LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gerhard Tutz & Andreas Groll

Variable Selection in Discrete Survival Models Including Heterogeneity

Technical Report Number 167, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Variable Selection in Discrete Survival Models Including Heterogeneity

Gerhard Tutz* & Andreas Groll#

Ludwig-Maximilians-Universität München

Akademiestraße 1, 80799 München

* gerhard.tutz@stat.uni-muenchen.de

groll@math.lmu.de

August 19, 2014

Abstract

Several variable selection procedures are available for continuous time-to-event data. However, if time is measured in a discrete way and therefore many ties occur models for continuous time are inadequate. We propose penalized likelihood methods that perform efficient variable selection in discrete survival modeling with explicit modeling of the heterogeneity in the population. The method is based on a combination of ridge and lasso type penalties that are tailored to the case of discrete survival. The performance is studied in simulation studies and an application to the birth of the first child.

Keywords: Variable selection, discrete survival, heterogeneity.

1 Introduction

In many applications time is measured on a discrete scale, for example, in days, months or weeks. One can consider the measurement as a discretized version of the underlying continuous time, but discrete time often is the natural way how observations are collected. For example, a natural measure for the time it takes a couple to conceive is the number of menstrual cycles, which is a truly discrete response, see Scheike and Jensen (1997) for an application of discrete survival models to model fertility.

Discrete survival models were already considered in the seminal paper of Cox (1972). The logistic model was later investigated by Thompson (1977) whereas

Prentice and Gloeckler (1978) focussed on the grouped Cox model. The representation of hazard models as binary Bernoulli trials was propagated by Brown (1975) and Laird and Olivier (1981). Extensions to more flexible models with nonparametric components were proposed, for example, by Fahrmeir (1994), Fahrmeir and Knorr-Held (1997), Fahrmeir and Kneib (2011), Tutz and Pritscher (1996) and Kauermann et al. (2005).

The basic discrete survival model does not account for heterogeneity in the population and is frequently too simple. The explicit modeling of heterogeneity in the form of frailty models was considered by Ham and Rea Jr (1987) when modeling unemployment duration. Scheike and Jensen (1997) used the clog-log model and, for convenience, assumed a gamma distributed heterogeneity component. Vermunt (1996) proposed a modified log linear model, which is restricted to categorical covariates, Land et al. (2001) extended the model to allow for metric covariates. Misspecified mixing distributions were investigated by Baker and Melino (2000) and, more recently, by Nicoletti and Rondinelli (2010). Nicoletti and Rondinelli (2010) found in their simulations that ignoring heterogeneity yields attenuated covariate coefficients for time invariant covariates. For covariates that vary over time attenuation was very weak.

When modeling the effects of covariates on duration one wants to identify those variables that actually have an effect. In particular, when many explanatory variables are available, the restriction to the relevant variables is important because simple maximum likelihood estimates typically do not exist in high-dimensional settings. Apart from simple forward, backward and forward/backward procedures not much is available to select variables in discrete survival models.

In the present paper, variable selection for discrete survival models with a frailty component models is obtained by using appropriately designed penalties to account for the special features of discrete survival. In Section 2 the framework of discrete survival is briefly considered. In Section 3 we use regularization techniques to enforce variable selection. In Section 4 the results of a simulation study are presented and in Section 5 the method is applied to real data.

2 Discrete Hazard Models Including Heterogeneity

2.1 Basic Models

Let time take values from $\{1, \ldots, k\}$. 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 $q = k - 1$. Discrete time $T \in \{1, \ldots, k\}$ means that $T = t$ is observed if failure occurs within the interval $[a_{t-1}, a_t)$.

The main tool in modeling survival data is the hazard function. In discrete

time it has the form

$$\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x}), \quad t = 1, \ldots, q,$$

which is the conditional probability for failure in interval $[a_{t-1}, a_t)$ given the interval is reached. The corresponding survivor function is

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

Models for discrete survival given covariates $\boldsymbol{x}$ have the form

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

where $h(.)$ is a fixed response function, which is assumed to be strictly monotonically increasing. The parameters γ_{0t} represent the baseline hazard, which is the same for all individuals. The contribution of the predictors is captured by the term $\boldsymbol{x}^T\boldsymbol{\gamma}$, where $\boldsymbol{x}^T = (x_1, \ldots, x_p)$ is a vector of predictors and $\boldsymbol{\gamma}^T = (\gamma_1, \ldots, \gamma_p)$ are the weights. The most prominent discrete survival model is the continuation ratio model

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

which uses the logistic distribution function $h(\eta) = \exp(\eta)/(1 + \exp(\eta))$. An alternative widely used model is the grouped proportional odds model

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

which uses the Gompertz distribution $h(\eta) = 1 - \exp(-\exp(\eta))$ as response function. The name refers to its derivation as a grouped version of Cox's proportional hazard model.

The model (1) includes an unspecified baseline hazard, which is assumed to be the same for all observations. This is frequently too strict an assumption because it ignores heterogeneity among individuals. Therefore we will consider the extended model, which includes potential heterogeneity. The corresponding basic frailty model for the ith observation has the form

$$\lambda(t|\boldsymbol{x}, b_i) = h(b_i + \gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}), \tag{2}$$

where b_i is a random effect that is assumed to follow a fixed mixture distribution with density $p(.)$, typically the normal distribution.

2.2 Estimation with Censoring

When modeling survival data censoring has to be taken into account. In the case of right censoring, which is considered here, for censored data it is only known

that T exceeds a certain value but the exact value is not known. Let C_i denote the censoring time and T_i the exact failure time for observation i. In random censoring it is assumed that T_i and C_i are independent random variables. The observed time is given by $t_i = \min(T_i, C_i)$ as the minimum of survival time T_i and censoring time C_i. It is often useful to introduce an indicator variable for censoring given by

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i, \\ 0 & \text{if } T_i > C_i, \end{cases}$$

where it is implicitly assumed that censoring occurs at the end of the interval.

Under random censoring the probability of observing (t_i, δ_i) is given by

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

It should be noted that the probability is defined given the random effect b_i, which is suppressed on the right hand side of the equation. In the simple survival model without heterogeneity b_i is omitted and the probability is $P(t_i, \delta_i | \boldsymbol{x}_i)$.

If one assumes that the censoring contributions do not depend on the parameters that determine the survival time (non-informative in the sense of Kalbfleisch and Prentice, 2002), one can separate the factor $c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ to obtain the simpler form

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

An important tool in discrete survival is that the probability and therefore the corresponding likelihood can be rewritten by using sequences of binary data. By defining for a non-censored observation ($\delta_i = 1$) the sequence $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0, 1)$ and for a censored ($\delta_i = 0$) the sequence $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0)$, the probability (omitting c_i) can be written as

$$P(t_i, \delta_i | \boldsymbol{x}_i, b_i) = \prod_{s=1}^{t_i} \lambda(s | \boldsymbol{x}_i, b_i)^{y_{is}} (1 - \lambda(s | \boldsymbol{x}_i, b_i))^{1-y_{is}}.$$

For the *model without heterogeneity* the random effect b_i is omitted and the corresponding log-likelihood is that of a binary response model, where $\lambda(s|\boldsymbol{x}_i)$ is the probability of the binary response, transition to the next category or not, and the corresponding linear predictor for $\lambda(s|\boldsymbol{x}_i)$ is given by $\gamma_{0s} + \boldsymbol{x}^T \boldsymbol{\gamma}$. By construction of an appropriate design matrix estimation methods and software for binary response models can be used. Early references for the embedding into the framework of binary regression are Brown (1975) and Laird and Olivier (1981), see also Fahrmeir and Tutz (2001).

In the *frailty model* the unconditional probability is given by

$$P(t_i, \delta_i | \boldsymbol{x}_i) = \int P(t_i, \delta_i | \boldsymbol{x}_i, b_i) p(b_i) db_i \tag{3}$$

and therefore by

$$P(t_i, \delta_i|\boldsymbol{x}_i) = \int P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} p(b_i) db_i.$$

One obtains

$$P(t_i, \delta_i|\boldsymbol{x}_i) = \int \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1 - \lambda(s|\boldsymbol{x}_i))^{1-y_{is}} p(b_i) db_i$$

This is the unconditional probability of a random effects model for structured binary data. Estimation can be based on integration techniques, in particular Gauss-Hermite integration can be used, see Hinde (1982), Anderson and Aitkin (1985). A procedure that may reduce the number of quadrature points is the adaptive Gauss-Hermite quadrature (Liu and Pierce, 1994, Pinheiro and Bates, 1995 Hartzel et al., 2001). An alternative way to estimate random effects models is penalized quasi-likelihood estimation, which uses the Laplace approximation (Breslow and Clayton, 1993). Simpler estimates can be obtained under specific assumptions. Scheike and Jensen (1997) used the grouped proportional hazard model and assumed for mathematical convenience that $\exp(b_i)$ is gamma distributed with mean 1. Then the probabilities for failure in t have an explicit form.

3 Variable Selection by Regularization

While several procedures for variable selection are available in the case of continuous time, in particular for the Cox model (Therneau and Grambsch, 2000, Goeman, 2010, Simon et al., 2011), discrete survival has been somewhat neglected.

In the following we consider variable selection for the general model

$$\lambda(t|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i) = h(\gamma_{0t} + \boldsymbol{x}_{it}^T \boldsymbol{\gamma} + \boldsymbol{z}_{it}^T \boldsymbol{b}_i), \tag{4}$$

with explanatory variables $\boldsymbol{x}_{it}, \boldsymbol{z}_{it}$, which can vary over time, and random effect vector $\boldsymbol{b}_i$. The model with random intercepts uses $\boldsymbol{z}_{it} = 1$. More general, in $\boldsymbol{z}_{it}$ also components from $\boldsymbol{x}_{it}$ can be included so that some predictors have subjects-specific slopes. For the distribution $p(\cdot)$ of the random effect it is assumed that $\boldsymbol{b}_i \sim \mathbb{N}(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\rho}))$, where $\boldsymbol{\rho}$ is a vector of unknown parameters that specifies the covariance matrix.

With the parameters that determine the baseline hazard collected in $\boldsymbol{\gamma}_0^T = (\gamma_{01}, \ldots, \gamma_{0k})$, the likelihood for the frailty model has the form

$$l(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho}) = \sum_{i=1}^{n} \log \left(\int \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i)^{y_{is}} (1 - \lambda(s|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i))^{1-y_{is}} p(b_i) db_i \right). \tag{5}$$

With $\boldsymbol{y}_i^T = (y_{i1}, \ldots, y_{it_i})$ and $f(\boldsymbol{y}_i|\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho}) = \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i)^{y_{is}}(1 - \lambda(s|\boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i))^{1-y_{is}}$ one obtains the simpler form

$$l(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho}) = \sum_{i=1}^{n} \log \left(\int f(\boldsymbol{y}_i|\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho}) p(b_i) db_i \right),$$

which is the log-likelihood of a specific random effects model for binary responses. Along the lines of Breslow and Clayton (1993) one can derive the approximation

$$l_{\text{app}}(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{b}) = \sum_{i=1}^{n} \log(f(\boldsymbol{y}_i|\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho})) - \frac{1}{2} \boldsymbol{b}^T \boldsymbol{Q}_{\boldsymbol{b}}(\boldsymbol{\rho})^{-1} \boldsymbol{b},$$

where $\boldsymbol{b}^T = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)$ collects all the random effects and $\boldsymbol{Q}_{\boldsymbol{b}}$ denotes the block-diagonal covariance matrix $\boldsymbol{Q}_{\boldsymbol{b}}(\boldsymbol{\rho}) = diag(\boldsymbol{Q}(\boldsymbol{\rho}), \ldots, \boldsymbol{Q}(\boldsymbol{\rho}))$. Based on the concept of penalized quasi-likelihood (PQL) as suggested by Breslow and Clayton (1993), Lin and Breslow (1996) and Breslow and Lin (1995) maximization of the log-likelihood $l_{\text{app}}(.)$ can be obtained by maximizing the log-likelihood separately with respect to the parameters $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{b}$ and $\boldsymbol{\rho}$. Given an estimate $\hat{\boldsymbol{\rho}}$ one maximizes the profile log-likelihood $l_{\text{app}}(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \hat{\boldsymbol{\rho}}, \boldsymbol{b})$ and separately the random effects parameter $\boldsymbol{\rho}$. For details of the algorithm see also Wolfinger and O'Connell (1993), Littell et al. (1996) and Vonesh (1996).

Variable selection in generalized linear models can be obtained by use of penalized log-likelihood concepts. In the famous lasso the log-likelihood is replaced by a penalized log-likelihood that includes a penalty term of the form $\lambda \sum_{i=1}^{p} |\gamma_i|$, see Tibshirani (1996), Park and Hastie (2007) and Zou (2006). As shown by Groll and Tutz (2014), for simple binary random effects models the inclusion of a lasso penalty can be used within the framework of penalized quasi-likelihood yielding selection procedures for random effects models. Groll and Tutz (2014) gave a detailed algorithm which is based on an approximate EM-method.

Although estimation in discrete survival analysis can be embedded into a framework for binary responses direct use of the method cannot be recommended. A specific feature of discrete survival models is that it contains the baseline hazard specified by the parameters in $\boldsymbol{\gamma}_0$. Even if one has only a moderate number of time intervals (or correspondingly discrete time points) estimation of the baseline hazard will be very unstable. Therefore one has to include an additional penalty term. The penalized approximate log-likelihood that is used has the form

$$l_p(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{b}) = \sum_{i=1}^{n} \log(f(\boldsymbol{y}_i|\boldsymbol{\gamma}_0, \boldsymbol{\gamma}, \boldsymbol{\rho})) - \frac{1}{2} \boldsymbol{b}^T \boldsymbol{Q}_{\boldsymbol{b}}(\boldsymbol{\rho})^{-1} \boldsymbol{b} - \lambda \sum_{i=1}^{p} |\gamma_i| - \lambda_s \text{Pen}(\boldsymbol{\gamma}_0).$$

The lasso penalty term $\lambda \sum_{i=1}^{p} |\gamma_i|$ enforces variable selection with the strength of selection determined by the size of λ. For $\lambda = 0$, no selection is obtained whereas

for $\lambda \to \infty$ all predictors are excluded from the model. The additional penalty term $\lambda_s \text{Pen}(\boldsymbol{\gamma}_0)$ is chosen to obtain smooth estimates of the baseline hazard. We consider two choices. One is the ridge type penalty, which is given by

$$\text{Pen}(\boldsymbol{\gamma}_0) = \sum_{t=1}^{q} \gamma_{0t}^2.$$

For $\lambda_s > 0$ it stabilizes estimates. In particular it prevents parameters γ_i to deteriorate, which is bound to happen for small to moderate sample sizes. A more elaborate penalty term is used after expanding the baseline hazard as a sum of basis functions. We used the expansion $\gamma_{0t} = \sum_{s=1}^{m} \alpha_s \phi_s(t)$, where $\phi_1(.), \ldots, \phi_m(.)$ are B-splines of order q on an equally spaced grid. For the definition of B-splines see, for example, Dierckx (1993). The corresponding penalty that is used within the framework of penalized splines (P-splines; Eilers and Marx, 1996) is

$$\text{Pen}(\boldsymbol{\gamma}_0) = \sum_{j=d+1}^{m} (\Delta^d \gamma_{0j})^2,$$

where Δ is the difference operator, operating on adjacent B-spline coefficients, that is, $\Delta\gamma_{0j} = \gamma_{0j} - \gamma_{0,j-1}, \Delta^2\gamma_{0j} = \Delta(\gamma_{0j} - \gamma_{0,j-1}) = \gamma_{0,j} - 2\gamma_{0,j-1} + \gamma_{0,j+2}$. The penalty has the effect that the parameters are estimated smoothly with the degree of smoothness determined by the tuning parameter λ_s. If a penalty of order d is used and the degree of the B-spline is higher than d, for large values of λ the fit will approach a polynomial of degree $d-1$.

By including an additional term that ensures that the baseline hazard parameters are estimable one obtains an additional tuning parameter. In addition to λ one has to select a value of λ_s. But in simulations we found that in general it is not worthwhile to select both parameters by cross-validation (though especially in very high-dimensional settings performance could be further improved, if also λ_s is selected on a grid of possible values). While some care should be taken to select λ, which determines the performance of the selection procedure, it suffices to choose a small value for λ_s. In the application we chose $\lambda_s = 0.1$ and in our simulations we chose $\lambda_s \in \{0.01, 0.05, 0.1, 0.5, 1\}$, depending on the choice of model parameters. Note here that in general for the ridge type penalty higher values of λ_s had to be chosen than for the expansion in basis functions.

4 Simulation Study

The underlying model is a random intercept logit model for discrete survival data with predictor

$$\eta_{it} = \gamma_{0t} + \sum_{j=1}^{p} x_{ij}\gamma_j + b_i, \quad i = 1, \ldots, n, \quad t = 1, \ldots, q.$$

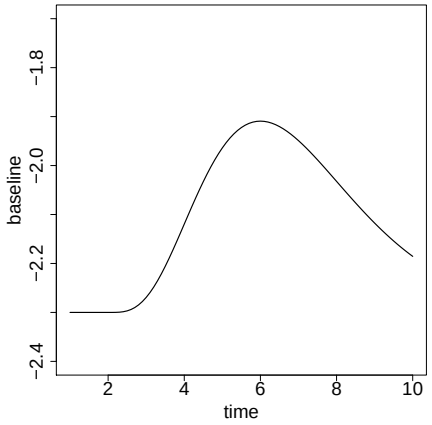

FIGURE 1: *Shape of the discrete baseline function* $\gamma_{01}, \ldots, \gamma_{0q}$

The baseline hazard is specified by $\gamma_{0t} = 2\xi_t - 2.3$, where $\xi_t = f_\Gamma(t-2)$ with $f_\Gamma(t)$ denoting the density of a Gamma distribution $\Gamma(\zeta, \theta)$. Shape and scale parameter were chosen as $\zeta = 5, \theta = 1$. This results in a baseline function with a moderate hump. It is shown in Figure 1.

The linear effects are given by $\gamma_1 = -4, \gamma_2 = -4, \gamma_3 = -4, \gamma_4 = 6, \gamma_5 = 6$ and $\gamma_j = 0, j = 6, \ldots, 500$. We chose the settings $p = 5, 50, 100, 500$. For $j = 1, \ldots, p$ the vectors $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{ip})$ were drawn independently with components following a uniform distribution within the interval $[0, 1]$. The number of clusters was set to $n = 100$ and the random effects were specified by $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b \in \{0, 1, 2\}$. The censoring probabilities were chosen as $\pi_{cens} \in \{0.05, 0.2\}$. Finally, for each subject $i = 1, \ldots, 100$, we conduct the following simulation scheme.
For $t = 1, \ldots, q$:

(a) simulate a binary response variable by $y_{it} \sim \mathrm{B}(1, \lambda(t|\boldsymbol{x}_i))$;

(b) (b1) if $y_{it} = 1$ stop and set $T_i = t$;

(b2) else, draw a censoring random variable $C \sim \mathrm{B}(1, \pi_{cens})$; if $C = 1$ stop and set $T_i = t$, else set $t = t + 1$;

All parameters and interval boundaries have been chosen such that the discrete distribution of the number of repeated measurements T_i (which represents the event or censoring time, respectively) is comparatively balanced.

In the following we use the `glmmLasso` algorithm (Groll and Tutz, 2014) for two different fitting approaches: in the first approach (denoted by $\texttt{glmmLasso}_{dis}$) we fit the model directly, using a small ridge penalty on the parameters $\gamma_{0t}, t =$

$1, \ldots, k$ and by specifying suitable design matrices. The second approach (denoted by `glmmLasso`$_{smooth}$) fits a GLMM together with a smooth baseline hazard based on penalized B-spline expansion (Eilers and Marx, 1996). For both L_1-regularized approaches the optimal tuning parameter λ has been determined using the Bayesian Information Criterion (BIC, see Schwarz, 1978).

Fitting of the model can be obtained by various functions. After constructing the appropriate design matrix, the R-functions `glmmPQL` (Venables and Ripley, 2002), `glmmML` (Broström, 2009) and `glmer` (Bates and Maechler, 2010) are able to fit the model. The `glmmPQL` routine is provided by the `MASS` library. It operates by iteratively calling the R-function `lme` from the `nlme` library and returns the fitted `lme` model object for the working model at convergence. For more details about the `lme` function, see Pinheiro and Bates (2000). The `glmmML` function is available in the `glmmML` package (Broström, 2009) and features two different methods of approximating the integrals in the log-likelihood function, Laplace and Gauss-Hermite, whereas for the first method the results coincide with the results of the `glmmPQL` routine. Unfortunately, for both functions no model testing methods are available, thus no subset selection procedures can be performed.

Fitting of Frailty Models with Variable Selection

With the focus on variable selection one can use the `glmer` function from the `lme4` package (Bates and Maechler, 2010), which provides model testing based on an analysis of deviance. We use forward subset selection in order to perform variable selection and compare the results with our `glmmLasso` algorithm, implemented in the corresponding R-package `glmmLasso` (Groll, 2011). We restrict consideration to forward procedures because forward/backward procedures imply huge computational costs. It should be mentioned that the `glmer` function also features two different methods of approximating the integrals in the log-likelihood function, Laplace and adaptive Gauss-Hermite. We focused on the former and call the corresponding forward selection procedure `glmer-select`.

An alternative that fits GLMM together with a smooth baseline hazard is provided by the `gamm4` function in the corresponding package `gamm4` (Wood and Scheipl, 2013). As the function also provides model testing based on an analysis of deviance, one can use forward subset selection in order to obtain variable selection. The corresponding procedure is called `gamm4-select`.

Fitting of Basic Model with Variable Selection

A function that is able to fit the model without random effects after specification of an appropriate design matrix is the `glmnet` function from the corresponding package `glmnet` (Friedman et al., 2010). The function allows to fit an ordinary GLM with Lasso regularization, and therefore selection of variables, but no random effects can be incorporated. Hence, in the method called `glmnet` no estimates for the random effects variance σ_b^2 are available.

A similar Lasso-based regularization approach was provided by Goeman (2010) and is implemented in the package `penalized` (Goeman, 2011). The function is also designed for ordinary GLMs and again, no estimates for the random effects variance σ_b^2 are available. It is simply called `goeman`.

Finally, a GLM together with a smooth baseline hazard can be fitted by use of the `gam` function in the R-package `mgcv` (Wood, 2006). Similar to `gamm4`, the function also provides model testing based on an analysis of deviance and one can use forward subset selection. The method is called `gam-select`.

Evaluation of Performance

The performance of estimators was evaluated separately for the structural components and the variance. By averaging across 50 training data sets we consider mean squared errors for $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0k})^T$, $\boldsymbol{\gamma}$ and σ_b given by $\text{mse}_0 := ||\boldsymbol{\gamma}_0 - \hat{\boldsymbol{\gamma}}_0||^2$, $\text{mse}_{\boldsymbol{\gamma}} := ||\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}||^2$, $\text{mse}_{\sigma_b} := (\sigma_b - \hat{\sigma}_b)^2$. The means of these quantities are presented in Tables 1 to 3 together with the corresponding standard errors in brackets. The results for varying proportions of censoring are rather similar. Therefore we give the results for the low censoring case ($\pi_{cens} = 0.05$) only.

It is seen that in terms of mse_0 the two versions of `glmmLasso` perform very well (except for single outliers in the $p = 500$ settings: here the performance could be considerably improved, if also λ_s would be selected on a grid of possible values). Among the three methods that do not account for heterogeneity `gam-select` works best for small p, but `glmnet` outperforms the others for larger p. Both methods that fit frailty models, i.e. `glmer-select` and `gamm4-select`, show nonsatisfying results with respect to all considered performance measures, with some exceptions in the $p = 5$ scenarios. Note that we do not present the `gam-select` and `gamm4-select` results for $p = 500$, because the methods were quite slow[1] and anyhow the results were not satisfactory.

In terms of $\text{mse}_{\boldsymbol{\gamma}}$, which is more important since it refers to the estimation of variable effects, the picture is very similar, with the exception of `goeman`, which now is among the best performers, but showed very bad results for mse_0. The comparison with the procedure $\texttt{glmmLasso}_{dis}$ is visualized in Figures 2 and 3. It shows that the bad performance of the procedures that fit frailty models is mainly due to dab fits in some data sets. The estimation procedures are rather unstable so that for some data sets the performance deteriorates. In contrast, the methods that ignore the frailty are more stable. The exception are the lasso based procedures, which are stable and show good performance. Overall it is seen that all methods suffer if the model contains a heterogeneity component.

In terms of mse_{σ_b} (Table 3), the `glmmLasso` versions provide very good results if no or only a small random component is in the model, but underestimate the heterogeneity in the case $\sigma_b = 2$.

[1] On a MacBook Pro with 2.5 GHz Intel Core i5 processor the methods managed about 10 simulation runs per week.

Additional information on the performance of the algorithm was collected in *falseneg* (f.n.), the mean over all 50 simulations of the number of variables $\gamma_j, j = 1, 2, 3, 4, 5$, that were not selected and in *falsepos* (f.p.), the mean over all 50 simulations of the number of variables $\gamma_j, j = 6, \ldots, p$, that were selected. In addition, we give the number of simulation runs, where the fitting procedure did not converge (or produced MSE-results bigger than 10^6; denoted by "n.c."), see Table 4. In terms of false positives and negatives none of the other procedures comes close to the performance of the lasso based fits. `goeman` and `gam-select` have very large false positive rates, `glmnet` shows very large false negative rates. The frailty models have too large false positive rates although false negative rates are low. Their performance is comparable to that of `gam-select`. It is surprising that also in cases with frailty in the data generating models the performance is only slightly better than for `gam-select`, which ignores the heterogeneity. But it should be noted that `gam-select` did not converge in several scenarios, in particular in the high censoring rate setting the number of failed fits is larger than for the models that include heterogeneity.

Finally, in order to make the different approaches comparable with respect to computational efficiency, we present results of the average computational times (in minutes; including the determination of the tuning parameter for the L_1-regularized approaches and including the forward selection procedure for `glmer`, `gamm4` and `gam`), exemplarily for $\sigma_b = 0$ and $\pi_{cens} = 0.2$ in Table 5. It is seen that for all forward selection procedures the computational time is disproportionally increasing with the number of noise variables in comparison to the other methods.

		`glmer-select`	`gamm4-select`	`glmnet`	`goeman`	`gam-select`	`glmmLasso`$_{dis}$	`glmmLasso`$_{smooth}$
σ_b	p	mse_0	mse_0	mse_0	mse_0	mse_0	mse_0	mse_0
0	5	618.47 (1674.29)	42.72 (151.20)	39.75 (50.83)	815.30 (657.87)	18.83 (30.31)	13.75 (10.71)	10.10 (10.32)
0	50	1319.15 (2127.41)	13272.82 (52518.12)	40.65 (48.85)	742.62 (689.83)	167.58 (460.30)	34.20 (75.87)	17.17 (20.57)
0	100	5540.79 (6907.83)	118966.60 (221940.20)	38.37 (47.78)	710.52 (629.62)	144.86 (251.92)	106.70 (161.11)	74.35 (145.34)
0	500	- (-)	- (-)	35.94 (50.58)	628.08 (577.70)	- (-)	139.01 (206.04)	347.70 (1409.34)
1	5	896.18 (2738.33)	334.98 (1889.53)	37.70 (57.11)	825.18 (829.33)	19.03 (33.75)	19.72 (35.50)	11.50 (11.95)
1	50	3447.94 (9975.25)	29597.94 (150606.60)	43.05 (65.74)	662.65 (552.47)	99.19 (242.47)	75.05 (150.14)	38.60 (55.22)
1	100	23559.33 (86306.28)	109559.10 (245934.90)	31.38 (54.13)	668.47 (616.66)	125.88 (177.88)	88.27 (147.48)	77.95 (124.79)
1	500	- (-)	- (-)	24.86 (50.41)	650.64 (754.18)	- (-)	132.41 (201.37)	105.67 (147.98)
2	5	552.59 (1326.73)	118.20 (667.26)	28.62 (59.42)	762.90 (802.95)	20.15 (26.89)	23.92 (19.88)	17.20 (23.43)
2	50	2432.41 (4937.37)	369.62 (1088.57)	32.53 (61.42)	566.27 (497.25)	107.94 (240.01)	54.81 (77.12)	31.64 (33.14)
2	100	5200.14 (7926.65)	128396.50 (244808.80)	29.51 (59.98)	589.93 (555.21)	269.84 (559.60)	82.42 (87.24)	80.68 (111.01)
2	500	- (-)	- (-)	27.34 (58.56)	532.20 (463.00)	- (-)	134.90 (295.95)	71.60 (93.96)

TABLE 1: *Results for* mse_0 *for* `glmmLasso` *and alternative approaches (standard errors in brackets) with low censoring rate* ($\pi_{cens} = 0.05$)

5 Application

In the following we will illustrate the proposed method on a real data set that is based on Germany's current panel analysis of intimate relationships and family

σ_b	p	`glmer-select` mse_γ	`gamm4-select` mse_γ	`glmnet` mse_γ	`goeman` mse_γ	`gam-select` mse_γ	`glmmLasso`$_{dis}$ mse_γ	`glmmLasso`$_{smooth}$ mse_γ
0	5	450.52 (1782.90)	55.64 (251.93)	77.86 (36.04)	45.28 (25.74)	15.19 (20.69)	5.81 (5.76)	6.73 (6.25)
0	50	6217.44 (21150.96)	23582.23 (52518.12)	83.19 (30.97)	52.25 (23.39)	148.84 (451.86)	15.30 (18.50)	13.87 (14.16)
0	100	11682.99 (13842.38)	125906.90 (157816.40)	84.04 (28.98)	54.94 (21.74)	224.83 (379.39)	27.10 (29.92)	25.27 (24.51)
0	500	- (-)	- (-)	97.86 (18.14)	68.12 (13.63)	- (-)	50.05 (25.72)	106.96 (357.44)
1	5	574.02 (1935.87)	48.19 (240.37)	78.13 (39.08)	55.69 (24.75)	16.72 (24.19)	7.46 (10.94)	6.05 (4.32)
1	50	4699.25 (8465.54)	14033.57 (70491.35)	87.37 (30.34)	59.93 (22.37)	91.81 (205.29)	26.49 (23.79)	24.84 (35.66)
1	100	4443.44 (163900.60)	122569.00 (195076.30)	94.37 (23.68)	62.52 (19.58)	117.87 (101.59)	31.30 (28.56)	41.63 (50.31)
1	500	- (-)	- (-)	104.17 (14.24)	74.94 (12.39)	- (-)	55.82 (29.12)	76.95 (104.04)
2	5	269.12 (904.73)	88.27 (322.12)	101.60 (20.44)	79.20 (19.93)	36.36 (24.95)	29.10 (17.57)	27.20 (16.78)
2	50	2894.16 (4972.34)	135.90 (335.34)	104.00 (17.07)	85.50 (13.38)	52.43 (45.95)	43.39 (22.91)	39.30 (22.63)
2	100	8534.39 (12730.98)	156636.10 (219303.00)	107.25 (15.59)	86.65 (12.84)	198.69 (419.13)	50.79 (24.68)	45.21 (24.68)
2	500	- (-)	- (-)	108.80 (10.40)	92.57 (10.80)	- (-)	77.06 (25.48)	74.47 (25.50)

TABLE 2: *Results for* mse_γ *for* `glmmLasso` *and alternative approaches (standard errors in brackets) with low censoring rate* ($\pi_{cens} = 0.05$)

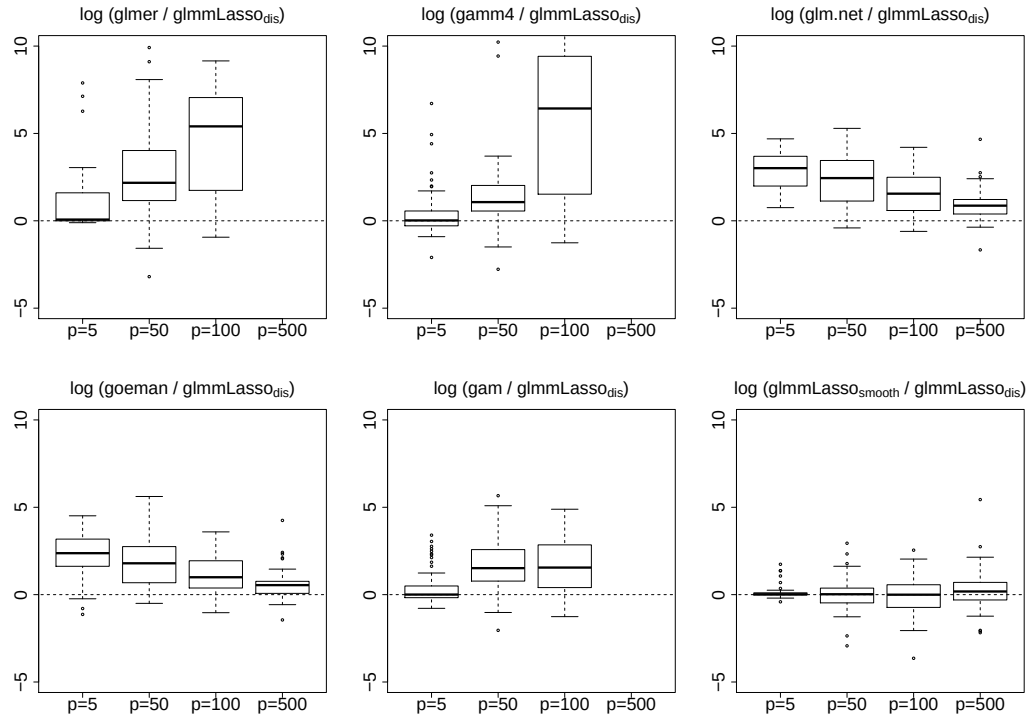

FIGURE 2: *Boxplots of* $\log(mse_\gamma(\cdot)/mse_\gamma(\texttt{glmmLasso}_{dis}))$ *for* `glmmLasso` *and alternative approaches for low censoring rate* ($\pi_{cens} = 0.05, \sigma_b = 0$)

dynamics (pairfam), release 4.0 (Nauck et al., 2013). The panel was started in 2008 and contains about 12.000 randomly chosen respondents from the birth cohorts 1971-73, 1981-83 and 1991-93. Pairfam follows the cohort approach, that is, the main focus is on an anchor person of a certain birth cohort, who provides detailed information, orientations and attitudes (mainly with regard to their family plans) of both partners in interviews that are conducted yearly. A

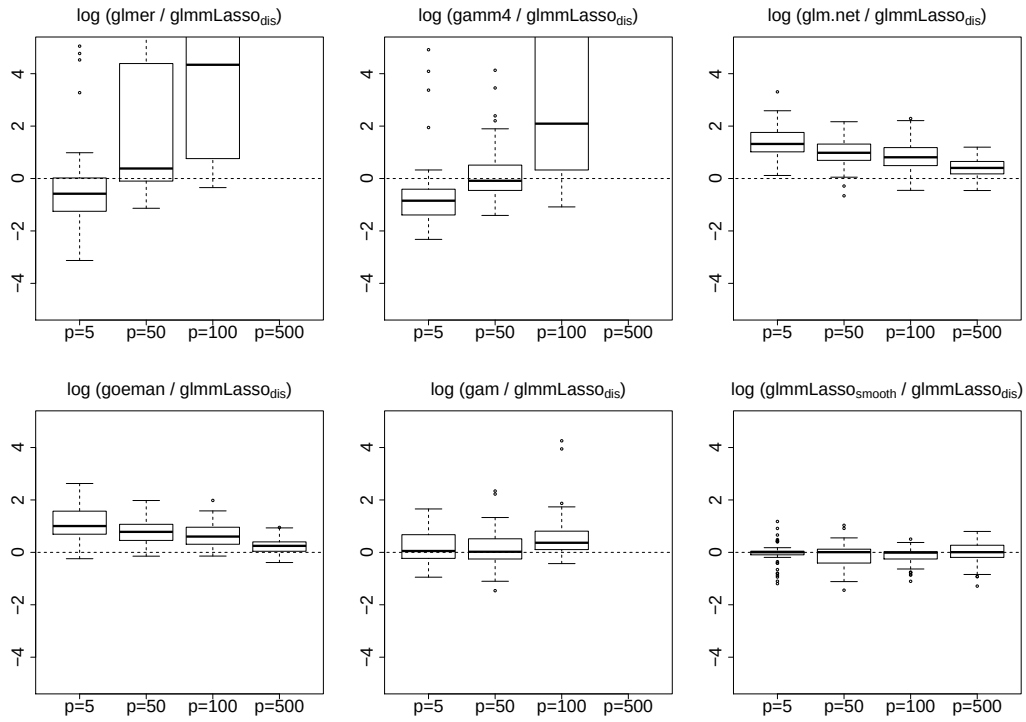

FIGURE 3: *Boxplots of* $\log(mse_{\boldsymbol{\gamma}}(\cdot)/mse_{\boldsymbol{\gamma}}(\texttt{glmmLasso}_{dis}))$ *for* `glmmLasso` *and alternative approaches for low censoring rate* ($\pi_{cens} = 0.05, \sigma_b = 2$)

		`glmer-select`	`gamm4-select`	$\texttt{glmmLasso}_{dis}$	$\texttt{glmmLasso}_{smooth}$
σ_b	p	mse_{σ_b}	mse_{σ_b}	mse_{σ_b}	mse_{σ_b}
0	5	22.93 (91.79)	.80 (1.86)	.12 (.19)	.17 (.27)
0	50	91.37 (179.33)	1003.398 (3985.08)	.15 (.22)	.15 (.29)
0	100	244.83 (251.55)	1458.40 (5951.23)	.11 (.15)	.14 (.29)
0	500	- (-)	- (-)	.20 (.28)	.09 (.01)
1	5	23.47 (81.63)	.48 (1.21)	.43 (.20)	.43 (.17)
1	50	119.75 (189.64)	464.56 (3275.09)	.34 (.18)	.47 (.13)
1	100	299.95 (308.86)	864.78 (3780.48)	.47 (.12)	.46 (.12)
1	500	- (-)	- (-)	.31 (.19)	.48 (.06)
2	5	18.10 (63.14)	1.12 (1.11)	2.49 (.79)	2.39 (.81)
2	50	118.84 (175.74)	8.40 (26.64)	2.17 (.95)	2.82 (.46)
2	100	212.45 (207.96)	1563.27 (8162.37)	2.88 (.13)	2.88 (.14)
2	500	- (-)	- (-)	2.58 (.53)	2.89 (.03)

TABLE 3: *Results for* mse_{σ_b} *for* `glmmLasso` *and alternative approaches (standard errors in brackets) with low censoring rate* ($\pi_{cens} = 0.05$)

detailed description of the study can be found in Huinink et al. (2011).

The present data set was created by Jasmin Abedieh during her master thesis, in order to analyze the relationship between the transition into parenthood and the leisure behavior of the parents. It consists of a sample of 1238 observations stemming from 690 anchor women living in Germany, who have participated in

		glmer-select			gamm4-select			glmnet			goeman			gam-select			glmmLasso$_{dis}$			glmmLasso$_{smooth}$		
σ_b	p	n.c.	f.p.	f.n.	n.c.	f.p.	f.n.	n.c.	f.p.	f.n.	n.c.	f.p.	f.n.	n.c.	f.p.	f.n.	n.c.	f.p.	f.n.	n.c.	f.p.	f.n.
0	5	0	0	0.06	0	0	0.16	1	0	1.96	1	0	0.18	0	0	0.44	0	0	0	0	0	0.04
0	50	1	3.69	0.12	1	3.27	0.14	1	1.18	2.02	2	4.31	0.25	1	4.33	0.20	0	0.98	0.26	0	0.72	0.14
0	100	24	5.12	0.23	4	5.72	0.17	1	2.61	1.96	0	5.74	0.28	29	6.10	0.10	0	1.96	0.52	0	1.26	0.34
0	500	-	-	-	-	-	-	1	1.04	2.69	0	8.60	0.44	50	-	-	0	1.80	1.80	0	2.74	1.40
1	5	0	0	0.06	0	0	0.16	0	0	2.16	2	0	0.25	0	0	0.44	0	0	0.12	0	0	0.04
1	50	1	3.96	0.16	0	4.82	0.16	0	1.62	2.32	2	5.50	0.27	1	4.53	0.16	0	1.26	0.78	0	1.26	0.50
1	100	14	5.89	0.39	2	11.52	0.21	0	0.76	2.60	2	6.54	0.31	26	6.58	0.25	0	2.04	0.70	0	1.92	0.56
1	500	-	-	-	-	-	-	0	0.76	3.14	1	9.65	0.55	50	-	-	0	2.88	1.92	0	3.36	1.54
2	5	0	0	0.30	0	0	0.28	0	0	3.04	0	0	0.80	0	0	0.70	0	0	0.58	0	0	0.44
2	50	0	4.24	0.46	0	4.08	0.36	0	0.44	3.20	0	2.30	0.98	0	5.12	0.44	0	1.42	1.26	0	0.78	1.14
2	100	15	7.29	0.54	6	12.80	0.45	0	0.74	3.58	0	3.74	1.06	23	7.56	0.44	0	1.52	1.60	0	1.86	1.20
2	500	-	-	-	-	-	-	0	0.56	3.62	1	7.82	1.51	50	-	-	0	2.84	1.94	0	1.64	2.72

TABLE 4: *Number of simulation runs, where the fitting procedures did not converge (n.c.) together with false positives (f.p.) and false negatives (f.n.) for* `glmmLasso` *and alternative approaches for low censoring rate ($\pi_{cens} = 0.05$)*

		glmer-select	gamm4-select	glmnet	goeman	gam-select	glmmLasso$_{dis}$	glmmLasso$_{smooth}$
σ_b	p	mse$_{\boldsymbol{\alpha}}$	mse$_{\boldsymbol{\alpha}}$	mse$_{\boldsymbol{\alpha}}$	mse$_{\boldsymbol{\alpha}}$	mse$_{\boldsymbol{\alpha}}$	mse$_{\boldsymbol{\alpha}}$	mse$_{\boldsymbol{\alpha}}$
0	5	0.85	0.51	0.11	0.16	0.02	1.30	1.4
0	50	33.92	31.26	0.19	0.31	0.62	1.27	1.65
0	100	116.59	140.09	0.24	0.39	1.49	1.75	1.75
0	500	> 300	> 300	0.37	1.11	-	25.49	24.96

TABLE 5: *Results for average computational times (in minutes) for* `glmmLasso` *and alternative approaches with high censoring rate ($\pi_{cens} = 0.2$)*

at least two of the first four pairfam waves and who fulfill the following criteria:

- they live in a relationship;
- their date of birth is available;
- both partners are generative, heterosexual, childless and not pregnant at the time of the interview.

In addition to several control variables explanatory variables that describe the leisure behavior of the women are included. A detailed description of all considered covariates is given in Table 6. The data set originally contained some missing values, which have been imputed by a simple last value carried forward method[2]. Since the empirical distributions of the variables *reldur* and *leisure* are quite skewed we have transformed them via the third square-root (which is similar to taking the logarithm).

[2]Also multiple imputation techniques have been used, which are implemented in the software R e.g. in the packages `mi` (Gelman et al., 2013) and `mice` (van Buuren and Groothuis-Oudshoorn, 2013). As in the work of Abedieh it was shown that all different imputation techniques lead to almost indistinguishable results, our analysis is based on the data set obtained via the last value carried forward method. For a very helpful description of the MICE-technique together with illustrative examples, see van Buuren and Groothuis-Oudshoorn (2011).

Variable	Description
child	a dummy ($\in \{0, 1\}$) indicating, if the woman gave birth to her first child within the regarded interval (or is currently pregnant);
age	age (in years) of the anchor woman;
page	age (in years) of the male partner;
sat6	degree of live satisfaction ($\in \{0, 1, ..., 10\}$) of the anchor woman;
reldur	duration of the relationship (in months);
relstat	status of relationship (categorical with three levels: "living apart together", "cohabit", "married");
yeduc	years of education ($\in [8, 20]$) of the anchor woman;
pyeduc	years of education ($\in [8, 20]$) of the male partner;
casprim	employment status of the anchor woman (categorical with five levels: "in education", "full-time employed", "part-time employed", "non-working", "other");
pcasprim	employment status of the male partner (categorical — see *casprim*);
siblings	number of siblings of the anchor woman
hlt7	average sleep length of the anchor woman (in hours)
leisure	(approx.) yearly leisure time of the anchor woman (in hours) spent for the following five major categories: 1) bar/cafe/restaurant; 2) sport; 3) internet/tv; 4) meet friends; 5) discotheque;
leisure.partner	relative proportion ($\in [0, 1]$) of *leisure* that the partner spends together with the anchor woman
holiday	time of the anchor woman (in weeks) spent for holiday

TABLE 6: *Description of covariates for the pairfam data: response (top), control (middle) and leisure variables (bottom).*

For each of the anchor women from the two age groups $[24; 30]$ and $[34; 40]$ it is known if she has given birth to her first child within the year between two interview dates; altogether, 137 events are observed. We consider years as the unit in our discrete-survival model starting with 24 years, which is the age of the youngest woman in the sample. The baseline hazard, which corresponds to the effect of age, is certainly non-linear. Therefore, it is included in the form of a penalized smooth effect, which should be able to model the observation gap within ages. Similarly, we allow for a non-linear effect of their male partner's age, simply by including higher potencies of this covariate. For the categorical variables *relstat*, *casprim* and *pcasprim* the reference levels "living apart together" and "non-working", respectively, are chosen. It should be noted that all variables can vary over time an are included as time-varying. Including random intercepts

b_i for heteroscedastic baseline hazards, the following model is fitted:

$$\begin{aligned}\lambda(t|\boldsymbol{x}_{it}, b_i) \;=\; h\Big(& f_0(age) + siblings_{it}\,\beta_1 + sat6_{it}\,\beta_2 + relstat.cohab_{it}\,\beta_3 \\ & + relstat.married_{it}\,\beta_4 + yeduc_{it}\,\beta_5 + pyeduc_{it}\,\beta_6 \\ & + casprim.educ_{it}\,\beta_7 + casprim.fulltime_{it}\,\beta_8 \\ & + casprim.parttime_{it}\,\beta_9 + casprim.other_{it}\,\beta_{10} \\ & + pcasprim.educ_{it}\,\beta_{11} + pcasprim.fulltime_{it}\,\beta_{12} \\ & + pcasprim.parttime_{it}\,\beta_{13} + pcasprim.other_{it}\,\beta_{14} \\ & + page_{it}\,\beta_{15} + page_{it}^2\,\beta_{16} + page_{it}^3\,\beta_{17} \\ & + page_{it}^4\,\beta_{18} + (reldur_{it}^{(1/3)})\,\beta_{19} + (leisure^{(1/3)})_{it}\,\beta_{20} \\ & + leisure.partner_{it}\,\beta_{21} + hlt7_{it}\,\beta_{22} + holiday_{it}\,\beta_{23} + b_i \Big)\,,\end{aligned} \tag{6}$$

In principle the model can be fitted by `gam`, `gamm4` and `glmmLasso`, but for `gamm4` there was a warning message saying that the procedure did not converge. In addition, the variance of the random effect was absurdly large (6.38). Therefore, we do not give the fitted values for this procedure. Due to the presence of categorical covariates, we abstained from performing forward subset selection for `gam` and instead refitted the model containing only those covariates that turned out to be significant in a first fit of the full model ($\alpha = 0.01$). To demonstrate the difference between AIC and BIC with regard to model sparsity, we used both criteria to select the tuning parameter λ in `glmmLasso`. The additional tuning parameter controlling the smoothness of the baseline hazard has been fixed to $\lambda_s = 0.1$. The results of the estimation of fixed effects and the level of heterogeneity $\hat{\sigma}_b$ are given in Table 7 and the corresponding coefficient built-ups (though before performing the final Fisher scoring re-estimation step implemented in the `glmmLasso` function) are illustrated in Figure 4.

All used approaches include the status of the relationship, with higher hazard rates obtained for stronger forms of the relationship. With `gam` and `glmmLasso` with AIC also the time spent for holidays was found to have a positive effect. When using `glmmLasso` with AIC in addition the employment status of the women was found to be influential, with all categories having a positive effect on the transition into motherhood when compared to the reference level "non-working". The effect is strongest for women with a full-time job. The fitting of models that account for heterogeneity support that the frailty should be included in the model.

	`gam`	`glmmLasso`$_{smooth}$ (AIC)	`glmmLasso`$_{smooth}$ (BIC)
intercept	-3.24 (0.43)	-1.43 (3.94)	-1.40 (3.92)
page	-	-	-
page^2	-	-	-
page^3	-	-	-
page^4	-	-	-
hlt7	-	-	-
sat6	-	-	-
reldur$^{(1/3)}$	-	-	-
siblings	-	-	-
relstat:cohab	1.03 (0.30)	0.50 (0.15)	0.52 (0.15)
relstat:married	1.95 (0.32)	0.81 (0.14)	0.82 (0.13)
yeduc	-	-	-
pyeduc	-	-0.21 (0.10)	-
leisure$^{(1/3)}$	-	-	-
leisure.partner	-	-	-
holiday	0.18 (0.08)	0.21 (0.09)	-
casprim:educ	-	0.39 (0.46)	-
casprim:fulltime	-	0.74 (0.50)	-
casprim:parttime	-	0.38 (0.26)	-
casprim:other	-	0.17 (0.21)	-
pcasprim:educ	-	-	-
pcasprim:fulltime	-	-	-
pcasprim:parttime	-	-	-
pcasprim:other	-	-	-
$\hat{\sigma}_b$	-	0.53	0.53

TABLE 7: *Estimated linear effects and standard deviation of the random intercept for the pairfam data with* `gam` *and* `glmmLasso` *(standard errors in brackets).*

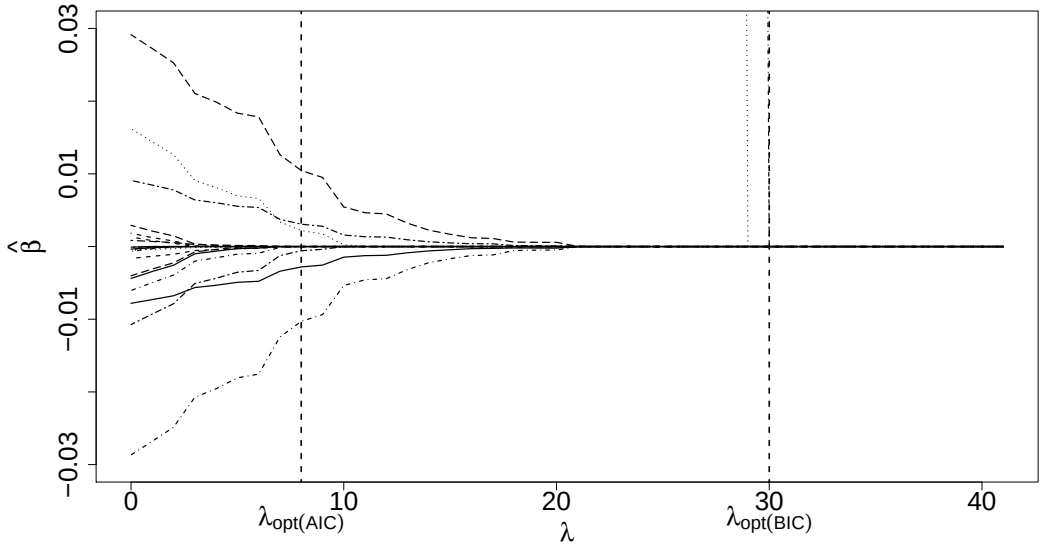

FIGURE 4: *Coefficient built-ups for* `glmmLasso` *for the pairfam data (before running the final Fisher scoring re-estimation step); the optimal value of the penalty parameter* λ *with AIC/BIC is shown by the two dashed vertical lines.*

In Figure 5 the estimates of the smoothed baseline hazard in terms of the variable "age" are shown. While for `gam` and `gamm4` straigth lines are fitted, the results of `glmmLasso` of manifest the typical non-linear and bell-shaped course with a maximum in the mid-twenties (grey line: AIC, black line: BIC).

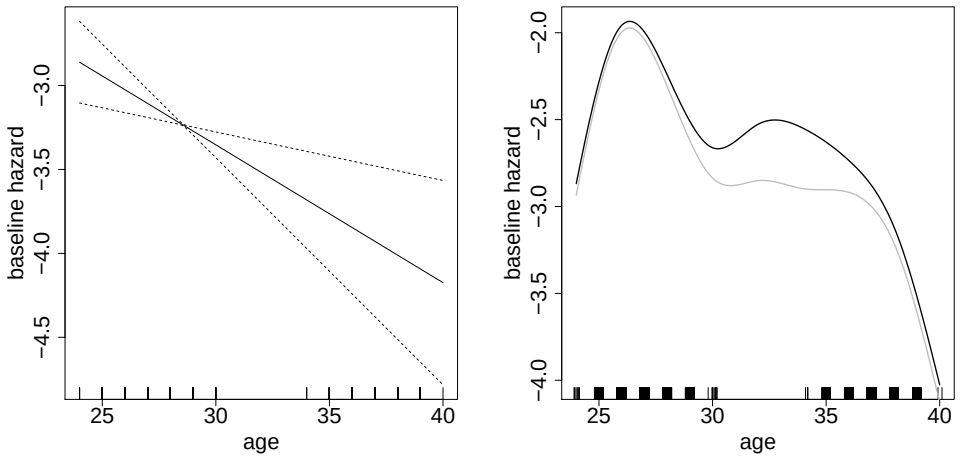

FIGURE 5: *Estimates of the smoothed baseline hazard in terms of "age" for* `gam` *(left)and* `glmmLasso` *(right; grey line: AIC, black line: BIC) for the pairfam data.*

6 Concluding Remarks

Procedures for efficient variable selection in discrete survival modeling including heterogeneity have been proposed that are based on a combination of ridge and lasso type penalties. The performance of the procedures was studied in simulation studies and an application to the birth of the first child. It turned out that the procedures yield stable estimates in cases where methods that do not include variable selection typically fail because of the complexity of the fitting task.

More precisely, the simulations have shown that also simple forward, backward and forward/backward procedures are clearly outperformed, regardless of whether the heterogeneity was accounted for in the fitting approaches or not. These simple attempts to variable selection only work well when few covariates are present. For a large number of covariates they are very time-consuming and either produce high false positive rates or exorbitant MSEs and hence, are not useful.

Lasso-based regularization approaches designed for ordinary GLMs and therefore ignoring heterogeneity, such as the `glmnet` and the `penalized` R-packages, performed unexpectedly well, even in scenarios with frailty in the data generating models. This is somewhat surprising as it has been shown that both the hazard rates (see for example Heckman and Singer, 1984) and the effects of the observed covariates (see for example Lancaster, 1990, Van den Berg, 2001) tend

to be biased, if unobserved heterogeneity is disregarded. However, `glmnet` produces clearly higher false negative rates than our proposed lasso approach, which includes heterogeneity, whereas the `penalized` package has problems to fit the baseline hazard. In summary, our proposed lasso approach represents a strong competitor for efficient variable selection in discrete survival modeling including heterogeneity.

As in several applications the influence of some covariates may change over time, a worthy extension could be to adapt the proposed variable selection procedures to models including time-varying effects. In these models, more sophisticated problems of model selection arise, as one has to determine which covariates should be included in the model, or, which of the covariates included have a time-varying effect. So a future objective is to develop penalization approaches for variable selection in discrete survival frailty models with time-varying coefficients, such that single varying effects are either included, are included in the form of a constant effect or are totally excluded. The main challenge is the construction of appropriate penalty terms that are able to distinguish between these effects.

Acknowledgements

This article uses data from the German family panel pairfam, coordinated by Josef Brüderl, Johannes Huinink, Bernhard Nauck, and Sabine Walper. Pairfam is funded as long-term project by the German Research Foundation (DFG). Besides, we are grateful to Jasmin Abedieh for providing the specific discrete survival data, which were constructed from the pairfam data and were part of her master thesis.

References

Anderson, D. A. and M. Aitkin (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society Series B 47*, 203–210.

Baker, M. and A. Melino (2000). Duration dependence and nonparametric heterogeneity: A monte carlo study. *Journal of Econometrics 96*, 357–393.

Bates, D. and M. Maechler (2010). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-0.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

Breslow, N. E. and X. Lin (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

Broström, G. (2009). *glmmML: Generalized linear models with clustering*. R package version 0.81-6.

Brown, C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics 31*, 863–872.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B 34*, 187–220.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Oxford: Oxford Science Publications.

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

Fahrmeir, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data. *Biometrika 81*, 317–330.

Fahrmeir, L. and T. Kneib (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Cambridge University Press.

Fahrmeir, L. and L. Knorr-Held (1997). Dynamic discrete-time duration models: Estimation via markov chain monte carlo. *Sociological Methodology 27*(1), 417–452.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gelman, A., J. Hill, Y. Su, M. Yajima, and M. G. Pittau (2013). *mi: Missing Data Imputation and Model Checking*. R package version 0.09-18.03.

Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal 52*, 70–84.

Goeman, J. J. (2011). *Penalized*. R package version 0.9-42.

Groll, A. (2011). *glmmLasso: Variable Selection for generalized linear mixed models by L_1-penalized estimation*. R package version 1.2.3.

Groll, A. and G. Tutz (2014). Variable selection for generalized linear mixed models by L_1-penalized estimation. *Statistics and Computing 24*(2), 137–154.

Ham, J. C. and S. A. Rea Jr (1987). Unemployment insurance and male unemployment duration in canada. *Journal of Labor Economics*, 325–353.

Hartzel, J., I. Liu, and A. Agresti (2001). Describing heterogenous effects in stratified ordinal contingency tables, with applications to multi-center clinical trials. *Computational Statistics & Data Analysis 35*(4), 429–449.

Heckman, J. J. and B. Singer (1984). Econometric duration analysis. *Journal of Econometrics 24*(1), 63–132.

Hinde, J. (1982). Compound poisson regression models. In R. Gilchrist (Ed.), *GLIM 1982 International Conference on Generalized Linear Models*, pp. 109–121. New York: Springer-Verlag.

Huinink, J., J. Brüderl, B. Nauck, S. Walper, L. Castiglioni, and M. Feldhaus (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Journal of Family Research 23*, 77–101.

Kalbfleisch, J. and R. Prentice (2002). *The Statistical Analysis of Failure Time Data (2ned.)*. New York: Wiley.

Kauermann, G., G. Tutz, and J. Brüderl (2005). The survival of newly founded firms: A case-study into varying-coefficient models. *Journal of the Royal Statistical Society A 168*, 145–158.

Laird, N. and D. Olivier (1981). Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. *Journal of the American Statistical Association 76*(374), 231–240.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. New York: Cambridge University Press.

Land, K. C., D. S. Nagin, and P. L. McCall (2001). Discrete-time hazard regression models with hidden heterogeneity the semiparametric mixed poisson regression approach. *Sociological methods & research 29*(3), 342–373.

Lin, X. and N. E. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*, 1007–1016.

Littell, R., G. Milliken, W. Stroup, and R. Wolfinger (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

Liu, Q. and D. A. Pierce (1994). A note on Gauss-Hermite quadrature. *Biometrika 81*, 624–629.

Nauck, B., J. Brüderl, J. Huinink, and S. Walper (2013). The german family panel (pairfam). *GESIS Data Archive, Cologne*. ZA5678 Data file Version 4.0.0.

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), 1–13.

Park, M. Y. and T. Hastie (2007). An l1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society B 69*, 659–677.

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.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-Plus*. New York: Springer.

Prentice, R. L. and L. A. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics 34*, 57–67.

Scheike, T. and T. Jensen (1997). A discrete survival model with random effects: an application to time to pregnancy. *Biometrics*, 318–329.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software 39(5)*, 1–13.

Therneau, T. and P. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag.

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 B 58*, 267–288.

Tutz, G. and L. Pritscher (1996). Nonparametric estimation of discrete hazard functions. *Lifetime Data Analysis 2*, 291–308.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(3), 1–67.

van Buuren, S. and K. Groothuis-Oudshoorn (2013). *mice: Multivariate Imputation by Chained Equations in R*. R package version 2.18.

Van den Berg, G. J. (2001). Duration models: specification, identification and multiple durations. *Handbook of econometrics 5*, 3381–3460.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S. Fourth edition*. New York: Springer–Verlag.

Vermunt, J. K. (1996). *Log-linear event history analysis: A general approach with missing data, latent variables, and unobserved heterogeneity*, Volume 8. Tilburg University Press Tilburg.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika 83*, 447–452.

Wolfinger, R. and M. O'Connell (1993). Generalized linear mixed models; a pseudolikelihood approach. *Journal Statist. Comput. Simulation 48*, 233–243.

Wood, S. and F. Scheipl (2013). *gamm4: Generalized additive mixed models using mgcv and lme4*. R package version 0.2-2.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall/CRC.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.