Variable Selection For Discrete Competing Risks Models

Jingwei Zhang[*], Lupengkong, Chunlong Luo, Quanfeng Yang, Yiqing Lu

June 29, 2016

1 Introduction

1.1 Applied problem

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.

1.2 Statistical problem

Modeling of the event times in the presence of multiple outcomes is usually referred to as competing risks modeling. Alternatively, one also speaks of competing events, competing causes or failures to convey that several events compete with each other to be observed.

1.3 Existing methods

Most of the literature for competing risks considers the case of continuous time. If time is discretely observed, 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, but without referring to the problem of variable selection. 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. 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. One of the oldest penalization methods is the ridge method. An alternative penalty approach that has become very popular is the lasso Tibshirani (1996) .Several improvements for the lasso method have been proposed in the last decade, for example the group lasso, the adaptive lasso, SCAD, the elastic net and the Dantzig selector.

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 model. More recently, alternatives that enforce variable selection instead of variable select in multiple response models were proposed.

[*] skykiny@outlook.com

1.4 New contribution

When modeling cause-specific hazard rates in competing risk model, each explanatory variable is linked to a group of parameters. A new penalization method is proposed in this article that 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.

We apply this method to one problems, the congressional careers of members of the US congress.

2 Competing Risk Model for Discrete Time

2.1 The Discrete Competing Risk Model

Let $\mathbf{x}$ be the covariates of one object. Competing risk model determines to analysis the influence of $\mathbf{x}$ on the happening of the several events(or causes of risks).

Let time take values from $\{1, \ldots, k\}$ and let $q = k - 1$. Discrete time $T \in \{1, \ldots, k\}$ means that $T = t$ is observed if failure occurs on time t. If time values result from intervals, one has k underlying intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{q1}, a_q), [a_q, \infty)$, where typically $a_0 = 0$ is assumed and a_q denotes the final follow-up. $T = t$ is observed if failure occurs within the time interval $[a_{t1}, a_t)$.

Let the distinct terminating causes be denoted by $R \in \{1, \ldots, m\}$. Then the cause-specific discrete hazard function resulting from cause r is determined by the conditional probability

$$\lambda_r(t|\mathbf{x}) = P(T = t, R = r | T \geq t, \mathbf{x}) \tag{1}$$

where $r = 1, \ldots, m$ and $t = 1, \ldots, q$. The m hazard functions sum up to an overall hazard function

$$\lambda(t|\mathbf{x}) = \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}) = P(T = t | T \geq t, \mathbf{x}) \tag{2}$$

The survival function, which indicates the unconditional probability of no event happening on time collection $\{1, \ldots, t\}$, is given by

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x}) = \prod_{j=1}^{t} (1 - \lambda(j|\mathbf{x})) \tag{3}$$

The unconditional probability of one event happening on time t is given by

$$P(T = t|\mathbf{x}) = \lambda(t|\mathbf{x}) \prod_{j=1}^{t-1} (1 - \lambda(j|\mathbf{x})) = \lambda(t|\mathbf{x}) S(t - 1|\mathbf{x}) \tag{4}$$

If a object reaches time $t-1$, there are $m+1$ possible outcomes, transition to one of the m target events or survival. The corresponding conditional probabilities are given by

$$\lambda_1(t|\mathbf{x}), \ldots, \lambda_m(t|\mathbf{x}), 1 - \lambda(t|\mathbf{x}) \tag{5}$$

where $1 - \lambda(t|\mathbf{x})$ is the probability for survival(no event happens).

Therefore, given an individual reaches time $t - 1$, the hazards can be modeled by the multinomial logit model given by

$$\lambda_r(t|\mathbf{x}) = \frac{\exp\left(\beta_{0tr} + \mathbf{x}^{\mathsf{T}} \gamma_{\mathbf{r}}\right)}{1 + \sum_{s=1}^{m} \exp\left(\beta_{0ts} + \mathbf{x}^{\mathsf{T}} \gamma_{\mathbf{s}}\right)} \tag{6}$$

where $t = 1, \dots, q$, and $r = 1, \dots, m$. We define $\eta_i tr = \beta_{0tr} + \mathbf{x}^\mathsf{T}\gamma_\mathbf{r}$. Then the parameters $\beta_{01r}, \dots, \beta_{0qr}$ determine the cause-specific baseline hazard functions and $\gamma_\mathbf{r}$ contains the cause-specific effects of covariates. Conditional probability of survival is implicitly determined by

$$\lambda_0(t|\mathbf{x}) = P(T > t|T \geq t, \mathbf{x}) = 1 - \sum_{r=1}^{m} \lambda_r(t|\mathbf{x}) = \frac{1}{1 + \sum_{s=1}^{m} \exp\left(\beta_{0ts} + \mathbf{x}^\mathsf{T}\gamma_\mathbf{s}\right)} \tag{7}$$

Let $R = 0$ denote the conditional survival, the conditional probabilities are given by $\lambda_0(t|\mathbf{x}) = P(T > t|T \geq t, \mathbf{x}), \lambda_1(t|\mathbf{x}), \dots, \lambda_m(t|\mathbf{x})$, which sum up to one.

2.2 Estimation

Let data be given by $(t_i, r_i, \delta_i, \mathbf{x}_i)$, $i = 1, \dots, n$, where $t_i = min(Ti, Ci)$ 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,$r_i \in \{1, \dots, m\}$ indicates the type of the terminating event, $\mathbf{x}_i$ is the covariate vector and δ_i denotes the censoring indicator with

$$\delta_i = \begin{cases} 1, & \text{if event occured on time } t_i \\ 0, & \text{if it censores on time } t_i \end{cases} \tag{8}$$

Under the assumption that censoring does not depend on the parameters that determine the survival time (non-informative censoring), the likelihood contribution of the i-th observation is

$$\begin{aligned} L_i &= P(T_i = t_i, R_i = r_i|\mathbf{x})^{\delta_i} P(T_i > t_i|\mathbf{x})^{1-\delta_i} \\ &= \lambda_{r_i}(t_i|\mathbf{x}_i)^{\delta_i}(1 - \lambda(t_i|\mathbf{x}_i))^{1-\delta_i} \prod_{t=1}^{t_i - 1}(1 - \lambda(t|\mathbf{x}_i)) \end{aligned} \tag{9}$$

Let $R_t = \{i : t \leq t_i\}$ be the risk set containing all objects who are at risk on time t. For an alternative form of the likelihood, indicators for the transition to the next period are defined by

$$y_{itr} = \begin{cases} 1, & \text{if event of type } r \text{ occured on time } t \\ 0, & \text{if event of type } r \text{ did not occure on time } t \end{cases} \tag{10}$$

and

$$y_{it0} = \begin{cases} 1, & \text{if no events occured on time } t \text{ (survive)} \\ 0, & \text{if one of the } m \text{ events occures on time } t \end{cases} \tag{11}$$

where $i \in R_t$ and $r = 1, \dots, m$. These indicator variables are gathered in the vector $\mathbf{y}_{it}^\mathsf{T} = (y_{it0}, y_{it1}, ..., y_{itm})$ denoting the response vector of object i, $i = 1, \dots, n$, $t = 1, \dots, t_i$. Then, the likelihood function Li can be rewritten as

$$L_i = \prod_{t=1}^{t_i}(\prod_{r=0}^{m} \lambda_r(t_i|\mathbf{x}_i)^{y_{itr}}) \tag{12}$$

The total log-liklihood is given by

$$l = \sum_{i=1}^{n}\sum_{t=1}^{t_i}\sum_{r=0}^{m} y_{itr} \log \lambda_r(t|\mathbf{x}) \tag{13}$$

where $\lambda_r(t|\mathbf{x})$ is given by the model (6) and (7). This ML estimates can be easily computed by using statistical software for multinomial regression models.

3 Choice of Penalty Term

3.1 The original penalty term

The model of the cause-specific hazard function $\lambda_{\mathrm{r}}(t|x_i)$ has the form of parameters:

$$\eta_{itr} = \beta_{0tr} + \mathbf{x_i}^T \gamma_r, t = 1, ...q; r = 1, ..., m, \tag{14}$$

where $\mathbf{x_i}^T = (x_{i1}, ..., x_{ip})$ and $\gamma_r^T = (\gamma_{r1}, ..., \gamma_{rp})$.

To obtain a sparse representation and in particular variable selection, the authors consider the penalized ML estimation:

$$l_{\zeta_1,\zeta_2}(\beta_0, \gamma) = l(\beta_0, \gamma) - J_{\zeta_1,\zeta_2}(\beta_0, \gamma), \tag{15}$$

where $\beta_0^T = (\beta_{01}^T, ..., \beta_{0m}^T)$ and $\gamma^T = (\gamma_1^T, ..., \gamma_m^T)$. $l(\beta_0, \gamma)$ denotes the ordinary log-likelihood, and $J_{\zeta_1,\zeta_2}(\beta_0, \gamma)$ stands for a penalty term that depends on scalar tuning parameters ζ_1 and ζ_2.

Authors adopt lasso for variable selection. The penalty term is given by

$$\begin{aligned} J_{\zeta_1,\zeta_2}(\beta_0, \gamma) &= \zeta_1 \sum_{r=1}^{m} \sum_{t=2}^{q} (\beta_{0tr} - \beta_{0,t-1,r})^2 + \zeta_2 \sum_{j=1}^{\mathrm{p}} \phi_{\mathrm{j}} ||\gamma_{.j}|| \\ &= \zeta_1 J_1(\beta_0) + \zeta_2 J_2(\gamma) \end{aligned} \tag{16}$$

where $||u|| = ||u||_2 = \sqrt{u^T u}$, $\phi_j = \sqrt{m}$ and $\gamma_{.j}^T = (\gamma_{1j}, ..., \gamma_{mj})$. The penalty item $\zeta_1 J_1(\beta_0)$ on the baseline parameters β_0 is used to ensure that the estimated hazard rates are sufficiently smooth over time. The penalty $\zeta_2 J_2(\gamma)$ is used for variable selection. A variable is removed from the model if and only if all of its effect parameters are set to zero simultaneously. Therefore, $J_2(\gamma)$ makes it if $\mathbf{x_i}$ is removed only when $||\gamma_{.j}|| = \sqrt{\gamma_{1j}{}^2 + \gamma_{2j}{}^2 + ... + \gamma_{mj}{}^2} = 0$ and then $\gamma_{ij} = 0$ where $i = 1, 2, ..., m$.

3.2 Improvement for penalty item for $J_1(\beta_0)$

To simplify the baseline effects and reduce the number of time periods q in $J_1(\beta_0)$, it is expanded in in an equidistant and low-rank B-spline basis function, resulting in

$$\beta_{0tr} = \sum_{s=1}^{d_r} \alpha_{0sr} B_s(t) \tag{17}$$

with $d_r < q$. The total penalty function can be rewritten as

$$J_{\zeta_1,\zeta_2}(\alpha_0, \gamma) = \zeta_1 \sum_{r=1}^{m} \sum_{s=2}^{d_r} (\alpha_{0\mathrm{s}r} - \alpha_{0,s-1,r})^2 + \zeta_2 \sum_{j=1}^{\mathrm{p}} \phi_{\mathrm{j}} ||\gamma_{.j}|| \tag{18}$$

3.3 The final penalty term

The choice of tuning parameter ζ_2 involves a tradeoff between fewer variable selection and unbiasedness. In the light of this conflict, adaptive weights for the penalties is an optimal choice.

$$\phi_j{}^a = \frac{\sqrt{m}}{||\gamma_{.j}^{Init}||} \tag{19}$$

$\gamma_{.j}^{Init}$ is the penalized estimator that results from application of penalty function above with $\zeta_2 = 0$. Then the penalty function be finally written as

$$J_{\zeta_1,\zeta_2}(\alpha_0, \gamma) = \zeta_1 \sum_{r=1}^{m} \sum_{s=2}^{d_r} (\alpha_{0\mathrm{s}r} - \alpha_{0,s-1,r})^2 + \zeta_2 \sum_{j=1}^{\mathrm{p}} \phi_{\mathrm{j}}{}^a ||\gamma_{.j}|| \tag{20}$$

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 for strong predictors and large for weak predictors. However, these penalized estimators with adaptive weights can provide consistent variable selection.

4 Computional Issues

To estimate the parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\gamma}$,the penalized log-likelihood $l_{\zeta_1,\zeta_2}(\boldsymbol{\beta_0},\boldsymbol{\gamma})$ can be formulated as

$$(\hat{\boldsymbol{\beta_0}},\hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta_0},\boldsymbol{\gamma}}{\operatorname{argmin}} \Big(- l(\boldsymbol{\beta_0},\boldsymbol{\gamma}) + \zeta_1 \boldsymbol{J_1(\beta_0)} + \zeta_2 \boldsymbol{J_2(\gamma)}\Big) \tag{21}$$

The alogrithm for solving this convex optimization problem used by author is based on proximal gradient algorithms.

Proximal gradient method belongs to a class of algorithms, called proximal algorithms, for solving convex optimization problems.Proximal methods sit at a higher level of abstraction than classical algorithms like Newton's method: the base operation is evaluating the proximal operator of a function, which itself involves solving a small convex optimization problem.

So in this section, we discribe some details of the estimate approach.First, some definition of proximal operator are introduced,then the proximal gradient method that applied in here is outlined.Finally, the tuning parameter selection for discrete competing risk models is presented.

4.1 Proximal operator

Let $f : \boldsymbol{R}^n \to \boldsymbol{R} \cup \{+\infty\}$ be a closed proper convex function,the proximal operator $\mathbf{prox}_f : \boldsymbol{R}^n \to \boldsymbol{R}^n$ of f is defined by

$$\mathbf{prox}_f(v) = \underset{x}{\operatorname{argmin}} \Big(f(x) + (1/2)\|x - v\|_2^2 \Big) \tag{22}$$

where $\|\cdot\|_2$ is L_2 norm.

The definition indicates that $\mathbf{prox}_f(v)$ is a point that compromises between minimizing f and being near to v. For this reason, $\mathbf{prox}_f(v)$ is sometimes called a proximal point of v with respect to f.

The proximal operator of f can also be interpreted as a kind of gradient step for the function f. In particular, we have

$$\mathbf{prox}_{\lambda f}(v) \approx v - \lambda \nabla f(v) \tag{23}$$

when λ is small and f is differentiable. This suggests that there is a close connection between proximal operators and gradient methods, and also indicates that the proximal operator may be useful in optimization. It also suggests that λ will play a role similar to a step size in a gradient method.

4.2 Proximal gradient method

Consider the unconstrained problem with cost function split in two components

$$\min\ f(x) + g(x) \tag{24}$$

where $f : \boldsymbol{R}^n \to \boldsymbol{R}$ and $g : \boldsymbol{R}^n \to \boldsymbol{R} \cup \{+\infty\}$are closed proper convex and f is differentiable.In this form, we split the objective into two terms, one of which is differentiable.

The proximal gradient method is

$$x^{(k+1)} := \mathbf{prox}_{\lambda^{(k)} g}\Big(x^{(k)} - \lambda^{(k)} \nabla f(x^{(k)}) \Big) \tag{25}$$

where $\lambda^{(k)} > 0$ is a step size,constant or determined by line search.

So it was exploited that log-likelihood term $-l(\boldsymbol{\beta_0},\boldsymbol{\gamma})$ is convex and differentiable ,both the overall penalty term $\boldsymbol{J}_{\zeta_1,\zeta_2}$ and the L_2^2-term in can be decomposed into nonoverlapping parts that only contain either $\boldsymbol{\beta}_0$ or γ.

So for $k = 0, 1, 2, \dots$ until convergence, the proximal gradient iterations are given by

$$\hat{\boldsymbol{\beta}}_0^{(k+1)} = \mathbf{Prox}_{\zeta_1/v_{(k)} J_1}\left(\mathbf{v}^{(k)} := \hat{\boldsymbol{\beta}}_0^{(k)} + \frac{1}{v^{(k)}} \cdot \frac{\partial l(\hat{\boldsymbol{\beta}}_0^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)})}{\partial \boldsymbol{\beta}_0} \right) \tag{26}$$

and

$$\hat{\boldsymbol{\gamma}}^{(k+1)} = \mathbf{Prox}_{\zeta_2/v_{(k)} J_2}\left(\mathbf{w}^{(k)} := \hat{\boldsymbol{\gamma}}^{(k)} + \frac{1}{v^{(k)}} \cdot \frac{\partial l(\hat{\boldsymbol{\beta}}_0^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)})}{\partial \boldsymbol{\gamma}} \right) \tag{27}$$

where $v^k > 0$ is an inverse stepsize parameter.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.

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 $J_2(\gamma) = \sum_{j=1}^{p} \phi_j \|\gamma_j\| = \sum_{j=1}^{p} J_{2j}$and let $\boldsymbol{w}$ be partitioned like $\boldsymbol{\gamma}$.Then, we can get the analytical solution

$$\mathbf{Prox}_{\zeta_2/v\cdot J_{2j}}(\boldsymbol{w}_{\cdot j}) = \left(1 - \frac{\zeta_2\phi_j/v}{\|\boldsymbol{w}_{\cdot j}\|}\right)_+ \boldsymbol{w}_{\cdot j}, \quad j = 1, \dots, p. \tag{28}$$

where $(u)_+ = \max(u, 0)$.

To derive a closed solution to, we rewrite the penalty on the baseline parameters:$\mathbf{J}_1(\boldsymbol{\beta}_0) = \sum_{r=1}^{m} \sum_{t=2}^{q} (\beta_{0tr} - \beta_{0,t-1,r})^2 = \sum_{r=1}^{m} \mathbf{J}_{1r}$.Let 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} =$\|\boldsymbol{D\beta_{0r}}\|_2^2$.Hence, the proximal operator only contains quadratic terms and thus,with $\boldsymbol{\Omega} = \boldsymbol{D}^T\boldsymbol{D}$ and identity matrix $\boldsymbol{I}$,admits an analytical solution:

$$\mathbf{Prox}_{\zeta_1/v\cdot J_{1r}}(\boldsymbol{v}_{\cdot r}) = (\boldsymbol{I} + \frac{\zeta_1}{v}\boldsymbol{\Omega})^{-1}\boldsymbol{v}_{\cdot r}, \quad r = 1, \dots, m \tag{29}$$

4.3 Tuning parameter selection

The tuning parameters ζ_1 and ζ_2 can be chosen by k-fold cross-validation with a modification to standard form.We let $\mathbf{I}_s$ denote the index set of observations that belog to fold s for folds $s = 1, \dots, k$.And the $\hat{\lambda}_r^{(-s)}$ denote the estimate for λ_r that is based on other folds except for $\mathbf{I}_s$.

We use the (predictive) deviance as the criterion to be cross-validated.So the cross-validated deviance is defined by

$$D_{CV} = 2\sum_{s=1}^{K}\sum_{i\in\mathbf{I}_s}\sum_{t=1}^{t_i}\sum_{r=0}^{m} y_{itr} log\left(\frac{y_{itr}}{\hat{\lambda}_r(t|\mathbf{x}_i)^{(-s)}}\right) \tag{30}$$

In which we can see that all $(m+1)\cdot t_i$ data points y_{itr} belong to the same original observation i assigned to the same cross-validation fold.And the closer the distance between D_{CV} with zero ,the better tuning parameters will be chosen.

5 Application

In this part, the proposed penalized competing risk model with discrete duration time applied to describe Congressional careers in the United States. 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).

This research aim is to seek for covariates probably influence the transition process of a Congressman from his first election up to one of the competing events general, primary, retirement or ambition . The used data set contains the career paths of 860 Congressmen. Several covariates in Table1 are available as predictors for the end of careers. 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.

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 $\lambda_r(t|\mathbf{x}) = \frac{\exp(\eta_{itr})}{1+\sum_{s=1}^{m}\exp(\eta_{its})}$ is specified by the cause-specific linear predictors $\eta_{itr} = \beta_{0tr} + \mathbf{x}^{\mathsf{T}}\gamma_{\mathbf{r}}$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.

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 district was substantially redistricted 0: no, 1: yes

Since the adaptive version of the penalty yielded better cross-validation score, adaptive weights were used. Tuning parameters ξ_1 and ξ_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 $\xi_1 = 6.0$ and $\xi_2 = 2.64$. For a fixed model, the corresponding cross validation score is shown in Figure 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.

Parameter estimates of the covariate effects are summarized in Table 2. 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 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.

In Figure 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 Figure 4 we show only the paths for the main effects. Each path indicates the penalized estimates subject to tuning parameter ξ_2, where the abscissa is transformed by $log(1 + \xi_2))$. The paths illustrate how the estimates changes towards zero for increasing ξ_2. Hence, they show the effects of covariates on the terminating events when penalization is increased. The dashed black line indicates the value of ξ_2 that was chosen via cross-validation.

6 Discussion

In competing risk models for discrete duration time, one is interested in the the causespecific 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.

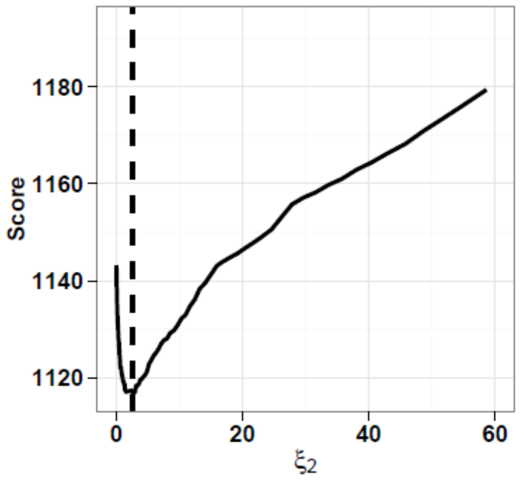


Figure 1: Cross validation score subject to penalty parameter ξ_2 for $\xi_1 = 6.0$ for the Congressional career data

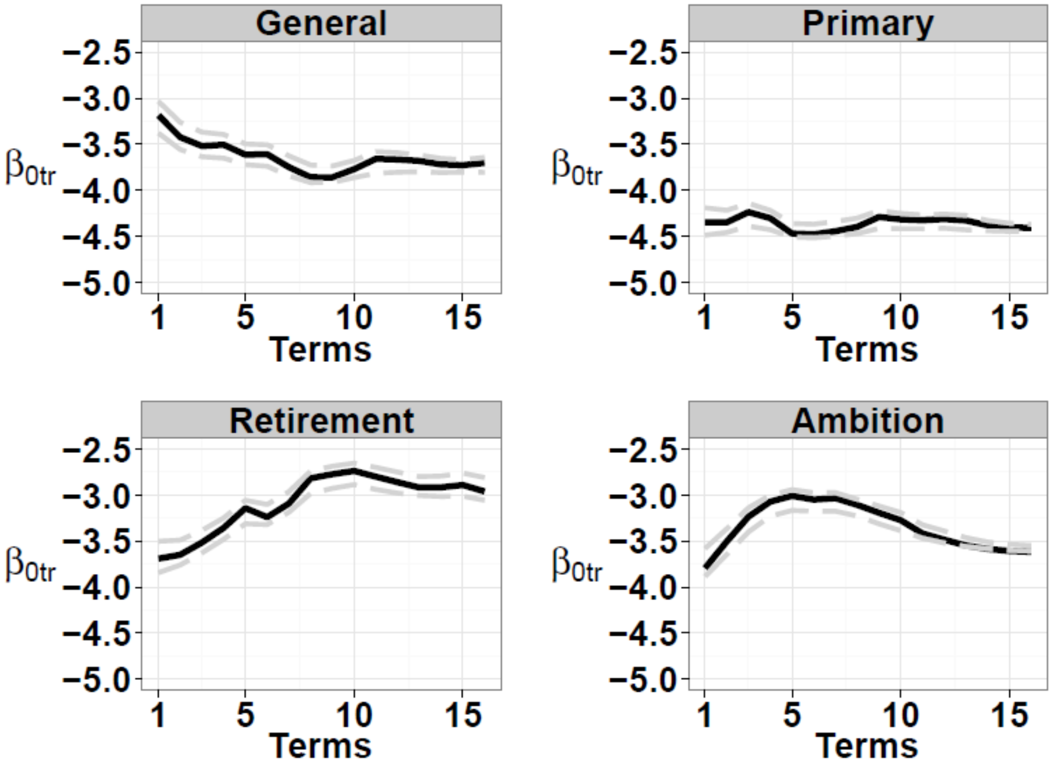
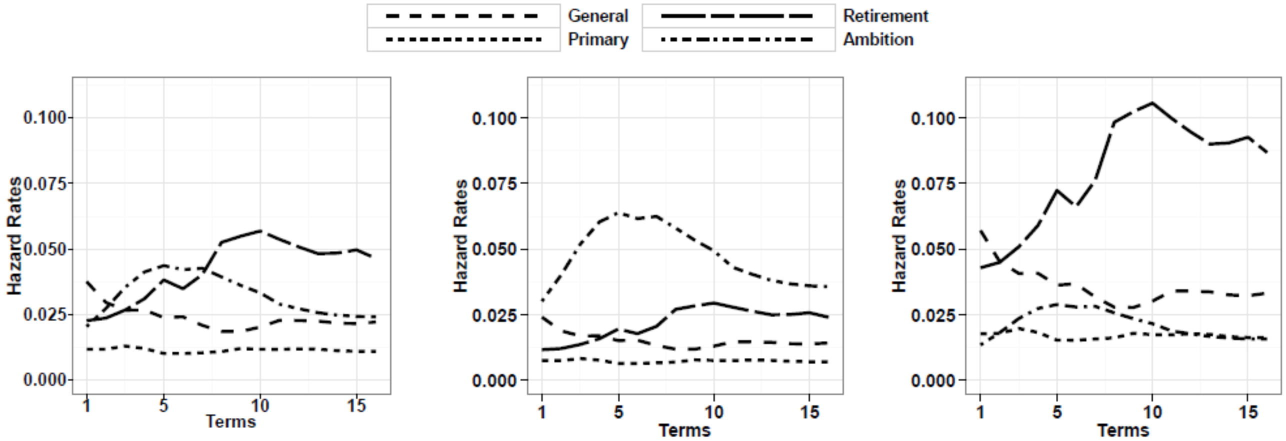


Figure 2: Parameter estimates of the cause-specific time-varying baseline effects for the Congressional careers data. Dashed lines represent the 95% pointwise bootstrap interval



(a) 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 3: Estimated cause-specific hazard rates over time for the Congressional careers data

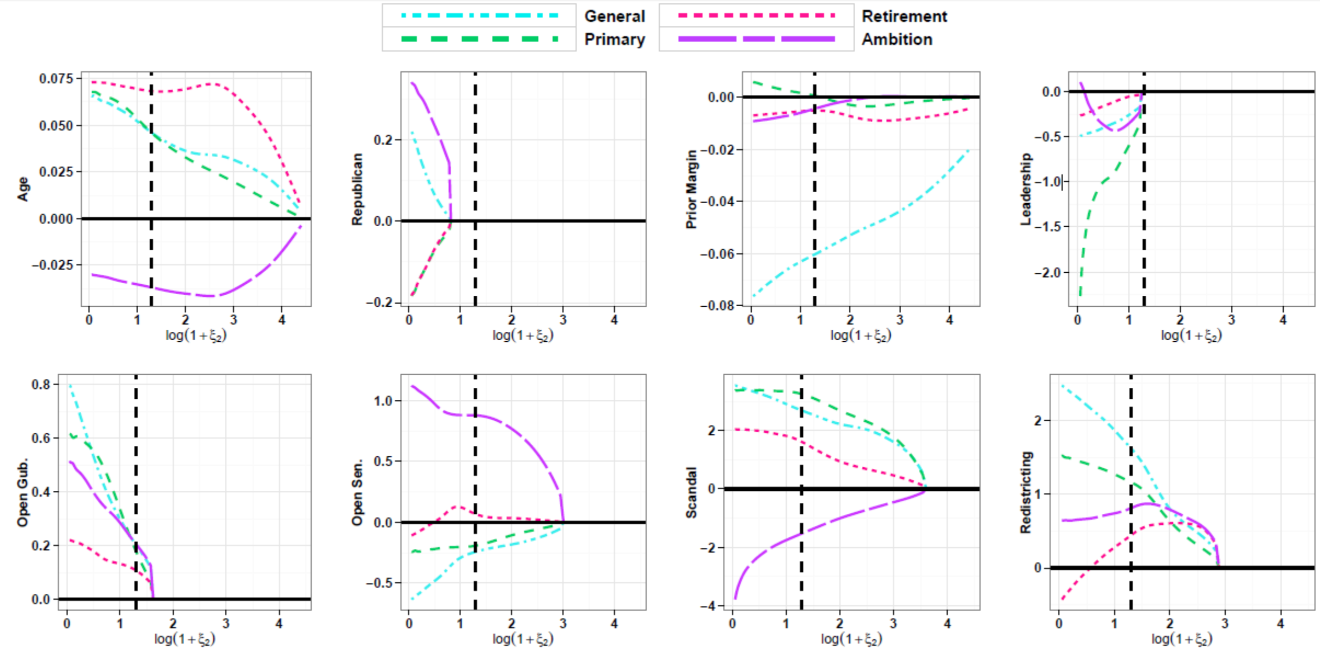


Figure 4: Coefficients paths of the main effects for the Congressional career data