Qual Quant (2016) 50:1589–1610
DOI 10.1007/s11135-015-0222-0

Variable selection for discrete competing risks models

Stephanie Möst [1] · Wolfgang Pößnecker[1] · Gerhard Tutz[1]

Published online: 2 June 2015

Abstract In competing risks models one distinguishes between several distinct target events that end duration. Since the effects of covariates are specific to the target events, the model contains a large number of parameters even when the number of predictors is not very large. Therefore, reduction of the complexity of the model, in particular by deletion of all irrelevant predictors, is of major importance. A selection procedure is proposed that aims at selection of variables rather than parameters. It is based on penalization techniques and reduces the complexity of the model more efficiently than techniques that penalize parameters separately. An algorithm is proposed that yields stable estimates. We consider reduction of complexity by variable selection in two applications, the evolution of congressional careers of members of the US congress and the duration of unemployment.

Keywords Competing risks · Event history · Discrete survival · Penalized likelihood · Regularization

1 Introduction

In survival or, more general, time-to-event regression analysis, one aims at quantifying the effects of explanatory variables on the duration time. Simple survival analysis considers one terminating event, for example death in disease studies. In many applications, however, duration can end by the occurrence of several possible events. For example, in unemployment studies the time of unemployment ends if an individual takes a full-time job, a part-time job, or retires. Modeling of the event times in the presence of multiple outcomes is usually referred to as competing risks modeling. Alternatively, one also speaks

✉ Wolfgang Pößnecker
wolfgang.poessnecker@stat.uni-muenchen.de

[1] Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 Munich, Germany

of competing events, competing causes or failures to convey that several events compete with each other to be observed.

Most of the literature for competing risks considers the case of continuous time, see, for example, Beyersmann et al. (2011), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003) and Kleinbaum and Klein (2013). If time is discretely observed, for example in months, ties may cause problems in the estimation procedure and the model might become inappropriate, especially for a low number of time periods. Competing risks models for discrete time have been considered, for example, by Han and Hausman (1990), Enberg et al. (1990) Narendranathan and Stewart (1993), Steele et al. (2004), Fahrmeir and Tutz (2001), Tutz (1995) and Fahrmeir and Wagenpfeil (1996), but without referring to the problem of variable selection.

When modeling the effects of covariates on duration one wants to identify those variables that actually have an effect. But variable selection in competing risks model differs from variable selection in models that allow for one terminating event only. While in simple survival models the impact of an explanatory variable is typically contained in one parameter, in competing risk models there is always a group of parameters that are linked to one predictor. This special feature calls for specific variable selection techniques.

Conventional variable selection methods are *forward-* and *backward-stepwise selection* (e.g. Hastie et al. 2009, pp. 57–58). However, these methods are frequently unstable and cannot be recommended. More current alternative model selection approaches use regularization techniques. In particular, penalization is nowadays widely used to regularize estimates by adding a penalty term to the log-likelihood. For suitably chosen penalties, stable and structured estimates are obtained. One of the oldest penalization methods is the *ridge* method, which uses a squared L_2-type penalty on the regression coefficients. However, it does not enforce variable selection. An alternative penalty approach that has become very popular is the *lasso* Tibshirani (1996) , which uses an L_1-type penalty on the regression coefficients and enforces variable selection. Several improvements for the lasso method have been proposed in the last decade, for example the *group lasso* proposed by Yuan and Lin (2006), which can handle categorical predictors efficiently. 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. Further extensions and alternatives are *SCAD* Fan and Li (2001), the *elastic net* Zou and Hastie (2005) and the *Dantzig selector* Candes and Tao (2007).

However, these methods are designed for models with univariate response. If used in multiple response models as the competing risks model they are not efficient in terms of variable selection because the effect of one predictor variable is represented by several parameters. Hence, there is a difference in providing variable selection and parameter selection. Variable selection is obtained only if all the parameters belonging to a variable are simultaneously set to zero. The available penalty techniques for multinomial logit models, which could be used in competing risks modeling, (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 belonging to a covariate are shrunk to zero. More recently, alternatives that enforce variable selection instead of variable select in multiple response models were proposed by Tutz (2012), Tutz et al. (2015) and Simon et al. (2013).

In the present paper, variable selection in competing risks models is obtained by extending these penalties to account for the special features of discrete survival. In Sect. 2 the framework of competing risks for discrete time is given. In Sect. 3 we introduce penalty

terms that enforce variable selection. Computational issues are treated in Sect. 4. In Sect. 5 we apply the method to two modeling problems, the congressional careers of members of the US congress and the duration of unemployment in Germany.

2 Competing risks models for discrete time

In this section a competing risk model for discrete duration time is considered. We define the model and embed maximum likelihood (ML) estimation into the framework of multivariate generalized linear models (GLMs).

2.1 The discrete competing risks model

Let time take values from $\{1, \ldots, k\}$ and let $q = k - 1$. If it results from intervals, one has k underlying intervals $[a_0, a_1), [a_1, a_2), \ldots \ldots, [a_{q-1}, a_q), [a_q, \infty)$, where typically $a_0 = 0$ is assumed and a_q denotes the final follow-up. Discrete time $T \in \{1, \ldots, k\}$ means that $T = t$ is observed if failure occurs within the interval $[a_{t-1}, a_t)$. If is intrinsically discrete, T is the original observation.

Let the distinct terminating causes be denoted by $R \in \{1, ..., m\}$. Then the *cause-specific discrete hazard function* resulting from cause or risk r is determined 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 $r = 1, ..., k$, $t = 1, ..., q$. The m hazard functions $\lambda_1(t|\boldsymbol{x}), ..., \lambda_m(t|\boldsymbol{x})$ sum up to an overall hazard function

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

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 and are given by

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

and

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

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

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

where $1 - \lambda(t|\boldsymbol{x})$ is the probability for survival.

Therefore, given an individual reaches interval $[a_{t-1}, a_t)$, a natural parametric model for the hazards is the multinomial logit model given by

$$\lambda_r(t|\boldsymbol{x}) = \frac{\exp(\beta_{0tr} + \boldsymbol{x}^T\boldsymbol{\gamma}_r)}{1 + \sum_{s=1}^{m}\exp(\beta_{0ts} + \boldsymbol{x}^T\boldsymbol{\gamma}_s)}, \tag{1}$$

where $t = 1, ..., q$, and $r = 1, ..., m$. 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 $1, ..., m$ since conditional survival corresponds to the reference category in the multinomial logit model. Conditional probability of survival is implicitly determined by

$$P(T > t|T \geq t, \boldsymbol{x}) = 1 - \sum_{r=1}^{m}\lambda_r(t|\boldsymbol{x}) = \frac{1}{1 + \sum_{s=1}^{m}\exp(\beta_{0ts} + \boldsymbol{x}^T\boldsymbol{\gamma}_s)}.$$

With $R \in \{1, ..., m\}$, where $R = 0$ denotes the conditional survival, the conditional probabilities are given by $\lambda_0(t|\boldsymbol{x}) = P(T > t|T \geq t, \boldsymbol{x}), \lambda_1(t|\boldsymbol{x}), ..., \lambda_m(t|\boldsymbol{x})$, which sum up to one.

2.2 Estimation

In this section the ML estimates for the multinomial logit model are given. Let data be given by $(t_i, r_i, \delta_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where $t_i = \min(T_i, C_i)$ is the observed discrete time, which is the minimum of survival time T_i and censoring time C_i. We always assume random censoring, that is, T_i and C_i are assumed to be independent. Moreover, $r_i \in \{1, ..., m\}$ indicates the type of the terminating event, $\boldsymbol{x}_i$ a covariate vector and δ_i denotes the censoring indicator with

$$\delta_i = \begin{cases} 1, & T_i \leq C_i, \text{ i.e. event of interest occured in interval } [a_{t_i-1}, a_{t_i}) \\ 0, & T_i > C_i, \text{ which means censoring in interval } [a_{t_i-1}, a_{t_i}). \end{cases}$$

This definition of the censoring indicator implicitly assumes that censoring occurs at the end of the interval. The likelihood contribution of the i-th observation for the model (1) is

$$L_i = P(T_i = t_i, R_i = r_i)^{\delta_i}P(T_i > t_i)^{1-\delta_i}P(C_i \geq t_i)^{\delta_i}P(C_i = t_i)^{1-\delta_i}, \tag{2}$$

where for notational simplicity, the conditioning on 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, yielding 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 $[a_{t_i-1}, a_{t_i})$. For an alternative form of the likelihood, indicators for the transition to the next period are defined by

$$y_{itr} = \begin{cases} 1, & \text{event of type } r \text{ occurs in interval } [a_{t_i-1}, a_{t_i}) \\ 0, & \text{no event of type } r \text{ occurs in interval } [a_{t_i-1}, a_{t_i}), \end{cases} \tag{3}$$

and

$$y_{it0} = \begin{cases} 0, & \text{event of type } r \text{ occurs in interval } [a_{t_i-1}, a_{t_i}) \\ 1, & \text{no event of type } r \text{ occurs in interval } [a_{t_i-1}, a_{t_i}), \end{cases} \tag{4}$$

where $i \in R_t$ and $r = 1, ..., m$. That means, the indicator variable (4) is derived from the indicator variable (3) by $y_{it0} = 1 - y_{it1} - ... - y_{itm}$. These indicator variables are gathered in the vector $\boldsymbol{y}_{it}^T = (y_{it0}, y_{it1}, ..., y_{itm})$ denoting the response vector of object i, $i = 1, ..., n$, $t = 1, ..., t_i$. By means of the indicator variables (3) and (4) the likelihood contribution of the i-th observation is given by

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

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. Given that an object reaches interval $[a_{t-1}, a_t)$, the response is multinomially distributed with $\boldsymbol{y}_{it}^T = (y_{it0}, y_{it1}, ..., y_{itm}) \sim \mathcal{M}\big(1, (1 - \lambda(t|\boldsymbol{x}_i), \lambda_1(t|\boldsymbol{x}_i), ..., \lambda_m(t|\boldsymbol{x}_i))\big)$. Therefore, the likelihood is that of the multicategorical model

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

with $\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_i^T \boldsymbol{\gamma}_r$. Accordingly, the total log-likelihood is given by

$$\begin{aligned} l &= \sum_{i=1}^{n} \sum_{t=1}^{t_i} \left(\sum_{r=1}^{m} y_{itr} \log \lambda_r(t|\boldsymbol{x}_i) + y_{it0} \log \left(1 - \sum_{r=1}^{m} \lambda_r(t|\boldsymbol{x}_i)\right) \right) \\ &= \sum_{t=1}^{q} \sum_{i \in R_t} \left(\sum_{r=1}^{m} y_{itr} \log \lambda_r(t|\boldsymbol{x}_i) + y_{it0} \log \left(1 - \sum_{r=1}^{m} \lambda_r(t|\boldsymbol{x}_i)\right) \right). \end{aligned} \tag{5}$$

Hence, ML estimates can be easily computed by using statistical software for multinomial regression models after construction of an appropriate design matrix, which we describe in the following. Let $\mathbb{1}_t = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ be a vector of length q with 1 in t-th position and zeros otherwise and let $\tilde{\boldsymbol{x}}_{it}^T = (\mathbb{1}_t^T, \boldsymbol{x}_i^T)$ denote a design vector that includes the baseline effect for time period t and the covariate vector $\boldsymbol{x}_i$. With corresponding parameter vectors $\tilde{\boldsymbol{\gamma}}_r^T = (\beta_{01r}, \ldots, \beta_{0qr}, \boldsymbol{\gamma}_r^T) = (\boldsymbol{\beta}_{0r}^T, \boldsymbol{\gamma}_r^T)$, one obtains for the linear predictors $\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_i^T \boldsymbol{\gamma}_r$ the closed form

$$\boldsymbol{\eta}_{it} = (\eta_{it1}, ..., \eta_{itm})^T = (\tilde{\boldsymbol{x}}_{it}^T \tilde{\boldsymbol{\gamma}}_1, ..., \tilde{\boldsymbol{x}}_{it}^T \tilde{\boldsymbol{\gamma}}_m)^T.$$

In compact matrix notation, the matrix of linear predictors for all artificial data points that belong to one real observation is then given by

$$\eta_i = \begin{bmatrix} \eta_{i1}^T \\ \vdots \\ \eta_{it_i}^T \end{bmatrix}_{t_i \times m} = \tilde{X}_i \tilde{\Gamma} = \begin{bmatrix} \tilde{x}_{i1}^T \\ \vdots \\ \tilde{x}_{it_i}^T \end{bmatrix}_{t_i \times (q+p)} [\,\tilde{\gamma}_1 | \cdots | \tilde{\gamma}_m\,]_{(q+p) \times m}.$$

Finally, for the whole dataset, one obtains with $\boldsymbol{\eta}^T = (\boldsymbol{\eta}_1^T | \cdots | \boldsymbol{\eta}_n^T)$ and $\tilde{\boldsymbol{X}}^T = (\tilde{\boldsymbol{X}}_1^T | \cdots | \tilde{\boldsymbol{X}}_n^T)$ the form $\boldsymbol{\eta} = \tilde{\boldsymbol{X}} \tilde{\boldsymbol{\Gamma}}$, so that unpenalized estimation and inference for our model is readily available via standard methods for multivariate GLMs.

Some applications, for example the congressional careers study in Sect. 5.1, involve covariates that vary over time. This case of time-varying predictor variables is easily handled within our framework because all affected formulas in this paper remain valid if '$\boldsymbol{x}_i$' is simply replaced by '$\boldsymbol{x}_{it}$'.

3 Penalization

3.1 Choice of the penalty term

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

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

where $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$ and $\boldsymbol{\gamma}_r^T = (\gamma_{r1}, ..., \gamma_{rp})$. Because there are m competing risks, each covariate adds m parameters, which increases the need for effective variable selection. Since the baseline hazard parameters β_{0tr} in addition vary over time, the number of parameters can be very large, rendering simple ML estimators unstable and difficult to interpret. To obtain a sparse representation and in particular variable selection, we consider penalized ML estimation, which uses a penalty term in the log-likelihood (5), yielding the penalized log-likelihood

$$l_{\zeta_1,\zeta_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) - J_{\zeta_1,\zeta_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}), \tag{6}$$

where $\boldsymbol{\beta}_0^T = (\boldsymbol{\beta}_{01}^T, \ldots, \boldsymbol{\beta}_{0m}^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\gamma}_1^T, ..., \boldsymbol{\gamma}_m^T)$ collect all the corresponding parameters. The first term, $l(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$, denotes the ordinary log-likelihood, written as a function of the model parameters, whereas the second term, $J_{\zeta_1,\zeta_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$, stands for a penalty term that depends on scalar tuning parameters ζ_1 and ζ_2. The choice of the penalty $J_{\zeta_1,\zeta_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$ determines the properties of the penalized estimator.

The penalty on the baseline parameters $\boldsymbol{\beta}_0$ must ensure that the estimated hazard rates are sufficiently smooth over time. Concerning the covariate effects $\boldsymbol{\gamma}$, our goal is variable selection, that is, finding those covariates that are influential at predicting hazard rates or survival. It immediately follows from (1) that the influence of a variable, say x_j, is only removed from the model if all of its effects $\boldsymbol{\gamma}_{\bullet j}^T = (\gamma_{1j}, ..., \gamma_{mj})$ are set to zero simultaneously. For example, if we have $\hat{\gamma}_{1j} = 0$ and $\hat{\gamma}_{2j} \neq 0$, then $\hat{\lambda}_1(t|\boldsymbol{x})$ would still be influenced by x_j.

A penalty that enforces such a structured and thus effective variable selection and that smooths the baseline hazards over time is given by

$$\begin{aligned} J_{\zeta_1,\zeta_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) &= \zeta_1 \sum_{r=1}^{m} \sum_{t=2}^{q} (\beta_{0tr} - \beta_{0,t-1,r})^2 + \zeta_2 \sum_{j=1}^{p} \phi_j \left\| \boldsymbol{\gamma}_{\bullet j} \right\| \\ &= \zeta_1 J_1(\boldsymbol{\beta}_0) + \zeta_2 J_2(\boldsymbol{\gamma}), \end{aligned} \tag{7}$$

where $\|\boldsymbol{u}\| = \|\boldsymbol{u}\|_2 = \sqrt{\boldsymbol{u}^T \boldsymbol{u}}$ denotes the L_2-norm and $\phi_j = \sqrt{m}$ is a weight that adjusts the penalty level on parameter vectors $\boldsymbol{\gamma}_{\bullet j}$ for their dimension. The first term in (7) uses that time intervals are 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 smooths 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 in the group lasso the grouping refers to the parameters that are linked to the dummies of a categorical predictor within a univariate regression model, while in the present model grouping arises from the multivariate model structure. The strength of the penalty terms is determined by the tuning parameters ζ_1 and ζ_2. Without a penalty, that is with $\zeta_1 = \zeta_2 = 0$, ordinary ML-estimation is obtained.

3.2 Complexity reduction by incorporating splines

Even though penalty (7) smooths the cause-specific baseline coefficients $\boldsymbol{\beta}_0$ and hence reduces their effective dimensionality, one might want to reduce their complexity *a priori*, for example if the number of time periods q is very large. To simplify the baseline effects, they can be expanded in basis functions, for example in an equidistant, low-rank B-spline basis, resulting in

$$\beta_{0tr} = \sum_{s=1}^{d_r} \alpha_{0sr} B_s(t)$$

with $d_r < q$. The incorporation of B-splines yields more parsimonious models, but requires a modification of the penalty term which is given by

$$J_{\zeta_1,\zeta_2}(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}) = \zeta_1 \sum_{r=1}^{m} \sum_{s=2}^{d_r} (\alpha_{0sr} - \alpha_{0,s-1,r})^2 + \zeta_2 \sum_{j=1}^{p} \phi_j \left\| \boldsymbol{\gamma}_{\bullet j} \right\|. \tag{8}$$

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

3.3 Adaptive penalties

Since penalization necessarily introduces bias (which grows with ζ_2), the choice of this tuning parameter involves a tradeoff between 'sharp' variable selection and unbiasedness. In the light of this conflict, the common penalty level ζ_2 in the penalties (7) and (8) cannot be an optimal choice. As was shown by Zou (2006) for the simple lasso and by Wang and Leng (2008) for the group lasso the methods are inconsistent if used with a common penalty parameter. The proposed remedy are so-called adaptive weights, which for the penalties (7) or (8) are obtained by replacing the weights ϕ_j by

$$\phi_j^a = \frac{\sqrt{m}}{||\hat{\boldsymbol{\gamma}}_{\bullet j}^{\text{Init}}||}, \tag{9}$$

where $\hat{\boldsymbol{\gamma}}_{\bullet j}^{\text{Init}}$ denotes an appropriate initial estimate. For our model, $\hat{\boldsymbol{\gamma}}_{\bullet j}^{\text{Init}}$ is the penalized estimate that results from application of penalties (7) or (8) with $\zeta_2 = 0$. Thus, the initial estimate uses unpenalized covariate effects, but an active smoothing penalty on the baseline effects.

The intuition behind this weighting procedure is rather straightforward. Assuming that all predictors are centered around zero and standardized to a common variance, the norm of unpenalized estimates for the parameter groups is rather large if they belong to strong predictors and small otherwise. Consequently, the corresponding penalization is small/large for strong/weak predictors, respectively. Zou (2006) and Wang and Leng (2008) prove that these penalized estimators provide consistent variable selection if used with adaptive weights. In Tutz et al. (2015), the improved performance of adaptive penalties was empirically confirmed for the multinomial logit model.

4 Computational issues

In the following, some details regarding the computation of numerical estimates are described. First, details of the estimation approach itself are outlined, then the tuning parameter selection for discrete competing risk models is presented.

4.1 Numerical estimates

To estimate the parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}$, the penalized log-likelihood $l_{\zeta_1,\zeta_2}(\boldsymbol{\beta}_0, \boldsymbol{\gamma})$ from (6) has to be maximized, which can also be formulated as

$$(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta}_0, \boldsymbol{\gamma}}{\operatorname{argmin}} \Big(- l(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) + \zeta_1 J_1(\boldsymbol{\beta}_0) + \zeta_2 J_2(\boldsymbol{\gamma}) \Big). \tag{10}$$

Our algorithm for solving (10) is based on proximal gradient algorithms, for an overview, see Parikh and Boyd (2013). The key building block is the so-called proximal operator, which for a generic search point $\boldsymbol{v}$ and a generic penalty $\zeta \boldsymbol{J}(\cdot)$ is defined as

$$\mathbf{Prox}_{\zeta J}(\boldsymbol{v}) = \underset{\boldsymbol{u}}{\operatorname{argmin}} \Big(||\boldsymbol{u} - \boldsymbol{v}||_2^2 + \zeta \boldsymbol{J}(\boldsymbol{u}) \Big). \tag{11}$$

For $s = 0, 1, 2, \ldots$ until convergence, the proximal gradient iterations are given by

$$\hat{\boldsymbol{\beta}}_0^{(s+1)} = \mathbf{Prox}_{\zeta_1 / v^{(s)} \cdot J_1} \left(\boldsymbol{v}^{(s)} := \hat{\boldsymbol{\beta}}_0^{(s)} + \frac{1}{v^{(s)}} \cdot \frac{\partial l(\hat{\boldsymbol{\beta}}_0^{(s)}, \hat{\boldsymbol{\gamma}}^{(s)})}{\partial \boldsymbol{\beta}_0} \right) \tag{12}$$

and

$$\hat{\boldsymbol{\gamma}}^{(s+1)} = \mathbf{Prox}_{\zeta_2 / v^{(s)} \cdot J_2} \left(\boldsymbol{w}^{(s)} := \hat{\boldsymbol{\gamma}}^{(s)} + \frac{1}{v^{(s)}} \cdot \frac{\partial l(\hat{\boldsymbol{\beta}}_0^{(s)}, \hat{\boldsymbol{\gamma}}^{(s)})}{\partial \boldsymbol{\gamma}} \right), \tag{13}$$

where $v^{(s)} > 0$ is an inverse stepsize parameter. In (12) and (13), it was exploited that both the overall penalty term J_{ζ_1,ζ_2} (see (7)) and the L_2^2-term in (11) can be decomposed into nonoverlapping parts that only contain either $\boldsymbol{\beta}_0$ or $\boldsymbol{\gamma}$. The search points $\boldsymbol{v}$ and $\boldsymbol{w}$ for $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}$, respectively, are obtained from a first order approximation of the log-likelihood term in (10) and can be considered a one-step approximation of the ML estimator, based on the current solution. Applying the proximal operator to these search points incorporates the penalties and ensures solutions with structured sparsity.

Since the penalty on the γ-parameters in (7) is a groupwise L_2-norm, the solution to (13) is obtained by blockwise application of the well-known group-soft-thresholding operator. Let $J_2(\boldsymbol{\gamma}) = \sum_{j=1}^{p} \phi_j ||\boldsymbol{\gamma}_{\bullet j}|| = \sum_{j=1}^{p} J_{2j}$ and let $\boldsymbol{w}$ be partitioned like $\boldsymbol{\gamma}$. Then, one obtains with $(u)_+ = \max(u, 0)$ the analytical solution

$$\mathbf{Prox}_{\zeta_2/\nu \cdot J_{2j}}(\boldsymbol{w}_{\bullet j}) = \left(1 - \frac{\zeta_2 \phi_j / \nu}{||\boldsymbol{w}_{\bullet j}||}\right)_+ \boldsymbol{w}_{\bullet j}, \quad j = 1, \ldots, p.$$

To derive a closed solution to (12), we rewrite the penalty on the baseline parameters: $J_1(\boldsymbol{\beta}_0) = \sum_{r=1}^{m} \sum_{t=2}^{q} (\beta_{0tr} - \beta_{0,t-1,r})^2 = \sum_{r=1}^{m} J_{1r}$. Let $\boldsymbol{D}$ denote the first-order difference matrix, that is,

$$D = \begin{pmatrix} -1 & 1 & & & 0 \\ & -1 & 1 & & \\ & & \ddots & & \\ 0 & & & -1 & 1 \end{pmatrix},$$

so that we have $J_{1r} = ||\boldsymbol{D}\boldsymbol{\beta}_{0r}||_2^2$. Hence, the proximal operator in (12) only contains quadratic terms and thus, with $\boldsymbol{\Omega} = \boldsymbol{D}^T\boldsymbol{D}$ and identity matrix $\boldsymbol{I}$, admits an analytical solution:

$$\mathbf{Prox}_{\zeta_1/\nu \cdot J_{1r}}(\boldsymbol{v}_{\bullet r}) = \left(\boldsymbol{I} + \tfrac{\zeta_1}{\nu}\boldsymbol{\Omega}\right)^{-1} \boldsymbol{v}_{\bullet r}, \quad r = 1, \ldots, m. \tag{14}$$

To the best of our knowledge, formula (14) has never been explicitly given in the literature.

The steps described above are crucial and sufficient for the computation of numerical estimates with proximal gradient algorithms. However, our implementation uses an accelerated version of proximal gradient, the so-called Fast Iterative Shrinkage and Thresholding Algorithm (FISTA) of Beck and Teboulle (2009). By a technical modification to the search points in (12) and (13), FISTA achieves quadratic convergence, which is optimal within the class of first-order algorithms. The proposed algorithm is an extension of the algorithm given in Tutz et al. (2015). The latter contains only one penalty term, which is enough when modeling multinomial logit models but not when modeling survival.

Alternatives to the proposed proximal gradient algorithm are, for example, the coordinate descent approach described in Zou and Hastie (2005). However, the corresponding implementation in the package `glmnet` Friedman et al. (2010) for the statistical software `R` R Development Core Team (2014) does not support the quadratic penalty on parameter differences that is used in the first part of (7) to smooth the baseline parameters. Therefore, we settled on the proximal gradient algorithm described above, using a self-written implementation based on the `R` package `MRSP` Pößnecker (2014).

4.2 Tuning parameter selection

The tuning parameters ζ_1 and ζ_2 are chosen by k-fold cross-validation (Hastie et al. 2009) over a two-dimensional grid of possible values. However, a modification to standard cross-validation is required due to the data blow up described in Sect. 2.2.

For folds $s = 1, \ldots, k$, let $\mathcal{I}_s$ denote the index set of observations that belong to fold s and let $\hat{\lambda}_r^{(-s)}$ denote the estimate for λ_r that is based on all observations except for those in $\mathcal{I}_s$. As the criterion to be cross-validated, we use the (predictive) deviance (Hastie et al. 2009, p. 378). With the introduced notation, the cross-validated deviance is defined by

$$D_{CV} = 2 \sum_{s=1}^{K} \sum_{i \in \mathcal{I}_s} \sum_{t=1}^{t_i} \sum_{r=0}^{m} y_{itr} \log \left(\frac{y_{itr}}{\hat{\lambda}_r(t|\boldsymbol{x}_i)^{(-s)}} \right). \tag{15}$$

Hence, all $(m+1) \cdot t_i$ data points y_{itr} that belong to the same original observation i are always assigned to the same cross-validation fold. The dependence of $\hat{\lambda}_r$ (and thus D_{CV}) on ζ_1 and ζ_2 has been suppressed in the notation to improve readability.

5 Applications

In this section, the proposed penalized competing risk model with 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.

5.1 Congressional careers

The first data example deals with careers of incumbent members of the U.S. Congress. A detailed description can be found in the book of Box-Steffensmeier and Jones (2004) and in Jones (1994). It can be downloaded from the website http://psfaculty.ucdavis.edu/bsjjones/eventhistory.html. 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*). The dependent variable is defined by the transition process of a Congressman from his first election up to one of the competing events *general, primary, retirement or ambition*. 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. 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 2-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. The data set covers all election cycles from 1952 up to 1992.

Originally, up to 20 terms occurred, however, only for very few Congressmen. Hence, due to stability reasons, durations that exceed 15 terms have been aggregated. Furthermore, only complete cases, that is, observations with no missing values for any covariate, have been incorporated in the analysis. The used data set contains the career paths of 860 Congressmen. Several covariates are available as predictors for the end of careers. The covariate *age* gives the incumbent's age at each election cycle and, to improve interpretability, is centered around 51 years (sample mean: 51.26). The incumbent's margin of victory in his or her previous election is collected in the variable *priorMargin*, which is

centered around a margin of 35 (sample mean: 35.21). The covariate *redistricting* indicates if the incumbent's district was substantially redistricted. The covariate *scandal* captures if an incumbent was involved in an ethical or sexual misconduct scandal or if 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, *leadership* describes if a member is in the House leadership and/or is a chair of a standing House committee. With the exception of 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 1.

We fitted a penalized multinomial logit model with risks defined by cause 1 (*General*), 2 (*Primary*), 3 (*Retirement*) and 4 (*Ambition*). The effect of covariates in the model $\lambda_r(t|\boldsymbol{x}) = \exp(\eta_{itr})/(1+\sum_{j=1}^{4}\exp(\eta_{itj}))$ is specified by the cause-specific linear predictors $\eta_{itr} = \beta_{0tr} + \boldsymbol{x}_{it}^T\boldsymbol{\gamma}_r$. All covariates described in Table 1 are incorporated in the predictors. Moreover, we included all pairwise interactions with the exception of *Republican:Leadership*, *Leadership:Redistricting*, *Opengub:Scandal*, *Scandal:Redistricting* because too few observations of the corresponding combinations are in the data. Such a high-dimensional interaction model cannot be properly handled by unpenalized ML estimation but stable estimation and efficient variable selection is obtained by using penalization.

Since the adaptive version of the penalty yielded better cross-validation score, adaptive weights were used. Tuning parameters ζ_1 and ζ_2 were chosen on a 2-dimensional grid by 5-fold cross validation with the predictive deviance as loss criterion. The resulting tuning parameters were $\zeta_1 = 6.0$ and $\zeta_2 = 2.64$. For a fixed $\zeta_1 = 6.0$, the corresponding cross

Table 1 Description of the variables of the congressional career data

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 criminal investigation criminal investigation district was substantially redistricted
	0: no, 1: yes

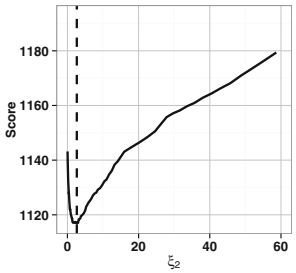

Fig. 1 Cross validation score subject to penalty parameter ζ_2 for $\zeta_1 = 6.0$ for the congressional career data

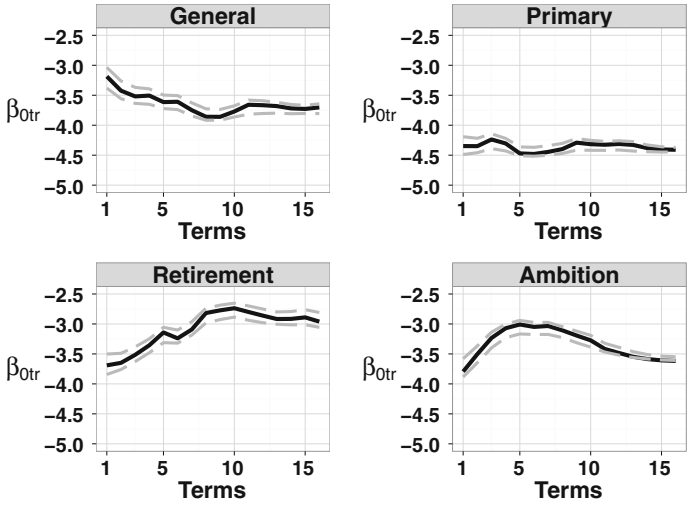

Fig. 2 Parameter estimates of the cause-specific time-varying baseline effects for the congressional careers data. *Dashed lines* represent the 95 % pointwise bootstrap interval

validation score is shown in Fig. 1, where the vertical black dashed line marks the chosen tuning parameter.

Figure 2 shows the parameter estimates for the cause-specific time-varying baseline effects. The corresponding pointwise confidence intervals, marked by light-gray dashed lines, have been estimated by a nonparametric bootstrap method as proposed by Efron (1979) with 1000 bootstrap replications of the fitted model (i.e. fixed tuning parameters across bootstrap samples). It can be seen that cause-specific baseline effects are necessary because the shapes are quite different. For retirement, the parameters are increasing over early terms and then become stable while for ambition there is an early peak at about five terms and then a decrease. Due to the penalization of adjacent coefficients $\beta_{0tr} - \beta_{0,t-1,r}$, the estimated baseline effects are rather smooth.

Parameter estimates of the covariate effects are summarized in Table 2b. It shows the ordinary ML estimates and the estimates resulting from the penalized competing risk model with their corresponding standard errors. The computation of the standard errors is based on the empirical standard deviation of the respective coefficient across 1000

Table 2 Parameter estimates for the congressional careers data

	General			Primary			Retirement			Ambition		
	ML	Pen.	SE	ML	Pen.	SE	ML	Pen.	SE	ML	Pen.	SE
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
OpenGub	0.815	0.205	0.116	0.598	0.181	0.097	0.227	0.109	0.077	0.528	0.208	0.121
OpenSen	−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

Table 2 continued

	General			Primary			Retirement			Ambition		
	ML	Pen.	SE	ML	Pen.	SE	ML	Pen.	SE	ML	Pen.	SE
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

ML Ordinary maximum likelihood estimates , *pen.* the penalized estimates, *SE* Estimated standard errors for the penalized model obtained by a bootstrap approach are given in the columns

nonparametric bootstrap samples. It is immediately seen that the penalization removes a considerable number of effects, that is, only 68 out of 128 parameters remain in the model, leading to a strong reduction of the model complexity. The selection procedure suggests that the main effects *Republican* and *Leadership* are not needed in the predictor. Moreover, a large number of interaction effects were not selected. Concerning interpretation, 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 loses a primary or general election or that he/she retires. In contrast, a scandal decreases the probability of seeking an alternative office as compared to reelection.

In Fig. 3a a selection of resulting hazard rates is depicted. It 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*. It can be seen that the probability of retirement tends to increase over early terms and then remains rather stable. The probability for seeking an alternative office as compared to reelection increases for early terms and then decreases. The hazard rates for losing either a primary or a general election are rather constant in the considered group. Figure 3b and c show the hazard rates respectively 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 larger than in the reference group.

The selection effect is visualized by coefficient paths. In Fig. 4 we show only the paths for the main effects. Each path indicates the penalized estimates subject to tuning parameter ζ_2, where the abscissa is transformed by $\log(1 + \zeta_2)$. The paths illustrate how the estimates changes towards zero for increasing ζ_2. Hence, they show the effects of covariates on the terminating events when penalization is increased. The dashed black line indicates the value of ζ_2 that was chosen via cross-validation.

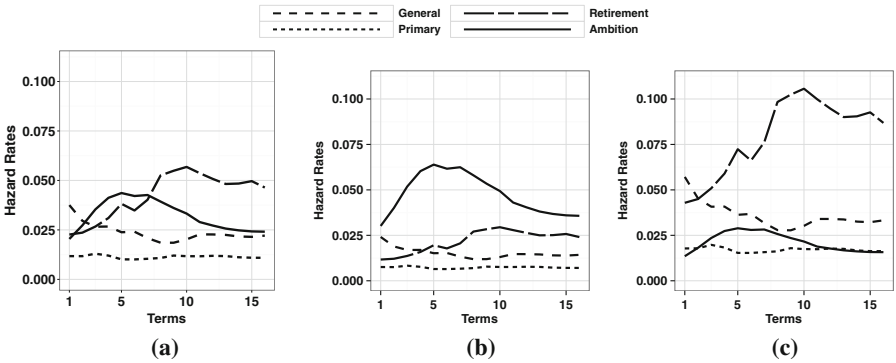

Fig. 3 Estimated cause-specific hazard rates over time for the congressional careers data. **a** Estimated rates for all predictors at reference: *age* 51, *prior margin* 35, no republican, no leadership, no open gubernatorial seat, no open senatorial seat, no scandal and no redistricting; **b** estimated 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

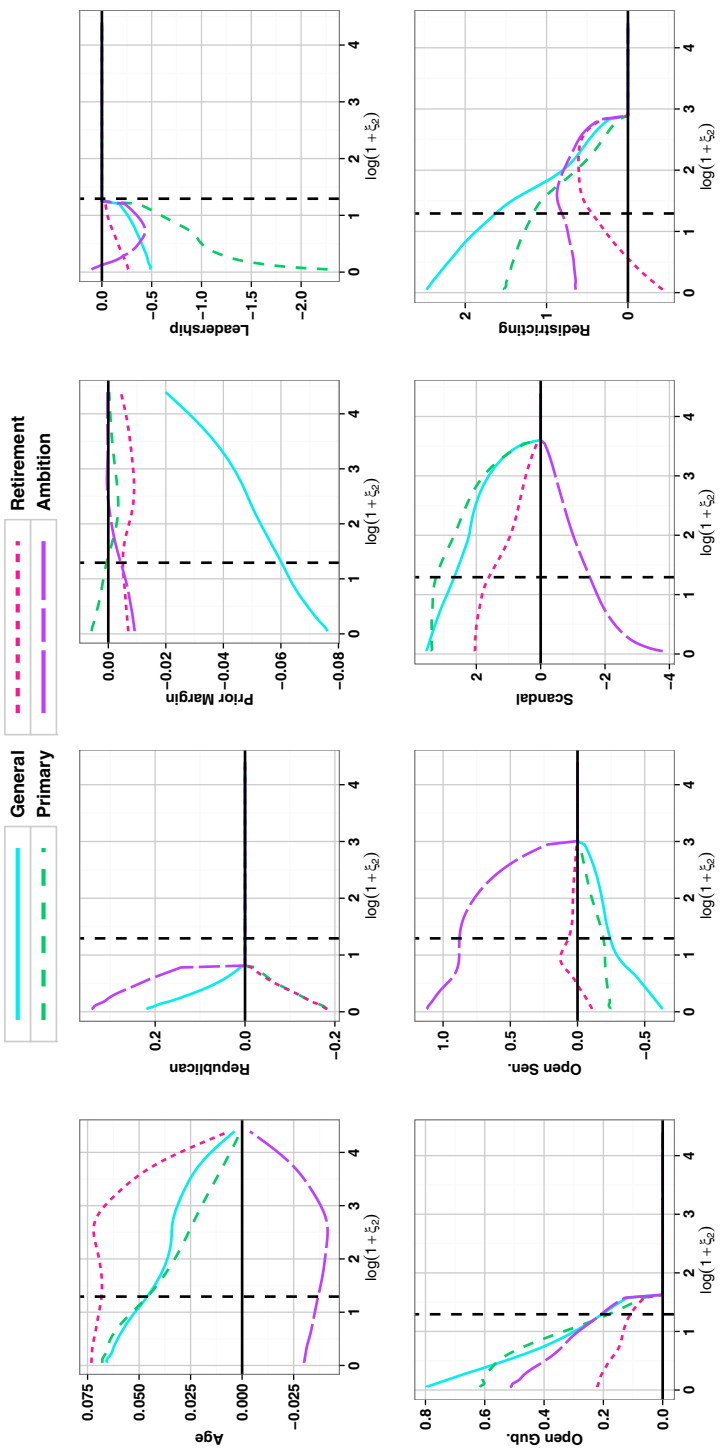

Fig. 4 Coefficients paths of the main effects for the congressional career data

5.2 Unemployment data

In this section, the proposed penalized competing risk 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 1990–2000 are considered. If more than one spell of unemployment occurred for an individual, only one spells was chosen to guarantee independence of the observations. The events that terminate an unemployment spell are *part-time* reemployment ($r = 1$) and *full-time* reemployment ($r = 2$). All other reasons for terminating unemployment are considered as censored.

The dependent variable is defined by the transition process of an individual up to one of the competing events *part-time* or *full-time* reemployment with the duration until the occurrence of one of the competing events measured in months. The maximal observation length in the data was 36 months. The available covariates, measured at the beginning of the unemployment spell, are *nationality*, *gender*, *age*, *education* and *training*. We use the publicly available version of the data that is part of the R add-on package `CompetingRiskFrailty`, which can be obtained from the CRAN archive. The explanatory variables that will be used for modeling are listed in Table 3.

The available data set consists of 500 unemployed persons. We use all the covariates described in Table 3 and included all pairwise interaction effects. The used penalty term is a version of (8a) and is given by

$$J(\boldsymbol{\beta}_0, \boldsymbol{\gamma}) = \zeta_1 \sum_{r=1}^{2} \sum_{t=2}^{36} (\beta_{0tr} - \beta_{0,t-1,r})^2 + \sqrt{2}\zeta_2 \sum_{j=1}^{20} \left\| \boldsymbol{\gamma}_{\bullet j} \right\|. \tag{16}$$

In analogy to the previous example the penalty term enforces smooth cause-specific baseline effects and variable selection of the covariate effects including the interactions. Since the adaptive version of the penalty did not show better performance we used the simpler version without weights. Tuning parameters ζ_1 and ζ_2 were chosen by 5-fold cross

Table 3 Description of the variables of the unemployment data

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

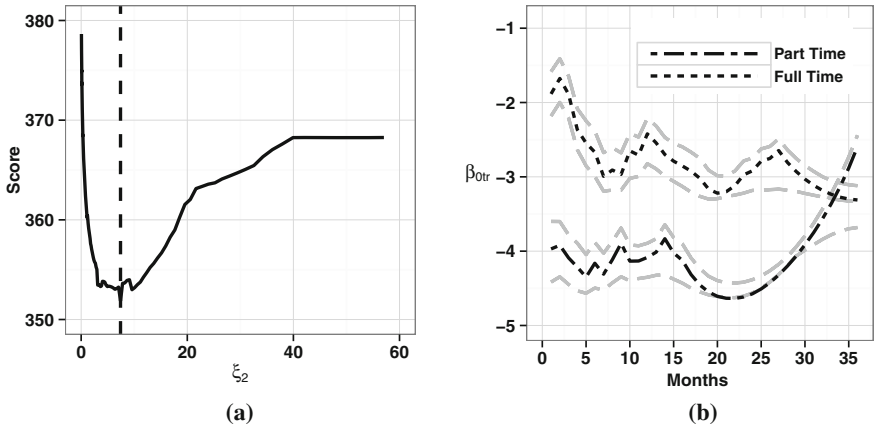

Fig. 5 *Plots* corresponding to the unemployment data. **a** Cross validation score subject to penalty parameter ζ_2 for $\zeta_1 = 1.0$, **b** parameter estimates of the cause-specific time-varying baseline effects. *Dashed lines* represent 95 % pointwise bootstrap confidence intervals

Table 4 Parameter estimates for the unemployment data

	Part-time			Full-time		
	ML	Pen.	SE	ML	Pen.	SE
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

ML Ordinary maximum likelihood estimates , *pen.* the penalized estimates, *SE* Estimated standard errors for the penalized model obtained by a bootstrap approach are given in the columns

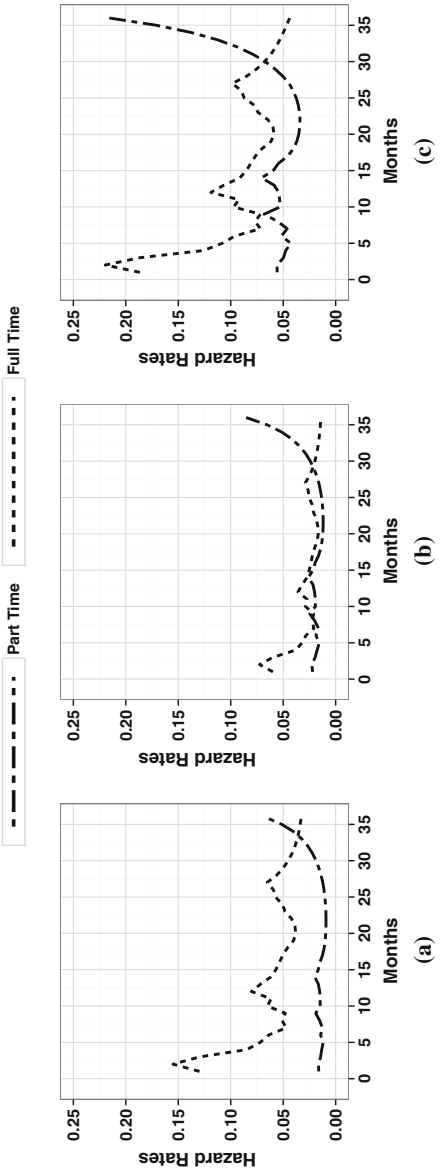

Fig. 6 Estimated cause-specific hazard rates over time for the transition to part-time reemployment and full-time reemployment. **a** Estimated rates for all predictorsat reference: German, male, middle age, training, no university, **b** estimated rates for German, female, middle age, training, nouniversity, **c** estimated rates for German, male, middle age, training, university

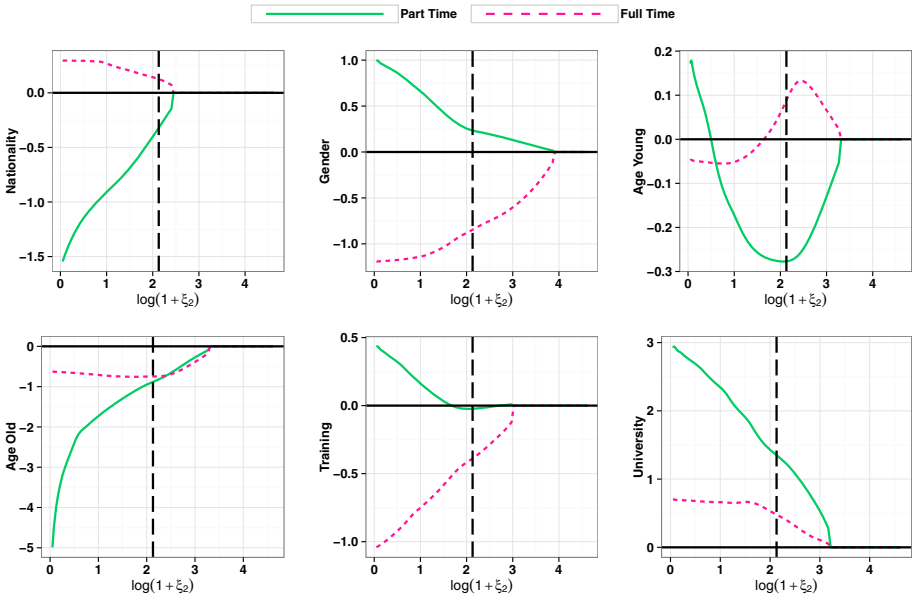

Fig. 7 Coefficients paths of the main effects for the unemployment data

validation based on the predictive deviance. The resulting tuning parameters were $\zeta_1 = 1.0$ and $\zeta_2 = 7.42$. The cross validation score for fixed value $\zeta_1 = 1.0$ is shown in Fig. 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 Fig. 5b. The corresponding confidence intervals have again been estimated by a non-parametric bootstrap method with 1000 bootstrap replications. Due to the small value of ζ_1 the baseline effects are less smooth than in the congressional careers example.

Table 4 shows the ordinary ML estimates and the estimates resulting from the penalized competing risk model with their corresponding standard errors. It is immediately seen that the penalization method removes a considerable number of effects, that is, 22 out of 40 parameters leading to a enormous reduction of the model complexity. But all main effects and three interaction effects remain in the model. One sees, for example, that for women it is more likely to get a part-time job and less likely to get a full-time job. For younger people, getting a full-time job is more likely to end unemployment than getting a part-time job.

Figure 6 depicts a selection of resulting hazard rates. In particular, Fig. 6a shows the 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 increases after a duration time of 25 months. It can be seen from Fig. 6b that fewer women than men get a full-time job, whereas slightly more women get a part-time job, holding everything else fixed. A transition to a university degree clearly increases the probability of getting a full-time or part-time job.

For illustration Fig. 7 shows the coefficient paths of the main effects. Each path indicates the penalized estimates subject to tuning parameter ζ_2 for $\zeta_1 = 1.0$. In particular, the

paths illustrate how the estimates change towards zero for increasing ζ_2. The dashed black line indicates the ζ_2 chosen via cross-validation and the resulting estimates.

6 Concluding remarks

In competing risk models for discrete duration time, one is interested in the cause-specific hazard rates. When modeling these cause-specific hazard rates, each explanatory variable is linked to a group of parameters. The proposed penalization method enforces the simultaneous shrinkage of parameters belonging to such a group. A parameter group even can be completely removed from the model yielding variable selection instead of parameter selection. 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, an additional penalty term is incorporated that steers the smoothness of the baseline effects.

References

Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
Beyersmann, J., Allignol, A., Schumacher, M.: Competing Risks and Multistate Models with R. Springer, New York (2011)
Box-Steffensmeier, J.M., Jones, B.S.: Event History Modeling: A Guide for Social Scientists. Cambridge University Press, New York (2004)
Candes, E., Tao, T.: The dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313–2351 (2007)
Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**, 1–26 (1979)
Eilers, P.H., Marx, B.D.: Flexible smoothing with b-splines and penalties. Stat. Sci. **11**, 89–121 (1996)
Enberg, J., Gottschalk, P., Wolf, D.: A random-effects logit model of work-welfare transitions. J. Econ. **43**, 63–75 (1990)
Fahrmeir, L., Tutz, G.: Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edn. Springer, New York (2001)
Fahrmeir, L., Wagenpfeil, S.: Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. J. Am. Stat. Assoc. **91**, 1584–1594 (1996)
Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
Gertheiss, J., Tutz, G.: Penalized regression with ordinal predictors. Int. Stat. Rev. **77**, 345–365 (2009)
Han, A., Hausman, J.A.: Flexible parametric estimation of duration and competing risk models. J. Appl. Econ. **5**(1), 1–28 (1990)
Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009)
Jones, B.: A Longitudinal Perspective on Congressional Elections. Ph. D. thesis, State University of New York at Stony Brook (1994)
Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data, 2nd edn. Wiley, New York (2002)
Kauermann, G., Khomski, P.: Full time or part time reemployment: a competing risk model with frailties and smooth effects using a penalty based approach. J. Comput. Gr. Stat. **18**, 106–125 (2009)
Klein, J., Moeschberger, M.: Survival Analysis: Statistical Methods for Censored and Truncated Data, 2nd edn. Springer, New York (2003)
Kleinbaum, D.G., Klein, M.: Survival Analysis: A Self-learning Text, 3rd edn. Springer, New York (2013)
Krishnapuram, B., Carin, L., Figueiredo, M.A., Hartemink, A.J.: Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 957–968 (2005)

Narendranathan, W., Stewart, M.B.: Modelling the probability of leaving unemployment: competing risks models with flexible base-line hazards. Appl. Stat. **42**(1), 63–83 (1993)
Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**, 123–231 (2013)
Pößnecker, W.: MRSP: Multinomial Response Models with Structured Penalties. R package version 0.4.3. (2014)
R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014)
Simon, N., Friedman, J., Hastie, T.: A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. arXiv preprint (2013)
Steele, F., Goldstein, H., Browne, W.: A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. Stat. Model. **4**(2), 145–159 (2004)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**, 267–288 (1996)
Tutz, G.: Competing risks models in discrete time with nominal or ordinal categories of response. Quality Quant. **29**, 405–420 (1995)
Tutz, G.: Regression for Categorical Data. Cambridge University Press, Cambridge (2012)
Tutz, G., Pößnecker, W., Uhlmann, L.: Variable selection in general multinomial logit models. Comput. Stat. Data Anal. **82**, 207–222 (2015)
Wang, H., Leng, C.: A note on adaptive group lasso. Comput. Stat. Data Anal. **52**, 5277–5286 (2008)
Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B **68**, 49–67 (2006)
Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**, 301–320 (2005)