LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Micha Schneider

# Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model

Technical Report Number 224, 2019
Department of Statistics
University of Munich

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

# Dealing with Heterogeneity in Discrete Survival Analysis using the Cure Model

Micha Schneider
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München

August 6, 2019

**Abstract**

Cure models are able to model heterogeneity which arises from two subgroups with different hazards. One subgroup is characterized as long-term survivors with a hazard equal to zero, while the other subgroup is at-risk of the event. While cure models for continuous time are well established, cure models for discrete time points are rarely prevalent. In this article I describe discrete cure models, how they are defined, estimated and can be applied to real data. I propose to use penalization techniques to stabilize the model estimation, to smooth the baseline and to perform variable selection. The methods are illustrated on data about criminal recidivism and applied to data about breast cancer. As one result patients with no positive lymph nodes, a very small tumor, which can be well differentiated from healthy cells and with ethnicity which is neither black or white have the best estimated chances to belong to the long-term survivors of breast cancer.

**Keywords:** Cure Model, Discrete, Survival Analysis, Variable Selection, lasso

## 1 Introduction

In traditional survival analysis it is assumed that all analyzed subjects may be affected by the event of interest at sometime. Thus all subjects are at-risk of that event. But it happens frequently that a certain subgroup of the population never experience the event of interest. This subjects are called “cured”, “long-term survivors” (LTS) or “not-at-risk”.

Traditional examples can be found in clinical studies where some patients are long-term survivors of a severe disease as cancer and never suffer from the recurrence of it. In the social sciences one could be interested in analyzing the recurrence of released prisoners (see Rossi et al., 1980). Some of the released prisoners will be arrested again and others never do. Another example can be found in the educational sphere. Some students may be never able to solve a certain task, because it is too difficult for them, while others can solve the problem.

While cure models for continuous time are widely used and described for example by Amico and Keilegom (2018), Sy and Taylor (2000), Kuk and Chen (1992) and Maller and Zhou (1996), cure models for discrete time points are rarely prevalent. Tutz and Schmid (2016) give an overview about discrete time modelling and Muthén and Masyn (2005) about discrete-time survival mixtures. Actually in a lot of settings the time is not measured in continuous time but in discrete time points. In most cases a study ask their participants at fixed time points as months or years if they are still cured by the disease or still not in jail. If it is a retrospective study, the respondents may have also difficulties in remembering the exact time, but give an approximated response. Furthermore discrete survival analysis has the advantage that the interpretation may be easier since the hazard can be interpreted as probability and time depended variables can be introduced quite easily. The model used in this article is not designed for re-occurrence of an event (see Willett and Singer, 1995) or competing events (see Tutz and Schmid, 2016).

In this article I describe discrete cure models, how they are defined, estimated and how variable selection and smoothing can be performed. Thus we get a very flexible and easy-to-interpret tool for understanding complex discrete survival data situations. The discrete cure model has been considered by Tutz and Schmid (2016). Steele (2003) also applied a discrete-time mixture model with long-term survivors, but uses a different estimation method.

The article is organized as follows: First the discrete cure model is described and an overview of the discrete data structure is given. Then the model is illustrated by an application about criminal recidivism (Section 4). In Section 5 variable selection with an adopted version of lasso is proposed, followed by the description of the estimation of the (penalized) discrete cure model. In Section 7 the proposed selection technique is used to improve the model for criminal recidivism, followed by a further application about breast cancer (in Section 8). After some comments to the identifiability of discrete cure models the article is concluded.

## 2 The Discrete Cure Model

The cure model is defined as a finite mixture of survival functions. Typically it consists of two latent classes: One sub-population at risk and one sub-population characterized as long-term survivors or “cured”. The survival function of the cured remains at 1 whereas the survival function of the non-cured population decrease over time $t$ so that the observed survival function of the cure model is defined as

$$S(t|\boldsymbol{x}) = \pi(\boldsymbol{z})S_1(t|\boldsymbol{x}) + (1 - \pi(\boldsymbol{z})) \cdot 1, \tag{1}$$

where $\pi(\boldsymbol{z})$ is the weight for the non-cured population determining for each observation the probability belonging to this group. The weights can be calculated using individual specific covariates $\boldsymbol{z}$ by

$$\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{z}^T\boldsymbol{\beta})}.$$

The discrete survival function is the probability that the event has not been occurred at time point $t$:

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

which can be expressed by the discrete hazard $\lambda(t|\boldsymbol{x})$. It is defined as the probability that an event occurs at time $T$, given that time $T$ is reached conditional on some covariables $\boldsymbol{x}$:

$$\begin{aligned}\lambda(t|\boldsymbol{x}) = P(T = t|T \geq t, \boldsymbol{x}) &= h(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma}) \\ &= \frac{\exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})}{1 + \exp(\gamma_{0t} + \boldsymbol{x}^T\boldsymbol{\gamma})}, \quad t = 1, \ldots, t^*.\end{aligned}$$

$\gamma_{0t}$ is the parameter of the so called baseline hazard. The logistic distribution function $h() = \exp()/(1 + \exp())$ leads to the logistic discrete hazard model. However, one may also choose other link functions as the clog-log link to obtain the group proportional hazard model (see Tutz and Schmid, 2016).

There are two covariable sets $\boldsymbol{x}$ and $\boldsymbol{z}$ in the cure model. They can be identical, overlap or completely different. But they have very different functions, $\boldsymbol{x}^T\boldsymbol{\gamma}$ is used to estimate the survival function of the non-cured population so that this predictor influence the probability of an event in the non-cured population. On the other hand $\boldsymbol{z}^T\boldsymbol{\beta}$ determine the probability of being cured or not. In Section 5 I propose variable selection via penalization to decide which variables should be included in which part of the model.

## 3 Data Structure in Discrete Survival Analysis

In discrete survival analysis a certain data structure is usually very helpful. Let $y_{is}$ be an indicator of the occurrence of an event so that

$$y_{is} = \begin{cases} 1, \text{if individual fails at time } s \\ 0, \text{if individual survives time } s \end{cases}$$

Thus each observation $i$ generates a specific vector $(y_{i1}, \dots, y_{it_i})$ with the entries 0 or 1 and the length $t_i$. For a non-censored observation the vector has the form $(0, \dots, 0, 1)$ because at time $t_i$ the event occurs. Censored observations can be individuals who drop out during the study without observing an event or the study concludes when some participants have not experienced an event yet. For the censored observations the vector contains only zeros until the individual is censored: $(0, \dots, 0)$. The length $t_i$ is variable and depends on how long each individual is observed. If the person drops out of the study in the first time interval the length of $y_{is}$ is one. Table 1 illustrates the data structure for $T = 3$ time points and three individuals $i$. The first individual is observed for all three time points and experience the event at time point 3. Consequently, $y_i$ has the form $(0, 0, 1)$ with $t_i = 3$. Each row contains the information about one specific person at one specific time point. Thus observations have as many rows as observed time points. The second observation $i = 2$ drops out of the study after two time points. Thus, there are only two rows for observation 2 and $y_i = (0, 0)$, because no event take place. Since $x_{i1}$ is a time-constant variable the value is the same for one person and different time points[1].

| $i$ | $y_i$ | $t=1$ | $t=2$ | $t=3$ | $x_{i1}$ = Age | $t_i$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 20 | $t_1 = 3$ |
| 1 | 0 | 0 | 1 | 0 | 20 | |
| 1 | 1 | 0 | 0 | 1 | 20 | |
| 2 | 0 | 1 | 0 | 0 | 30 | $t_2 = 2$ |
| 2 | 0 | 0 | 1 | 0 | 30 | |
| 3 | 0 | 1 | 0 | 0 | 55 | $t_3 = 1$ |

Table 1: *Example for data structure in long format*

In Section 6.1 it will be shown that the likelihood by using $y_{is}$ is equivalent to the likelihood of a binary response model with observations $y_{is}$.

To include time-varying covariables for the population under risk in the discrete cure model we just have to add a new column $x_{i2}$ to the data structure. While the value of the time-constant covariables is repeated for observation $i$ for each row, the values of time-varying covariables can change with each row of the same observation $i$. In Table 2 the time-varying covariable "employment" is added by $x_{i2}$. If the person has a job at time $t$ the value is one otherwise zero. For example person 1 is unemployed at time $t = 1$ and gets hired at $t = 2$. At time $t = 3$ person 1 is unemployed again.

[1] Note that this data structure may be adjusted for the need of the software which is used. For example MRSP by Pößnecker (2019) requires that $y_i$ has always the length $T$ and missing values are filled up with $NA$. In this case $y_2$ would be $(0, 0, NA)$ and $y_3 = (0, NA, NA)$

| $i$ | $y_i$ | $t = 1$ | $t = 2$ | $t = 3$ | $x_{i1}$ = Age | $x_{i2}$ = Emp | $t_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 20 | 0 | $t_1 = 3$ |
| 1 | 0 | 0 | 1 | 0 | 20 | 1 | |
| 1 | 1 | 0 | 0 | 1 | 20 | 0 | |
| 2 | 0 | 1 | 0 | 0 | 30 | 1 | $t_2 = 2$ |
| 2 | 0 | 0 | 1 | 0 | 30 | 1 | |
| 3 | 0 | 1 | 0 | 0 | 55 | 0 | $t_3 = 1$ |

TABLE 2: *Example for data structure with time-depending covariable*

## 4 Illustrative Example: Criminal Recidivism

For illustration I use data about criminal recidivism, which is available in the R-package RcmdrPlugin.survival by Fox and Carvalho (2012). The data was generated within the scope of the "Transitional Aid Research Project" and described by Rossi et al. (1980). The aim of this project was to reduce the recidivism of prisoners and to examine the effect of financial aid. The data set used here consist of 432 released prisoners, who were observed during one year after release.

We know for each week if the person has been rearrested or not, which leads to 52 time points. Since there are not events at every time point, the time is reduced to 49. Half of the convicts received financial aid. Other variables are the age of the person at the time of release, the race ("black", "others"), the marital status ("married", "not married") and the level of education ("6th grade or less", "7th to 9th grade", "10th to 11th grade", "12th grade or higher"). Furthermore it was reported if the convicts worked full-time before incarceration ("no", "yes"), if they were released on parole ("no", "yes") and the number of convictions prior to the current incarceration. An overview of the available variables can be found in Table 3 and Table 4.

| | Minimum | 1st Quantile | Median | Mean | 3rd Quantile | Maximum |
|---|---|---|---|---|---|---|
| Age (at release) | 17 | 20 | 23 | 25 | 27 | 44 |
| Prior convictions | 0 | 1 | 2 | 3 | 4 | 18 |

TABLE 3: *Descriptive statistics of quantitative explanatory variables for the recidivism data*

First I will focus on a few important variables which are included in both parts of the model. In Section 5 we will see how this model can be further improved by using variable selection and smoothing techniques. Financial aid is one of the main variables in this setting. If financial aid has a positive effect, one can assume that it increases the probability of being cured and decreases the probability of an event. If someone has enough money for his/her basic needs, it may be less

| | Category | observations | Proportions (in %) |
|---|---|---|---|
| Financial aid | No | 216 | 50 |
| | Yes | 216 | 50 |
| Race | Black | 379 | 88 |
| | Others | 53 | 12 |
| Work experience | No | 185 | 43 |
| | Yes | 247 | 57 |
| Married | Yes | 53 | 12 |
| | No | 379 | 88 |
| On parole | No | 165 | 38 |
| | Yes | 267 | 62 |
| Education | ≤6th | 239 | 55 |
| | 7-9th | 24 | 6 |
| | 10-11th | 119 | 28 |
| | 12th+ | 50 | 12 |

TABLE 4: *Descriptive statistics of discrete explanatory variables for the Recidivism data*

probable that the person commits a crime. Similar applies for work experience. Someone, who has work experience, should be hired easier than someone without any work experience. So the hypothesis is that work experience reduces the probability of being arrested. In contrary the number of prior convictions may increase the probability of being non-cured and the probability of an event after release, since multiple offender may have more difficulties than first offender to change their lifestyle. Finally, age is included to account for demographic effects.

The result of the model, which includes these variables, can be found in Table 5. The standard errors are calculated by 600 bootstrap samples. Although the same variables are used for both parts of the model the meaning is completely different. The parameters in the upper part correspond with the probability that the person is part of the non-long-term survivors. If the person received financial aid the chance to be non-cured compared to be cured is reduced by the multiplicative factor $\exp(-0.2147) = 0.8068$. Thus the probability to be long-term survivor seems to be increased by financial aid. The number of prior convictions shows a positive effect so that the more prior convictions someone has committed the higher the probability of being non-cured. However, none of the estimates are statistically significant, since all confidence intervals include zero, so that the coefficients need to be interpreted with care.

In the lower part of the table the effects on the hazard function are displayed. Positive values correspond with a higher (and earlier) risk of arrest while negative values reduce the risk of recidivism. Here financial aid and prior work experience seem to coincide with a lower risk of an event. The number of prior convictions and a greater age seem to increase the probability of recidivism at any time t compared to an event later than t. Although these effects are again statistically non-significant the results are consistent with the hypotheses.

| | Estimates | BS.sd | BS.2.5 | BS.97.5 | |
|---|---|---|---|---|---|
| Intercept | 0.1319 | 0.5188 | -0.0753 | 1.4505 | $\hat{\beta}$ |
| Financial aid: yes | -0.2147 | 0.1312 | -0.3167 | 0.1794 | |
| Age | -0.0522 | 0.0240 | -0.0538 | 0.0355 | |
| Work experience: yes | 0.2426 | 0.2079 | -0.1257 | 0.7782 | |
| Number prior convictions | 0.1023 | 0.0556 | -0.0154 | 0.1793 | |
| Financial aid: yes | -0.1186 | 0.2605 | -0.8841 | 0.1261 | $\hat{\gamma}$ |
| Age | 0.0154 | 0.0362 | -0.1237 | 0.0306 | |
| Work experience: yes | -0.9839 | 0.4536 | -1.7538 | 0.1102 | |
| Number prior convictions | 0.0412 | 0.0444 | -0.0167 | 0.1615 | |

TABLE 5: *Model 1 - Estimates for recidivism. First group of estimates indicates effects on being non-long-term survivor, second group indicates effects on the event fall-back. BS.sd, BS.2.5, BS.97.5 refer to the bootstrap standard error and the quantiles for 2.5% and 97.5%, respectively.*

Figure 1 illustrates some parameter estimates. On the left hand side the effect of "financial aid" and "work experience" is displayed in the two-dimensional space of non-cured on the $y$ axis and risk of an event on the $x$ axis. The stars correspond to 0.95 confidence intervals using the 2.5% and 97.5% quantiles of the bootstrap samples. At the dashed lines no effect is found, because $\exp(0) = 1$. Since each confidence intervals cover this lines, it is easy to see that none of the effects is statistically significant. However, since the effect of financial aid is in both dimensions below 1, it indicates that there might be reduction of the chances in both dimensions.

On the right hand side of Figure 1 the estimated effect of financial aid on the survival function of the cure model is displayed. In this figure the variable work experience is set to "no" and the other two variables to their median value of 23 for age and 2 for prior convictions. Thus financial aid increases the survival function and leads to a higher survived proportion at the end of the study.

The discrete cure model is a very helpful tool to gain better insights in this complex data situation and can be easily interpreted. In contrast to cure models for continuous time the hazard can be always interpreted as probability. However, there may be also some challenges. First the variable selection is an crucial point and it might be difficult to decide which variables should be included in which part of the model. Second the baseline hazard may need very much parameters and may result in a quite rough function. Furthermore time points where no event take place may cause difficulties in the estimation process since the corresponding intercept should be minus infinity. All this issues can be addressed by the proposed penalization technique in the next section.

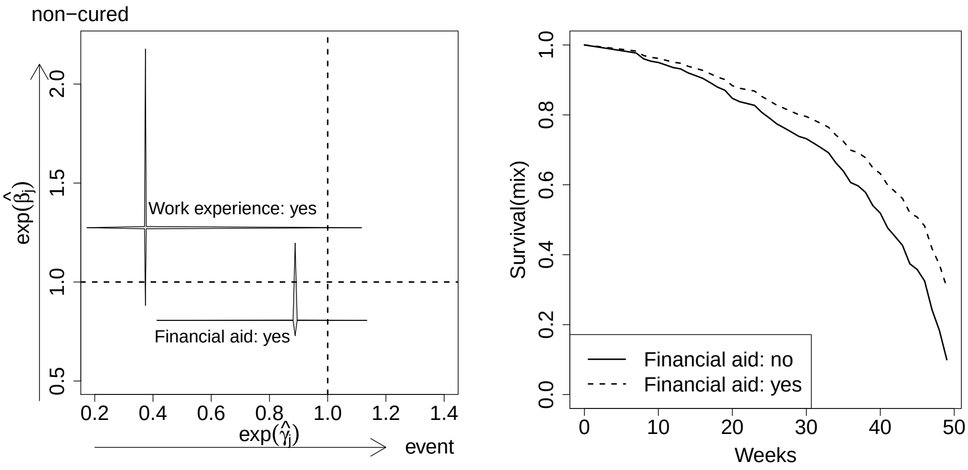

FIGURE 1: *Illustration of parameters estimates in model 1*

# 5 Penalization for Variable Selection and Smoothing

Penalization in discrete cure models can fulfill two main goals. First it is possible to select variables in a data driven way. Usually it is not obvious which covariates should be included in which part of the model. Using the proposed version of lasso (Tibshirani, 1996) for cure models can solve this issue. Second penalization can reduce the degrees of freedom concerning the intercepts. In discrete cure models there are intercepts for each transition from time $t$ to $t+1$. This may result in a large number of parameters which may not be necessary, in a quite rough baseline function and in computational difficulties if no event take place. Thus it is proposed to penalize the squared distances of two neighbouring intercepts. The penalized likelihood is given by

$$l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where $l(\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the unpenalized log-likelihood and $J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ a specific penalty term.

Let the vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ refer to the effect of $j$-th variable so that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_g)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_h)$. The corresponding vectors $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ are partitioned into $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \dots, \boldsymbol{z}_{ig})$ and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{ih})$ such that each components refer to a single variable. For example $\boldsymbol{x}_{ij}$ can represent for observation $i$ all dummy variables that are linked to the $j$-th variable. $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$ are defined as the number of parameters collected in the corresponding parameter vector $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$, respectively. So if the $j$-th $\boldsymbol{x}$-variable is marital status with the 4 categories “single”, “married”, “divorced” and “widowed”, the length of $\boldsymbol{x}_{ij}$ and the degrees of freedom $df_{\boldsymbol{\beta}_j}$ would be both 3. To ensure that the selection does not depend on the scale of the variables, all continuous and categorical variables need to be standardized.

The proposed penalty term is given by

$$J_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_{\boldsymbol{\beta}} \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \left\| \boldsymbol{\beta}_j \right\|_2 + \lambda_{\boldsymbol{\gamma}} \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \left\| \boldsymbol{\gamma}_j \right\|_2 \tag{2}$$

$$+ \lambda_0 \sum_{t=1;s>t}^{t^*} \left\| \gamma_{0t} - \gamma_{0s} \right\|_2^2. \tag{3}$$

It consists of three summands connected to the parameters $\boldsymbol{\beta}$ of the mixture weights, $\boldsymbol{\gamma}$ of the hazard function and $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0(t^*)})$ of the baseline hazard. Each component posseses its own tuning parameter $\lambda_{\boldsymbol{\beta}}$, $\lambda_{\boldsymbol{\gamma}}$ and $\lambda_0$, which regulate the amount of shrinkage. $\|\|_2$ is the unsquared $L_2$-Norm so that the penalty enforces the selection of variables in the spirit of the group lasso (Yuan and Lin, 2006) rather than selection of single parameters. A large $\lambda$ value corresponds with large shrinkage, which may also lead to more parameters set to zero. On the other hand a $\lambda$ value closer to zero results in a an estimate closer to the unpenalized ML-estimate with low shrinkage and less variable selection since less parameter groups are set to zero.

The first two penalty terms are constructed to shrink and select variables for the model components and refer to the values of each parameter vector. The aim of the third penalty term is the smoothing of the baseline hazard so that this term penalizes the squared distances of two neighbouring intercepts and not the intercepts itself. This penalty term can be also defined by matrices, which leads to

$$\lambda_0 \sum_{t=1;s>t}^{t^*} \left\| \gamma_{0t} - \gamma_{0s} \right\|_2^2 = \lambda_0 (R \cdot \boldsymbol{\gamma}_0)^T (R \cdot \boldsymbol{\gamma}_0)$$

and with $t^* = 4$ one obtains the following matrix:

$$R = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

An alternative strategy for smoothing the baseline hazard may be the use of splines as illustrated for example by Berger and Schmid (2018). However, using the squared distances is purely discrete and does not need any underlying continuous assumption about time. Furthermore there is no limitation at the borders of time space.

Since cross validation can be computational time consuming in mixture models it is proposed to use AIC or BIC as selection criteria. To account for the fit as well as for the complexity of the model it is necessary to define them in an appropriate way. Parameters, which are shrank should counted less than unpenalized parameters, which is captured by the effective degrees of freedom. The AIC and BIC are defined as

$$\begin{aligned} AIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + 2edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}), \\ BIC(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) + \log(n)edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}), \end{aligned}$$

where $edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ is the effective degrees of freedoms of the cure model. For each parameter set $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ the effective degrees of freedoms are calculated separately by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) &= edf(\hat{\boldsymbol{\beta}}) + edf(\hat{\boldsymbol{\gamma}}) \\ &= 1 + \sum_{j=1}^{g} edf(\hat{\boldsymbol{\beta}}_j) + edf(\hat{\boldsymbol{\gamma}}_0) + \sum_{j=1}^{h} edf(\hat{\boldsymbol{\gamma}}_j), \end{aligned}$$

where 1 refers to the intercept $\beta_0$ and $edf(\hat{\boldsymbol{\beta}}_j)$ to the effective degrees of freedom of the j-th parameter group of $\hat{\boldsymbol{\beta}}$. $edf(\hat{\boldsymbol{\gamma}}_0)$ denotes the effective degrees of freedom of the baseline and $edf(\hat{\boldsymbol{\gamma}}_j)$ to the j-th parameter group of $\hat{\boldsymbol{\gamma}}$. Following Yuan and Lin (2006) the effective degrees of freedom of each parameter group are given by

$$\begin{aligned} edf(\hat{\boldsymbol{\beta}}_j) &= \mathbb{1}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + (df_{\boldsymbol{\beta}_j} - 1)\frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\hat{\boldsymbol{\beta}}_j^{ML}\|_2}, \\ edf(\hat{\boldsymbol{\gamma}}_j) &= \mathbb{1}(\|\hat{\boldsymbol{\gamma}}_j\|_2 > 0) + (df_{\boldsymbol{\gamma}_j} - 1)\frac{\|\hat{\boldsymbol{\gamma}}_j\|_2}{\|\hat{\boldsymbol{\gamma}}_j^{ML}\|_2} \\ edf(\hat{\boldsymbol{\gamma}}_0) &= 1 + (df_{\boldsymbol{\gamma}_0} - 1)\frac{(R \cdot \hat{\boldsymbol{\gamma}}_0)^T(R \cdot \hat{\boldsymbol{\gamma}}_0)}{(R \cdot \hat{\boldsymbol{\gamma}}_0^{ML})^T(R \cdot \hat{\boldsymbol{\gamma}}_0^{ML})}, \end{aligned}$$

The idea is to relate the penalized estimates to the unpenalied maximum likelihood estimates (ML). For example, if the baseline parameters $\boldsymbol{\gamma}_0$ are not penalized, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_0^{ML}$ will be identical, which lead to $edf(\hat{\boldsymbol{\gamma}}_0) = df_{\boldsymbol{\gamma}_0}$. If the baseline parameters are penalized at most, the baseline hazard is almost constant and only one degree of freedom remains. In general if a variable is not penalized the *edf* are identical to $df_{\boldsymbol{\beta}_j}$ and $df_{\boldsymbol{\gamma}_j}$, respectively.

Since there are three independent tuning parameters there would be a three-dimensional grid for selection the best combination of tuning parameters. Since the smoothing is less crucial it can be recommended to fix $\lambda_{\mathbf{0}}$ at some medium level to reduce the model complexity and computing time.

# 6 Estimation

## 6.1 Construction of the log-likelihood

The likelihood of the discrete cure model can be derived from the unconditional probability of the occurrence of an event

$$P(T = t|\boldsymbol{x}) = \lambda(t|\boldsymbol{x}_i) \prod_{s=1}^{t-1} (1 - \lambda(s|\boldsymbol{x}_i))$$

If an observation is not censored and no event is observed the contribution is $(1 - \lambda(s|\boldsymbol{x}_i))$ for at least $t_i - 1$ time points. If an event take place the contribution is $\lambda(t|\boldsymbol{x}_i)$. Using the information provided by $y_{is}$ (introduced in Section 3) the likelihood of the discrete survival model of one specific observation $i$ can be written as

$$L_i^{disc} = \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1 - \lambda(s|\boldsymbol{x}_i))^{1-y_{is}}$$

This likelihood is equivalent to the likelihood of a binary response model with observations $y_{is}$. As long as $y_{is} = 0$ the contribution to the likelihood function is $1 - \lambda(s|\boldsymbol{x}_i)$. If an event is observed $\lambda(s|\boldsymbol{x}_i)$ is added to the log-likelihood. In the cured population the probability of an event is zero so that the likelihood of the long-term survivors can be simplified to[2]

$$L_i^{LTS} = \prod_{s=1}^{t_i} 0^{y_{is}} (1 - 0)^{1-y_{is}}$$

The likelihood of the cure model combines $L_i^{LTS}$ and $L_i^{disc}$ to

$$\begin{aligned} L_i = \pi(\boldsymbol{z}_i) & \left( \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1 - \lambda(s|\boldsymbol{x}_i))^{1-y_{is}} \right) \\ & + (1 - \pi(\boldsymbol{z}_i)) \left( \prod_{s=1}^{t_i} 0^{y_{is}} 1^{1-y_{is}} \right) \end{aligned} \tag{4}$$

Note that this equation only holds for modelling the failure time. One could also include the contribution of the censoring process itself as shown in Tutz and Schmid (2016).

[2] Note that $0^0 := 1$, $1^0 := 1$ and $\log(0) \to -\infty$

The complete log-likelihood for all observations is given by

$$
\begin{aligned}
l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \Bigg( \log(\pi(\boldsymbol{z}_i)) + \log\Big( \prod_{s=1}^{t_i} \lambda(s|\boldsymbol{x}_i)^{y_{is}} (1-\lambda(s|\boldsymbol{x}_i))^{1-y_{is}} \Big) \\
&\quad + \log(1-\pi(\boldsymbol{z}_i)) + \log\Big( \prod_{s=1}^{t_i} 0^{y_{is}} 1^{1-y_{is}} \Big) \Bigg) \\
&= \sum_{i=1}^{n} \Bigg( \log(\pi(\boldsymbol{z}_i)) + \sum_{s=1}^{t_i} \Big( \log\big(1-\lambda(s|\boldsymbol{x}_i)\big) + y_{is} \log\Big( \frac{\lambda(s|\boldsymbol{x}_i)}{1-\lambda(s|\boldsymbol{x}_i)} \Big) \Big) \\
&\quad + \log(1-\pi(\boldsymbol{z}_i)) + \sum_{s=1}^{t_i} \big( y_{is} \log(0) \big) \Bigg) \\
&:= \sum_{i=1}^{n} \Bigg( \log(\pi(\boldsymbol{z}_i)) + \log(S(y_i|\boldsymbol{x}_i)) + \log(1-\pi(\boldsymbol{z}_i)) + \log(S^{LTS}(y_i)) \Bigg), \quad (5)
\end{aligned}
$$

where $\boldsymbol{\theta}$ includes all parameters. $l_c(\boldsymbol{\theta})$ can be estimated using the EM-algorithm described in the next section. For readability reason only the last line is used for the further description.

## 6.2 Estimation via EM-Algorithm

The EM-algorithm by Dempster et al. (1977) is used to estimate $l_c(\boldsymbol{\theta})$ by treating the unknown class membership as a problem with incomplete data. $\zeta_i$ denote the unknown mixture component that indicate whether observation $i$ belongs to the non-cured population

$$
\zeta_i = \begin{cases} 1, \text{observation } i \text{ is from the non-cured population} \\ 0, \text{observation } i \text{ is from the cured population} \end{cases}
$$

With equation 5 follows

$$
l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big( \zeta_i \big\{ \log(\pi(\boldsymbol{z}_i)) + \log(S(y_i|\boldsymbol{x}_i)) \big\} + (1-\zeta_i) \big\{ \log(1-\pi(\boldsymbol{z}_i)) + \log(S^{LTS}(y_i)) \big\} \Big)
$$

In case of penalization the proposed penalty terms are added to $l_c(\boldsymbol{\theta})$. The penalized log-likelihood is

$$
\begin{aligned}
l_p(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \Big( \zeta_i \big\{ \log(\pi(\boldsymbol{z}_i)) + \log(S(y_i|\boldsymbol{x}_i)) \big\} + (1-\zeta_i) \big\{ \log(1-\pi(\boldsymbol{z}_i)) + \log(S^{LTS}(y_i)) \big\} \Big) \\
&\quad - \lambda_{\boldsymbol{\beta}} \sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \big\|\boldsymbol{\beta}_j\big\|_2 - \lambda_{\boldsymbol{\gamma}} \sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \big\|\boldsymbol{\gamma}_j\big\|_2 - \lambda_0 \sum_{t=1;s>t}^{t^*} \big\|\gamma_{0t} - \gamma_{0s}\big\|_2^2 .
\end{aligned}
$$

If the estimation is not penalized, the penalty terms can be omitted.

Within the EM algorithm the log-likelihood is iteratively maximized by using an expectation and a maximization step. During the E-step the conditional expectation of the complete log-likelihood given the observed data $\boldsymbol{y}$ and the current estimate $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)})$,

$$M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \mathrm{E}(l_p(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}^{(s)})$$

has to be computed. Because $l_p(\boldsymbol{\theta})$ is linear in the unobservable data $\zeta_i$, it is only necessary to estimate the current conditional expectation of $\zeta_i$. From Bayes's theorem follows

$$\begin{aligned} E(\zeta_i|\boldsymbol{y}, \boldsymbol{\theta}) &= P(\zeta_i = 1|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= P(y_i|\zeta_i = 1, \boldsymbol{x}_i, \boldsymbol{\theta})P(\zeta_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta})/P(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \pi_i S(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})/\{\pi_i S(y_i|\boldsymbol{x}_i) + (1-\pi_i)S^{LTS}(y_i)\} = \hat{\zeta}_i. \end{aligned}$$

This is the posterior probability that the observation $y_i$ belongs to the non-long-term survivor component of the mixture. In general it is permitted that an observation, for which an event is observed, might have a $\hat{\zeta}_i$ lower than one to account for all possible data structures including events by mistake. However, since the log-likelihood contribution of $S^{LTS}(y_i)$ would be close to minus infinity if an event take place this would occur very rarely and the algorithm usually avoids to assign such values for observations with observed events.

For the s-th iteration one obtains

$$\begin{aligned} M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) \quad & \left.\begin{aligned} &= \textstyle\sum_{i=1}^{n} \left\{ \hat{\zeta}_i^{(s)} \log(\pi_i) + (1-\hat{\zeta}_i^{(s)}) \log(1-\pi_i) \right\} \\ &-\lambda_\beta \textstyle\sum_{j=1}^{g} \sqrt{df_{\boldsymbol{\beta}_j}} \|\boldsymbol{\beta}_j\|_2 \end{aligned}\right\} M_1 \\ & \left.\begin{aligned} &+ \textstyle\sum_{i=1}^{n} \hat{\zeta}_i^{(s)} \log(S(y_i|\boldsymbol{x}_i)) \\ &-\lambda_\gamma \textstyle\sum_{j=1}^{h} \sqrt{df_{\boldsymbol{\gamma}_j}} \|\boldsymbol{\gamma}_j\|_2 - \lambda_0 \textstyle\sum_{t=1;s>t}^{t^*} \left\|\gamma_{0t} - \gamma_{0s}\right\|_2^2 \end{aligned}\right\} M_2 \\ & \left. + \textstyle\sum_{i=1}^{n} (1-\hat{\zeta}_i^{(s)}) \log(S^{LTS}(y_i)) \right\} M_3 \end{aligned}$$

$M_1$, $M_2$ and $M_3$ can be estimated independently from each other. The R-package MRSP by Pößnecker (2019) contains functions to estimate $M_1$ and $M_2$ including the mentioned penalty terms. Not every package would be suitable since the derivatives of $M_1$ and $M_2$ do not exist because of the group lasso penalty term. This problem can be solved with the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009) which is implemented in the MRSP package and is used for the maximisation problem of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. It can be generally formulated as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmax}}\, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}}\, l_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}}\; -l(\boldsymbol{\beta}, \boldsymbol{\gamma}) + J_\lambda(\boldsymbol{\beta}, \boldsymbol{\gamma}). \quad (6)$$

FISTA belongs to the class of proximal gradient methods in which only the unpenalized log-likelihood and its gradient is necessary. A detailed description can be found in Schneider et al. (2019). For given $\boldsymbol{\theta}^{(s)}$ one computes in the E-step the weights $\hat{\zeta}_i^{(s)}$ and in the M-step maximizes $M(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)})$. The E- and M-steps are repeated alternatingly until the relative tolerance

$$\Big|\frac{l_p(\boldsymbol{\theta}^{(s+1)}) - l_p(\boldsymbol{\theta}^{(s)})}{rel.tol/10 + |l_p(\boldsymbol{\theta}^{(s+1)})|}\Big| < rel.tol$$

is small enough to assume convergence. $\lambda_\beta$, $\lambda_\gamma$ and $\lambda_0$ span a three-dimensional grid of tuning parameter space. Dempster et al. (1977) showed that under weak conditions the EM algorithm finds a local maximum of the likelihood function. Hence it is always advisable to use meaningful start values to find a good solution of the maximization problem.

# 7 Illustrative Example: Penalization for Recidivism Data

Here I demonstrate, how the proposed penalization technique from Section 5 works and how it can improve the model of recidivism of prisoners. In Section 4 the chosen variables are used in both parts of the model. While most estimates were in line with the hypotheses, none of them were statistically significant. Now all those variables mentioned in Table 3 and 4 are included in the selection process. In addition to the previous variables marital status, race, released on parole and the level of education are available. The penalty terms ensure that only complete variables can be chosen but not single categories of one variable. The tuning parameter for the baseline hazard of the non-long-term survivor component $\lambda_0$ is set to 2, while the other two tuning parameters span a two-dimensional grid with $\lambda_\beta$ and $\lambda_\gamma$ range from 150 to 0.01 using 15 discrete values, respectively. $\lambda_\beta$ and $\lambda_\gamma$ are transformed by $\tilde{\lambda} = \log(\lambda + 1))$ to obtain a logarithm scale.

Figure 2 shows the results of the selection process using $15 \times 15 = 225$ tuning parameter combinations. If $\tilde{\lambda}_\beta = \tilde{\lambda}_\gamma \approx 5$ a pure intercept model is fitted. If both tuning parameters are close to zero an almost unpenalized model is estimated. The highest BIC values are detected in the corners of the graph in which at least one $\tilde{\lambda}$ is close to zero. That implies that models where all available variables are included in at least one component are not an appropriate choice according to BIC. It is possible to detect a clear region of very low BIC values. The minimum is found for $\tilde{\lambda}_\beta \approx 2.48$ and $\tilde{\lambda}_\gamma \approx 1.89$ at BIC = 1372.70. This is a strong reduction compared to the unpenalized model 1 with BIC value of 1673.36.

To get more insights in the mechanism of the variable selection Figure 2 is cut into slices and we look at the development of both coefficient sets $\beta$ and $\gamma$. For that matter one $\lambda$ value is fixed at the chosen value while the other $\lambda$ varies from high penalty (5.02) to low penalty (0.01). Each line type in the coefficient path represent one parameter group. In the first row of Figure 4 the $\gamma$ estimates

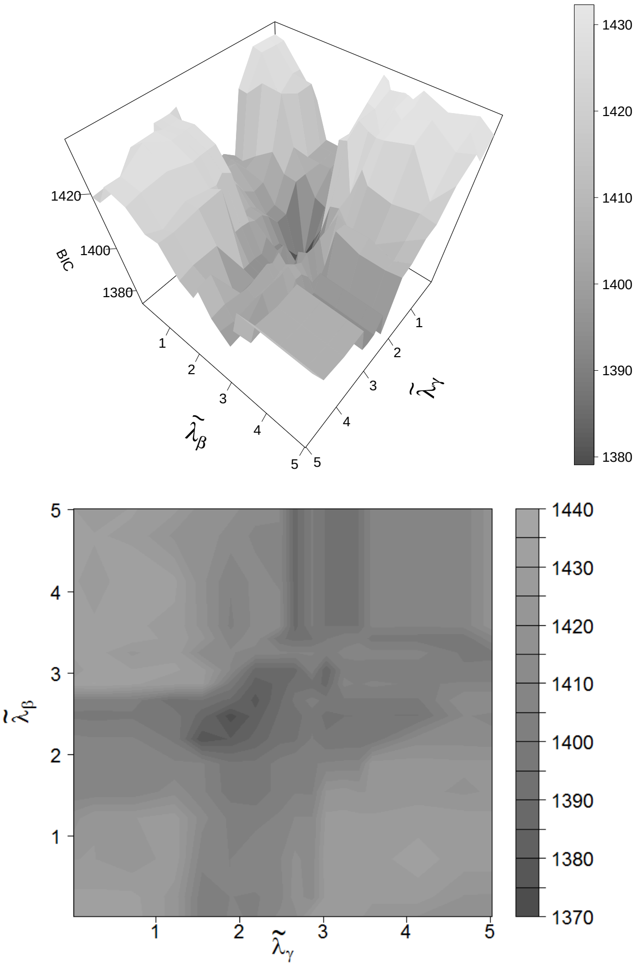

FIGURE 2: *Criminal recidivism: Grid of $\lambda$ values to find the best tuning parameter combination according to BIC*

for the standardized covariates are displayed. In the second row the $\beta$ coefficients can be found and the last row contains the boxplots of the estimated $\pi$. On the left hand side $\tilde{\lambda}_\gamma$ is set to 1.89 and $\tilde{\lambda}_\beta$ varies. On the right hand side $\tilde{\lambda}_\beta$ is fixed at 2.48 and $\tilde{\lambda}_\gamma$ is changing.

Usually the lower the tuning parameter the more coefficients are different from zero. But one should keep into mind that the estimates of $\beta$ and $\gamma$ are not completely independent from each other. Looking at the left hand side of Figure 4 one can see that the $\gamma$ estimates are quite unsteady although the corresponding $\tilde{\lambda}_\gamma$ is fixed. But the mixture weights determined by the $\beta$ coefficients change. At $\tilde{\lambda}_\beta = 5.02$ no $\beta$ coefficient is selected and for all observations a constant $\pi$ around 0.38 is estimated. Then $\pi$ increases to 0.53, before the first $\beta$ coefficient is selected and the weights become more and more individual specific. In this case a high variation in the $\pi$ boxplots can be seen as a higher individual differentiation which is desirable. However, each additional coefficient not only needs to improve the model fit but also reduces the BIC value to be selected. Thus the BIC is used to find a trade-off between model fit and number of parameters.

On the right hand side of Figure 4 $\tilde{\lambda}_\beta$ is fixed and $\tilde{\lambda}_\gamma$ varies. Here the weights $\pi$ and $\beta$ coefficients are almost constant. At the time when "not married" is selected in the $\gamma$ part of the model two $\beta$ coefficients are set to zero. Thus the interdependence works in both ways.

Since the graph illustrates the coefficient paths for the standardized covariates, we can also compare the absolute values of the estimates. In case of the $\beta$ coefficients at the left hand side of Figure 4 it is obvious that age and "prior convictions" are the first parameters which are selected. At $\tilde{\lambda}_\beta \approx 2.48$ the parameter "financial aid" is the smallest one out of the three coefficients. Thus age and "prior convictions" have a stronger impact than "financial aid". On the right "work experience" seem to have the greatest effect in the $\gamma$ dimension followed by "not married".

| | Penalized | | Refit | |
|---|---|---|---|---|
| Covariates | Non-LTS($\beta$) | Hazard($\gamma$) | Non-LTS($\beta$) | Hazard($\gamma$) |
| Constant | 0.0075 | | 0.1067 | |
| Financial aid: yes | -0.1543 | | -0.2857 | |
| Age | -0.0410 | | -0.0477 | |
| No prior convictions | 0.0556 | | 0.1064 | |
| Work experience: yes | | -0.6216 | | -0.8843 |
| Married: No | | 0.5461 | | 0.9952 |

Table 6: *Comparison of penalized and upenalized coefficients of the cure model for recidivism data*

| Covariates | Estimates | BS.sd | BS.2.5 | BS.97.5 | |
|---|---|---|---|---|---|
| Constant | 0.1067 | 0.3518 | 0.0323 | 1.3252 | |
| Financial aid: yes | -0.2857 | 0.2667 | -0.9565 | 0.1124 | $\hat{\beta}$ |
| Age | -0.0477 | 0.0202 | -0.0939 | -0.0162 | |
| Number prior convictions | 0.1064 | 0.0607 | 0.0291 | 0.2550 | |
| Work experience: yes | -0.8843 | 0.3422 | -1.4758 | -0.0793 | $\hat{\gamma}$ |
| Married: No | 0.9952 | 0.4367 | 0.1772 | 1.8567 | |

Table 7: *Model 2: Refit of the cure model for recidivism data with penalized intercepts*

Table 6 gives the estimates of the selected model. For $\beta$ only "financial aid", age and "prior convictions" are selected. "Work experience" and "not married" are chosen for modeling the hazard. It is an coincidence that in this case none of the variables is selected in both parts of the model. The first two estimation columns show the penalized estimates while the last two column contain the unpenalized estimates of the refit. The disadvantage of penalized estimates is that they are not unbiased but on the other hand they may lead to a smaller variance. Usually the unpenalized absolute estimates for one parameter group are larger than the penalized. However, if someone wants to have traditional standard errors and confidence intervals, it is plausible to refit the model without penalization to obtain unpenalized estimates and to be able to calculate standard errors. Table 7 contains the unpenalized estimates with Bootstrap standard errors and confidence intervals. They are obtained by 600 non-parametric samples of the data. Someone should keep into mind that these Bootstrap results ignore the model search and that the intercepts $\gamma_0$ are not displayed but their differences are still penalized. Now all coefficients are statistically significant to 5% level except of "financial aid". I would recommend to use the calculated bootstrap confidence intervals determined by the 2.5% and 97.5% quantiles of the bootstap distribution, because according to my experience the sampled distributions are often very skewed and the estimated coefficient value do not need to be in the middle of the sampled distribution. If this interval contains zero the corresponding coefficient is non-significant to the level 5%, which only applies for "financial aid".

The interpretation of the coefficients is the same as in Section 4. Age and financial aid reduce the probability to be non-cured while the number of prior convictions increase the probability. If someone is married and has work experience the probability of an event in the non-cured population is reduced.

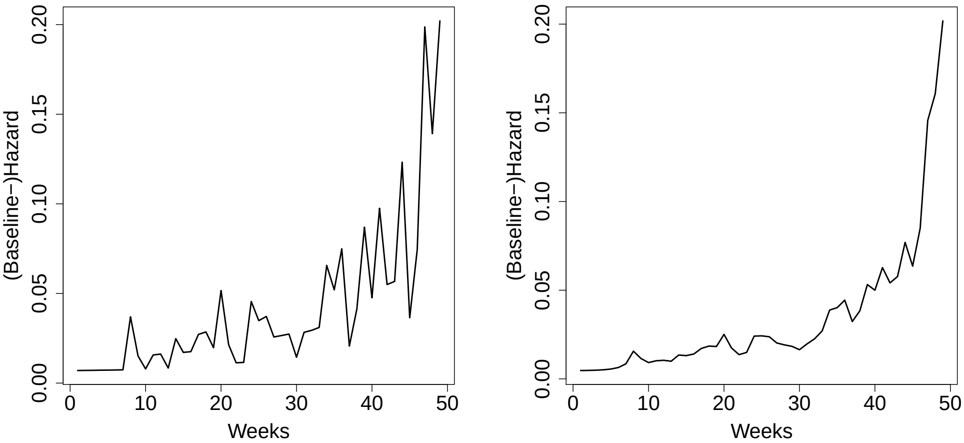

Figure 3: *Criminal recidivism: Comparison of the baseline hazard of the unpenalized model and the refitted penalized model*

Finally Figure 3 illustrates the effect of smoothing the baseline by penalizing the difference between neighbouring intercepts. On the left the baseline hazard of the unpenalized model is displayed. It is a quite rough function with many ups and downs. On the right the penalized baseline hazard is shown, which is much smoother, but keep the nature of the original function at the same time. The tuning parameter for smoothing can be enlarged to get a even smoother curve.

The proposed penalization technique could improve the original model substantially and results in an easy-to-interpret model.

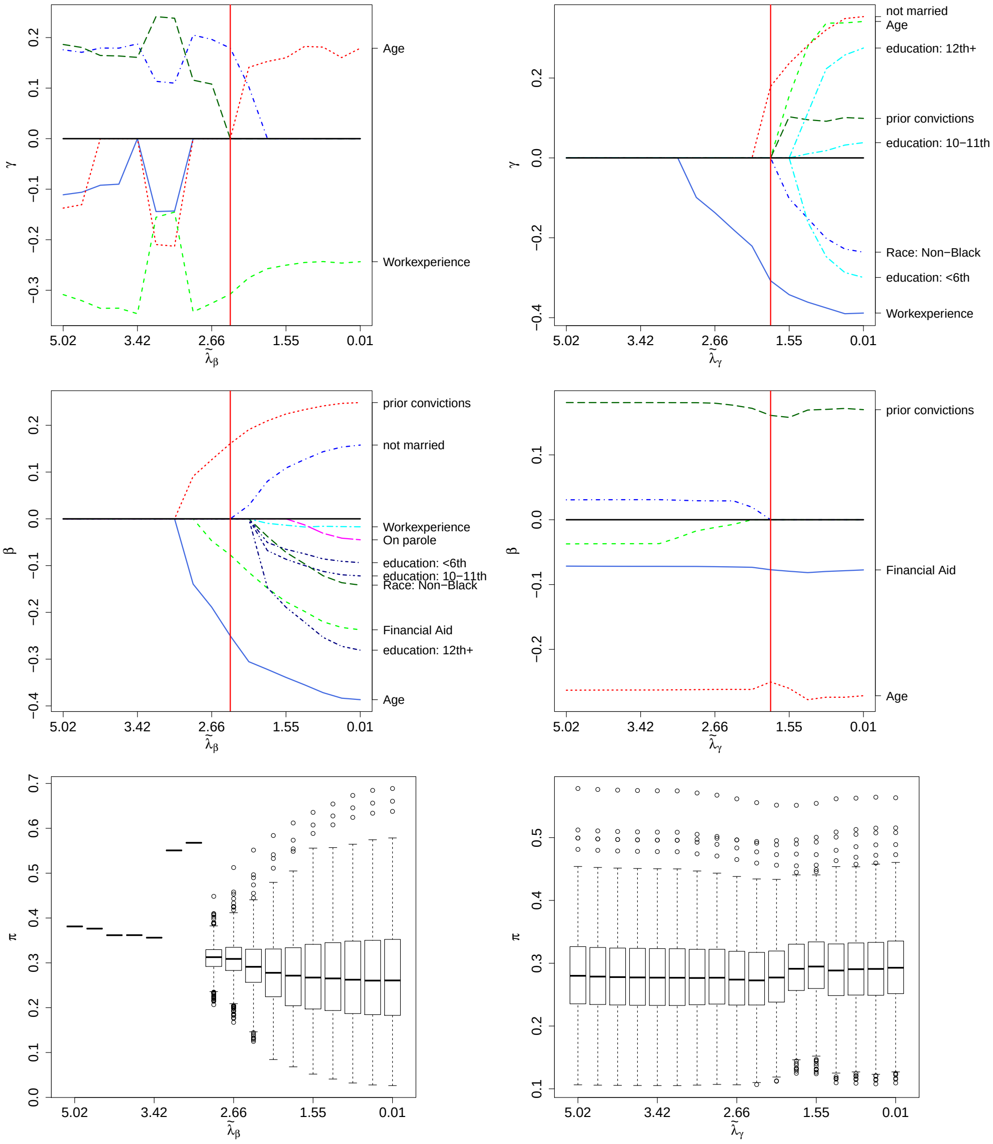

FIGURE 4: *Criminal recidivism: Standardized coefficient paths of $\beta$ and $\gamma$ and $\pi$ for fixed $\lambda_\gamma$ (left) and fixed $\lambda_\beta$ (right) in the cure model*

# 8 Application: Breast Cancer

Breast cancer is the most common cancer for women in developed countries. The average risk for a American woman to develop breast cancer sometime in her life is around 12% (see Akram et al., 2017). Thus, it is extremely relevant, which variables may be associated with being a long-term survivor from breast cancer and how variables are associated with the survival time of the patients. I use data of the SEER data base and the proposed methods to evaluate these questions.

SEER is the “Surveillance, Epidemiology, and End Results” Program (`www.seer.cancer.gov`), which collects information on cancer in the U.S. population on an individual basis. The time from diagnosis to death from breast cancer in years is given and I draw a random sample of $6,000$ breast cancer patients who entered the SEER data base between 1997 and 2011 (using SEER $1973-2011$ Research Data, version of November 2013). Since only the time span matters, the year of diagnosis can vary between the persons. The observed time may be also right-censored, when an event has not been observed (yet). Furthermore only female patients, younger than 76 years with first malignant tumor and without distant metastases were included so that there is a realistic chance to be a long-term survivor. Events can take place from the first until the 15th year.

| | Minimum | 1st Quantile | Median | Mean | 3rd Quantile | Maximum |
|---:|---:|---:|---:|---:|---:|---:|
| Age at diagnosis (years) | 18 | 48 | 56 | 56 | 64 | 75 |
| Tumorsize (mm) | 1 | 10 | 16 | 21 | 25 | 230 |
| Number examined nodes | 1 | 3 | 7 | 9 | 14 | 57 |

Table 8: *Descriptive statistics of quantitative explanatory variables for the breast cancer data (SEER)*

Table 8 and 9 shows the covariates, which might be selected. Most of the variables are related to the medical data. The primary site denotes where the breast cancer was found. The most frequent locations are C504 which is the upper outer quadrant of the breast and C508 which is the overlapping lesion of breast. The tumor grade specifies how well the tumor can be differentiated from healthy cells ranging from “well” over “moderately” to “poorly”. It is known which radiation therapy and in which order was applied. Then it is reported how many lymph nodes were examined and how many positive lymph nodes were found. The latter variable has four categories: None, one to three, four to six and seven or more positive lymph nodes. The T-stage variable classify the tumor according to AJCC 6th in four categories relying mainly on the size of the tumor and its extension. Further variables are the hormone receptor status (positive or negative) of estrogen (ER) and progesterone (PR), the laterality (right or left), the tumorsize (in mm), the age at diagnosis (in years), the race (white, black, others) and the marital status (single, married, separated, divorced, widowed).

|  | Category | Observations | Proportions (in %) |
|---|---|---|---|
| Marital status | single | 803 | 13 |
|  | married | 3898 | 65 |
|  | separated | 50 | 1 |
|  | divorced | 717 | 12 |
|  | widowed | 532 | 9 |
| Race | white | 4836 | 81 |
|  | black | 536 | 9 |
|  | others | 628 | 10 |
| Primary Site | C500 areolar | 27 | 0 |
|  | C501 subareolar | 289 | 5 |
|  | C502 Upper inner | 718 | 12 |
|  | C503 Lower inner | 356 | 6 |
|  | C504 Upper outer | 2201 | 37 |
|  | C505 Lower outer | 437 | 7 |
|  | C506 Axillary tail | 41 | 1 |
|  | C508 Overlapping lesion | 1203 | 20 |
|  | C509 Entire breast | 728 | 12 |
| Laterality | right | 2877 | 48 |
|  | left | 3123 | 52 |
| Tumor Grade | 1 well | 1300 | 22 |
|  | 2 moderately | 2569 | 43 |
|  | 3 poorly | 2131 | 36 |
| Radiation therapy | 1 None | 2106 | 35 |
|  | 2 Beam | 3715 | 62 |
|  | 3 Implants | 82 | 1 |
|  | 4 Combinations | 42 | 1 |
|  | 5 Other | 55 | 1 |
| Radiation Sequence | 1 None | 2170 | 36 |
|  | 2 Other | 37 | 1 |
|  | 3 Rad. after surgery | 3793 | 63 |
| ER status | positive | 4760 | 79 |
|  | negative | 1240 | 21 |
| PR status | positive | 4241 | 71 |
|  | negative | 1759 | 29 |
| Number positive nodes | 0 | 4018 | 67 |
|  | 1-3 | 1416 | 24 |
|  | 4-6 | 274 | 5 |
|  | 7+ | 292 | 5 |
| T-Stage | I | 3922 | 65 |
|  | II | 1701 | 28 |
|  | III | 281 | 5 |
|  | IV | 96 | 2 |

Table 9: *Descriptive statistics of discrete explanatory variables for the breast cancer data (SEER)*

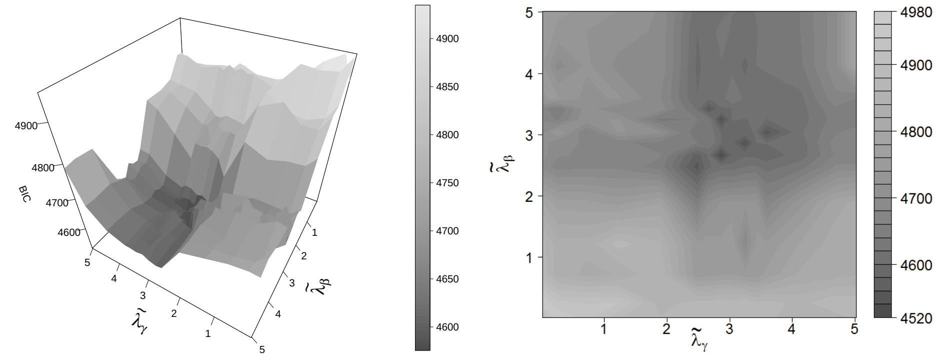

FIGURE 5: *Breast cancer: Grid of $\lambda$ values to find the best tuning parameter combination according to BIC*

Figure 5 shows the result of the grid search for 15 × 15 tuning parameter combinations. On the left the surface is illustrated and on the right the corresponding contour plot. As in the illustrative example the tuning parameter for the baseline hazard is fixed at 2. The transformed tuning parameters for the other two dimensions $\tilde{\lambda} = log(\lambda + 1)$ vary between 5.02 (high penalty) and 0.01 (low penalty). Including all variables in both parts of the model using a very low penalty leads to the highest BIC displayed in the right corner of the surface in Figure 5. But also using a very high penalty for both dimensions (left corner of the surface) does not lead to a desirable result. Although both tuning parameters are important to find the lowest BIC value, a too low $\tilde{\lambda}_\beta$ leads to higher BIC values regardless of $\tilde{\lambda}_\gamma$. Thus specifying the probability of being a long-term survivor seems to be more relevant.

The lowest BIC was found at 4525.62 with the tuning parameters $\tilde{\lambda}_\beta \approx 2.66$ and $\tilde{\lambda}_\gamma \approx 2.87$. After selecting the variables the model was refitted using only a penalized baseline, but no penalization term for the other coefficients. The parameter estimates of this refitted model are displayed in Table 10. The bootstrap confidence intervals rely on the bootstap 2.5% (BS.2.5) and 97.5% (BS.97.5) quantiles of 600 non-parametric bootstrap samples. Note that these bootstrap samples do not account for the selection process since only the selected variables are included.

The result of the proposed variable selection is a selection of only 20 out of 68 possible coefficients related to covariates. Moreover it can be decided which covariate effects the probability of being a non-long-term survivor captured by $\beta$, which covariate is important for the occurrence of an event modeled by $\gamma$ and which covariates are necessary in both components. Here only the race and the number of positive nodes are selected for modeling non-LTS. The tumor grade, size of tumor and T-stage are chosen in both components and the laterality, ER and PR status are only chosen for the event occurrence.

Positive estimates in the upper part of the table are related with an increase

| Covariates | Estimates | BS.sd | BS.2.5 | BS.97.5 | |
|---:|---:|---:|---:|---:|---|
| Constant | -2.7980 | 0.1133 | -3.0630 | -2.6120 | |
| Race: Black | 0.3157 | 0.0928 | 0.1446 | 0.5036 | |
| Race: Others | -0.4800 | 0.0857 | -0.6744 | -0.3351 | |
| Number of pos. nodes: 1-3 | 0.5516 | 0.0938 | 0.3994 | 0.7743 | |
| Number of pos. nodes: 4-6 | 0.9564 | 0.1524 | 0.7339 | 1.3237 | |
| Number of pos. nodes: 7+ | 1.8370 | 0.1791 | 1.5765 | 2.2828 | |
| Tumor Grade: II | 0.1317 | 0.1069 | 0.0243 | 0.4291 | $\hat{\beta}$ |
| Tumor Grade: III | 0.5968 | 0.1012 | 0.4458 | 0.8494 | |
| Size of tumor | 0.0141 | 0.0031 | 0.0076 | 0.0197 | |
| T-Stage: II | 0.2178 | 0.0712 | 0.0688 | 0.3496 | |
| T-Stage: III | -0.1057 | 0.0875 | -0.2926 | 0.0533 | |
| T-Stage: IV | 0.5393 | 0.1146 | 0.3595 | 0.8198 | |
| Tumor Grade: II | 0.3625 | 0.2727 | -0.1719 | 0.8821 | |
| Tumor Grade: III | 0.9388 | 0.2791 | 0.4019 | 1.4753 | |
| Size of tumor | 0.0051 | 0.0057 | -0.0022 | 0.0204 | |
| T-Stage: II | 0.3293 | 0.1640 | -0.0188 | 0.6223 | |
| T-Stage: III | 0.5239 | 0.4156 | -0.5143 | 1.1599 | $\hat{\gamma}$ |
| T-Stage: IV | 1.4500 | 0.3698 | 0.6414 | 2.2179 | |
| Laterality: Left | 0.3762 | 0.1279 | 0.2050 | 0.7069 | |
| ER status: negative | 0.8662 | 0.1617 | 0.4472 | 1.0762 | |
| PR status: negative | 0.5684 | 0.1485 | 0.3306 | 0.8956 | |
| $1-\bar{\pi}$ | 0.8498 | 0.0066 | 0.8313 | 0.8571 | |

Table 10: *Parameter estimates of the refitted cure model for breast cancer. Only the baseline is penalized. The standard errors and confidence intervals are obtained by bootstrap samples*

of the probability of being a non-LTS and in the lower part with an increase of the probability of an event namely death by breast cancer. Thus the number of positive nodes have an positive effect of being a non-LTS. The more positive nodes are found the higher the probability that the person is non-cured. If one to three nodes are positive the chance to be non-LTS compared to be LTS is increased by the factor $\exp(0.5516) = 1.74$ compared to patients without positive nodes. If the number of positive nodes are seven or more the multiplicative factor is with $\exp(1.8370) = 6.28$ much higher. Compared to white ethnic black people have a higher chance of being non-LTS while “others” have a lower chance.

Figure 6 illustrates the effect of tumor grade and T-stage in both dimensions. On the $y$ axis the effect of being non-LTS is marked. The $x$ axis shows the effect of an event. Generally the higher the category of tumor grade and T-stage the higher the chances in both dimensions. The only exception is tumor grade III which reduce the chance of non-LTS compared to tumor grade I. However the main driven factor of the T-stages categories I to III is the size of the tumor so that the negative effect of T-stage III can be compensated to some extend by the effect of tumorsize. The highest category of T-stage and tumor grade show the

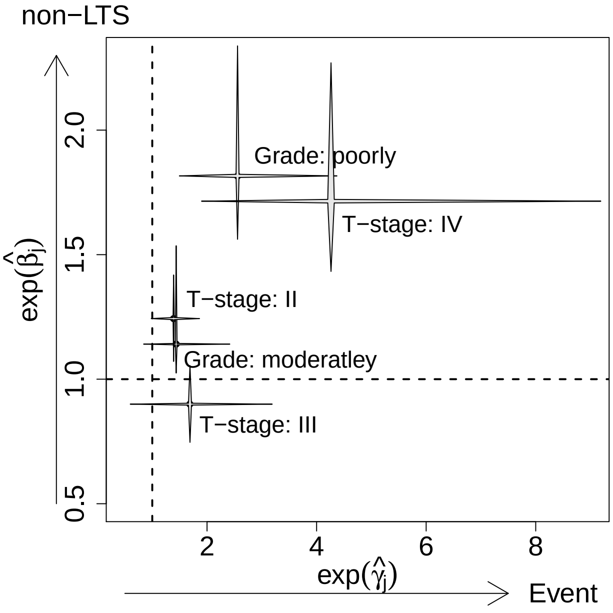

FIGURE 6: *Breast cancer: Effect of tumor grade and T-stage including confidence intervals computed by the* 2.5% *and* 97.5% *quantiles of non-parametric bootstrap samples*

strongest effects in both dimensions.

If the cancer is detected on the left the chance of the event "death by breast cancer" at time $t$ (compared to an event death by breast cancer later than $t$) is increased by $\exp(0.3762) = 1.46$ compared to laterality right. One reason may be that it is more difficult to treat cancer on the left side since the cancer is closer to the heart so that the radiation therapy for example need to be applied with more care than on the right. The negative status of both hormone receptors ER and PR increase the risk of the event, too. As long as the status of one of the hormone receptors is positive, it is possible to use drugs to fight the cancer. If the status is negative, hormone therapy does not work. The so called triple-negative breast cancers are defined by negative ER, PR and HER2. This type of cancer usually grows and spreads faster than other types of breast cancer and hormone therapy can not be applied. Because HER2 is only reported for observations from the year of diagnosis of 2010 onwards, the parameter could not be considered in this application.

According to the model the best chances of belonging to the long-term survivor group have patients with no positive lymph nodes, a very small tumor, which can be well differentiated from healthy cells and with ethnicity which is neither black or white. If the person does not belong to the long-term survivors the best survival chances are estimated for patients with a small tumor, which can be well differentiated from healthy cells, located at the right hand side and characterized by a positive ER and PR status. However, one should keep in mind that these results are not based on a randomized trial.

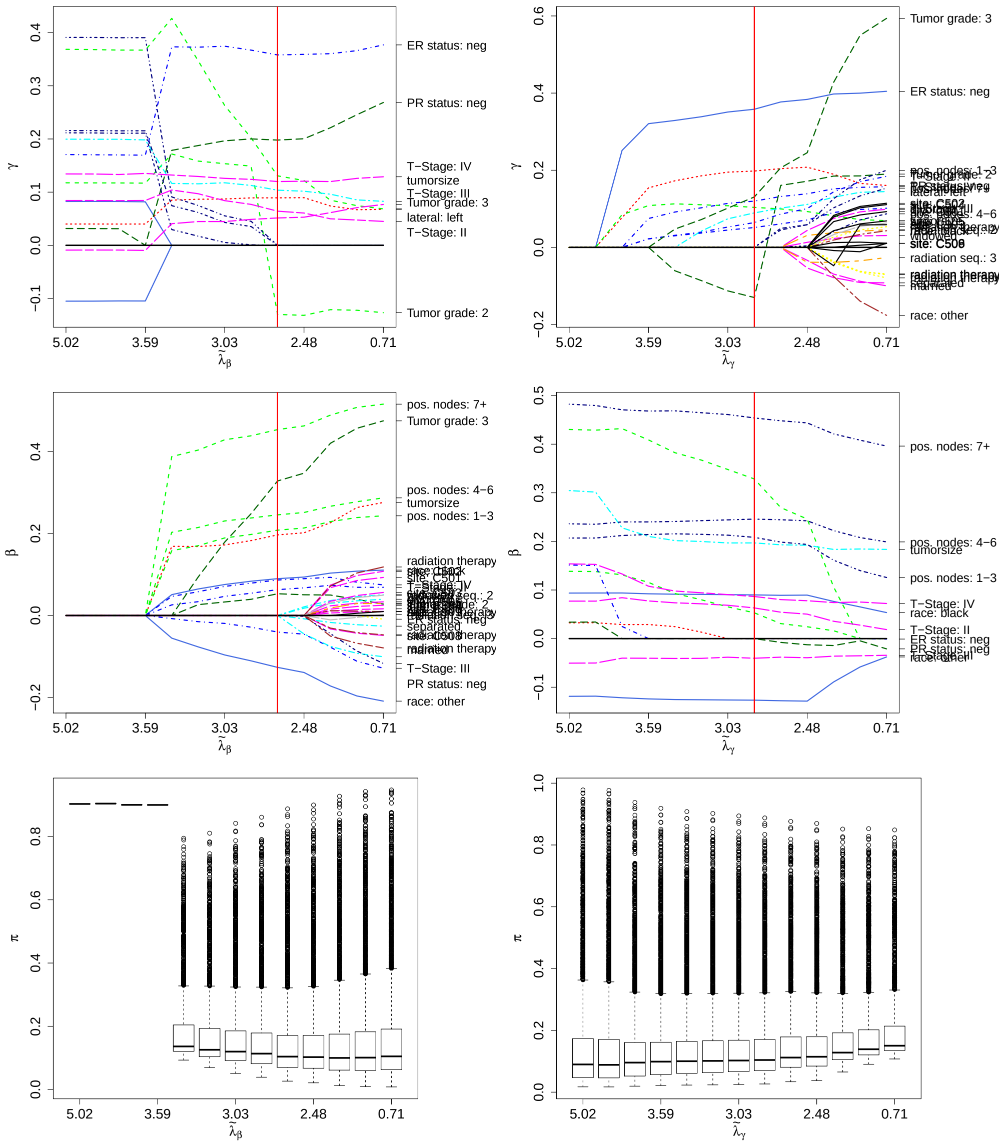

FIGURE 7: *Breast cancer: Standardized coefficient paths of $\beta$ and $\gamma$ and $\pi$ for fixed $\lambda_\gamma$ (left) and fixed $\lambda_\beta$ (right) in the cure model*

Figure 7 illustrates the standardized coefficient paths for this model. Since there are two varying tuning parameters it is necessary to introduce some constraints. On the left the coefficient paths are displayed when $\tilde{\lambda}_\gamma$ is hold constant at 2.87. On the right $\tilde{\lambda}_\beta$ is fixed at 2.66 and $\tilde{\lambda}_\gamma$ varies. The first row contains the estimates of $\gamma$, the second the estimates of $\beta$ and the last one the boxplots of $\pi$. Each line type correspond with one covariable which can consist of more than one coefficient as T-Stage for example. On the left the effect of entering $\beta$ coefficients is remarkable. At $\tilde{\lambda}_\beta = 3.42$ some $\beta$ coefficients enter the model so that the median $\pi$ drops dramatically. From there on the $\pi$ are calculated for each observation individually. The values of $\gamma$ coefficients change as well at this point although $\tilde{\lambda}_\gamma$ is kept constant. The estimated weights defined by $\beta$ may have a strong influence on the $\gamma$ estimates. On the right hand side $\tilde{\lambda}_\beta$ is fixed and $\tilde{\lambda}_\gamma$ varies. Here the effect of changing $\gamma$ has less effect on $\beta$ because $\gamma$ has no direct relation to $\pi$ which stay almost constant. However, it can be seen that the values of $\gamma$ and $\beta$ sometimes change the sign or become smaller with smaller penalty. This might be caused by inter dependencies between $\gamma$ and $\beta$ or by the data structure, when covariates influence each other by correlation.

# 9 Identifiability

Identifiability of cure models for continuous time was shown by Li et al. (2001) and Hanin and Huang (2014). It is assumed that there are at least three discrete time points ($t \geq 3$) and there is an effect $\gamma \neq 0$ of a continuous covariate $x$. Let the cure model be represented by two parameterizations

$$\pi_\beta S(\gamma_{0t} + \boldsymbol{x}^T\gamma) + (1 - \pi_\beta) = \pi_{\tilde{\beta}} S(\tilde{\gamma}_{0t} + \boldsymbol{x}^T\tilde{\gamma}) + (1 - \pi_{\tilde{\beta}})$$

There are values $\delta_{0r}$, $\delta$ such that $\tilde{\gamma}_{0r} = \gamma_{0r} + \delta_{0r}$, $\tilde{\gamma} = \gamma + \delta$. With $\eta_r(x) = \gamma_{0r} + x\gamma$ one obtains for all $x$ and $r$

$$\pi S(\eta_r(x)) - \tilde{\pi} S(\eta_r(x) + \delta_{0r} + x\delta) = (\pi - \tilde{\pi}).$$

Let us consider now the specific values $x_z = -\gamma_{0r}/\gamma + z/\gamma$ yielding for all values $z$ and $r$

$$\pi S(z) - \tilde{\pi} S(z + \delta_{0r} + x_z\delta) = (\pi - \tilde{\pi}).$$

By building the difference between these equations for values $z$ and $z - 1$ one obtains for all values $z$

$$\pi(S(z) - S(z-1)) = \tilde{\pi}(S(z + \delta_{0r} + x_z\delta) - S(z - 1 + \delta_{0r} + x_z\delta)).$$

The equation has to hold in particular for values $z = 1, 2, \ldots$. Since the logistic distribution function $F(\eta) = \exp(\eta)/(1 + \exp(\eta))$ is strictly monotonic and the derivative is different for all values $\eta$ it follows that $\delta_{0r} = \delta = 0$ and $\pi = \tilde{\pi}$.

# 10 Concluding Remarks

It has been shown that the discrete cure model can be used to model heterogeneity which arises from long-term survivors and patients at-risk in a discrete time setting. In the discrete survival analysis the hazard can be always interpreted as probability which makes any interpretation more intuitive. The instabilities of the model as no event occurrence at a certain time point or the number of parameters to estimate a rather rough baseline hazard can be overcome by the proposed penalization techniques. Furthermore it is possible to carry out variable selection so that there is a data driven way to decide which variable should be included in which part of the model. The variables can be chosen for one of the model components as well as for both model components. The proposed methods show stable and easy-to-interpret results in the applications. Thus it is possible to reduce the number of coefficients substantially and evaluate which covariates are associated with long-term survivors and the event of risk.

In case of breast cancer patients with no positive lymph nodes, a very small tumor, which can be well differentiated from healthy cells and with ethnicity which is neither black or white have the best chances to belong to the long-term survivors. The best survival chances in the group of non-LTS are estimated for patients with a small tumor, which can be well differentiated from healthy cells, located at the right hand side and characterized by a positive ER and PR status.

However, further research is necessary to evaluate the effect of the smoothing parameter on the general results and to develop computational efficient bootstrap samples which take the model search into account. In general, discrete cure models are the appropriate method, if the time is discrete and if there are two subgroups where one is characterized as long-term survivors.

*Acknowledgements*: Thanks to Gerhard Tutz and Paul Fink for fruitful discussions.

## References

Akram, M., M. Iqbal, M. Daniyal, and A. U. Khan (2017). Awareness and current knowledge of breast cancer. *Biological Research 50*(33), 1–23.

Amico, M. and I. V. Keilegom (2018). Cure models in survival analysis. *Annual Review of Statistics and Its Application 5*(1), 311–342.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Berger, M. and M. Schmid (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling 18*(3-4), 322–345.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B 39*, 1–38.

Fox, J. and M. S. Carvalho (2012). The rcmdrplugin.survival package: Extending the r commander interface to survival analysis. *Journal of Statistical Software 49*(7), 1–32.

Hanin, L. and L.-S. Huang (2014). Identifiability of cure models revisited. *Journal of Multivariate Analysis 130*, 261–274.

Kuk, A. Y. C. and C.-H. Chen (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika 79*(3), 531–541.

Li, C.-S., J. M. G. Taylor, and J. P. Sy (2001). Identifiability of cure models. *Statistics & Probability Letters 54*(4), 389–395.

Maller, R. A. and X. Zhou (1996). *Survival analysis with long-term survivors.* Wiley New York.

Muthén, B. and K. Masyn (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics 30*(1), 27–58.

Pößnecker, W. (2019). MRSP: Multinomial response models with structured penalties. R package version 0.6.11,
https://github.com/WolfgangPoessnecker/MRSP.

Rossi, P. H., R. A. Berk, and K. J. enihan (1980). *Money, Work, and Crime: Some Experimental Results.* New York: Academic Press.

Schneider, M., W. Pößnecker, and G. Tutz (2019). Variable selection in mixture models with an uncertainty component. Technical Report 225, Department of Statistics, Ludwig-Maximilians-Universität München.

Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in bangladesh. *Lifetime Data Analysis 9*(2), 155–174.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) (2014). *Research Data (1973-2011).* National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the November 2013 submission.

Sy, J. P. and J. M. G. Taylor (2000). Estimation in a cox proportional hazards cure model. *Biometrics 56*(1), 227–236.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tutz, G. and M. Schmid (2016). *Modeling Discrete Time-to-Event Data.* Springer.

Willett, J. B. and J. D. Singer (1995). Its déja vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics 20*(1), 41–67.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.