}_i)).$$

Let $R_t = \{i : t \leq t_i\}$ be the risk set containing all objects who are at risk in interval $[a_{t-1}, a_t)$. For an alternative form of the likelihood, indicators for the transition to the next period are defined by

$$y_{itr} = \begin{cases} 1, & \text{failure of type } r \text{ occurs in interval } [a_{t-1}, a_t) \\ 0, & \text{no failure of type } r \text{ occurs in interval } [a_{t-1}, a_t) , \end{cases} \tag{6.2}$$

and

$$y_{it0} = \begin{cases} 0, & \text{failure of type } r \text{ occurs in interval } [a_{t-1}, a_t) \\ 1, & \text{no failure of type } r \text{ occurs in interval } [a_{t-1}, a_t) , \end{cases} \tag{6.3}$$

where $i \in R_t$ and $r = 1, ..., k$. That means, the indicator variable (6.3) is derived from the indicator variable (6.2) by $y_{it0} = 1 - y_{it1} - ... - y_{itk}$. These indicator variables are gathered in the vector $\boldsymbol{y}_{it}^T = (y_{it0}, y_{it1}, ..., y_{itk})$ denoting the response vector of object i, $i = 1, ..., n$, $t = 1, ..., t_i$. By means of the indicator variables (6.2) and (6.3) the likelihood contribution of the i-th observation is given by

$$\begin{aligned} L_i &= \prod_{t=1}^{t_i} \left(\prod_{r=1}^{k} \lambda_r(t|\boldsymbol{x}_i)^{y_{itr}} \right) \left(1 - \lambda(t|\boldsymbol{x}_i)\right)^{y_{it0}} \\ &= \prod_{t=1}^{t_i} \left(\prod_{r=1}^{k} \lambda_r(t|\boldsymbol{x}_i)^{y_{itr}} \right) \left(1 - \sum_{r=1}^{k} \lambda_r(t|\boldsymbol{x}_i)\right)^{y_{it0}}. \end{aligned}$$

That means, the likelihood for the i-th observation is identical to that for the t_i observations $\boldsymbol{y}_{i1}, ..., \boldsymbol{y}_{it_i}$ of a multinomial response model. The indicator variables actually represent the distributions, given a specific interval is reached. Thus, given that an object reaches interval $[a_{t-1}, a_t)$, the response is multinomially distributed with $\boldsymbol{y}_{it}^T = (y_{it0}, y_{it1}, ..., y_{itk}) \sim \mathcal{M}(1, (1 - \lambda(t|\boldsymbol{x}), \lambda_1(t|\boldsymbol{x}), ..., \lambda_k(t|\boldsymbol{x})))$. Therefore, the likelihood is that of the multicategorical model

$$P(Y_{it} = r|\boldsymbol{x}_i) = P(y_{itr} = 1|\boldsymbol{x}_i) = \frac{\exp(\eta_{itr})}{1 + \sum_{j=1}^{k} \exp(\eta_{itj})},$$

with $\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_i^T \boldsymbol{\gamma}_r$. Hence, ML estimates can be easily computed by using statistical software for multinomial regression models after construction of an appropriate design matrix. By introducing the design vector $\tilde{\boldsymbol{x}}_i$ (that includes the baseline effects and the covariate vector $\boldsymbol{x}_i$) and the corresponding parameter vector $\tilde{\boldsymbol{\gamma}}_t^T = (\beta_{0t1}, ..., \beta_{0tk}, \boldsymbol{\gamma}_1^T, ..., \boldsymbol{\gamma}_k^T)$, for the linear predictor follows

$$\boldsymbol{\eta}_{it} = (\eta_{it1}, ..., \eta_{itk})^T = (\tilde{\boldsymbol{x}}_{i1}^T \tilde{\boldsymbol{\gamma}}_t, ..., \tilde{\boldsymbol{x}}_{ik}^T \tilde{\boldsymbol{\gamma}}_t) = \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{\gamma}}_t,$$

with

$$\tilde{\boldsymbol{X}}_i = \begin{bmatrix} 1 & & 0 & \boldsymbol{x}_i^T & 0 \\ & \ddots & & \vdots & \\ 0 & & 1 & 0 & \boldsymbol{x}_i^T \end{bmatrix}.$$

The matrix $\tilde{\boldsymbol{X}}_i$ represents the design matrix for object i for time period t. Consequently, the matrix $\tilde{\boldsymbol{X}}_i$ has to be stacked t_i times to get the design matrix for object i. The collection of all individual design matrices yield the global design matrix $\tilde{\boldsymbol{X}}$ resulting in the following equation

$$\underset{(\sum_{i=1}^{n} t_i \cdot k) \times 1}{\boldsymbol{\eta}} = \underset{(\sum_{i=1}^{n} t_i \cdot k) \times (\sum_{r=1}^{k} q \cdot k + p \cdot k)}{\tilde{\boldsymbol{X}}} \underset{(\sum_{r=1}^{k} q \cdot k + p \cdot k) \times 1}{\tilde{\boldsymbol{\gamma}}}.$$

Finally, the total log-likelihood is given by

$$\begin{aligned} l &= \sum_{i=1}^{n} \sum_{t=1}^{t_i} \left(\sum_{r=1}^{k} y_{itr} \log \lambda_r(t|\boldsymbol{x}_i) + y_{it0} \log \left(1 - \sum_{r=1}^{k} \lambda_r(t|\boldsymbol{x}_i) \right) \right) \\ &= \sum_{t=1}^{q} \sum_{i \in R_t} \left(\sum_{r=1}^{k} y_{itr} \log \lambda_r(t|\boldsymbol{x}_i) + y_{it0} \log \left(1 - \sum_{r=1}^{k} \lambda_r(t|\boldsymbol{x}_i) \right) \right). \end{aligned} \tag{6.4}$$

If β_{0tr} is expanded in basis functions the design matrix $\tilde{\boldsymbol{X}}$ and the parameter vector $\tilde{\boldsymbol{\gamma}}$ have to be adopted.

6.3. Penalization

The linear predictor for modeling the cause-specific hazard function $\lambda_r(t|\boldsymbol{x}_i)$ has the form

$$\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_i^T \boldsymbol{\gamma}_r, \quad t = 1, ...q,\ r = 1, ..., k,$$

where $\boldsymbol{x}_i^T = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{ip})$, $\boldsymbol{\gamma}_r^T = (\gamma_{r1}, ..., \gamma_{rp})$. Because each covariate adds k parameters and the baseline hazard parameters β_{0tr} vary over time, the number of parameters can be very large rendering simple ML estimators unstable. To obtain a sparse representation, and in particular variable selection, penalized estimators are considered. Penalized ML estimators are obtained by adding a penalty term to the log-likelihood (6.4) yielding the penalized log-likelihood

$$l_{\xi_1,\xi_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) - J_{\xi_1,\xi_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}), \tag{6.5}$$

where $\boldsymbol{\beta}_0^T = (\beta_{011}, ..., \beta_{01k}, ..., \beta_{0q1}, ..., \beta_{0qk})$, $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_k)$ collect all corresponding parameters. Thereby, $l(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$ denotes the ordinary log-likelihood and $J_{\xi_1,\xi_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$ stands for a penalty term that depends on scalar tuning parameters ξ_1 and ξ_2. The tuning parameters ξ_1 and ξ_2 control the strength of penalization, whereas the choice of $J_{\xi_1,\xi_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$ determines the properties of the penalized estimator. As the objective is variable selection, the penalty should enforce that for variables that are not influential all corresponding parameters are set to zero simultaneously. Therefore, let all effects of the j-th covariate be collected in $\boldsymbol{\gamma}_{\bullet j}^T = (\gamma_{1j}, ..., \gamma_{kj})$. A penalty that enforces variable selection and smoothness of the baseline hazards then is given by

$$\begin{aligned} J_{\xi_1,\xi_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) &= \xi_1 \sum_{r=1}^{k} \sum_{t=2}^{q} (\beta_{0tr} - \beta_{0,t-1,r})^2 + \xi_2 \sum_{j=1}^{p} \phi_j \left\| \boldsymbol{\gamma}_{\bullet j} \right\| \\ &= \xi_1 J(\boldsymbol{\beta}_0) + \xi_2 J(\boldsymbol{\gamma}), \end{aligned} \tag{6.6}$$

where $\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_2 = \sqrt{\boldsymbol{u}^T \boldsymbol{u}}$ denotes the L_2-norm and $\phi_j = \sqrt{k}$ is a weight that adjusts the penalty level on parameter vectors $\boldsymbol{\gamma}_{\bullet j}$ for their dimension.

The first penalty term uses that time intervals are naturally ordered. Therefore, for each cause r, differences between coefficients of adjacent time periods are penalized in a similar way as in penalized splines (Eilers and Marx, 1996) and regression with ordered predictors (Gertheiss and Tutz, 2009). The penalty controls how quickly hazard rates can change and hence smoothes them over time. The second term enforces variable selection, that means, all parameters collected in $\boldsymbol{\gamma}_{\bullet j}$ are simultaneously shrunk towards zero. It is strongly related to the group lasso method (Yuan and Lin, 2006), but as stated in Tutz et al. (2012), in the group lasso the grouping refers to the parameters that are linked to a categorical predictor with respect to a univariate regression model, while in the present model grouping arises from the multivariate response structure. Without a penalty, that is with $\xi_1 = \xi_2 = 0$, ordinary ML-estimation is obtained.

The cause-specific baseline hazards may also be expanded in basis functions. This can be carried out by using equally spaced B-splines resulting in the linear predictor

$$\eta_{tr} = \sum_{m=1}^{m_r} \alpha_{0m}^{(r)} B_m(t) + \boldsymbol{x}^T \boldsymbol{\gamma}_r.$$

The incorporation of B-splines requires a modification of the penalty term given by

$$J_{\xi_1,\xi_2}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \xi_1 \sum_{r=1}^{k} \sum_{m=2}^{m_r} (\alpha_{0m}^{(r)} - \alpha_{0,m-1}^{(r)})^2 + \xi_2 \sum_{j=1}^{p} \phi_j \left\| \boldsymbol{\gamma}_{\bullet j} \right\|. \tag{6.7}$$

Again, the first term of the penalty steers the smoothness of the baseline effects, whereas the second term enforces variable selection.

6.4. Computational Issues

In the following, some details regarding the estimation approach are described. Some details on the estimation approach itself are outlined and the modification to adaptive penalties is described. Finally, the tuning parameter selection for discrete competing risks models is presented.

Estimation

To estimate the parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}$, the penalized log-likelihood (6.5) has to be maximized. For a simple penalized multinomial logit model, Tutz et al. (2012) have shown how a corresponding maximization problem can be solved by means of the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009) and called the resulting method Categorically Structured Lasso (CATS Lasso). FISTA is a proximal gradient method, where only the log-likelihood and its gradient, but no higher derivatives are used. FISTA combines quick convergence with cheap iterates that are well-suited for the specific challenges of multinomial logit models.
In a simple multinomial regression model the intercept is often considered to be category-specific but usually the intercept is not assumed to be time-varying. As Tutz et al. (2012) considered only category-specific intercepts they do not regard the special structure of the competing risks model (6.1). In model (6.1), the parameters $\beta_{01r}, ..., \beta_{0qr}$ represent the time-varying cause-specific baseline hazard functions. In addition to the variable selection of all parameters belonging to one predictor variable, the objective is smooth baseline hazard functions that can be obtained by using the penalty term (6.6). This penalty term uses that time intervals are naturally ordered by penalizing differences between coefficients of adjacent time periods for each cause r. Hence, the existing CATS Lasso is extended to allow for smooth cause-specific baseline effects.

All details on the algorithm regarding penalized multinomial logit models can be found in Tutz et al. (2012). An implementation of the algorithm is provided by the R package `MLSP`. The current version 0.1 can be downloaded from `http://www.statistik.lmu.de/~poessnecker/software.html` and will be available as a proper R add-on package via CRAN (see `http://cran.r-project.org`) in the near future.

Adaptive Penalties

Similar to Section 3.2.1, the idea of adaptive penalties, that is, weighting the penalty terms by the inverse of the respective unpenalized parameter estimates, can be used. This modification is applied due to the inconsistency of simple lasso or group lasso penalties (Zou, 2006; Wang and Leng, 2008). Given penalty terms (6.6) or (6.7), the adaptive version is obtained by replacing the weights ϕ_j by

$$\phi_j^a = \frac{\sqrt{k}}{||\hat{\boldsymbol{\gamma}}_{\bullet j}^{\mathrm{ML}}||}, \tag{6.8}$$

where $\hat{\boldsymbol{\gamma}}_j^{\mathrm{ML}}$ denote the according ML-estimates. The intuition behind this weighting procedure is rather straightforward. With very large data sets, unpenalized point estimates can be expected to be rather accurate. Thus, the norm of ML-estimates of parameter groups belonging to relevant predictors is rather large. Consequently, the corresponding penalization term is small. In contrast, a strong penalization is obtained for parameter groups belonging to irrelevant predictors and, hence, leading to a small norm of the ML-estimates. Moreover, the ML-estimates in definition (6.8) can be replaced by any $\sqrt{n}$-consistent estimates (Tutz et al., 2012).

Tuning Parameter Selection

The tuning parameters ξ_1 and ξ_2 are chosen by K-fold cross-validation (see also Section 3.2.2 and Section 5.3.2). The choice of the tuning parameters ξ_1 and ξ_2 is based on a two-dimensional grid of possible parameters on which the optimal parameter combination is determined by cross-validation. In analogy to the previous chapters, the splitting of the cross-validation approach refers to objects instead of individual data points leading to the inclusion of the whole information of an object.

A possible approach of assessing the predictive performance of a model is the predictive deviance. For a new observation $(t_i^{pred}, r_i^{pred}, \delta_i^{pred}, \boldsymbol{x}_i^{pred})$, it is defined by

$$\mathrm{D}_i = 2 \sum_{t=1}^{t_i} \sum_{r=0}^{k} y_{itr}^{pred} \log \frac{y_{itr}^{pred}}{\hat{\lambda}_r(t|\boldsymbol{x}_i^{pred})},$$

where $\lambda_r(t|\boldsymbol{x}_i) = P(T_i = t, R_i = r | T_i \geq t, \boldsymbol{x}_i^{pred})$ and $(y_{i11}^{pred}, ..., y_{i1k}^{pred}, ..., y_{it_ik}^{pred}, ..., y_{it_ik}^{pred})$ denotes the transitions over periods and risks of object i.

6.5. Applications

In this section, the proposed penalized competing risks model for discrete duration time is applied to two real data problems. The first data set describes Congressional careers in the United States. Unemployment data taken from the German socioeconomic panel constitute the second data set.

6.5.1. Congressional Careers

The first data example deals with the careers of incumbent members of the U.S. Congress and is used in the book of Box-Steffensmeier and Jones (2004). Therein, the data set is employed to compute unpenalized discrete survival models. It is available on the corresponding website of the book `http://psfaculty.ucdavis.edu/bsjjones/eventhistory.html`. A detailed description can be found in Jones (1994).
A congressman can end his legislative career in four different ways. He might retire (*Retirement*), he might be ambitious and seek an alternative office (*Ambition*), he might lose a primary election (*Primary*) or he might lose a general election (*General*). Career path data were collected on every member of the House of Representatives from each freshman class elected from 1950 to 1976. Each incumbent in the data set was tracked from the first reelection bid until the last term served in office. A member initially elected in 1950 does not enter the risk set until the election cycle of 1952 as the members of the House of Representatives serve two-year terms. At each subsequent election, a terminating event or reelection is observed. Once a terminating event is experienced, the incumbent is no longer observed. All election cycles from 1952 up to 1992 are covered in this data set. The last freshman class on which data were collected was 1976.
The dependent variable defines the transition process of a Congressman from his first election up to one of the competing events *General, Primary, Retirement or Ambition*. Thereby, the duration until the occurrence of one of the competing events is measured as terms served, where a maximum of 16 terms can be reached. Originally, up to 20 terms occurred, however, only for very few Congressmen. Hence, due to stability reasons, durations that exceed 15 terms are aggregated. Furthermore, only complete cases, that is, observations with no missing values for any covariate are incorporated in the analysis. Several covariates are used as predictors of career termination. Thereby, the covariate *Age* constitutes the incumbent's age at each election cycle and is centered around 51 years (sample mean: 51.26) to improve interpretability. The incumbent's margin of victory in his or her previous election is collected in the variable *PriorMargin*, that is centered around a margin of 35 (sample mean: 35.21). The covariate *Redistricting* indicates if the incumbent's district was substantially redistricted. By means of the covariate *Scandal* it is captured if an incumbent was involved in an ethical or sexual misconduct scandal or when the incumbent was under criminal investigation. The covariates *OpenGub* and *OpenSen* indicate if there is an open gubernatorial and/or open Senatorial seat available in the incumbent's state. The data set considers members of the Republican and the Democratic Party. Whether the

Congressman is a member of the Republican party is gathered in the variable *Republican*. Finally, in *Leadership* it is described if a member is in the House leadership and/or is a chair of a standing House committee. Except the predictor *Republican* all covariates are time-varying, that is, the covariate values per object may vary over the duration time. An overview of the used predictors is shown in Table 6.1.

Variable	Description
Duration	Time (in terms served) the incumbent has spent in Congress prior to the election cycle
Age	Incumbent's age (in years) at each election cycle, centered around 51
Republican	Member of the Republican party 0: no, 1: yes
PriorMargin	The incumbent's margin of victory in his or her previous election, centered around 35
Leadership	Prestige position 0: otherwise, 1: member is in the House leadership and/or is a chair of a standing House committee
OpenGub	Open gubernatorial seat available in the incumbent's state 0: no, 1: yes
OpenSen	Open Senatorial seat available in the incumbent's state 0: no, 1: yes
Scandal	Incumbent was involved in an ethical or sexual misconduct scandal or when the incumbent was under criminal investigation 0: no, 1: yes
Redistricting	The incumbent's district was substantially redistricted 0: no, 1: yes

Table 6.1. Description of the variables of the Congressional career data.

The used data set contains the career paths of 860 Congressmen and is already available in the long format, that means that each row of the data set represents an individual's observation for a specific time period (see also Section 6.2.2). To be on comparable scales, all covariates are standardized to have equal variance, in order to avoid that coefficient values are scale dependent. A penalized multinomial logit model is applied with k=4 risks defined by cause 1 (*General*), 2 (*Primary*), 3 (*Retirement*) and 4 (*Ambition*). Thus, the model is

$$\lambda_r(t|\boldsymbol{x}) = \frac{\exp(\eta_{itr})}{1 + \sum_{j=1}^{4} \exp(\eta_{itj})}, \qquad r = 1, 2, 3, 4,$$

with cause-specific linear predictors $\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_{it}^T \boldsymbol{\gamma}_r$. All covariates described in Table 6.1 are incorporated in the model. Moreover, the model considers all pairwise interactions except for *Republican:Leadership*, *Leadership:Redistricting*, *Opengub:Scandal*, *Scandal:Redistricting* since too few observations of those covariate combinations appear in the data. The use of interactions increases the model's complexity and its interpretation. Such a high-dimensional interaction model cannot be properly handled by unpenalized ML-estimation. The task of model stabilization and efficient variable selection is tackled by

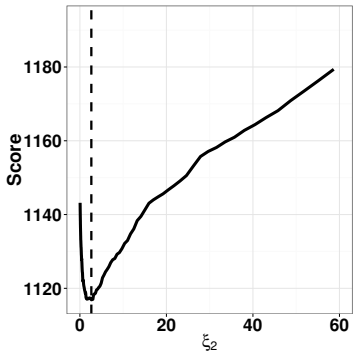


Figure 6.1. Cross-validation score subject to penalty parameter ξ_2 for $\xi_1 = 6.0$ for the Congressional career data.

using penalization (see Section 6.3). Referring to the penalty term (6.6), the employed penalty term is given by

$$J(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) = \xi_1 \sum_{r=1}^{4} \sum_{t=2}^{16} (\beta_{0tr} - \beta_{0,t-1,r})^2 + 2\xi_2 \sum_{j=1}^{32} \left\| \boldsymbol{\gamma}_{\bullet j} \right\| .$$

It allows for smooth cause-specific baseline effects β_{0tr} and a variable selection of the covariate effects and the interaction effects collected in the vectors $\boldsymbol{\gamma}_{\bullet 1}, ..., \boldsymbol{\gamma}_{\bullet 32}$. The adaptive version of the penalty has shown an improvement referring to the model assessment (e.g. cross-validation score) and the variable selection. The latter means, that also coefficients with an extremely small estimate for the non-adaptive case were removed from the model when using adaptive weights in the penalty. Hence, adaptive weights (6.8) are included in the penalty. Tuning parameters ξ_1 and ξ_2 are chosen on a 2-dimensional grid by 5-fold cross-validation with the predictive deviance as loss criterion. This results in tuning parameters $\xi_1 = 6.0$ and $\xi_2 = 2.64$. For a fixed $\xi_1 = 6.0$, the corresponding cross-validation score is shown in Figure 6.1, where the vertical black dashed line marks the chosen tuning parameter.

In Figure 6.2, parameter estimates for the cause-specific time-varying baseline effects are shown. The corresponding pointwise confidence intervals, marked by light-gray dashed lines, have been estimated by a nonparametric bootstrap method proposed by Efron (1979) with 1000 bootstrap replications (see also Section 3.3). It can be seen that it is justified to allow for cause-specific baseline effects as their run is quite different. Due to the penalization of differences between coefficients of adjacent time periods $\beta_{0tr} - \beta_{0,t-1,r}$, the estimated baseline effects are quite smooth.

Parameter estimates of the covariate effects are summarized in Table 6.2. Therein, the ordinary ML-estimates and the estimates resulting from the penalized competing risks model with their corresponding standard errors are shown. The computation of the standard er-

rors is based on a nonparametric bootstrap approach with 1000 bootstrap replications. It can be immediately seen, that the penalization method removes a considerable amount of effects, that is 68 out of 128 parameters, from the model, leading to an enormous reduction of the model complexity. The incorporated selection procedure suggests that the main effects *Republican* and *Leadership* are not needed in the predictor. Moreover, a large number of interaction effects are not selected. For example, the absolute values of the covariate *Scandal* indicates a strong effect. If a Congressman became embroiled in a scandal it is more likely that he/she lose a primary or general election or to retire when compared to being reelected. In contrast, a scandal decreases the probability of seeking an alternative office compared to reelection. The exact interpretation of the parameter estimates is analogous to the multinomial logit model. Especially, the use of log odds facilitates the interpretation. Thereby, the log odds between cause r and the category of being reelected corresponds to the linear predictor η_{itr} for a fixed time period t.

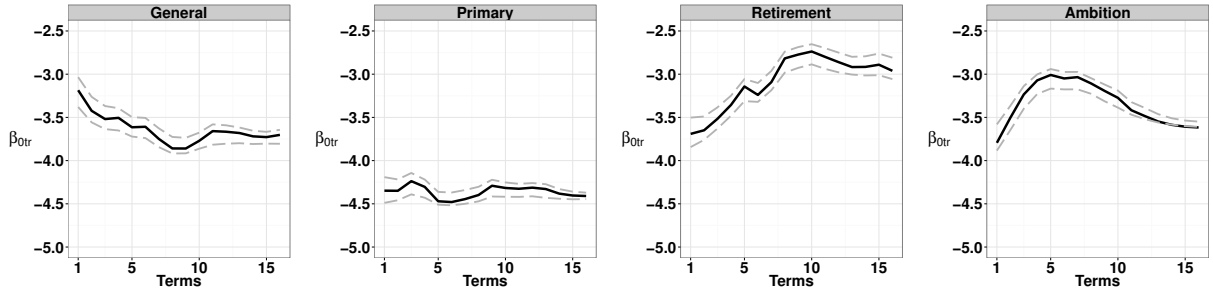


Figure 6.2. Parameter estimates of the cause-specific time-varying baseline effects for the Congressional careers data. Dashed lines represent the 95% pointwise bootstrap interval.

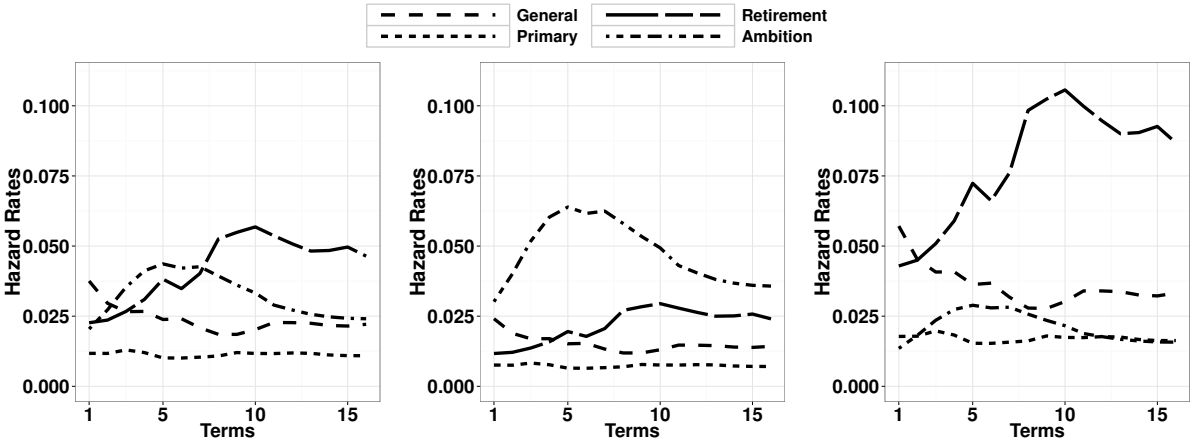
A selection of resulting hazard rates is depicted in Figure 6.3. In particular, Figure 6.3a shows hazard functions for the following covariate characteristics: Age=51, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting for the transitions to *General*, *Primary*, *Retirement* and *Ambition*. That means, all covariate characteristics are set at reference. It can be seen, that the probability of retirement tends to increase continuously over time. The probability for seeking an alternative office compared to reelection increases for small terms and decreases slowly. The hazard rate for losing either a primary or a general election is rather constant in the reference group. Figures 6.3b and 6.3c show respectively the hazard rates for younger (Age=41) and older (Age=61) Congressmen compared to the reference group (Age=51), while everything else remains unchanged. Younger Congressmen prefer to seek an alternative office and they do not intend to retire. For older Congressmen, the probability of retirement compared to reelection strongly increases. Moreover, the probability of losing either a primary or a general election is enhanced compared to the reference group. Further plots of estimated cause-specific hazard rates can be found in Figure A.5 in Appendix A.4.

The visualization of the shrinkage and the selection effect is carried out by coefficient paths. For the main effects these coefficient paths are summarized in Figure 6.4, whereas those of

	General			Primary			Retirement			Ambition		
	ML	pen.	sd	ML	pen.	sd	ML	pen.	sd	ML	pen.	sd
Age	0.069	0.046	0.008	0.071	0.046	0.011	0.070	0.068	0.008	-0.034	-0.037	0.007
Republican	0.255	0	0.005	-0.188	0	0.002	-0.201	0	0.009	0.343	0	0.018
PriorMargin	-0.078	-0.060	0.005	0.006	0.001	0.005	-0.007	-0.005	0.003	-0.010	-0.004	0.002
Leadership	-0.272	0	0.087	-2.779	0	0.081	-0.393	0	0.065	0.033	0	0.080
Open Gub.	0.815	0.205	0.116	0.598	0.181	0.097	0.227	0.109	0.077	0.528	0.208	0.121
Open Sen.	-0.638	-0.243	0.125	-0.215	-0.193	0.134	-0.086	0.062	0.125	1.136	0.878	0.134
Scandal	3.750	2.689	0.370	3.215	3.272	0.428	1.921	1.611	0.441	-3.118	-1.532	0.073
Redistricting	2.548	1.617	0.447	1.465	1.149	0.499	-0.563	0.431	0.251	0.574	0.801	0.309
Age:Republican	0.007	0.011	0.007	-0.045	-0.010	0.007	0.041	0.030	0.009	-0.038	-0.029	0.009
Age:PriorMargin	0.001	0.000	0.000	-0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0000
Age:Leadership	0.014	0	0.002	-0.117	0	0.002	0.018	0	0.002	-0.269	0	0.001
Age:OpenGub.	-0.006	0	0	0.034	0	0	-0.016	0	0	-0.011	0	0
Age:OpenSen.	-0.005	0	0.001	-0.074	0	0.001	-0.039	0	0.004	-0.015	0	0.002
Age:Scandal	-0.106	0	0	0.022	0	0	0.090	0	0	0.009	0	0
Age:Redistricting	-0.001	0.007	0.016	-0.066	-0.039	0.018	0.174	0.097	0.031	0.037	0.018	0.016
Republican:PriorMargin	0.016	0.005	0.004	-0.041	-0.016	0.005	-0.008	-0.004	0.004	0.015	0.012	0.004
Republican:OpenGub.	-0.532	-0.342	0.200	-4.282	-1.337	0.147	-0.147	-0.233	0.201	-0.063	0.294	0.184
Republican:OpenSen.	0.323	0	0.001	-0.092	0	0.002	0.802	0	0.010	-0.260	0	0.011
Republican:Scandal	0.007	0	0.021	2.121	0	0.054	0.182	0	0.005	-1.418	0	0.001
Republican:Redistricting	-1.833	0	0.076	0.447	0	0.059	1.247	0	0.050	-0.276	0	0.051
PriorMargin:Leadership	0.025	0	0	-0.009	0	0	-0.008	0	0.001	0.057	0	0
PriorMargin:OpenGub.	0.020	0	0	-0.001	0	0.001	0.008	0	0.001	0.009	0	0.001
PriorMargin:OpenSen.	-0.016	0	0.001	-0.019	0	0.002	0.013	0	0.002	0.011	0	0.004
PriorMargin:Scandal	0.006	0.007	0.005	-0.017	-0.010	0.004	-0.071	-0.019	0.006	-0.028	-0.001	0
PriorMargin:Redistricting	0.066	0.037	0.019	0.000	-0.002	0.003	0.030	0.010	0.006	-0.013	-0.009	0.007
Leadership:OpenGub.	-5.168	0	0.117	-1.693	0	0.087	1.054	0	0.359	-5.402	0	0.116
Leadership:OpenSen.	-4.513	0	0	-0.941	0	0	1.001	0	0	-6.053	0	0
Leadership:Scandal	-0.213	-0.029	0.594	-4.212	-1.803	0.733	-8.621	-1.925	0.756	-0.897	-0.108	0.047
OpenGub.:OpenSen.	-0.436	0	0	0.124	0	0	-0.280	0	0	-0.429	0	0
OpenGub.:Redistricting	-0.175	0.172	0.663	-4.274	-0.415	0.125	-5.297	-0.666	0.237	2.751	2.126	0.932
OpenSen.:Scandal	-2.277	0	0.307	-1.482	0	0.206	-8.270	0	0.266	-3.311	0	0.058
OpenSen.:Redistricting	0.914	0	0.052	-4.560	0	0.006	-0.522	0	0.031	1.771	0	0.147

Table 6.2. Parameter estimates for the Congressional careers data. Ordinary maximum likelihood estimates are denoted by "ML", the penalized estimates are denoted by "pen.". Estimated standard errors for the penalized model obtained by a bootstrap approach are given in the columns denoted by "sd".

the interaction effects can be found in Figure A.6 in Appendix A.4. Each path indicates the penalized estimates subject to tuning parameter ξ_2. In particular, the paths illustrate how the estimates changes towards zero for increasing ξ_2. Hence, they show the development of the covariates for the terminating events when penalization is increased. The dashed black line indicates the ξ_2 chosen via cross-validation and the corresponding estimates. For simplicity reasons the abscissa is transformed by applying $\log(1 + \xi_2)$.



(a) Estimated rates for all predictors at reference: Age=51, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

(b) Estimated rates for Age=41, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

(c) Estimated rates for Age=61, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

Figure 6.3. Estimated cause-specific hazard rates over time for the Congressional careers data.

6.5.2. Unemployment Data

In this section, the proposed penalized competing risks model is applied to unemployment data. The data set has originally been analyzed by Kauermann and Khomski (2009). Based on the German socio economic panel (SOEP; see `www.diw.de`), individuals who have been unemployed at least once during the years 1990 to 2000 are considered. Generally, for individuals in the panel with more than one spell of unemployment, one of their spells is chosen randomly and the others are ignored. This guarantees independence of the observations. Each unemployment spell terminates due to different competing reasons. It is focused on two competing risks (or in terminology of unemployment better called chances), that is, *part-time* reemployment (r = 1) and *full-time* reemployment (r = 2). All other reasons for terminating unemployment are taken as censoring.

The dependent variable defines the transition process of an individual up to one of the competing events *part-time* or *full-time* reemployment. Thereby, the duration until the occurrence of one of the competing events is measured in months, where a maximum of 36 months can be reached. Several covariates are used as predictors of reemployment like *Nationality*, *Gender*, *Age*, *Education* and *Training*, measured respectively at the beginning of the unemployment spell. In the following analysis, the focus is on the publicly available version of the data that is part of the R add-on package `CompetingRiskFrailty`. In the meantime, the package was removed from the CRAN repository, but formerly available versions, including the data set, can be obtained from the archive. The list of explanatory variables that will be used for modeling is presented in Table 6.3.

The available data set consists of 500 unemployed persons. Restructuring of the data was executed according to Section 6.2.2, to fit a penalized multinomial logit regression model

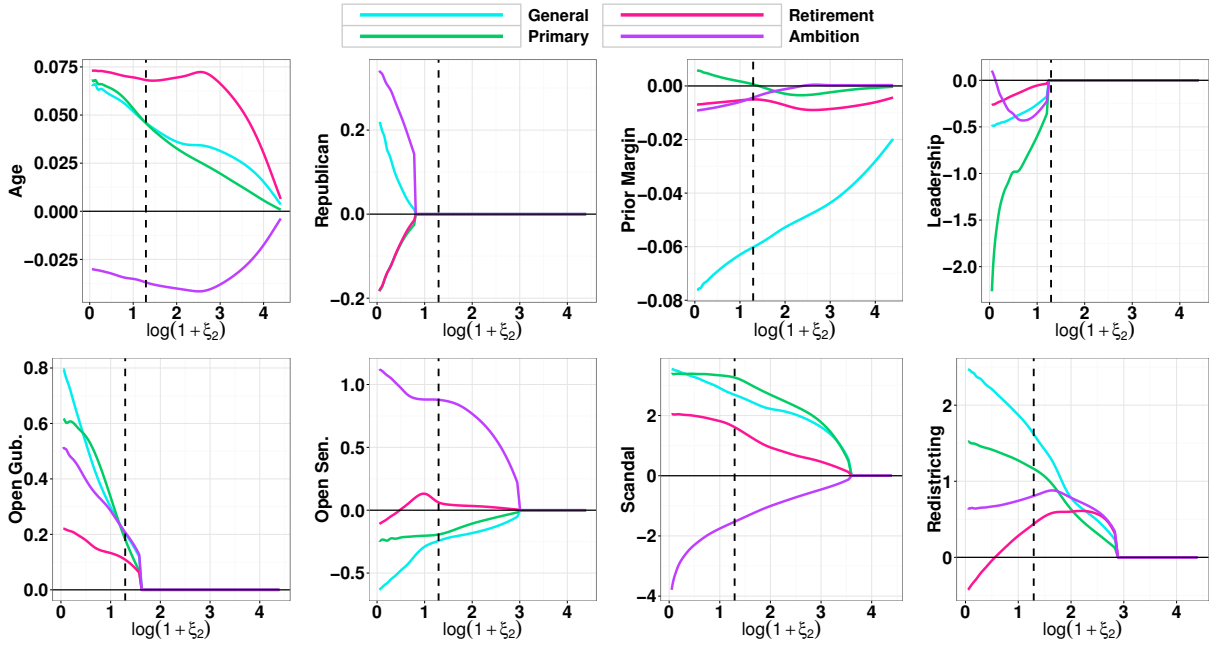


Figure 6.4. Coefficient paths of the main effects for the Congressional career data.

with k=2 risks defined by cause 1 (*part-time*) and 2 (*full-time*). To be on comparable scales, all covariates are standardized to have equal variance. The used model is given by

$$\lambda_r(t|\boldsymbol{x}) = \frac{\exp(\eta_{itr})}{1 + \sum_{j=1}^{2} \exp(\eta_{itj})}, \qquad r = 1, 2,$$

with cause-specific linear predictors $\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_i^T \boldsymbol{\gamma}_r$. All covariates described in Table 6.3 are incorporated in the model. Moreover, the model considers all pairwise interaction effects. Referring to the penalty term (6.6), the used penalty term is given by

$$J(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) = \xi_1 \sum_{r=1}^{2} \sum_{t=2}^{36} (\beta_{0tr} - \beta_{0,t-1,r})^2 + \sqrt{2}\xi_2 \sum_{j=1}^{20} \left\| \boldsymbol{\gamma}_{\bullet j} \right\|.$$

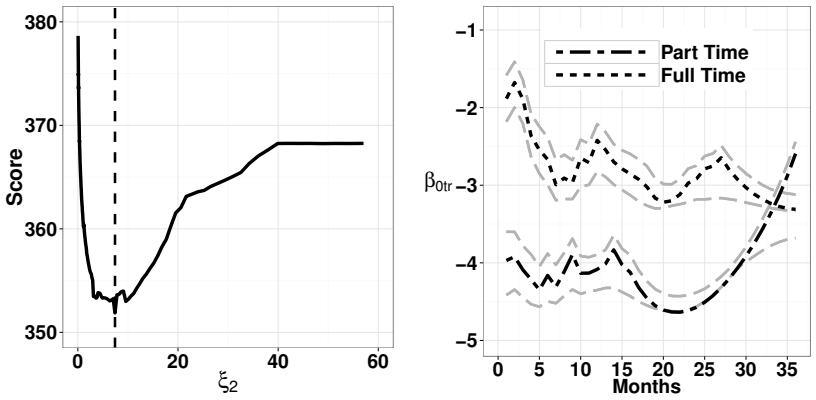
In analogy to the previous example the penalty term allows for smooth cause-specific baseline effects and a variable selection of the covariate effects and the interaction effects. The adaptive version of the penalty has not shown a clear improvement regarding model assessment (e.g. cross-validation score) and the variable selection, hence, it is omitted. Tuning parameters ξ_1 and ξ_2 are chosen by 5-fold cross-validation with regard to the predictive deviance. This results in tuning parameters $\xi_1 = 1.0$ and $\xi_2 = 7.42$. For a fixed $\xi_1 = 1.0$, the corresponding cross-validation score is shown in Figure 6.5a, where the vertical black dashed line marks the chosen tuning parameter.
Parameter estimates for the cause-specific time-varying baseline effects are shown in Figure 6.5b. The corresponding pointwise confidence intervals have been estimated by a nonpara-

Variable	Description
Time	Time spent in the unemployment spell, measured in months.
	The spells which lasted more than 36 months have been truncated on
	36 months and denoted as censored
Nationality	Nationality of the unemployed person
	0: German, 1: Foreigners
Gender	Gender of the unemployed person
	0: Male, 1: Female
	Age of the unemployed person at the beginning of the unemployment spell
Age young	0: no, 1: yes ($\leq$ 25 years)
Age old	0: no, 1: yes ($>$ 50 years)
Training	Unemployed individual has successfully completed a professional training
	0: yes, 1: no
University	Unemployed individual has an university degree or equivalent qualification
	0: no, 1: yes

Table 6.3. Description of the variables of the unemployment data.

metric bootstrap method proposed by Efron (1979) with 1000 bootstrap replications (see also Section 3.3). As the run of the baseline functions is quite different, it is justified to allow for cause-specific baseline effects. The run of the baseline effects over time is not as smooth as in the Congressional careers example. This is due to the fact that ξ_1, steering the smoothness of the baseline effects, is chosen to be equal to one.

In Table 6.4, the parameter estimates of the covariate effects are summarized. Therein, the ordinary ML-estimates, the estimates resulting from the penalized competing risks model with their corresponding standard errors are shown. The computation of the standard errors is based on a nonparametric bootstrap approach with 1000 bootstrap replications. It can be immediately seen, that the penalization method removes a considerable amount of effects, that is 22 out of 40 parameters, from the model, leading to a enormous reduction of the model complexity. All main effects remain in the models, whereas only three interaction effects are selected. For example, for women it is more likely to get a part-time job and less likely to get a full-time job and for younger people, getting a full-time job is more likely than getting a part-time job. The exact interpretation of the parameter estimates is analogous to the multinomial logit model. Especially, the use of log odds facilitates the interpretation. Thereby, the log odds between cause r and the category of being unemployed corresponds to the linear predictor η_{itr} for a fixed time period t. Figure 6.6 depicts a selection of resulting hazard rates. In particular, Figure 6.6a gives hazard functions for a middle-aged German men with a professional training and no university degree for the transitions to *part-time* reemployment and *full-time* reemployment. That means, that all characteristics are set at reference. For a transition to *full-time* reemployment the hazard rate shows the typical pattern of unemployment data with a short increase and slow decrease. The hazard rate for the transition to *part-time* reemployment is rather constant at the beginning of the observation period but from a duration time of 25 months it increases. In Figure 6.6b, it can be observed that fewer women get a full-time job than men, whereas slightly more women

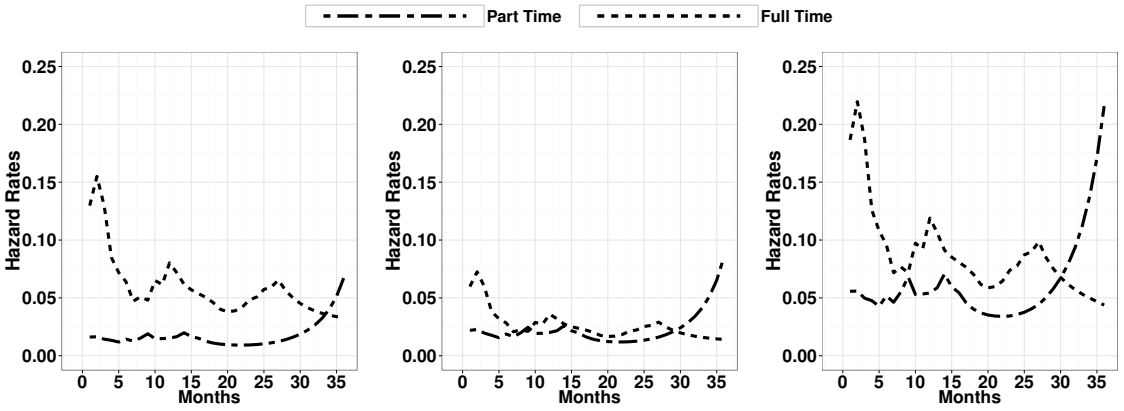


(a) Cross-validation score subject to penalty parameter ξ_2 for $\xi_1 = 1.0$

(b) Parameter estimates of the cause-specific time-varying baseline effects. Dashed lines represent pointwise 95% bootstrap confidence intervals.

Figure 6.5. Plots corresponding to the unemployment data.

get a part-time job, while everything else remains unchanged, that is, in the reference group. A transition to a university degree clearly increases the probability of getting a full-time or part-time job. The remaining plots of the estimated cause-specific hazard rates can be found in Figure A.7 in Appendix A.4. The visualization of the shrinkage and the



(a) Estimated rates for all predictors at reference: German, Male, Middle Age, Training, No University

(b) Estimated rates for German, Female, Middle Age, Training, No University

(c) Estimated rates for German, Male, Middle Age, Training, University

Figure 6.6. Estimated cause-specific hazard rates over time for the transition to *part-time* reemployment and *full-time* reemployment.

selection effect is carried out by coefficient paths. For the main effects, these coefficient paths are summarized in Figure 6.7, whereas those of the interaction effects can be found in Figure A.8 in Appendix A.4. Each path indicates the penalized estimates subject to tuning parameter ξ_2 for $\xi_1 = 1.0$. In particular, the paths illustrate how the estimates change towards zero for increasing ξ_2. The dashed black line indicates the ξ_2 chosen via cross-

	Part-time			Full-time		
	ML	pen.	sd	ML	pen.	sd
Nationality	-1.569	-0.317	0.115	0.269	0.125	0.079
Gender	1.115	0.236	0.126	-1.207	-0.847	0.129
Age young	0.371	-0.274	0.133	-0.042	0.088	0.124
Age old	-3.501	-0.879	0.156	-0.642	-0.746	0.179
Training	0.547	-0.023	0.069	-1.058	-0.389	0.142
University	3.043	1.360	0.380	0.757	0.483	0.189
Nationality:Gender	0.428	0	0.028	-0.251	0	0.047
Nationality:Age young	-2.851	0	0.007	-0.029	0	0.033
Nationality:Age old	-1.534	0	0.007	-5.104	0	0.021
Nationality:Training	0.414	0	0.015	0.299	0	0.016
Nationality:University	-0.343	-0.350	0.230	-2.618	-0.898	0.393
Gender:Age young	-1.135	-0.090	0.052	0.468	0.222	0.122
Gender:Age old	-2.278	-0.188	0.080	-0.711	-0.166	0.091
Gender:Training	-0.645	0	0.010	0.777	0	0.019
Gender:University	-1.324	0	0.069	-1.020	0	0.093
Age young:Training	-0.204	0	0.024	0.432	0	0.041
Age old:Training	0.977	0	0.040	-1.407	0	0.107
Age young:University	-5.885	0	0.045	0.876	0	1.073
Age old:University	0.671	0	0.063	-0.959	0	0.125
Training:University	-0.822	0	0.049	0.959	0	0.033

Table 6.4. Parameter estimates for the unemployment data. Ordinary maximum likelihood estimates are denoted by "ML", the penalized estimates are denoted by "pen.". Estimated standard errors for the penalized model obtained by a bootstrap approach are given in the columns denoted by "sd".

validation and the resulting estimates. For simplicity reasons the abscissa is transformed by applying $\log(1 + \xi_2)$.

6.6. Concluding Remarks

In this chapter, a penalized competing risks model for discrete duration time is proposed. The embedding of discrete competing risks models into the framework of a multinomial regression model allows for the application of the CATS lasso method introduced by Tutz et al. (2012). In competing risks models for discrete duration time, the cause-specific hazard rates are of main interest. When modeling these cause-specific hazard rates, several coefficients for each explanatory variable exist forming their own group. The proposed penalization method enables the simultaneous shrinkage of parameters belonging to such a group. A parameter group even can be completely removed from the model. Hence, true variable selection is performed. Moreover, the proposed method allows that parameters representing the cause-specific baseline hazards vary over time. In order to avoid that parameters of adjacent time periods of the baseline effects have completely different values, a further penalty term is incorporated steering the smoothness of the baseline effects.

Another interesting question would be the incorporation of time-varying covariate effects. By expanding the time-varying coefficients in basis functions, smooth curves over time can be obtained. This can be carried out, for example, by using equally spaced B-splines. To control for unobserved heterogeneity, a further advancement might be the inclusion of frailty

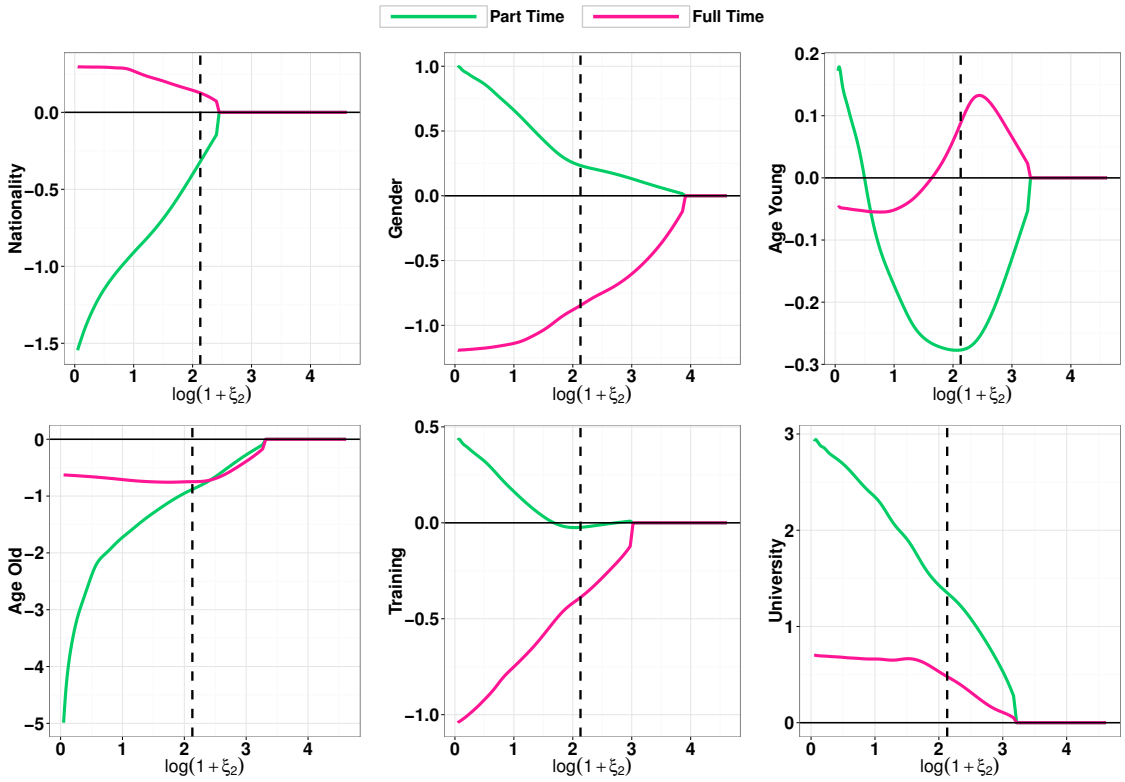


Figure 6.7. Coefficient paths of the main effects for the unemployment data.

effects in analogy to Chapter 5. By accounting for these two issues, the linear predictor is given by

$$\eta_{itr} = b_{ir} + \beta_{0tr} + \sum_j x_{itj}\beta_{jtr} = b_{ir} + \beta_{0tr} + \sum_j \tilde{x}_{itj}\alpha_{jr},$$

with $\beta_{jtr} = \sum_{m=1}^{m_{jr}} \alpha_{jm}^{(r)} B_m(t)$ and b_{ir} is considered to be an individual-specific random effect that is appropriately distributed. Furthermore, multivariate frailties are required here, with a separate component for each competing risk. This leads to a frailty vector for each individual. Note, that $\tilde{x}_{itj}$ differs from x_{itj}. The $\tilde{x}_{itj}$ contain the design of the interaction of the according covariates and the evaluations of the appropriate basis functions. The incorporation of B-splines requires a modification of the penalty term. Moreover, a weighting has to be introduced that accounts for smoothing and selection of the time-varying covariate effects. If frailties are considered in the linear predictor, the estimation algorithm has to be completely revised to maximize the marginal log-likelihood.

7. Conclusion and Outlook

In this thesis, penalization methods for survival models for discrete duration time are proposed. Single spell discrete-time survival models can be embedded into the framework of generalized linear models leading to a data situation with a large number of observations that are only rarely observed, especially when many time periods are considered. This data situation becomes even more difficult when time-varying covariate effects are incorporated. However, as survival data are measured over time, in this thesis, it is mainly focused on the incorporation of time-varying covariate effects.

To cope with this special data situation, in Chapter 3 penalty methods for discrete survival models with time-varying coefficients are considered. Penalization means, to add a penalty term to the log-likelihood. This penalty term determines the properties of the penalized estimator. Thereby, in Chapter 3, the incorporation of penalties is not restricted to a special type of penalty terms, but it is allowed for any combination of penalty terms. Hence, it is possible to add a specific penalty term referring to the baseline effects and adding further penalty terms that only affect the covariate effects. One further benefit of this approach is the predominantly smooth temporal variation of time-varying covariate effects. Furthermore, the proposed method can perform variable selection leading to interpretable and parsimonious models. Hence, the resulting procedure is considerably flexible and can be applied to a variety of applications. However, the results of the simulation study of Chapter 3 have shown that caution is recommended in the case of correlated variables due to poor variable selection.
The estimates of the penalization approach are obtained by a penalized pseudo Fisher scoring that is quite time-consuming. Hence, a future objective would be the optimization of the algorithm with the aim to be more efficient. Furthermore, in Chapter 3, the proposed variable selection technique, regarding time-varying coefficients, removes whole parameter groups belonging to one time-varying coefficient, that is, the interaction of time and the corresponding predictor, from the model. That means, grouping refers to the parameters that are linked to this interaction and finally, only single parameters can be removed from the model. However, in terms of categorical predictors an additional grouping of the categories would be of interest representing a further idea of future research.

The strength of penalization is steered by tuning parameters. The tuning parameters of the penalization approach in Chapter 3 are chosen by cross-validation with the predictive deviance as loss function. In Chapter 4, it is systematically investigated if the performance

of penalized discrete survival regression models can be improved by modifying the loss function. Therefore, several prediction measures used in the area of continuous survival outcomes are adapted to discrete time survival outcomes. It has been shown that the predictive deviance cannot be outperformed by other loss functions. This can be explained by the predictive deviance being a likelihood-based measure. Since the corresponding optimization criterion maximizes the penalized likelihood, it is based on the likelihood as well. Hence, an interesting aspect would be that the predictive measures of interest are already used in the optimization approach. This can be carried out, for example, by a componentwise gradient boosting algorithm (Bühlmann and Hothorn, 2007) that uses the predictive measure as optimization criterion.
Furthermore, the results of the simulation study of Chapter 4 indicate that the predictive deviance regarding complete cases per object with respect to the cross-validation splits has to be preferred to the predictive deviance using individual data points for splitting.

Chapter 5 extends the method proposed in Chapter 3, where lasso-type penalties are treated, by an incorporation of frailty effects. Thereby, it is controlled for unobserved heterogeneity since ignoring unobserved heterogeneity may lead to biased estimates. Hence, in the context of survival models for discrete duration time, complex penalty terms are combined with random effects. Moreover, time-varying coefficients are regarded in the linear predictor. The incorporation of these different issues may lead to very complex and difficult data situations. The proposed method is even able to yield stable estimates for those cases, whereas existing methods provided by the functions `gam` and `gamm` of the R add-on package `mgcv` (Wood, 2006) or the function `glmmPQL` of the R add-on package `MASS` (Venables and Ripley, 2002), typically fail. However, it has to be mentioned that the computation of `fpendsm` is very time-consuming. In analogy to Chapter 3, a further challenge is the optimization of the estimation algorithm with the aim to be more efficient. A possible alternative algorithm might be based on FISTA that is used in Chapter 6. It belongs to the class of proximal gradient methods in which only the log-likelihood and its gradient, but no higher-order derivatives are used.

Penalized competing risks models for discrete duration time are proposed in Chapter 6. Competing risks models are considered when more than one terminating event is of interest. Discrete competing risks models can be embedded into the framework of multinomial regression models. Due to the large amount of parameters that arise with the use of this model type, a penalization technique for discrete-time competing risks models is proposed. This penalization technique is based on the CATS lasso, introduced by Tutz et al. (2012). In competing risks models for discrete duration time, the cause-specific hazard rates are of main interest. When modeling these cause-specific hazard rates, several coefficients for each explanatory variable are existent forming their own group. The proposed penalization method enables the simultaneous shrinkage of parameters belonging to such a group. A parameter group even can be completely removed from the model. Hence, true variable selection is performed. Moreover, the proposed method allows that parameters representing the cause-specific baseline hazards vary over time. In order to avoid that adjacent parameters of the baseline effects have completely different values, a further penalty term

is incorporated steering the smoothness of the baseline effects.
Another interesting question for future research would be the incorporation of time-varying covariate effects. By expanding the time-varying coefficients in basis functions, smooth curves over time are obtained. This can be carried out, for example, by using equally spaced B-splines. To control for unobserved heterogeneity, a further advancement might be the inclusion of frailty effects in analogy to Chapter 5. By accounting for these two issues, the linear predictor is given by

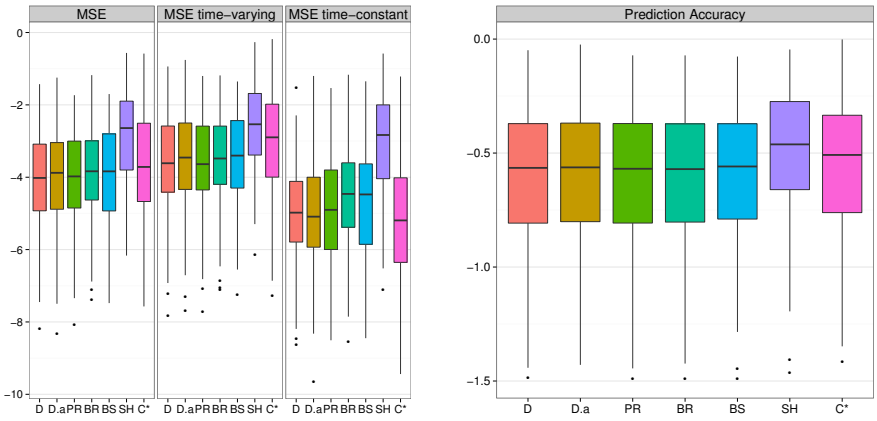
$$\eta_{itr} = b_{ir} + \beta_{0tr} + \sum_j x_{itj}\beta_{jtr} = b_{ir} + \beta_{0tr} + \sum_j \tilde{x}_{itj}\alpha_{jr},$$

with $\beta_{jtr} = \sum_{m=1}^{m_{jr}} \alpha_{jm}^{(r)} B_m(t)$ and b_{ir} is considered to be a individual-specific random effect that is appropriate distributed. Furthermore, the frailties are multivariate, with a separate component for each competing risk. This leads to a frailty vector for each individual. Note, that $\tilde{\boldsymbol{x}}_{itj}$ differs from x_{itj}. The $\tilde{\boldsymbol{x}}_{itj}$ contain the design of the interaction of the according covariates and the evaluations of the appropriate basis functions. The incorporation of B-splines requires a modification of the penalty term. Moreover, a weighting has to be introduced that accounts for smoothing and selection of the time-varying covariate effects. If frailties are considered in the linear predictor, the estimation algorithm has to be completely revised to maximize the marginal log-likelihood.

Finally, this thesis provides an overview of the benefit of using penalization techniques for survival models for discrete duration time. However, the mentioned ideas for extensions show that there is still need for further research in this field.

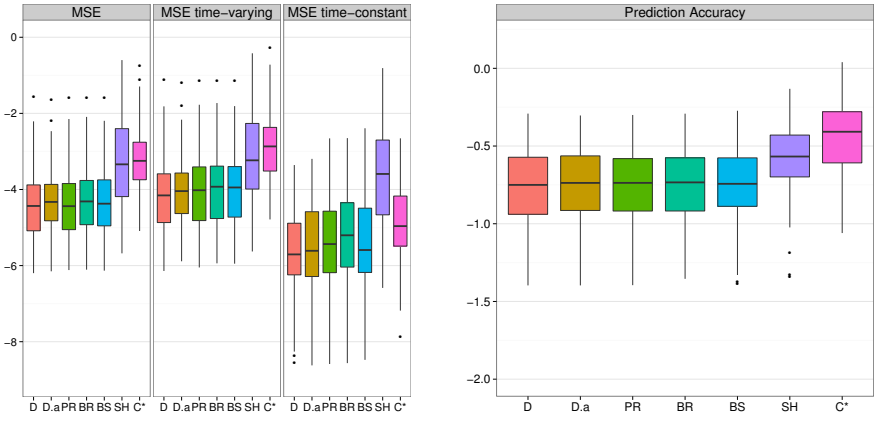
A. Appendix

A.1. Additional Figures for Section 4.3.2



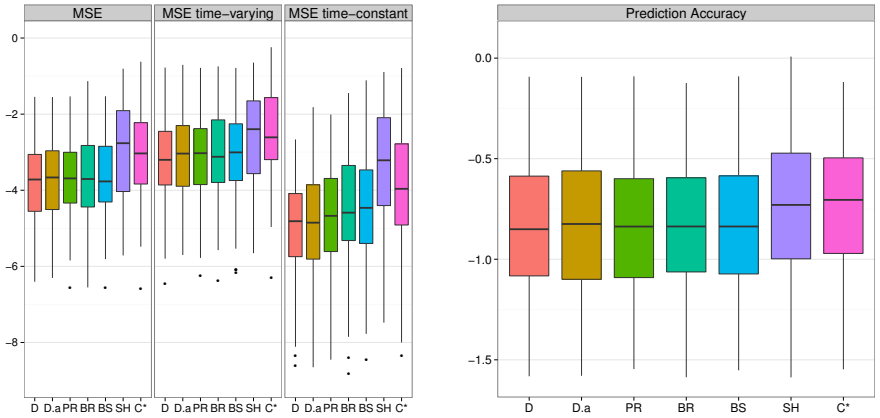
(a) Mean squared errors **(b)** Predictive accuracy

Figure A.1. Boxplots of the mean squared errors and predictive accuracy for setting 2 using a regression censoring model for IPCW estimators



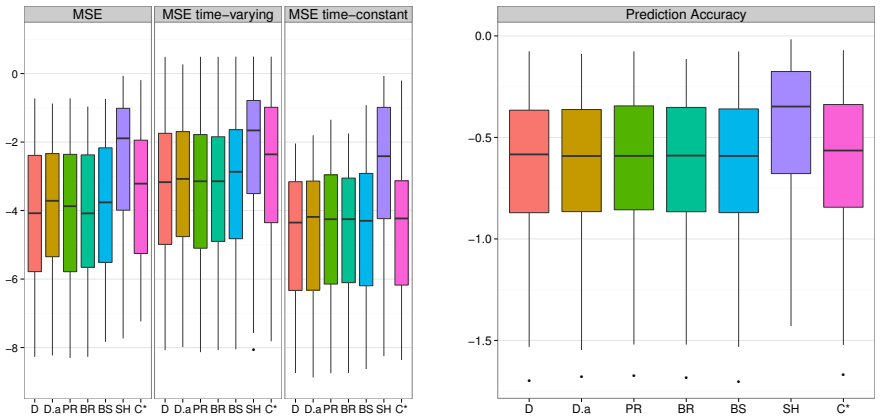
(a) Mean squared errors **(b)** Predictive accuracy

Figure A.2. Boxplots of the mean squared errors and predictive accuracy for setting 3 using a regression censoring model for IPCW estimators



(a) Mean squared errors **(b)** Predictive accuracy

Figure A.3. Boxplots of the mean squared errors and predictive accuracy for setting 4 using a marginal regression model for IPCW estimators



(a) Mean squared errors **(b)** Predictive accuracy

Figure A.4. Boxplots of the mean squared errors and predictive accuracy for setting 5 using a marginal regression model for IPCW estimators

A.2. Laplace Approximation

Based on Tutz (2012) and Groll (2011), in this section the numerical integration by means of the Laplace approximation is summarized. The Laplace approximation provides an approximation for integrals of the form $\int e^{nl(\theta)}d\theta$ when n is large (e.g. De Bruijn, 1981). For unidimensional θ it holds

$$\int e^{nl(\theta)}d\theta \approx \exp(nl(\hat{\theta}))\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{2\pi},$$

where $\hat{\theta}$ is the unique maximum of $l(\theta)$ and $\hat{\sigma}^2 = 1/(\partial^2 l(\hat{\theta})/\partial\theta^2)$. This result can be obtained by using the Taylor approximation of second order

$$l(\theta) \approx l(\hat{\theta}) + \frac{\partial l(\hat{\theta})}{\partial\theta}(\theta-\hat{\theta}) + \frac{1}{2}\frac{\partial^2 l(\hat{\theta})}{\partial\theta^2}(\theta-\hat{\theta})^2,$$

resulting in (using $\frac{\partial l(\hat{\theta})}{\partial\theta} = 0$, as $\hat{\theta}$ is the unique maximum of $l(\theta)$)

$$\begin{aligned}
\int e^{nl(\theta)}d\theta &\approx \int \exp\left(nl(\hat{\theta}) + n\frac{\partial l(\hat{\theta})}{\partial\theta}(\theta-\hat{\theta}) + \frac{1}{2}n\frac{\partial^2 l(\hat{\theta})}{\partial\theta^2}(\theta-\hat{\theta})^2\right)d\theta \\
&= \exp(nl(\hat{\theta}))\int \exp\left(\frac{1}{2}n\frac{\partial^2 l(\hat{\theta})}{\partial\theta^2}(\theta-\hat{\theta})^2\right)d\theta \\
&= \exp(nl(\hat{\theta}))\int \exp\left(-\frac{1}{2}\frac{(\theta-\hat{\theta})^2}{-\frac{1}{n}\frac{1}{\frac{\partial^2 l(\hat{\theta})}{\partial\theta^2}}}\right)d\theta \\
&= \exp(nl(\hat{\theta}))\int \exp\left(-\frac{1}{2}\frac{(\theta-\hat{\theta})^2}{\frac{\hat{\sigma}^2}{n}}\right)d\theta \\
&= \exp(nl(\hat{\theta}))\int \frac{\sqrt{2\pi}\hat{\sigma}}{\sqrt{n}}\frac{\sqrt{n}}{\sqrt{2\pi}\hat{\sigma}}\exp\left(-\frac{1}{2}\frac{(\theta-\hat{\theta})^2}{\frac{\hat{\sigma}^2}{n}}\right)d\theta \\
&= \exp(nl(\hat{\theta}))\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{2\pi}\int \frac{1}{\sqrt{2\pi}\frac{\hat{\sigma}}{\sqrt{n}}}\exp\left(-\frac{1}{2}\frac{(\theta-\hat{\theta})^2}{\frac{\hat{\sigma}^2}{n}}\right)d\theta \\
&= \exp(nl(\hat{\theta}))\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{2\pi}.
\end{aligned}$$

By substituting $g(\theta) = e^{nl(\theta)}$ it follows

$$\int g(\theta)d\theta \approx g(\hat{\theta})\hat{\sigma}_g\sqrt{2\pi}, \tag{A.1}$$

where $\hat{\sigma}_g = \frac{\hat{\sigma}}{n} = \left(-\frac{\partial^2 \log g(\hat{\theta})}{\partial \theta^2}\right)^{-1}$ and $\hat{\theta}$ is the unique maximum of $g(\theta)$. The i-th contribution to the marginal likelihood for a discrete frailty survival model is given by

$$l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_b^2) = \log\left(\int f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i) p(b_i; \sigma_b^2) db_i\right).$$

For normally distributed $b_i \sim \mathcal{N}(0, \sigma_b^2)$ it follows

$$l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_b^2) = \log\left\{\frac{1}{\sigma\sqrt{2\pi}} \int \exp\left[\log(f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i)) - \frac{1}{2}\left(\frac{b_i}{\sigma}\right)^2\right] db_i\right\}.$$

With $\kappa_\beta(b_i) = -\log(f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i)) - \frac{1}{2}\left(\frac{b_i}{\sigma}\right)^2$ the integrand is $\exp(-\kappa_\beta(b_i))$ and univariate Laplace approximation (Equation A.1) yields

$$l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_b^2) \approx \log\left\{\exp(-\kappa_\beta(\tilde{b}_i))\frac{\sigma}{n}\right\},$$

where $\tilde{b}_i$ minimizes $\kappa_\beta(b_i)$. Further simplification yields

$$\begin{aligned} l_i(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_b^2) &\approx -\kappa_\beta(\tilde{b}_i) + \log\frac{\sigma}{n} \\ &= \log(f(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i)) - \frac{1}{2}\left(\frac{b_i}{\sigma}\right)^2 + \log\frac{\sigma}{n}. \end{aligned}$$

In addition, Breslow and Clayton (1993) ignore the last term. Since $\tilde{b}_i$ minimizes $\kappa_\beta(b_i)$, it is also defined as the maximum of the i-th penalized log-likelihood

$$l_i^{app}(\boldsymbol{\omega}) = \log f(\boldsymbol{y}_i|\boldsymbol{\omega}) - \frac{\sigma_b^2}{2} b_i^2.$$

A.3. Inversion of Pseudo Fisher Matrix

To keep the notation simple, the argument σ_b^2 is omitted in the following, for example writing $l(\boldsymbol{\omega}_\alpha)$ instead of $l(\boldsymbol{\omega}_\alpha, \sigma_b^2)$. Furthermore, the vector $\tilde{\boldsymbol{\beta}}$ represents the vector $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. According to Tutz (2012), the inversion of the penalized pseudo-Fisher matrix $\boldsymbol{F}_p(\boldsymbol{\omega}_\alpha)$ can be simplified by partitioning of the matrix. Hence, the used partitioning is given by

$$\boldsymbol{F}_p(\boldsymbol{\omega}_\alpha) = \begin{bmatrix} \boldsymbol{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} & F_{\tilde{\boldsymbol{\beta}}1} & F_{\tilde{\boldsymbol{\beta}}2} & \cdots & F_{\tilde{\boldsymbol{\beta}}n} \\ F_{1\tilde{\boldsymbol{\beta}}} & F_{11} & & & 0 \\ F_{2\tilde{\boldsymbol{\beta}}} & & F_{22} & & \\ \vdots & & & \ddots & \\ F_{n\tilde{\boldsymbol{\beta}}} & 0 & & & F_{nn} \end{bmatrix}$$

with

$$\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} = -E\left(\frac{\partial^2 l(\boldsymbol{\omega}_\alpha)}{\partial\tilde{\boldsymbol{\beta}}\partial\tilde{\boldsymbol{\beta}}^T}\right),$$
$$\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}i} = \boldsymbol{F}_{i\tilde{\boldsymbol{\beta}}} = -E\left(\frac{\partial^2 l(\boldsymbol{\omega}_\alpha)}{\partial\tilde{\boldsymbol{\beta}}\partial b_i}\right),$$
$$\boldsymbol{F}_{ii} = -E\left(\frac{\partial^2 l(\boldsymbol{\omega}_\alpha)}{\partial b_i\partial b_i}\right).$$

Note, that apart from the Matrix $\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}$ the remaining components are scalar quantities due to the fact that only random intercepts are incorporated in the model. By using standard formulas for inverting partitioned matrices (see e.g. Magnus and Neudecker, 2007), the inverse can be easily computed by

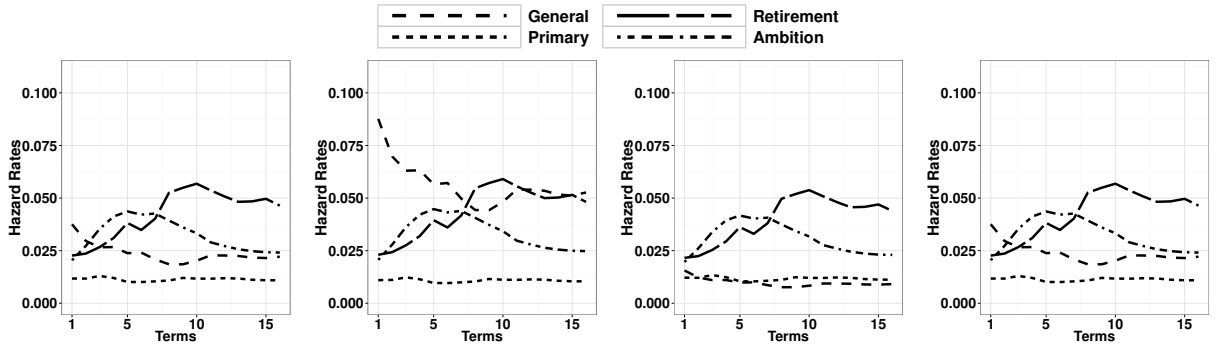
$$\boldsymbol{F}_p^{-1}(\boldsymbol{\omega}_\alpha) = \begin{bmatrix} \boldsymbol{V}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} & V_{\tilde{\boldsymbol{\beta}}1} & V_{\tilde{\boldsymbol{\beta}}2} & \cdots & V_{\tilde{\boldsymbol{\beta}}n} \\ V_{1\tilde{\boldsymbol{\beta}}} & V_{11} & V_{12} & \cdots & V_{1n} \\ V_{2\tilde{\boldsymbol{\beta}}} & V_{21} & V_{22} & \cdots & V_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ V_{n\tilde{\boldsymbol{\beta}}} & V_{n1} & V_{n2} & \cdots & V_{nn}, \end{bmatrix}$$

with

$$\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} = (\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} - \sum_{i=1}^{n} F_{\tilde{\boldsymbol{\beta}}i}F_{ii}^{-1}F_{i\tilde{\boldsymbol{\beta}}})^{-1}, \quad V_{\tilde{\boldsymbol{\beta}}i} = V_{i\tilde{\boldsymbol{\beta}}} = -\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}F_{\tilde{\boldsymbol{\beta}}i}F_{ii}^{-1},$$
$$V_{ii} = F_{ii}^{-1} + F_{ii}^{-1}F_{i\tilde{\boldsymbol{\beta}}}\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}F_{\tilde{\boldsymbol{\beta}}i}F_{ii}^{-1}, \quad V_{ij} = V_{ji} = F_{ii}^{-1}F_{i\tilde{\boldsymbol{\beta}}}\boldsymbol{V}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}F_{\tilde{\boldsymbol{\beta}}j}F_{jj}^{-1}, i \neq j.$$

A.4. Additional Figures for Section 6

Congressional Careers

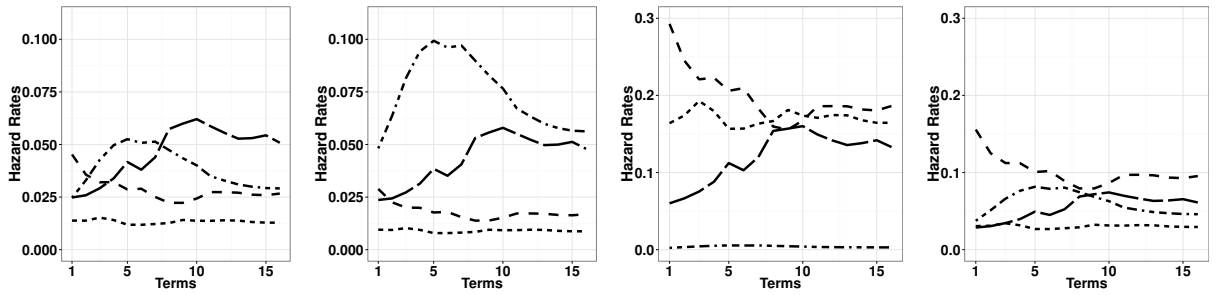


(a) Estimated rates for Age=51, Prior Margin=35, Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

(b) Estimated rates for Age=51, Prior Margin=25, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

(c) Estimated rates for Age=51, Prior Margin=45, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

(d) Estimated rates for Age=51, Prior Margin=35, no Republican, Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.



(e) Estimated rates for Age=51, Prior Margin=35, no Republican, no Leadership, open Gubernatorial seat, no open Senatorial seat, no Scandal and no Redistricting.

(f) Estimated rates for Age=51, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, open Senatorial seat, no Scandal and no Redistricting.

(g) Estimated rates for Age=51, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, Scandal and no Redistricting.

(h) Estimated rates for Age=51, Prior Margin=35, no Republican, no Leadership, no open Gubernatorial seat, no open Senatorial seat, no Scandal and Redistricting.

Figure A.5. Estimated cause-specific hazard rates over time for the Congressional career data.

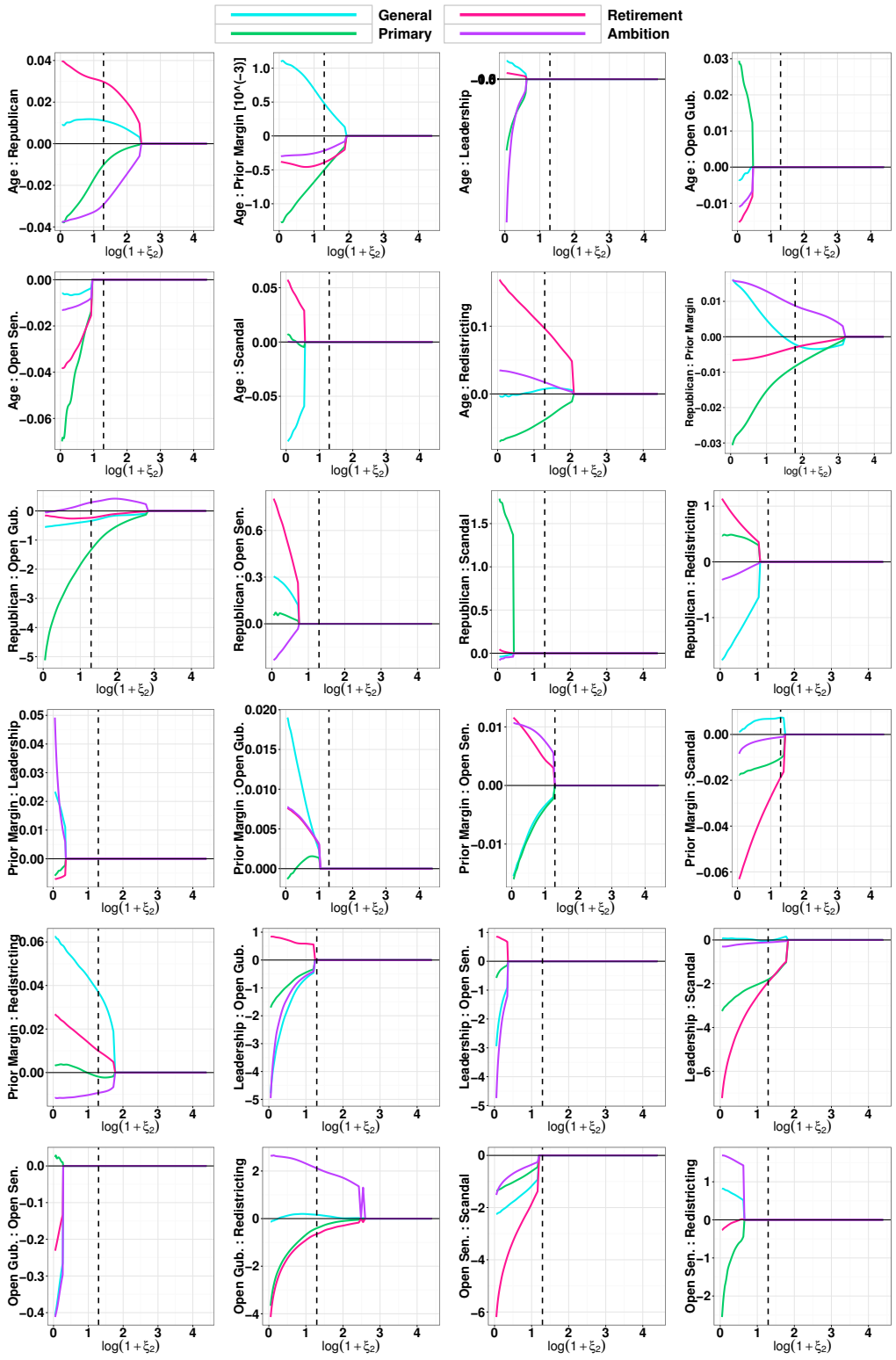
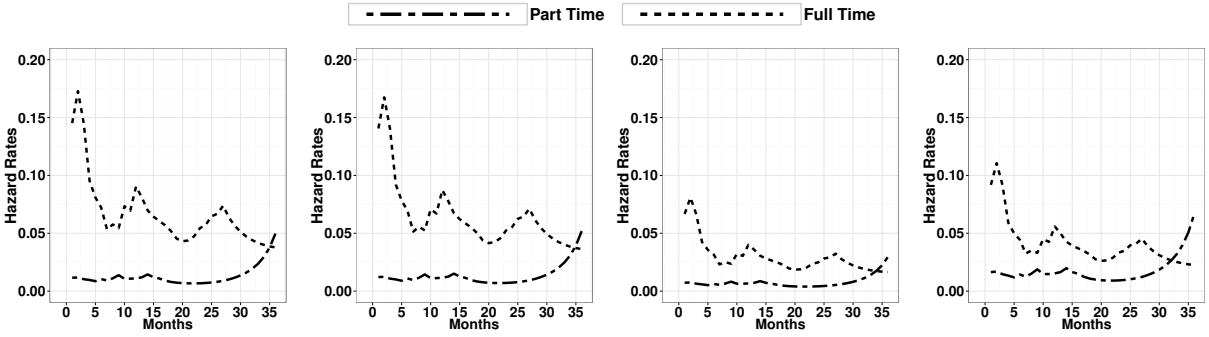


Figure A.6. Coefficients paths of the interaction effects for the Congressional career data.

Unemployment Data



(a) Estimates rates for Foreigner, Male, Middle Age, Training, No University)

(b) Estimates rates for German, Male, Young Age, Training, No University

(c) Estimates rates for German, Male, Old Age, Training, No University

(d) Estimates rates for German, Male, Middle Age, No Training, No University

Figure A.7. Estimated cause-specific hazard rates over time for the transition to part-time reemployment and full-time reemployment.

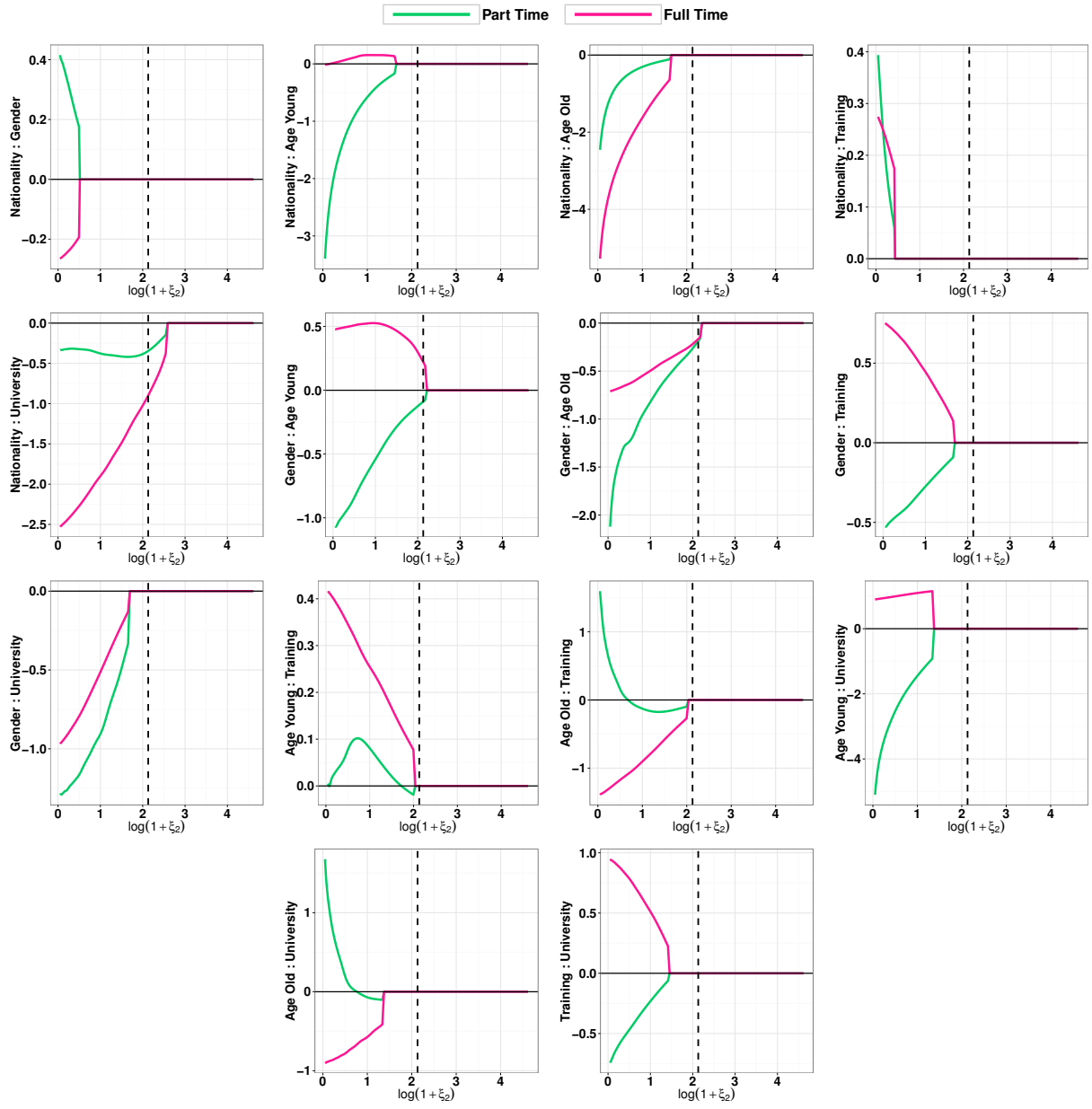


Figure A.8. Coefficients paths of the interaction effects for the unemployment data.

References

Aalen, O. O. (1975). *Statistical Inference for a Family of Counting Processes*. Ph. D. thesis, University of California, Berkeley.

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine 7*, 1121–1137.

Abedieh, J. (2013). Erwerbstätigkeit und Fertilität: Replikation und Erweiterung. Term Paper, Ludwig-Maximilians-Universität München.

Agresti, A. (2013). *Categorical Data Analysis* (3 ed.). New York: Wiley.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology 13*, 61–98.

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

Aranda-Ordaz, F. J. (1983). An extension of the proportional-hazards model for grouped data. *Biometrics 39*, 109–117.

Baker, M. and A. Melino (2000). Duration dependence and nonparametric heterogeneity: A Monte Carlo study. *Journal of Econometrics 96*, 357–393.

Bates, D., M. Maechler, B. Bolker, and S. Walker (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-6.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*, 183–202.

Beyersmann, J., A. Allignol, and M. Schumacher (2011). *Competing Risks and Multistate Models with R*. New York: Springer.

Bock, R. D. and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika 46*, 443–459.

Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B61*, 265–285.

Box-Steffensmeier, J. M. and B. S. Jones (2004). *Event History Modeling: A Guide for Social Scientists*. New York: Cambridge University Press.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics 24*, 2350–2383.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*, 89–99.

Breslow, N. and J. Crowley (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics 2*, 437–453.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Brillinger, D. R. and H. K. Preisler (1983). Maximum likelihood estimation in a latent variable problem. In T. Amemiya, S. Karlin, and T. Goodman (Eds.), *Studies in Econometrics*, pp. 31–65. New York: Academic Press.

Brown, C. C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics 31*, 863–872.

Brüderl, J., P. Preisendörfer, and R. Ziegler (1992). Survival chances of newly founded business organizations. *American Sociological Review 57*, 227–242.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science 22*, 477–505.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer.

Cai, T., M. S. Pepe, Y. Zheng, T. Lumley, and N. S. Jenny (2006). The sensitivity and specificity of markers for event times. *Biostatistics 7*, 182–197.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*, 2313–2351.

Clayton, D. G. (1996). Generalized linear mixed models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall. 275–301.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B34*, 187–220.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. London: Chapman & Hall.

De Bruijn, N. G. (1981). *Asymptotic Methods in Analysis*. New York: Dover.

Diermeier, D. and R. T. Stevenson (1999). Cabinet survival and competing risks. *American Journal of Political Science 43*, 1051–1068.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association 72*, 557–565.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*, 1–26.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

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

Elbers, C. and G. Ridder (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies 49*, 403–409.

Enberg, J., P. Gottschalk, and D. Wolf (1990). A random-effects logit model of work-welfare transitions. *Journal of Econometrics 43*, 63–75.

Fahrmeir, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data. *Biometrika 81*, 317–330.

Fahrmeir, L., T. Kneib, and S. M. Lang (2009). *Regression* (2 ed.). Berlin: Springer.

Fahrmeir, L., T. Kneib, S. M. Lang, and B. Marx (2013). *Regression*. Berlin: Springer.

Fahrmeir, L. and L. Knorr-Held (1997). Dynamic discrete-time duration models: Estimation via Markov chain Monte Carlo. *Sociological Methodology 27*, 417–452.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer.

Fahrmeir, L. and S. Wagenpfeil (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association 91*, 1584–1594.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fisher, R. A. (1919). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh 52*, 399–433.

Fleming, T. R. and D. P. Harrington (2005). *Counting Processes and Survival Analysis*. New York: Wiley.

Flinn, C. and J. Heckman (1982). Models for the analysis of labor force dynamics. In R. Basmann and G. Rhodes (Eds.), *Advances in Econometrics*, Volume 1, pp. 35–95. Greenwich: Conn.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Gaure, S., K. Røed, and T. Zhang (2007). Time and causality: A Monte Carlo assessment of the timing-of-events approach. *Journal of Econometrics 141*, 1159–1195.

Gerds, T. A. and M. Schumacher (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal 48*, 1029–1040.

Gerds, T. A. and M. Schumacher (2007). Efron-type measures of prediction error for survival analysis. *Biometrics 63*, 1283–1287.

Gertheiss, J., S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society C60*, 377–395.

Gertheiss, J. and G. Tutz (2009). Penalized regression with ordinal predictors. *International Statistical Review 77*, 345–365.

Gertheiss, J. and G. Tutz (2012). Regularization and model selection with categorical effect modifiers. *Statistica Sinica 22*, 957–982.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Gorfine, M. and L. Hsu (2011). Frailty-based competing risks model for multivariate survival data. *Biometrics 67*, 415–426.

Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine 18*, 2529–2545.

Groll, A. (2011). *Variable Selection by Regularization Methods for Generalized Linear Models*. Göttingen: Cuvillier.

Groll, A. and G. Tutz (2014). Variable selection for generalized linear mixed models by $L1$-penalized estimation. *Statistics and Computing 24*, 137–154.

Hamerle, A. and G. Tutz (1989). *Diskrete Modelle zur Analyse von Verweildauer und Lebenszeiten*. Frankfurt am Main: Campus.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B55*, 757–796.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning* (2 ed.). New York: Springer.

Heagerty, P. J., T. Lumley, and M. S. Pepe (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics 56*, 337–344.

Heagerty, P. J. and Y. Zheng (2005). Survival model predictive accuracy and ROC curves. *Biometrics 61*, 92–105.

Heckman, J. J. and B. Singer (1982). The identification problem in econometric models for duration data. In W. Hildebrand (Ed.), *Advances in Econometrics: Proceedings of World Meeting of the Econometric Society*, Cambridge. 39-77.

Heckman, J. J. and B. Singer (1984a). Econometric duration analysis. *Journal of Econometrics 24*, 63–132.

Heckman, J. J. and B. Singer (1984b). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*, 271–320.

Hess, W. and M. Persson (2012). The duration of trade revisited. *Empirical Economics 43*, 1083–1107.

Hess, W., M. Persson, S. Rubenbauer, and J. Gertheiss (2014). Using lasso-type penalties to model time-varying covariate effects in panel data regressions. *Journal of Applied Econometrics*, submitted.

Hinde, J. (1982). Compound poisson regression models. In R. Gilchrist (Ed.), *GLIM 82 International Conference on Generalised Linear Models*, pp. 109–121. New York: Springer.

Hoerl, A. and R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*, 809–822.

Hosmer, Jr., D. W., S. Lemeshow, and S. May (2011). *Applied Survival Analysis: Regression Modeling of Time to Event Data* (2nd ed.). New York: Wiley.

Hougaard, P., P. Myglegaard, and K. Borch-Johnsen (1994). Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. *Biometrics 50*, 1178–1188.

Hsieh, F. Y. (1995). A cautionary note on the analysis of extreme data with Cox regression. *The American Statistician 49*, 226–228.

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*, 111–128.

James, G. M., P. Radchenko, and J. Lv (2009). DASSO: Connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society B71*, 127–142.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics 57*, 129–136.

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*.

Jones, B. (1994). *A Longitudinal Perspective on Congressional Elections*. Ph. D. thesis, State University of New York at Stony Brook.

Kalbfleisch, J. D. and R. L. Prentice (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika 60*, 267–278.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). New York: Wiley.

Kauermann, G. and P. Khomski (2009). Full time or part time reemployment: A competing risk model with frailties and smooth effects using a penalty based approach. *Journal of Computational and Graphical Statistics 18*, 106–125.

Kent, J. T. and J. O'Quigley (1988). Measures of dependence for censored survival data. *Biometrika 75*, 525–534.

Klein, J. and M. Moeschberger (2003). *Survival Analysis: Statistical Methods for Censored and Truncated Data* (2nd ed.). New York: Springer.

Kleinbaum, D. G. and M. Klein (2013). *Survival Analysis: A Self-learning Text* (3nd ed.). New York: Springer.

Kolonko, M. (2006). *Stochastische Simulation - Grundlagen, Algorithmen und Anwendungen*. London: Vieweg+Teubner.

Korn, E. L. and R. Simon (1990). Measures of explained variation for survival data. *Statistics in Medicine 9*, 487–503.

Krishnapuram, B., L. Carin, M. A. T. Figueiredo, and A. J. Hartemink (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 957–968.

Laird, N. and D. Olivier (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association 76*, 231–240.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. New York: Cambridge University Press.

Lancaster, T. and S. Nickell (1980). The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society A143*, 141–165.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Loomis, A. L., E. N. Harvey, and G. Hobart (1937). Cerebral states during sleep, as studied by human brain potentials. *Journal of Experimental Psychology 21*, 127–144.

Magnus, J. R. and H. Neudecker (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (3nd ed.). London: Wiley.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports 50*, 163–170.

Marx, B. D. and P. H. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis 28*, 193–209.

McCall, B. P. (1994). Testing the proportional hazards assumption in the presence of unmeasured heterogeneity. *Journal of Applied Econometrics 9*, 321–334.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society B42*, 109–142.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*, 162–170.

Meier, L. (2013). *grplasso: Fitting user specified models with Group Lasso penalty.* R package version 0.4-3.

Meier, L., S. Van De Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B70*, 53–71.

Meier, L., S. van de Geer, and P. Bühlmann (2009). High-dimensional additive modeling. *The Annals of Statistics 37*, 3779–3821.

Muthén, B. and K. Masyn (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics 30*, 27–58.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika 78*, 691–692.

Nauck, B., J. Brüderl, J. Huinink, and S. Walper (2012). The German family panel (pairfam). GESIS data archive, Cologne. ZA5678 data file version 3.0.0.

Nicoletti, C. and C. Rondinelli (2010). The (mis)specification of discrete duration models with unobserved heterogeneity: A Monte Carlo study. *Journal of Econometrics 159*, 1–13.

Oelker, M.-R., J. Gertheiss, and G. Tutz (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modeling 14*, 157–177.

Oelker, M.-R. and R Development Core Team (2013). *gvcm.cat: Regularized Categorial Effects/Categorial Effect Modifiers in GLMs*. R package version 1.6.

Oelker, M.-R. and G. Tutz (2013). A general family of penalties for combining differing types of penalties in generalized structured models. Technical Report 139, Department of Statistics, Ludwig-Maximilians-Universität München.

O'Quigley, J., R. Xu, and J. Stare (2005). Explained randomness in proportional hazards models. *Statistics in Medicine 24*, 479–489.

Pepe, M. S., Y. Zheng, Y. Jin, Y. Huang, C. R. Parikh, and W. C. Levy (2008). Evaluating the ROC performance of markers for future events. *Lifetime data analysis 14*, 86–113.

Pinheiro, J. C. and D. M. Bates (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics 4*, 12–35.

Prentice, R. L. and L. A. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics 34*, 57–67.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*, 719–727.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning 5*, 197–227.

Scheike, T. H. and T. K. Jensen (1997). A discrete survival model with random effects: An application to time to pregnancy. *Biometrics 53*, 318–329.

Scheike, T. H. and Y. Sun (2007). Maximum likelihood estimation for tied survival data under Cox regression model via EM-algorithm. *Lifetime Data Analysis 13*, 399–420.

Schemper, M. and R. Henderson (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics 56*, 249–255.

Schemper, M. and J. Stare (1996). Explained variation in survival analysis. *Statistics in Medicine 15*, 1999–2012.

Schmid, M., T. Hielscher, T. Augustin, and O. Gefeller (2011). A robust alternative to the schemper–henderson estimator of prediction error. *Biometrics 67*, 524–535.

Schmid, M., H. A. Kestler, and S. Potapov (2014). On the validity of time-dependent AUC estimators. *Briefings in Bioinformatics*, accepted for publication.

Schoop, R., E. Graf, and M. Schumacher (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics 64*, 603–610.

Schröder, J. and J. Brüderl (2008). Der Effekt der Erwerbstätigkeit von Frauen auf die Fertilität: Kausalität oder Selbstselektion? *Zeitschrift für Soziologie 37*, 117–136.

Singer, J. D. and J. B. Willett (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin 110*, 268–290.

Singer, J. D. and J. B. Willett (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics 18*, 155–195.

Stiratelli, R., N. Laird, and J. H. Ware (1984). Random-effects models for serial observations with binary response. *Biometrics 40*, 961–971.

Stroud, A. H. and D. Secrest (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall.

Sueyoshi, G. T. (1995). A class of binary response models for grouped duration data. *Journal of Applied Econometrics 10*, 411–431.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending The Cox Model*. New York: Springer.

Thompson, W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics 33*, 463–470.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B58*, 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B67*, 91–108.

Tuerlinckx, F., F. Rijmen, G. Verbeke, and P. Boeck (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology 59*, 225–255.

Tuma, N. B., M. T. Hannan, and L. P. Groeneveld (1979). Dynamic analysis of event histories. *American Journal of Sociology 84*, 820–854.

Tutz, G. (1995). Competing risks models in discrete time with nominal or ordinal categories of response. *Quality and Quantity 29*, 405–420.

Tutz, G. (2012). *Regression for Categorical Data*. New York: Cambridge University Press.

Tutz, G. and H. Binder (2004). Flexible modelling of discrete failure time including time-varying smooth effects. *Statistics in Medicine 23*, 2445–2461.

Tutz, G. and G. Kauermann (2003). Generalized linear random effects models with varying coefficients. *Computational Statistics & Data Analysis 43*, 13–28.

Tutz, G., W. Pößnecker, and L. Uhlmann (2012). Variable selection in general multinomial logit models. Technical Report 126, Department of Statistics, Ludwig-Maximilians-Universität München.

Ulbricht, J. (2010). *Variable Selection in Generalized Linear Models*. Ph. D. thesis, Department of Statistics, Ludwig-Maximilians-Universität München.

Uno, H., T. Cai, L. Tian, and L. Wei (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association 102*, 527–537.

van den Berg, G. J. (2001). Duration models: Specification, identification, and multiple durations. In J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5. Amsterdam: North-Holland. 3381-3460.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*, 439–454.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis 52*, 5277–5286.

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*, 747–757.

Wei, F., J. Huang, and H. Li (2011). Variable selection and estimation in high-dimensional varying-coeffient models. *Statistica Sinica 21*, 1515–1540.

Wolfinger, R. and M. O'connell (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation 48*, 233–243.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Wood, S. (2014). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation*. R package version 1.7-29.

Wood, S. and F. Scheipl (2013). *gamm4: Generalized additive mixed models using mgcv and lme4*. R package version 0.2-2.

Xu, R. and J. O'Quigley (1999). A measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics 12*, 83–107.

Xue, L. and A. Qu (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *The Journal of Machine Learning Research 13*, 1973–1998.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B68*, 49–67.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association 86*, 79–86.

Zhang, H. and W. Lu (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika 94*, 691–703.

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

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B67*, 301–320.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 29.4.14

Stephanie Möst