A hierarchical mixture cure model with unobserved heterogeneity for credit risk

Lore Dirick[1], **Gerda Claeskens**[1], **Andrey Vasnev**[2], **Bart Baesens**[3]

[1] *ORSTAT and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium*; loredirick@gmail.com; gerda.claeskens@kuleuven.be

[2] *University of Sydney Business School, Abercrombie Building (H70), NSW 2006, Sydney, Australia*; andrey.vasnev@sydney.edu.au.

[3] *LIRIS, Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium*; bart.baesens@kuleuven.be

Abstract

The specific nature of credit loan data requires the use of mixture cure models within the class of survival analysis tools. The constructed models allow for competing risks such as early repayment and default, and for incorporating maturity, expressed as an unsusceptible part of the population. A novel further extension of such models incorporates unobserved heterogeneity within the risk groups. A hierarchical expectation-maximization algorithm is derived to fit the models and standard errors are obtained. Simulations and a data analysis illustrate the applicability and benefits of these models, and in particular an improved event time estimation.

Key words: Credit risk modeling; Competing risks; EM-algorithm; Mixture cure model; Survival analysis; Unobserved heterogeneity.

1 Introduction

The analysis of credit risks via survival analysis takes advantage of the nature of time-to-event data, in particular by its ability to naturally capture the specifics of default, prepayment and maturity events. While those credit risk events were first examined and modeled individually, see for example Banasik et al. (1999), Stepanova and Thomas (2002), Andreeva (2006) and Bellotti and Crook (2009), these models were soon extended by allowing for a cured fraction while modeling early repayment or default, known as mixture cure models, see Tong et al. (2012) and Dirick et al. (2015). The simultaneous analysis of all different events is evident in Deng et al. (2000), Pavlov (2001), Ciochetti et al. (2002), Dirick et al. (2015) and Watkins et al. (2014). Dirick et al. (2019) studied the inclusion of macro-economic effects in mixture cure models with time-varying covariates.

For a recent overview about mixture cure models, the different types of models and a review of the literature, see Amico and Van Keilegom (2018). A general class of mixture cure models is theoretically studied in Patilea and Keilegom (2020).

The extension proposed in this paper of multiple event models for credit risk data acknowledges the fact that there are different kinds of customers. For example, some people are risk-averse, others might be risk-neutral or even risk-seeking. While this characteristic is not observed, our new approach makes it possible to incorporate unobserved heterogeneity in the models. More specifically, we construct a novel expectation-maximization (EM) algorithm that simultaneously deals with the mixture cure model with multiple events and with a number of subgroups for each of the modeled events. We explain the implementation of such a hierarchical EM algorithm for the credit risk models. Note that the presence of unobserved heterogeneity is structurally different from the multilevel mixture cure models (Lai and Yau, 2009; Tawiah et al., 2020) that incorporate observed grouping structures, such as institutions, hospitals, etc. via random effects. Heterogeneity has also been studied in the context of bivariate, twin data via frailty models and a cured fraction (Wienke et al., 2007).

This paper gives for credit risk modeling the first simulation study of the mixture cure models with unobserved heterogeneity. An application of the model for personal loan data from a UK bank reiterates the importance of the unobserved heterogeneity for credit risk. In the simultaneous modeling of competing events, similar to Watkins et al. (2014) we find for credit risk that the explanatory variables can act in different directions upon incidence and duration; and, variables exist that are statistically significant in explaining only incidence or duration.

Data collection processes are never complete and many real-life credit risk data sets are characterized by unobserved, yet potentially important predictive variables which typically reflect heterogeneity in the credit population. Deng et al. (2000) show that there exists significant heterogeneity among mortgage borrowers which generated discussion in this area, though not many researchers followed their lead. Hanson et al. (2008) find that ignoring heterogeneity in firm returns and default thresholds may lead to an underestimation of the expected loss and that there is an effect on portfolio risk too. Burda et al. (2015) employ an approach to build a semiparametric competing risk model with unobserved heterogeneity for the analysis of unemployment in the US. Their Bayesian method does not involve the EM-algorithm, and introduces unobserved heterogeneity through an infinite mixture of generalized inverse Gaussian densities. Djeundje and Crook (2018) work with multi-state delinquency models for credit cards and incorporate unobserved heterogeneity between accounts by including account-level random effects.

Despite the fact that in this paper the focus is on competing risks for loan data, the model is not restricted to these types of data and is applicable in a large range of situations where competing risks (and a possibility of not undergoing the risk or an "unsusceptible" part of the population) and a substantial amount of censoring are present. In the biomedical context, many disease-related research uses these models when there are several possible death causes (for example Lunn and McNeil, 1995), when there is a cured fraction of the patients (see, e.g., Bremhorst and Lambert, 2016) or the combination of both (e.g. Ng et al., 2002). In the economic context, an interesting example is given in Berrington and Diamond (2000), where first-partnership formation (competing risks are cohabitation and marriage) of males and females born in 1958 in Britain is studied. An unsusceptible population part can then be defined as the subjects that will never marry or cohabitate, however, censoring is present through the subjects that have not yet entered the first-partnership at the moment of the study, but will afterwards. Burda et al. (2015) use another application where the time to moment of exit from unemployment is modeled (to the same versus another industry where they had been employed previously). In portfolio allocation, our approach can extend Gambacciani and Paolella (2017) to take into account bankruptcies. In the educational context of Forcina (2017), the benefits could be derived by including the school leavers cohort into the analysis.

Section 2 gives the hierarchical mixture cure model with unobserved heterogeneity. Section 3 details the EM-algorithm. The simulation study is summarized in Section 4 and the empirical application is in Section 5. Concluding comments are in Section 6 followed by the theoretical derivations in the appendix.

2 The hierarchical mixture cure model

We observe life times T_i and a set of covariates for observation $i \in \{1, \dots, n\}$. The life times T_i represent the time until an event $j \in \{1, \dots, J\}$ takes place, or until the observation is censored. In credit risk, the events are default, early repayment, and maturity. For event J, the general censoring indicator δ_i for observation i is equal to 0, indicating that none of the competing events was observed. Additionally, each observation has J event-specific censoring indicators, denoted by $\delta_{j,i}$. As it is assumed that events are mutually exclusive, the rationale is that the occurrence of a certain event causes the observation to be censored from any other event type. Note that $\delta_i = \sum_{y=1}^{J} \delta_{y,i}$. For censored observations ($\delta_{j,i} = 0$ for every j and, consequently, $\delta_i = 0$), it is unknown which of the event types will be experienced eventually, or in other words, the event "group" that a censored observation belongs to is unknown. This group membership is represented by a partially observed variable $Y \in \{1, \dots, J\}$, with Y being observed only when $\delta_i = 1$.

Denote by $\pi_j(z, b) = P(Y = j \mid z, b)$ the probability of belonging to a certain group j, with $j \in \{1, \dots, J\}$, given the covariate vector z and the vector of coefficients b. For this discrete distribution it holds that (for a fixed z) $0 \leq \pi_j(z, b) \leq 1$ and that $\sum_{j=1}^{J} \pi_j(z, b) = 1$. The estimation of $\pi_j(z, b)$ is done through a multinomial logistic regression model with covariate vector z and corresponding parameter vector b. For $j = J$ it holds that $\pi_J(z, b) = 1 - \sum_{y=1}^{J-1} P(Y = y \mid z, b)$ with for $j \in \{1, \dots, J-1\}$,

$$\pi_j(z, b) = P(Y = j \mid z, b) = \frac{\exp(z^\top b_j)}{1 + \sum_{y=1}^{J-1} \exp(z^\top b_y)}. \tag{1}$$

The survival probabilities $S(t \mid Y = j, x; \beta_j)$ (or the probability of not having experienced any event by time t) use a covariate vector x, which may be different from, overlapping with, or be identical to the covariate vector z. In group J, the subjects are unsusceptible to any of the considered events, or in other words "cured", which is originating from medical studies considering cured patients, see e.g. Kuk and Chen (1992), Sy and Taylor (2000), Peng and Dear (2000). In the model, the cured or unsusceptible group has a survival probability $S(t \mid Y = J) = 1$ for every t and does not depend on x or on any parameters. In credit risk, the cured group consists of matured loans. The probability of not having experienced any event by time t is then given by

$$S(t; x, z, b, \beta) = \sum_{y=1}^{J-1} \pi_y(z, b) S(t \mid Y = y, x; \beta_y) + \pi_J(z, b), \tag{2}$$

where $S(t \mid Y=j, x; \beta_j)$ is the probability of not having experienced the event j by time t. In Theorem 1, see A.1, we prove the identifiability of (2).

To incorporate heterogeneity, in a hierarchical model we assume that all $J-1$ main groups, thus except for the "unsusceptible" Jth group, may be further divided into K_j subgroups, of which observations experience the same event and have a similar covariate structure but differ with regard to their event time structure. So instead of immediately modeling the survival function $S(t \mid Y=j, x; \beta_j)$ which depends on main group membership only, the survival structure depends on the subgroups as well. The probability of not having experienced event j at time t when belonging to subgroup k is modeled by a semi-parametric Cox proportional hazards model and given by

$$S_{T|\widetilde{Y}_j,Y}(t\,|\,\widetilde{Y}_j=k, Y=j, x, \beta_{jk}) = \exp\bigl\{-\exp(x^T\beta_{jk})\int_0^t h_0(u\,|\,\widetilde{Y}_j=k, Y=j)du\bigr\}, \qquad (3)$$

with h_0 the unspecified baseline hazard function, estimated using Breslow's estimator. The latent variable $\widetilde{Y}_j$ which takes values in $\{1,\dots,K_j\}$ represents the subgroup membership for group $j \in \{1,\dots,J-1\}$ and β_{jk} is the parameter vector related to the survival function of subgroup k in main group j. For identifiability reasons, the vectors β_{jk} do not contain an intercept. For further use, we define the probability to belong to subgroup k as $\tau_{k|j} = P(\widetilde{Y}_j = k \mid Y = j)$. We follow McLachlan and Peel (2000, Sec. 1.14), see also Aitkin and Rubin (1985), to carry out the estimation without constraints but order the estimated subgroup probabilities $\tau_{k|j}$ in a predetermined order to guarantee identifiability of the parameters β_{jk}, $j=1,\dots,J-1$, $k=1,\dots,K_j$. The identifiability of the hierarchical mixture cure model is investigated in simulations in Section 4.

3 The joint likelihood and EM-algorithm for the hierarchical model

The likelihood contribution of an observation $i=1,\dots,n$ is given by

$$f_{T,\widetilde{Y},Y}(t_i \mid \tilde{y}, y) = f_{T_i|\widetilde{Y}_i,Y_i}(T_i \mid \widetilde{Y}_i, Y_i, x_i, \delta_i, \beta_{Y,\widetilde{Y}}) \cdot \tau_{\widetilde{Y}_i|Y_i} \cdot \pi_{Y_i}(z_i, b),$$

where we use that, for a given j and k, the conditional likelihood contribution is

$$\begin{aligned} &f_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j, x_i, \delta_{j,i}, \beta_{jk}) \\ &= h_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j, x_i, \beta_{jk})^{\delta_{j,i}} S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j, x_i, \beta_{jk}), \end{aligned} \qquad (4)$$

with h the hazard function, formally,

$$h_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j, x_i, \beta_{jk}) = h_0(t \mid \widetilde{Y}_j = k, Y = j)\exp(x^T\beta_{jk}).$$

The joint hierarchical log-likelihood of $(T_i, Y_i, \tilde{Y}_i)$ now takes the form

$$\begin{aligned} &\mathcal{L}_n^h(b, \beta_{Y,\widetilde{Y}}; Y_1,\dots,Y_n, \widetilde{Y}_1,\dots,\widetilde{Y}_n, T_1,\dots,T_n, \delta_1,\dots,\delta_n) \\ &= \log\prod_{i=1}^n \bigl\{f_{T_i|\widetilde{Y}_i,Y_i}(T_i \mid \widetilde{Y}_i, Y_i, x_i, \delta_i, \beta_{Y,\widetilde{Y}}) \cdot \tau_{\widetilde{Y}_i|Y_i} \cdot \pi_{Y_i}(z_i, b)\bigr\}. \end{aligned} \qquad (5)$$

The expected complete-data log likelihood

Since the main group indicators Y_i as well as the subgroup indicators $\widetilde{Y}_i$ are not fully observed, the EM-algorithm (Dempster et al., 1977) is used in order to maximize the log likelihood. In this iterative procedure, the parameter estimates of the r-th EM-iteration are used along with the observed information to compute the expected complete-data log likelihood of the $(r+1)$-th EM-iteration, formally,

$$\begin{aligned} Q^h\{(b,\beta_{Y,\widetilde{Y}})^{(r+1)} \mid (b,\beta_{Y,\widetilde{Y}})^{(r)}\} &= E_{(b,\beta_{Y,\widetilde{Y}})^{(r)}}[\mathcal{L}_n^h\{(b,\beta_{Y,\widetilde{Y}})^{(r+1)}\} \mid T_1,\ldots,T_n] \\ = \sum_{i=1}^{n} \Big\{ E[\log \pi_{Y_i}(z_i, b^{(r+1)}) \mid T_i, b^{(r)}] &+ E[\log \tau^{(r+1)}_{\widetilde{Y}_i|Y_i} \mid T_i, \beta^{(r)}_{Y_i,\widetilde{Y}_i}] \\ +E[\log f_{T_i|Y_i,\tilde{Y}_i}(T_i \mid Y_i, \widetilde{Y}_i, x_i, \delta_i, \beta^{(r+1)}_{Y,\tilde{Y}}) &\mid T_i, \beta^{(r)}_{Y_i,\widetilde{Y}_i}]\Big\}. \end{aligned} \tag{6}$$

Rewriting the first term gives

$$E[\log \pi_{Y_i}(z_i, b^{(r+1)}) \mid T_i, b^{(r)}] = \sum_{y=1}^{J} P(Y_i = y \mid T_i = t_i, x_i) \log \pi_y(z_i, b^{(r+1)}),$$

with $P(Y_i = j \mid T_i = t_i, x_i)$ the probability of belonging to group j, conditional on the censoring or event time, which for uncensored cases is either 1 or 0. It is a weighted average of the time densities in the censored case,

$$\begin{aligned} &P(Y_i = j \mid T_i = t_i, x_i) \equiv w_j(\beta^{(r)}; t_i, x_i) \\ &= \begin{cases} \dfrac{\pi_j(z_i, b^{(r)}) f_{T|Y=j}(t_i \mid Y = j, x_i, \beta_j^{(r)})}{\sum_{y=1}^{J} \pi_y(z_i, b^{(r)}) f_{T|Y=y}(t_i \mid Y = y, x_i, \beta_y^{(r)})} & \text{if } \delta_i = 0, \\ 1 \quad \text{if } \delta_i = 1 \text{ and } Y_i = j, \\ 0 \quad \text{if } \delta_i = 1 \text{ and } Y_i \neq j. \end{cases} \end{aligned}$$

As the density $f_{T|Y=j}(t_i \mid Y = j, x_i, \beta_j^{(r)})$ itself is composed of the subgroup time densities with their respective subgroup probabilities, $\tau_{j|k}$,

$$\begin{aligned} &w_j(\beta^{(r)}; t_i, x_i) \\ &= \begin{cases} \dfrac{\pi_j(z_i, b^{(r)}) \sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j} f_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = \tilde{y}, Y = j, x_i, \beta_{j\tilde{y}}^{(r)})}{\sum_{y=1}^{J} \pi_y(z_i, b^{(r)}) \sum_{\tilde{y}=1}^{K_y} \tau_{\tilde{y}|y} f_{T|\widetilde{Y}_y,Y}(t_i \mid \widetilde{Y}_y = \tilde{y}, Y = y, x_i, \beta_{y\tilde{y}}^{(r)})} & \text{if } \delta_i = 0, \\ 1 \quad \text{if } \delta_i = 1 \text{ and } Y_i = j, \\ 0 \quad \text{if } \delta_i = 1 \text{ and } Y_i \neq j, \end{cases} \end{aligned}$$

and using (4) along with the fact that $\delta_{j,i} = 0$ when $\delta_i = 0$,

$$\begin{aligned} &w_j(\beta^{(r)}; t_i, x_i) \\ &= \begin{cases} \dfrac{\pi_j(z_i, b^{(r)}) \sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j} S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = \tilde{y}, Y = j, x_i, \beta_{j\tilde{y}}^{(r)})}{\sum_{y=1}^{J} \pi_y(z_i, b^{(r)}) \sum_{\tilde{y}=1}^{K_y} \tau_{\tilde{y}|y} S_{T|\widetilde{Y}_y,Y}(t_i \mid \widetilde{Y}_y = \tilde{y}, Y = y, x_i, \beta_{y\tilde{y}}^{(r)})} & \text{if } \delta_i = 0, \\ 1 \quad \text{if } \delta_i = 1 \text{ and } Y_i = j, \\ 0 \quad \text{if } \delta_i = 1 \text{ and } Y_i \neq j. \end{cases} \end{aligned} \tag{7}$$

For the second term in (6) we get that

$$E[\log \tau^{(r+1)}_{\widetilde{Y}_i|Y_i} \mid T_i, \beta^{(r)}_{Y_i,\widetilde{Y}_i}] = \sum_{y=1}^{J}\sum_{\tilde{y}=1}^{K_y} P(\widetilde{Y}_{y,i} = \tilde{y}, Y_i = y|T_i = t_i, x_i) \log \tau^{(r+1)}_{\tilde{y}|y}$$

$$= \sum_{y=1}^{J}\sum_{\tilde{y}=1}^{K_y} P(\widetilde{Y}_{y,i} = \tilde{y} \mid Y_i = y, T_i = t_i, x_i) P(Y_i = y|T_i = t_i, x_i) \log \tau^{(r+1)}_{\tilde{y}|y},$$

with $P(\widetilde{Y}_{j,i} = k \mid Y_i = j, T_i = t_i, x_i)$ the probability of belonging to subgroup k, given the event type j and the censoring or event time. Similarly to (7), we get

$$P(\widetilde{Y}_{j,i} = k \mid Y_i = j, T_i = t_i, x_i) \equiv v_{k|j}(\beta^{(r)}_j; t_i, x_i)$$

$$= \frac{\tau^{(r)}_{k|j} f_{T|Y=j,\tilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k, x_i, \beta^{(r)}_{jk})}{\tau^{(r)}_{\tilde{y}|j} \sum_{\tilde{y}=1}^{K_j} f_{T|Y=j,\tilde{Y}_j=\tilde{y}}(t_i \mid Y = j, \widetilde{Y}_j = \tilde{y}, x_i, \beta^{(r)}_{j\tilde{y}})},$$

for $j \in \{1, \dots, J-1\}$ and $k \in \{1, \dots, K_j\}$. By consequence,

$$E[\log \tau^{(r+1)}_{\widetilde{Y}_i|Y_i} \mid T_i, \beta^{(r)}_{Y_i,\widetilde{Y}_i}] = \sum_{y=1}^{J-1}\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\beta^{(r)}_y; t_i, x_i) w_y(\beta^{(r)}; t_i, x_i) \log \tau^{(r+1)}_{\tilde{y}|y}.$$

As opposed to the main groups where $\delta_{j,i}$ gives partial information on membership, no prior information is available for subgroup membership thus τ does not depend on a covariate vector. In the first iteration of the EM-algorithm, a value for $\tau_{k|j}$ is chosen for each k such that $\sum_{k=1}^{K} \tau_{k|j} = 1$ for each j. Without prior information, a logical starting value for $\tau_{k|j}$ is $1/K_j$ with K_j the total number of subgroups in main group j. In the next steps of the EM-algorithm, $\tau_{k|j}$ is updated as follows (see A.2 for details),

$$\tau^{(r+1)}_{k|j} = \tau^{(r+1)}_{k|j}(x_1, \dots, x_n) = P(\widetilde{Y}_j = k \mid Y = j) = \frac{\sum_{i=1}^{n} v^{(r)}_{k|j_i}(\beta^{(r)}_j; t_i, x_i)}{n}. \tag{8}$$

Similarly, the third term of (6) is given by

$$E[\log f_{T_i|Y_i,\tilde{Y}_i}(T_i \mid Y_i, \widetilde{Y}_i, x_i, \beta^{(r+1)}_{Y,\tilde{Y}}) \mid T_i, \beta^{(r)}_{Y,\tilde{Y}}]$$

$$= \sum_{y=1}^{J}\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\beta^{(r)}_y; t_i, x_i) w_y(\beta^{(r)}; t_i, x_i) \log f_{T_i,y_i,\tilde{y}_i}(T_i \mid Y_i = y, \widetilde{Y}_i = \tilde{y}, x_i, \beta^{(r+1)}_{y\tilde{y}})$$

$$= \sum_{y=1}^{J-1}\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\beta^{(r)}_y; t_i, x_i) w_y(\beta^{(r)}; t_i, x_i) \log f_{T_i,y_i,\tilde{y}_i}(T_i \mid Y_i = y, \widetilde{Y}_i = \tilde{y}, x_i, \beta^{(r+1)}_{y\tilde{y}}).$$

The resulting hierarchical expected complete-data log likelihood is then given by

$$
\begin{aligned}
&Q^h\{(b,\beta_{Y,\widetilde{Y}})^{(r+1)} \mid (b,\beta_{Y,\widetilde{Y}})^{(r)}\} \\
&= \sum_{i=1}^{n} \Big[w_J(\beta^{(r)};t_i,x_i)\log \pi_{J_i}(z_i,b^{(r+1)}) + \sum_{y=1}^{J-1} \Big\{ w_y(\beta^{(r)};t_i,x_i)\log \pi_{y_i}(z_i,b^{(r+1)}) \\
&\quad + \sum_{\tilde{y}=1}^{K_j} w_y(\beta^{(r)};t_i,x_i) v_{\tilde{y}|y}(\beta_y^{(r)};t_i,x_i) \times \\
&\quad \Big[\log \tau_{\tilde{y}|y}^{(r+1)} + \log f_{T_i,y_i,\tilde{y}_i}(T_i \mid Y_i = y, \widetilde{Y}_i = \tilde{y}, x_i, \beta_{y\tilde{y}}^{(r+1)}) \Big] \Big\} \Big].
\end{aligned}
\tag{9}
$$

Initialization and iterative E- and M-step

The three main steps of the computational algorithm are performed as follows:

a) Initialization stage

1) Determine the *number of subgroups* K_j for each of the J-1 main groups. Whereas the number of main groups J is fixed by the data structure, the number of subgroups is not. It is suggested to try several values for K_j, and evaluate the results (see also Section 5).

2) *Initialization of w:* Set $w_j^{(0)}(\beta^{(0)};t_i,x_i) = \delta_{j,i}$ for every j. Hence, the initial value is 1 for an observed event of category j and is 0 otherwise.

3) *Initialization of b:* Fit a multinomial logit model to $w^{(0)}$ using covariate vector z, in order to retrieve an initial estimate $\hat{b}^{(0)}$.

4) *Initialization of β:* Obtain estimates $\hat{\beta}_{j,k}^{(0)}$ at each subgroup level. The parameter estimates of the multiple event mixture cure model (Dirick et al., 2015) without heterogeneity can be used to set the initial values for the $\sum_{y=1}^{J-1} K_y$ parameter vectors. *Remark:* The K_j initial values for every $j \in \{1,\ldots,J-1\}$ should be different for the algorithm to work more efficiently. Hence, we start by specifying different $\beta_{j,k}$ parameters for each group.

5) *Initialization of τ:* $\tau_{k|j}^{(0)} = 1/K_j$ if no information about subgroups.

6) *Initialization of densities:* Compute density $f_{T|Y=j,\tilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k, x_i, \beta_{j,k})$, baseline hazard $h_0(u \mid \widetilde{Y}_j = k, Y = j)$ through Breslow's estimator and the survival function $S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j, x_i, \beta_{j,k})$ values (using Formula 3) for each observation, using the $\hat{\beta}_{j,k}^{(0)}$-estimates obtained from step 4.

b) E-step

1) Compute $\pi_j^{(1)}(z_i,b)$ for each j, using $\hat{b}^{(0)}$.

2) Compute $w_j^{(1)}(\beta;t_i,x_i)$ for each j, using $\hat{\beta}^{(0)}$.

3) Compute $v_{k|j}^{(1)}(\beta_j;t_i,x_i)$ for each k and each j, using $\hat{\beta}_j^{(0)}$.

c) M-step

1) *Update b:* Obtain a new estimate $\hat{b}^{(1)}$ using the $w_j^{(1)}(\beta; t_i, x_i)$'s of the E-step.

2) *Update β_{jk}:* Obtain a new estimate $\hat{\beta}_{j,k}^{(1)}$ using mixture weights $w_j^{(1)}(\beta; t_i, x_i)$ and $v_{k|j}^{(1)}(\beta_j; t_i, x_i)$. *Method:* The likelihood contribution corresponding to the event times can be written as

$$\sum_{j=1}^{J-1}\sum_{k=1}^{K_j}\sum_{i=1}^{n} w_j(\beta; t_i, x_i) v_{k|j}(\beta_j; t_i, x_i) \log\left\{h_{T,j,k}(t_i \mid x_i, \beta_{jk})^{\delta_{j,i}} S_{T,j,k}(t_i \mid x_i, \beta_{jk})\right\}$$

$$\begin{aligned} = & \sum_{j=1}^{J-1}\sum_{k=1}^{K_j}\sum_{i=1}^{n} w_j(\beta; t_i, x_i) v_{k|j}(\beta_j; t_i, x_i)\delta_{j,i} \log h_{T,j,k}(t_i \mid x_i, \beta_{jk}) \\ & + w_j(\beta; t_i, x_i) v_{k|j}(\beta_j; t_i, x_i) \log S_{T,j,k}(t_i \mid x_i, \beta_{jk}) \\ = & \sum_{j=1}^{J-1}\sum_{k=1}^{K_j}\sum_{i=1}^{n} v_{k|j}(\beta_j; t_i, x_i)\Big(\delta_{j,i} \log\left\{w_j(\beta; t_i, x_i) h_{T,j,k}(t_i \mid x_i, \beta_{jk})\right\} \\ & + w_j(\beta; t_i, x_i) \log S_{T,j,k}(t_i \mid x_i, \beta_{jk})\Big). \end{aligned}$$

For the last step, we used $\log w_j(\beta; t_i, x_i)\delta_{j,i} = 0$ and $\delta_{j,i} w_j(\beta; t_i, x_i) = \delta_{j,i}$.
Due to this, β can be estimated using standard software for fitting Cox proportional hazards models, such as the `coxph`-function in `R`, with an additional offset variable $\log(w_j(\beta; t_i, x_i))$ and weights equal to $v_{k|j}(\beta_j; t_i, x_i)$. A similar reasoning has been used by Cai et al. (2012).

3) *Update densities:* Obtain a new estimate of $f_{T|Y=j,\tilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k, x_i, \beta_{jk})$, $h_0(u \mid \widetilde{Y}_j = k, Y = j)$ and $S_{T|\widetilde{Y}_j,Y}(t_i \mid \widetilde{Y}_j = k, Y = j, x_i, \beta_{jk})$, using $\hat{\beta}_{j,k}^{(1)}$.

Repeat the E- and M-step with all updated estimates, until parameter convergence. The algorithm stops when the sum of the absolute value of the relative differences between $\big(\hat{\beta}^{(r+1)}, \hat{b}^{(r+1)}\big)$ and $\big(\hat{\beta}^{(r)}, \hat{b}^{(r)}\big)$ is smaller than 10^{-6}.

Note that for some observations i, $f_{T|Y=j,\tilde{Y}_j=k}(t_i \mid Y = j, \widetilde{Y}_j = k, x_i, \beta_{jk})$ for all subgroups $k \in \{1, \ldots, K_j\}$ in one of the main groups $j \in \{1, \ldots, (J-1)\}$ can be very small or even 0. As a result, $v_{k|j}(\beta_j; t_i, x_i)$ can go to infinity for all K_j subgroups of group j. As $\sum_k^{K_j} v_{k|j}$ should be equal to 1, this issue is solved by putting $v_{k|j}(\beta_j; t_i, x_i) = 1/K_j$ for every k when the denominator of $v_{k|j}(\beta_j; t_i, x_i) < 10^{-10}$ for a certain observation i. Intuitively, when a certain observation is seemingly found to not belong to any of the subgroups of a main group j, its subgroup probabilities should be equal. As a result of the equal weights, for such a situation the estimates for all groups will be approximately the same. This may be interpreted as a case where different subgroups are not needed.

Standard errors through the SEM-algorithm

The typical execution of the EM-algorithm does not automatically produce standard errors of the parameter estimates. The supplemented EM-algorithm, introduced by Meng and

Rubin (1991), is widely used in various applications for standard error estimation when applying the EM-algorithm, see for example Segal et al. (1994), Cai and Lee (2009), Cai (2008). For a discussion on standard errors for EM estimators, see Jamshidian and Jennrich (2000). Meng and Rubin (1991) show that a numerically stable asymptotic variance matrix for the estimators can be obtained using the supplemented EM-algorithm, more specifically, $V = I_{oc}^{-1}(I_d - DM)^{-1}$, where I_{oc} is the negative second Hessian matrix of the expected complete-data log likelihood, with $\Theta = (b, \beta_{Y,\widetilde{Y}})$, $I_{oc} = -\partial^2 Q^h\big(\widehat{\Theta} \mid \widehat{\Theta}\big)\backslash(\partial\Theta \cdot \partial\Theta^\top)$. The matrix I_d is the identity matrix with dimension $d \times d$, with d equal to the length of the parameter vector Θ. The $d \times d$-matrix DM can be interpreted as the matrix rate of convergence of the EM-algorithm. The idea behind this is that, implicitly, a mapping $\Theta \to M(\Theta) = (M_1(\Theta), \ldots, M_d(\Theta))^\top$ is defined by the EM-algorithm from the parameter space to itself such that $M(\hat{\Theta}^{(r)}) = \hat{\Theta}^{(r+1)}$ for $r = 0, 1, \ldots$. A Taylor series expansion in the neighborhood of $\hat{\Theta}$ yields that

$$(\widehat{\Theta}^{(r+1)} - \widehat{\Theta})^\top \approx (\widehat{\Theta}^{(r)} - \widehat{\Theta})^\top DM, \text{ where } DM = \left(\frac{\partial M_l(\Theta)}{\partial \Theta_m}\right)\Bigg|_{\Theta=\widehat{\Theta}},$$

a $d \times d$-matrix evaluated at $\Theta = \widehat{\Theta}$. In practice, DM is obtained by numerically differentiating $M(\Theta)$. For more information on the computation of DM, we refer to Meng and Rubin (1991, section 3.3). For the calculation of the Hessian matrix, we used the exact expressions as provided by the Cox proportional hazard models and the logistic regression models, respectively. We have implemented this procedure to obtain standard errors of the estimators of b_j and $\beta_{j,k}$ for all considered j and k.

Alternative methods for standard error calculations include the bootstrap. A bootstrap approach similar to that of Cai et al. (2012) in the R package smcure is applied in the data analysis, see Section 5, to the hierarchical mixture cure models.

In some cases one can compute the estimator's covariance matrix theoretically. For mixture cure models with a single main group and one cured group, Sy and Taylor (2001) obtained an expression for the standard errors and suggested approximations in case of tied event times. Unlike mixtures of all t-distributions, for example, for which approximations to the information matrix have been studied, see Wang and Lin (2016), mixture cure models are more challenging to handle due to the survival component, censored data and the nonparametric estimation of the baseline hazard. We expect that calculations similar to those of Sy and Taylor (2001) could be performed too for the hierarchical mixture cure model with multiple events of interest.

4 Simulation study

Simulation settings

To validate the model, a simulation study was conducted using the software R (R Core Team, 2020). Since mixture cure models as such have already been thoroughly investigated in the literature, we particularly focus attention to the model's performance regarding the unobserved grouping structure. In the first scenario in each simulation run, a dataset of

					Setting I		Setting II		Setting III	
	β_{A1}	β_{A2}	β_{B1}	β_{B2}	b_A	b_B	b_A	b_B	b_A	b_B
(intercept)					0.7	0.2	-0.6	0.3	-0.7	-1.1
λ_j	0.7	0.9	0.5	0.8	-	-	-	-	-	-
ν_j	0.6	1	0.7	1.1	-	-	-	-	-	-
$x_{j,1}$	-0.2	0	0.3	1.3	0.7	-0.4	0.8	-1.2	0.5	-1.5
$x_{j,2}$	0.1	0.4	-0.5	-0.1	0.6	1.2	-0.1	-0.3	-1.2	-0.4
$x_{j,3}$	-0.5	-0.3	0.1	0.3	0.2	-1	0.6	-1	0.7	-1.3

Table 1: Generating values for the parameter vectors β and b in the simulation study. For identifiability in the survival part of the model, there is no intercept in the vectors β.

size $n = 7500$ was constructed with three event groups A, B and C, and two subgroups for groups A and B (C is the "cured" group). This sample size resembles that of a typical credit risk study. There are three variables, generated respectively by $x_1 \sim \mathcal{N}(-2, 1)$, $x_2 \sim \mathcal{N}(2.6, 1.2)$ and $x_3 \sim \mathcal{N}(3, 2)$. These variables in combination with b-parameters shape the event group memberships of the observations when using a multinomial logit transformation. As a result, we get the probability of belonging to group $j \in \{A, B\}$:

$$\pi_{i,j}(z_i, b) = P(Y = j \mid b) = \frac{\exp(z_i^\top b_j)}{1 + \exp(z_i^\top b_A) + \exp(z_i^\top b_B)},$$

in this simulation study we took $z = x$. The general model formulation (Section 2) allows for z and x to be different. The value $\pi_{i,C} = 1 - \pi_{i,A} - \pi_{i,B}$.

In total three settings were explored which differ solely with regard to the values of b (which can be found in Table 1), resulting in different group sizes. In Setting I, there is a low number of "cured" (group C) cases of around 12%. Setting II contained around 35% cured cases, and Setting III around 67%. For all settings, the remainder of cases was approximately evenly split over groups A and B.

The event times differ for the two subgroups of A and B. Because of this, there are different parameter vectors for the survival functions of the subgroups. See Table 1 for the generating parameters β. To model the subgroups, for each main group the observations are randomly split into two groups of equal size, with event times generated using a Weibull distribution each with a covariate vector β_{Ak} and corresponding scale parameter λ and shape parameter $1/\nu$, see Table 1 for the parameter values for each group. More precisely, $T = [-\log(U)/\{\lambda \exp(x^\top \beta_{Ak})\}]^{1/\nu}$ with U a uniform distribution on $(0, 1)$. The event times for observations in the unsusceptible group C are all the same, for these simulations we put these equal to 10^3. Finally, some censoring was introduced. For both simulation settings, the censoring distribution was Censor $\sim$ Unif$(0, 200)$, leading to $15 - 20\%$ censoring in the main groups A and B.

In the second scenario we reduce the sample size to 3000 and increase the censoring by taking Censor $\sim$ Unif$(0, 50)$. This leads to 34-40% censoring. The other settings remain the same.

Simulation results

According to the simulation settings in Section 4, 1000 simulation runs were conducted. Three types of models are fitted and compared, (i) a misspecified homogeneous model with three main groups and no subgroups, K_j=1, (ii) a model with heterogeneity assuming two subgroups for both groups A and B, which corresponds to the data generation, and (iii) a misspecified model with too much heterogeneity as compared to how the data were generated, we use three subgroups for both main groups A and B.

Model without heterogeneity

We evaluate the estimation of the parameters in a misspecified model, assuming that there is no heterogeneity. Table 2 shows the mean of the parameter estimates over all simulation runs along with their standard deviations. Since this model is misspecified, we only compute an empirical bias for the parameters of the logistic model. Due to the misspecification, such results should be interpreted with care. We observe that the parameters of the incidence model are well-estimated. The main interest, however, lies in the estimates $\hat{\beta}$, since the model without heterogeneity forces to estimate only one β per main group, whereas the data are generated with two parameters β for both A and B. It is observed that abandoning the heterogeneity might lead to undetected effects. An example for this is $\beta_{B,1}$: the model estimates are between 0.542 and 0.644 when using no subgroups, whereas the two generating values $\beta_{B1,1} = 0.3$ and $\beta_{B2,1} = 1.3$. It seems that the model favors weaker relationships, which results in estimated βs that are relatively small in absolute value. A comparison between the two simulation scenario's leads to the expected finding of smaller biases and standard errors for scenario 1 where the sample size is larger and less censoring is present.

Model with heterogeneity: two subgroups

The mean parameter estimates for all settings over 1000 simulation runs for a model with heterogeneity can be found in Table 3. The parameter estimates for b in the right panel are close to the estimates for the model without heterogeneity, resulting in good estimates with respect to the simulated parameters and similar standard errors for all settings. The estimates $\hat{\beta}$ in the left panel show that the two subgroups are well-identified for groups A and B. When comparing the estimates $\hat{\beta}$ with the true parameter values used for simulation (see Table 1), we note that higher cure does in general not disturb the estimation of β. However, it should be noted that standard errors are larger for subgroup $A1$ when comparing with other subgroups. Additionally, standard errors tend to go up when the cured fraction is larger. This first phenomenon can be explained by the fact that the parameters of both subgroups in A lie closer to each other as compared to the subgroup parameters of B. This possibly makes estimation more difficult. The second phenomenon is explained by the smaller number of cases in groups A and B for Setting III in comparison with Setting I. Because approximately 2/3 of the observations is cured in Setting III, only around 600 observations belong to each subgroup in this setting. This leads to less accuracy and higher variation in parameter estimation. For most parameters, under the more challenging scenario 2, an increase in bias and standard error is observed as compared

		$\hat{\beta}_A$		$\hat{\beta}_B$		$\hat{b}_A$		$\hat{b}_B$	
Setting I	int.	Scenario 1: $n = 7500$				0.86	(0.16,0.17)	0.39	(0.19,0.19)
	$x_{j,1}$	-0.08	(0.02)	0.54	(0.03)	0.64	(-0.06,0.09)	-0.43	(-0.03,0.10)
	$x_{j,2}$	0.13	(0.02)	-0.28	(0.02)	0.41	(-0.19,0.07)	0.98	(-0.23,0.09)
	$x_{j,3}$	-0.27	(0.02)	0.15	(0.02)	0.19	(-0.01,0.08)	-0.98	(0.02,0.09)
	int.	Scenario 2: $n = 3000$				1.01	(0.31,0.41)	0.61	(0.41,0.42)
	$x_{j,1}$	-0.07	(0.05)	0.67	(0.08)	0.71	(0.01,0.23)	-0.26	(0.14,0.27)
	$x_{j,2}$	0.19	(0.05)	-0.23	(0.05)	0.05	(-0.55,0.19)	0.56	(-0.65,0.21)
	$x_{j,3}$	-0.25	(0.04)	0.21	(0.04)	0.30	(0.10,0.23)	-0.79	(0.21,0.24)
Setting II	int.	Scenario 1: $n = 7500$				-0.48	(0.12,0.13)	0.38	(0.08,0.15)
	$x_{j,1}$	-0.08	(0.03)	0.62	(0.04)	0.76	(-0.04,0.05)	-1.16	(0.04,0.08)
	$x_{j,2}$	0.13	(0.02)	-0.26	(0.03)	-0.13	(-0.03,0.03)	-0.35	(-0.05,0.04)
	$x_{j,3}$	-0.26	(0.02)	0.16	(0.02)	0.55	(-0.05,0.04)	-0.97	(0.03,0.05)
	int.	Scenario 2: $n = 3000$				-0.48	(0.12,0.21)	0.45	(0.15,0.27)
	$x_{j,1}$	-0.07	(0.06)	0.71	(0.09)	0.76	(-0.04,0.10)	-0.91	(0.29,0.20)
	$x_{j,2}$	0.17	(0.04)	-0.20	(0.06)	-0.13	(-0.03,0.06)	-0.39	(-0.09,0.07)
	$x_{j,3}$	-0.23	(0.04)	0.21	(0.04)	0.50	(-0.10,0.08)	-0.86	(0.14,0.11)
Setting III	int.	Scenario 1: $n = 7500$				-0.59	(0.11,0.15)	-0.95	(0.15,0.18)
	$x_{j,1}$	-0.06	(0.04)	0.64	(0.05)	0.46	(-0.04,0.05)	-1.43	(0.07,0.09)
	$x_{j,2}$	0.16	(0.04)	-0.26	(0.04)	-1.18	(0.02,0.05)	-0.45	(-0.05,0.05)
	$x_{j,3}$	-0.24	(0.03)	0.17	(0.03)	0.63	(-0.07,0.04)	-1.25	(0.05,0.05)
	int.	Scenario 2: $n = 3000$				-0.65	(0.05,0.25)	-0.71	(0.39,0.34)
	$x_{j,1}$	-0.04	(0.08)	0.68	(0.12)	0.44	(-0.06,0.09)	-1.09	(0.41,0.20)
	$x_{j,2}$	0.22	(0.07)	-0.19	(0.08)	-1.14	(0.06,0.09)	-0.49	(-0.09,0.09)
	$x_{j,3}$	-0.21	(0.05)	0.21	(0.06)	0.57	(-0.13,0.06)	-1.09	(0.21,0.12)

Table 2: Mean (standard error) and mean, (bias,standard error) of the parameter estimates for the misspecified model without heterogeneity over 1000 simulation runs, for Setting I (with cure around 12%), Setting II (35% cure) and Setting III (67% cure). The true generating model has two subgroups for the main groups A and B. For this reason there is no true β_A and β_B to compare with.

to scenario 1.

Table 4 compares the standard errors obtained via the supplemented EM algorithm with those obtained from computing the standard deviation over the simulated estimates. For the larger sample size, the estimated standard errors come closer to their empirical counterparts. The results are in particular good for the estimators of the incidence parameters in the logistic regression models. There is more underestimation for the parameters of the Cox regression model. The results become better for the larger sample size in scenario 1.

The aim of the analysis is to determine the main group for the censored observations. In this simulation study, we perform such classification to main groups using either one, two or three subgroups. In Table 5, the percentages of correctly classified observations for the model with two subgroups are listed on the diagonal of the middle part. The percentages of observations belonging to each main group, that are presented in the table, have been

(1)	$\hat{\beta}_{A1}$	$\hat{\beta}_{A2}$	$\hat{\beta}_{B1}$	$\hat{\beta}_{B2}$	$\hat{b}_A$	$\hat{b}_B$
int.	Setting I				0.88 (0.18,0.19)	0.39 (0.19,0.22)
$x_{jk,1}$	-0.51 (-0.31,0.38)	-0.06 (-0.06,0.04)	0.36 (0.06,0.17)	1.05 (-0.25,0.21)	0.64 (-0.06,0.11)	-0.47 (-0.07,0.12)
$x_{jk,2}$	0.01 (-0.09,0.30)	0.17 (-0.23,0.08)	-0.39 (0.11,0.11)	-0.08 (0.02,0.13)	0.49 (-0.11,0.09)	1.06 (-0.14,0.11)
$x_{jk,3}$	-1.00 (-0.50,0.44)	-0.26 (0.04,0.03)	0.14 (0.04,0.07)	0.19 (-0.11,0.11)	0.20 (-0.00,0.11)	-0.99 (0.01,0.11)
int.	Setting II				-0.48 (0.12,0.14)	0.35 (0.05,0.16)
$x_{jk,1}$	-0.51 (-0.31,0.41)	-0.05 (-0.05,0.09)	0.58 (0.28,0.20)	1.52 (0.22,0.32)	0.78 (-0.02,0.06)	-1.21 (-0.01,0.09)
$x_{jk,2}$	0.09 (-0.01,0.31)	0.15 (-0.25,0.05)	-0.29 (0.21,0.07)	-0.13 (-0.03,0.18)	-0.14 (-0.04,0.04)	-0.35 (-0.05, 0.04)
$x_{jk,3}$	-0.99 (-0.49,0.36)	-0.23 (0.07,0.06)	0.16 (0.06,0.07)	0.31 (0.01,0.20)	0.60 (-0.00,0.07)	-0.98 (0.02,0.06)
int.	Setting III				-0.57 (0.13,0.17)	-1.03 (0.07,0.20)
$x_{jk,1}$	-0.32 (-0.12,0.45)	-0.03 (-0.03,0.13)	0.58 (0.28,0.13)	1.64 (0.34,0.41)	0.47 (-0.03,0.05)	-1.50 (0.00,0.11)
$x_{jk,2}$	0.28 (0.18,0.40)	0.21 (-0.19,0.23)	-0.29 (0.21,0.07)	-0.11 (-0.01,0.24)	-1.22 (-0.02,0.06)	-0.45 (-0.05,0.05)
$x_{jk,3}$	-0.67 (-0.17,0.39)	-0.21 (0.09,0.08)	0.15 (0.05,0.05)	0.39 (0.09,0.27)	0.67 (-0.03,0.06)	-1.27 (0.03,0.06)
(2)	$\hat{\beta}_{A1}$	$\hat{\beta}_{A2}$	$\hat{\beta}_{B1}$	$\hat{\beta}_{B2}$	$\hat{b}_A$	$\hat{b}_B$
int.	Setting I				1.25 (0.55,0.53)	0.81 (0.61,0.51)
$x_{jk,1}$	-0.54 (-0.34,0.51)	-0.03 (-0.03,0.07)	0.51 (0.21,0.15)	1.61 (0.31,0.45)	0.74 (0.04,0.26)	-0.25 (0.15,0.29)
$x_{jk,2}$	0.18 (0.08,0.40)	0.25 (-0.15,0.09)	-0.30 (0.20,0.13)	-0.14 (-0.04,0.36)	0.07 (-0.53,0.23)	0.59 (-0.61,0.24)
$x_{jk,3}$	-0.82 (-0.32,0.48)	-0.22 (0.08,0.04)	0.19 (0.09,0.06)	0.50 (0.20,0.40)	0.32 (0.12,0.25)	-0.79 (0.21,0.26)
int.	Setting II				-0.37 (0.23,0.28)	0.51 (0.21,0.29)
$x_{jk,1}$	-0.51 (-0.31,0.57)	-0.03 (-0.03,0.09)	0.63 (0.33,0.12)	2.04 (0.74,0.67)	0.78 (-0.02,0.14)	-0.97 (0.23,0.23)
$x_{jk,2}$	0.14 (0.04,0.39)	0.19 (-0.21,0.07)	-0.24 (0.26,0.08)	-0.02 (0.08,0.43)	-0.15 (-0.05,0.07)	-0.40 (-0.10,0.08)
$x_{jk,3}$	-0.80 (-0.30,0.41)	-0.19 (0.11,0.04)	0.18 (0.08,0.05)	0.61 (0.31,0.42)	0.55 (-0.05,0.12)	-0.87 (0.13,0.13)
int.	Setting III				-0.56 (0.14,0.30)	-0.68 (0.42,0.36)
$x_{jk,1}$	-0.28 (-0.08,0.49)	0.00 (0.00,0.18)	0.58 (0.28,0.16)	1.79 (0.49,0.75)	0.45 (-0.05,0.10)	-1.16 (0.34,0.23)
$x_{jk,2}$	0.38 (0.28,0.45)	0.29 (-0.11,0.25)	-0.24 (0.26,0.12)	0.04 (0.14,0.52)	-1.18 (0.02,0.10)	-0.49 (-0.09,0.10)
$x_{jk,3}$	-0.55 (-0.05,0.36)	-0.17 (0.13,0.08)	0.18 (0.08,0.09)	0.57 (0.27,0.43)	0.60 (-0.10,0.08)	-1.12 (0.18,0.13)

Table 3: Mean (bias,standard error) of the parameter estimates for the correctly specified model with two subgroups for both A and B over 1000 simulation runs, for Setting I with low cure, Setting II with medium cure and Setting III with high cure. Top panel: Scenario 1 with $n = 7500$, bottom panel: Scenario 2 with $n = 3000$.

obtained by averaging the weights w_j over all observations and further averaging over all simulation runs. When comparing these results with the percentages of correctly classified observations of the model without heterogeneity (on the left part of the same table), we observe a high percentage of correctly classified observations for all settings. Note that classification is better for groups A and B in the model with two subgroups. This is offset by a worse classification of the cured cases C, especially in Setting I. In general, we note that classification on the main group level does not incontestably favor one model over another. This is not an unexpected result, as classification is driven by the parameters b, which only marginally change when changing the number of subgroups (see the b-parameters in Tables 2 and 3). A similar result was observed when looking at the b-parameters of a model when having three subgroups (see also Section 4).

For the correctly specified models for settings I and II it never happened in the 1000 simulated datasets that the parameter vectors for the two subgroups were estimated with nearly identical values. For setting III (high cure percentage) it only happened in 1 out of 1000 simulated datasets for the parameter vector related to early repayment. Thus in that single case the model suggests a single group. This comparison for equality of the estimates is reconsidered for estimation with three subgroups, see Table 7.

$n = 3000$	b_p				b_d				β_{p1}			β_{p2}			β_{d1}			β_{d2}		
I sem	0.28	0.09	0.07	0.06	0.31	0.10	0.09	0.07	0.14	0.11	0.11	0.04	0.03	0.02	0.05	0.04	0.03	0.18	0.11	0.10
empstd	0.53	0.26	0.23	0.25	0.51	0.29	0.24	0.26	0.51	0.40	0.48	0.07	0.09	0.04	0.15	0.13	0.06	0.45	0.36	0.40
II sem	0.22	0.07	0.05	0.04	0.23	0.08	0.06	0.06	0.16	0.11	0.12	0.04	0.03	0.03	0.06	0.04	0.03	0.34	0.16	0.15
empstd	0.28	0.14	0.07	0.12	0.29	0.23	0.08	0.13	0.57	0.39	0.41	0.09	0.07	0.04	0.12	0.08	0.05	0.67	0.43	0.42
III sem	0.25	0.08	0.08	0.05	0.27	0.09	0.08	0.08	0.15	0.14	0.11	0.08	0.08	0.05	0.08	0.06	0.05	0.39	0.20	0.19
empstd	0.30	0.10	0.10	0.08	0.36	0.23	0.10	0.13	0.49	0.45	0.36	0.18	0.25	0.08	0.16	0.12	0.09	0.75	0.52	0.43
$n = 7500$	b_p				b_d				β_{p1}			β_{p2}			β_{d1}			β_{d2}		
I sem	0.17	0.06	0.05	0.03	0.19	0.07	0.06	0.04	0.11	0.08	0.10	0.02	0.02	0.01	0.03	0.02	0.02	0.05	0.04	0.03
empstd	0.19	0.10	0.09	0.10	0.20	0.12	0.11	0.10	0.33	0.27	0.37	0.03	0.04	0.02	0.08	0.09	0.03	0.20	0.17	0.13
II sem	0.13	0.04	0.03	0.02	0.13	0.05	0.03	0.04	0.12	0.08	0.10	0.02	0.02	0.01	0.03	0.02	0.02	0.14	0.07	0.06
empstd	0.13	0.06	0.04	0.06	0.16	0.09	0.04	0.05	0.40	0.28	0.31	0.04	0.03	0.02	0.06	0.04	0.02	0.28	0.20	0.22
III sem	0.15	0.05	0.05	0.03	0.16	0.06	0.04	0.05	0.11	0.10	0.08	0.04	0.04	0.02	0.04	0.03	0.02	0.18	0.10	0.09
empstd	0.16	0.05	0.06	0.06	0.20	0.10	0.05	0.06	0.36	0.31	0.33	0.08	0.12	0.04	0.08	0.05	0.03	0.33	0.22	0.25

Table 4: Averages of estimated standard deviations via SEM and empirical standard deviations over 1000 simulation runs for setting I with low cure, II with medium cure and III with high cure percentage. Top: scenario 2 with $n = 3000$ and high censoring. Bottom: scenario 1 with n=7500 and less censoring.

			Classified as (in %)								
			no heterogeneity			two subgroups			three subgroups		
			A	B	C	A	B	C	A	B	C
I	Reality	group A	94.74	0.56	4.70	95.55	0.65	3.79	95.63	0.66	3.71
		group B	0.66	94.70	4.64	0.80	95.80	3.40	0.82	95.82	3.36
		group C	12.89	9.50	77.62	17.95	13.95	68.09	18.60	14.06	67.34
II	Reality	group A	92.64	0.15	7.21	93.58	0.15	6.26	93.66	0.16	6.18
		group B	0.16	91.03	8.80	0.21	91.66	8.14	0.21	91.74	8.04
		group C	4.87	5.61	89.52	7.43	6.85	85.72	7.72	7.03	85.25
III	Reality	group A	89.08	0.05	10.87	90.07	0.06	9.87	95.63	0.66	3.71
		group B	0.05	88.65	11.30	0.06	89.45	10.48	0.82	95.82	3.36
		group C	1.36	2.21	96.43	2.38	2.78	94.84	18.60	14.06	67.34

Table 5: Percentage of observations classified to each of the groups over 1000 simulation runs for scenario 1 and settings I–III. The left part shows this for the model without heterogeneity, the middle part for the correct model with two subgroups for A and B, and the right part gives the classification percentages for the overspecified model with three subgroups for A and B. The percentages of correct classification are on the diagonals of each part.

Comparison of model with and without heterogeneity

The main advantage of the use of a heterogeneous model when there is some heterogeneity in the data lies in the possibility to model more accurate event times. To illustrate this, we investigate the estimates of the survival probabilities of the subjects using (i) a model without heterogeneity and (ii) with two subgroups per main group, and compare these estimates to the true survival probabilities of the simulated data. For the 1000 simulation runs, we considered the true survival rate in each subgroup (A1, A2, B1, B2), for each decile from 0.2 to 0.7 of all 7500 event times. These are compared to the average estimated survival

		decile rank=2				decile rank=3				decile rank=4			
		$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$
I	1 group	0.177	0.164*	0.008	0.005*	0.253	0.234	0.029	0.020*	0.281	0.262	0.062	0.045*
	2 subgroups	0.075*	0.300	0.006*	0.031	0.099*	0.192*	0.023*	0.067	0.101*	0.100*	0.051*	0.121
II	1 group	0.254	0.228*	0.008*	0.013*	0.296	0.268	0.022	0.008*	0.250	0.226	0.066	0.034*
	2 subgroups	0.118*	0.242	0.009	0.083	0.115*	0.112*	0.018*	0.188	0.086*	0.062*	0.057*	0.335
III	1 group	0.282	0.218	0.046	0.014*	0.142	0.054*	0.071	0.018*	0	0	0	0
	2 subgroups	0.148*	0.123*	0.038*	0.341	0.111*	0.058	0.056*	0.504	0	0	0	0
		decile rank=5				decile rank=6				decile rank=7			
		$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$	$A1$	$A2$	$B1$	$B2$
I	1 group	0.256	0.241	0.102	0.074*	0.201	0.188	0.132	0.096*	0.139	0.124	0.132	0.096*
	2 subgroups	0.092*	0.054*	0.085*	0.193	0.099*	0.035*	0.108*	0.262	0.121*	0.023*	0.103*	0.302
II	1 group	0.172	0.147	0.091	0.051*	0.087*	0.047	0.067	0.022*	0	0	0	0
	2 subgroups	0.097*	0.041*	0.080*	0.461	0.132	0.020*	0.053*	0.458	0	0	0	0
III	1 group	0	0	0	0	0	0	0	0	0	0	0	0
	2 subgroups	0	0	0	0	0	0	0	0	0	0	0	0

Table 6: The mean of the absolute differences between the population survival rate and the estimated survival probabilities for the model without heterogeneity (1 group) and with heterogeneity (2 groups) for censoring settings I–III and scenario 1. Six different time points are analyzed, using the real event-time deciles. The asterisk indicates where the performance was better, each time using a pairwise comparison between the model with 1 or 2 (sub)groups). Note that zeroes in this table are exact and not rounded, as here both estimated survival probabilities and population survival rate are equal (both zero).

probability of all observations in groups A1, A2, B1 and B2, first using a homogeneous model, and secondly using a model with two subgroups. The result can be found in Table 6.

For each simulation setting at each of the six listed time decile ranks, the estimates of the proportion of the populations that experienced the event not later than time t_d, $S_{jk}(t_0) - S_{jk}(t_d) = 1 - S_{jk}(t_d)$, with t_d equal to the generated time decile rank d and t_0 the starting time of the study (hence, before any event has occurred) are compared to the true proportions. Table 6 lists the absolute differences between the estimated and the true proportions, either using a model with or without heterogeneity. When the model without heterogeneity performs better than the model with heterogeneity (this is, when the absolute difference for "1 group" is lower than for "2 subgroups"), the numbers in Table 6 for 1 group received an asterisk. Similar for the numbers for two subgroups where the model with heterogeneity performed better. Note that, for high censoring rates and for bigger time decile ranks, the survival probability estimate goes to 0 for both models, and both models perform about equally (regular text font). Note that in a large majority of the cases (34 out of 52, not considering the zeros), the model with heterogeneity performed either equally well or better than the model without heterogeneity.

Model with heterogeneity: three subgroups

The third model that was fit to the simulated data was a model with three subgroups for both main groups A and B. For this model, it appears that in many of the simulation runs, the $\hat{\beta}_{jk}$ for several subgroups are converging to (approximately) equal values, see Table 7. This is a result of putting several subgroup probabilities equal to $1/K_j$, as discussed in Section 3. In 606, 737 and 756 out of the 1000 simulation runs (for Setting I to III respectively), at least two out of three β_{jk} for groups A, B or both were estimated to be equal, with equal shares of observations classified in those equally estimated groups. The occurrence of equal parameter estimates is a strong indication that too many subgroups were modeled and that the number of subgroups K_j should be decreased. In the data example, we use this in combination with a version of Akaike's Information criterion to determine the number of subgroups.

group A	group B	Setting I	Setting II	Setting III
$\hat{\beta}_{A1} \neq \hat{\beta}_{A2} \neq \hat{\beta}_{A3}$	$\hat{\beta}_{B1} \neq \hat{\beta}_{B2} \neq \hat{\beta}_{B3}$	394	263	244
$\hat{\beta}_{A1} = \hat{\beta}_{A2} \neq \hat{\beta}_{A3}$	$\hat{\beta}_{B1} \neq \hat{\beta}_{B2} \neq \hat{\beta}_{B3}$	171	125	164
$\hat{\beta}_{A1} \neq \hat{\beta}_{A2} \neq \hat{\beta}_{A3}$	$\hat{\beta}_{B1} = \hat{\beta}_{B2} \neq \hat{\beta}_{B3}$	315	438	363
$\hat{\beta}_{A1} = \hat{\beta}_{A2} \neq \hat{\beta}_{A3}$	$\hat{\beta}_{B1} = \hat{\beta}_{B2} \neq \hat{\beta}_{B3}$	120	174	229

Table 7: Analysis of the parameter estimates for heterogeneity with 3 subgroups for groups A and B, for censoring settings I–III, scenario 1. In the majority of the simulation runs, there were equal estimates for 2 out of the 3 β_{jk} for A or B, or both.

5 Credit risk data example

Data description

We analyze a credit loan data set, with the main interest in the prediction of defaults and early loan repayments. The cured or unsusceptible group is given by the matured loans with the loan repayment on the predefined end date (maturity). The data concern personal loans and are from a UK bank, previously used in Stepanova and Thomas (2002), Tong et al. (2012) and Dirick et al. (2015). We use the data set consisting of 7521 observations with loan term 36 months. Table 8 lists the eight variables that were used to build our model. The event of early repayment was observed 2992 times, default 376 times and maturity 269 times. The remaining 3884 other observations were censored.

Decision on the number of subgroups

To determine the value of K_j for each of the $J-1$=2 main groups, or rephrased to this data set, the K_d subgroups for default and the K_p subgroups for early repayment, we use a version of Akaike's information criterion (AIC) that accounts for incomplete data. Introduced by Cavanaugh and Shumway (1998), the so-called "complete-data AIC" (AIC$_{cd}$) makes use of the expected complete-data log likelihood instead of the observed log likelihood. Dirick et al. (2015) obtained the AIC$_{cd}$ for multiple event mixture cure models. In this context,

$$\text{AIC}_{cd} = -2Q^h\big(\widehat{\Theta} \mid \widehat{\Theta}\big) + 2d + 2\,\text{trace}\{DM(I_d - DM)^{-1}\}.$$

An additional selection restriction is that increasing the number of subgroups K_j should be stopped when the estimates $\hat{\beta}_{jk}$ of different components are the same.

	Description	Type	**default**	**early repayment**	**cured**	**unknown (censored)**
x_1	The gender of the customer (1=M, 0=F)	cat	75.0%	71.2%	71.7%	73.6%
x_2	Amount of the loan	cont	3588.8 (1842.4)	3607.4 (1756.8)	3740.9 (1900.2)	3630.7 (1805.9)
x_3	Number of years at current address	cont	6.0 (6.9)	7.2 (7.5)	8.2 (8.0)	8.2 (7.9)
x_4	Number of years at current employer	cont	4.5 (4.8)	7.6 (7.4)	9.2 (9.0)	8.1 (7.7)
x_5	Amount of insurance premium	cont	342.3 (293.0)	219.3 (260.0)	200.9 (260.0)	231.8 (272.8)
x_6	Home phone or not (1=N,0=Y)	cat	6.9%	3.7%	3.7%	4.2%
x_7	Own house or not (1=N, 0=Y)	cat	46.5%	31.4%	26.8%	32.9%
x_8	Payment frequency (1=low/unknown, 0=high)	cat	56.6%	69.5%	66.9%	67.0%

Table 8: Data on credit risk. Description of the variables, continuous (cont) or categorical (cat), stratified by failure event. For continuous variables, the observed mean (and standard deviation) is given, for categorical variables (which are all binary) the proportion of one-values.

Since the classical Akaike information criterion and Bayesian information criterion are not defined for these models, we imitate their construction and use 'AIC' $= -2Q^h\big(\widehat{\Theta} \mid \widehat{\Theta}\big) + 2d$ and 'BIC' $= -2Q^h\big(\widehat{\Theta} \mid \widehat{\Theta}\big) + 2\log(n)$. Note that no theoretical guarantee is present for these imitations and in particular we do not expect these criteria to have the same properties as their studied counterparts in the literature (see, e.g. Claeskens and Hjort, 2008).

For the credit risk data, we considered values for both K_p and K_d in $\{1, 2, 3\}$, with the value of one representing homogeneity. Using all combinations for K_p and K_d gave rise to 9 models. Table 9 presents the values of AIC_{cd}, and the imitations of AIC and BIC. The top three models for all three criteria contain three heterogeneity groups for early repayment. The best model has a single group for default. When looking at the estimated values of the β_{pk}, however, we received (up to rounding) equal estimates for two out of three parameter vectors. This was also the case for β_{pk} estimates for (K_p=3, $K_d = 2$) and (K_p=3, $K_d = 3$), which have the next lowest AIC_{cd}-values. The next lowest AIC_{cd}-value has K_p=2, $K_d = 1$, which is the preferred value as there we have the model with minimal AIC_{cd} without equal estimates for β_{pk}-parameters. The situation with approximately equal estimates is a consequence of assigning equal weights to the subgroups in case the weights $v_{k|j}(\beta_j; t_i, x_i)$ would go to infinity. The reduction of the number of subgroups by setting the weights equal and hence also leading to approximately equal estimates solves such a situation where no different subgroups should be identified. Hence, the suggested final model is one with only one group for default, and two subgroups for early repayment.

The runtime of the EM algorithm depends on the choice of the starting values as well as on the number of subgroups. For the homogeneous model with $K_p = K_d = 1$ on the full dataset with 7521 observations, the algorithm using the software R needed 98 iterations to convergence and this took 3.1 minutes on a computer with Processor Intel Core i7-4600

K_p	K_d	AICcd	‘AIC’	‘BIC’
1	1	61997.10	61978.41	62213.88
1	2	62470.99	62440.43	62731.30
1	3	62496.09	62458.78	62805.05
2	1	58041.93 (4)	58022.49 (4)	58313.36 (4)
2	2	58479.48	58448.51	58794.78
2	3	58487.28	58449.92	58851.60
3	1	55795.23 (1)	55775.60 (1)	56121.87 (1)
3	2	56220.73 (2)	56189.64 (3)	56591.31 (2)
3	3	56224.52 (3)	56187.28 (2)	56644.36 (3)

Table 9: Data on credit risk. For each model the first column states first the number of subgroups for early repayment (K_p), and next for default (K_d). Criterion values of AIC$_{cd}$, ‘AIC’ and ‘BIC’ which use Q^h instead the maximized log likelihood are shown together with the ranks for the four best models.

CPU @ 2.10 GHz 2.70 GHz, 64-bit operating system. No parallel computing was used. With the very same starting values (thus not optimized from estimates of other models) the model for $K_p = 2$ and $K_d = 1$ needed 499 iterations and this took 12.0 minutes. These calculations include the extra calculations of the supplemented EM algorithm to provide standard errors. Of course, this could have been sped up by using smarter starting values and by using parallel computing.

Final result

	$\hat{\beta}_{d1}$	$\hat{\beta}_{p1}$	$\hat{\beta}_{p2}$	$\hat{b}_d$	$\hat{b}_p$
τ	1	0.587	0.413		
int.				-1.290 (0.25)	0.535 (0.16)
x_1	0.245 (0.19)	0.022 (0.10)	0.156 (0.10)	-0.120 (0.16)	-0.190 (0.10)
x_2	$-7 \cdot 10^{-5}$ $(5 \cdot 10^{-5})$	$2 \cdot 10^{-5}$ $(3 \cdot 10^{-5})$	$-9 \cdot 10^{-5}$ $(4 \cdot 10^{-5})$	$-3 \cdot 10^{-5}$ $(5 \cdot 10^{-5})$	$2 \cdot 10^{-5}$ $(2 \cdot 10^{-5})$
x_3	-0.022 (0.01)	-0.013 $(7 \cdot 10^{-3})$	-0.009 (0.01)	-0.030 (0.01)	-0.013 (0.005)
x_4	-0.051 (0.02)	-2$\cdot 10^{-4}$ (0.01)	-0.004 (0.01)	-0.067 (0.02)	-0.011 (0.005)
x_5	$4 \cdot 10^{-4}$ $(3 \cdot 10^{-4})$	$-3 \cdot 10^{-4}$ $(2 \cdot 10^{-4})$	-10^{-4} $(2 \cdot 10^{-4})$	0.001 $(3 \cdot 10^{-4})$	10^{-4} $(2 \cdot 10^{-4})$
x_6	0.560 (0.31)	-0.234 (0.20)	-0.420 (0.25)	0.393 (0.31)	0.123 (0.27)
x_7	-0.042 (0.18)	-0.020 (0.09)	-0.192 (0.13)	0.408 (0.15)	0.106 (0.08)
x_8	-0.032 (0.22)	0.090 (0.08)	0.220 (0.10)	-0.380 (0.16)	-0.066 (0.10)

Table 10: Data on credit risk. Parameter estimates (bootstrap standard errors) for the hierarchical mixture cure model with $K_d = 1$, $K_p = 2$. The value τ represents the proportion of the population belonging to a respective subgroup, given the main group, ‘int.’ stands for intercept.

The parameter estimates of the final model are given in Table 10, along with their standard errors. Setting III of the simulation study is similar with regard to the approximately 50% censoring present in this dataset. Hence, we seem to get consistent estimates for β, but for b, there might be small deviations due to the relatively high censoring percentage.

For the latter parameter group, mainly the sign and the relative magnitude should be used in the analysis.

First, we discuss $\hat{b}_d$ and $\hat{b}_p$ which affect the probabilities of default and prepayment, which are smaller for men than for women. The effect on prepayment is stronger and statistically significant. The residential stability (x_3) and employment stability (x_4) reduce significantly both probabilities, but the effect on the probability of default is much stronger. Having no home phone (x_6) and no own home (x_7) greatly increase both probabilities, the relative effect on the default probabilities is even stronger here. However, only the coefficient of x_7 for default is statistically significant. The low frequency of payment reduces significantly the default probability. To summarize, the variables affect the probability of defaults more than the probability of prepayment and the signs of the coefficients are reasonable.

Second, we discuss the estimated parameters of the survival function, $\hat{\beta}_{d1}$, $\hat{\beta}_{p1}$ and $\hat{\beta}_{p2}$. When men go into default and prepayment the time to event is much shorter, though it is statistically significant only for the second prepayment group. The residential stability (x_3) and employment stability (x_4) delay the timing of defaults and prepayments and mostly statistically significant. Having no home phone (x_6) accelerates defaults, but delays the prepayments. Having no own home (x_6) significantly delays the prepayment of the second default group. The low frequency of payments delays the defaults, but accelerates the prepayments. The variables can have different effects in the probability of the event and the conditional survival function. The gender dummy variable (x_1) decreases the probability of prepayment or default for men, but if men experience the event, it happens faster. No home phone (x_6) and no own home (x_7) increases the prepayment probability, but delays the timing of prepayment. Finally, when two prepayment groups are compared, all coefficients for the second group are larger. The first group can be called the 'base group', while the second group can be called the 'sensitive' group. All factors have much stronger effects in the 'sensitive' group. To illustrate the difference in terms of early repayment probabilities between the base group and the sensitive group, the two curves representing early repayment for two random observations are plotted in Figure 1 (for one subject in the left panel and for the other in the right panel). The solid and short-dashed curves represent early repayment groups 1 and 2 respectively. The long-dashed curve represents the early repayment curve for the two random observations fitted using a model without heterogeneity. This figure illustrates that quite different results are obtained when not using models with heterogeneity. Note that the two curves can be further apart for some observations (as on the right panel), and closer for other observations (as on the left panel).

We considered typical measures for model evaluation such as mean absolute error (MAE) and mean squared error (MSE) to predict the time of default. Hereby we followed Dirick et al. (2017), see also Zhang and Thomas (2012). The dataset is divided in a test set consisting of 1/3rd of the observations and a training set consisting of the remaining 2/3rd of the observations. We want to obtain an MSE and MAE comparing the "observed" event times with the event times produced by our model. The model, however, does not produce one single event time, but a survival function. Instead of calculating one MSE and one MAE, we calculated 1000 MSEs and MAEs for the training set, each one corresponding

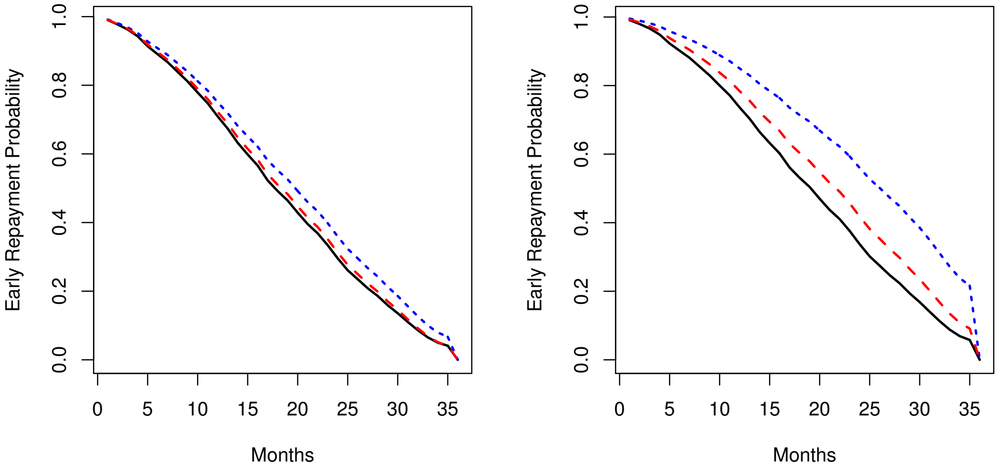

Figure 1: Credit loan data. Estimated survival curves for two observations. The solid and short-dashed lines are the survival curves (estimated through formula (3) and the Breslow-type estimator for the baseline hazard) for early repayment group 1 and group 2, for the final model where heterogeneity is present in the early repayment-group. The long-dashed line represents the estimated survival curve fitted with a model with no heterogeneity (hence $\widehat{S}(t \mid \tilde{Y} = p, x_i, \hat{\beta}_p)$, assuming no subgroups).

with each permille of the survival function. The permille value that resulted in the best training set MSE and MAE were then withheld to compute the training set MSE and MAE for just that permille. Averaged MSE and MAEs over all the test set observations are shown in Table 11.

MSE	$K_d = 1$	$K_d = 2$	$K_d = 3$
$K_p = 1$	129.34	133.16	115.68
$K_p = 2$	129.42	131.58	117.91
$K_p = 3$	129.42	131.80	115.07

MAE	$K_d = 1$	$K_d = 2$	$K_d = 3$
$K_p = 1$	8.66	8.57	8.10
$K_p = 2$	8.67	8.57	8.10
$K_p = 3$	8.67	8.52	8.07

Table 11: Mean squared error and mean absolute error for predictions on a test set consisting of 1/3rd of the data, based on estimates obtained from a training set consisting of 2/3rd of the data.

Both MAE and MSE have the lowest value for the largest model with three subgroups for early repayment and three subgroups for default. Note that this method is not developed for model selection and thus is not expected to give detailed information on the model to be used for the full dataset. All models with a single group for default, including the homogeneous model, are the worst in this comparison.

6 Conclusion

This paper highlights the importance of unobserved heterogeneity for credit risk models. It derives a hierarchical EM algorithm for estimation and provides the first simulation study in this area, which reveals that not using models that allow heterogeneity when heterogeneity is present can lead to distorted conclusions on the magnitude of the parameters related to the timing of a certain event.

An application of the model on loan data from a UK bank finds that the explanatory variables can act in different directions upon incidence and duration; and, variables exist that are statistically significant in explaining only incidence or duration.

While the model proposed is general and can be used in many contexts, there are still many aspects that would be interesting for future research. First of all, it would be interesting to consider other survival analysis techniques for cure models for extension towards heterogeneity, such as the promotion time cure model, or incorporating nonparametric effects. The presence of censoring and its effects on estimation also requires more investigation. One possibility is to explicitly incorporate censoring as missing information in the EM-algorithm.

Finally, alternative distributions can be used to model the subgroup densities, e.g., the skew-t distribution employed in Murray et al. (2017), and the approach of Hasnat et al. (2017) can be used to investigate the evolution of subgroups over time.

Acknowledgements

We acknowledge the support of the Research Foundation Flanders and KU Leuven grant C1/16/20/002. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government – department EWI. We thank the reviewers and editor for their constructive remarks.

A Appendix

A.1 Identifiability of the main groups of the hierarchical mixture cure model.

Similar as in Heckman and Honoré (1989); Peng and Zhang (2008), we start from (2). For a model with three main groups, without subgroups, the unconditional survival function can be written as

$$S(t \mid \boldsymbol{x}, \boldsymbol{z}, b, \beta) = \sum_{j=1}^{2} \pi_j(\boldsymbol{z}, b_j) S(t \mid Y_j = 1, \boldsymbol{x}, \beta_j) + 1 - \sum_{j=1}^{2} \pi_j(\boldsymbol{z}, b_j), \tag{10}$$

where

$$S(t \mid Y_j = 1, \boldsymbol{x}, \beta_j) = \exp\big\{ - \exp(\boldsymbol{x}^T \beta_j) \int_0^t h_0(u \mid \widetilde{Y}_j = 1) du \big\} \equiv \exp\left(-\phi_j(\boldsymbol{x}) H_0^j(t)\right),$$

with $\exp(\boldsymbol{x}^\top \beta_j) = \phi_j(\boldsymbol{x})$ for $j = 1, 2$. Omitting parameters for notational convenience,

$$S(t \mid \boldsymbol{x}, \boldsymbol{z}) = 1 + \sum_{j=1}^{2} \{\pi_j(\boldsymbol{z}) \exp\{-\phi_j(\boldsymbol{x}) H_0^j(t)\} - \pi_j(\boldsymbol{z})\}.$$

Theorem 1. *Assume that for $j = 1, 2$,*
(A1) The cumulative hazard function satisfies $\lim_{t \to 0} H_0^j(t) = H_0^j(0) = 0$,
(A2) $\pi_j(\boldsymbol{z})$ is non-negative and non-constant,
(A3) $\phi_j(\boldsymbol{x})$ is non-negative, differentiable and non-constant with $\phi_j(0) = 1$.
Then, $S(t \mid \boldsymbol{x}, \boldsymbol{z})$ as in (10) is identifiable if and only if for any sets $\{\pi_j(\boldsymbol{z}), \phi_j(\boldsymbol{x}), H_0^j; j = 1, 2\}$, and $\{\pi_j^*(\boldsymbol{z}), \phi_j^*(\boldsymbol{x}), H_0^{j*}; j = 1, 2\}$ such that

$$\sum_{j=1}^{2} \{\pi_j(\boldsymbol{z}) \exp\left(-\phi_j(\boldsymbol{x}) H_0^j(t)\right) - \pi_j(\boldsymbol{z})\} = \sum_{j=1}^{2} \{\pi_j^*(\boldsymbol{z}) \exp\left(-\phi_j^*(\boldsymbol{x}) H_0^{*j}(t)\right) - \pi_j^*(\boldsymbol{z})\},$$

it follows that $\pi_j(\boldsymbol{z}) = \pi_j^*(\boldsymbol{z})$, $\phi_j(\boldsymbol{x}) = \phi_j^*(\boldsymbol{x})$, $H_0^j = H_0^{*j}$, for $j = 1, 2$.

Proof. We define 'crude' survival functions $K_1(t)$ and $K_2(t)$ as the probability of not experiencing event type 1, resp. 2, before time t,

$$\begin{aligned} K_1(t \mid \boldsymbol{x}, \boldsymbol{z}) &= P((T_1 > t) \cap (T_2 > T_1) \mid \boldsymbol{x}, \boldsymbol{z}), \\ K_2(t \mid \boldsymbol{x}, \boldsymbol{z}) &= P((T_2 > t) \cap (T_1 > T_2) \mid \boldsymbol{x}, \boldsymbol{z}). \end{aligned}$$

Tsiatis (1975) obtained that for $j = 1, 2$,

$$\frac{\partial K_j}{\partial t}(t \mid \boldsymbol{x}, \boldsymbol{z}) = \left[\frac{\partial S}{\partial t_j}(t \mid \boldsymbol{x}, \boldsymbol{z})\right]_{t_1 = t_2 = t}. \tag{11}$$

Fix $j \in \{1, 2\}$. It follows that

$$\frac{\partial K_j}{\partial t}(t) = -\pi_j(\boldsymbol{z}) \phi_j(\boldsymbol{x}) h_0^j(t) \exp\left(-\phi_j(\boldsymbol{x}) H_0^j(t)\right) \equiv s^j(t \mid \boldsymbol{x}, \boldsymbol{z}).$$

First we show that there exists a constant c_j such that

$$h_0^{*j}(t) \exp\left(-\phi_j^*(\boldsymbol{x}) H_0^{*j}(t)\right) = c_j h_0^j(t) \exp\left(-\phi_j(\boldsymbol{x}) H_0^j(t)\right). \tag{12}$$

<u>*Case 1: $x = z$*</u>. Take any constant x_0 in the domain of $s^1(\cdot \mid \boldsymbol{x})$. Dividing $s^j(t \mid \boldsymbol{x})$ by $s^j(t \mid \boldsymbol{x_0})$ and letting $t \to 0$ we get by assumption (A1)

$$-\pi_j(\boldsymbol{x_0}) \phi_j(\boldsymbol{x_0}) = -\pi_j(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \lim_{t \to 0} \frac{s^j(t \mid \boldsymbol{x})}{s^j(t \mid \boldsymbol{x_0})}.$$

Likewise, $-\pi_j^*(\boldsymbol{x_0}) \phi_j^*(\boldsymbol{x_0}) = -\pi_j^*(\boldsymbol{x}) \phi_j^*(\boldsymbol{x}) \lim_{t \to 0} \frac{s^j(t \mid \boldsymbol{x})}{s^j(t \mid \boldsymbol{x_0})}$. Consequently,

$$\frac{\pi_j(\boldsymbol{x_0}) \phi_j(\boldsymbol{x_0})}{\pi_j^*(\boldsymbol{x_0}) \phi_j^*(\boldsymbol{x_0})} = \frac{\pi_j(\boldsymbol{x}) \phi_j(\boldsymbol{x})}{\pi_j^*(\boldsymbol{x}) \phi_j^*(\boldsymbol{x})} \equiv c_j,$$

hence $\pi_j(\boldsymbol{x})\phi_j(\boldsymbol{x})$ can be determined upon a constant. By differentiating $S(t \mid \boldsymbol{x})$ with respect to t, (12) results.

Case 2: $x \neq z$. By dividing $s^j(t \mid \boldsymbol{x}, \boldsymbol{z})$ by $s^j(t \mid 0, \boldsymbol{z})$ and letting $t \to 0$ gives by assumptions A1, and A3, that $\lim_{t\to 0} \frac{s^j(t \mid \boldsymbol{x}, \boldsymbol{z})}{s^j(t \mid 0, \boldsymbol{z})} = \phi_j(\boldsymbol{x})$. Hence $\phi_j(\boldsymbol{x})$ is uniquely determined. Consider such another set $\{\pi_j^*(\boldsymbol{z}), \phi_j(\boldsymbol{x}), H_0^{j*}; j = 1, 2\}$. Then, for any value $\boldsymbol{z}_0$ in the domain of s^j, dividing $s^j(t \mid \boldsymbol{x}, \boldsymbol{z})$ by $s^j(t \mid \boldsymbol{x}, \boldsymbol{z_0})$ and letting $t \to 0$ leads to $\lim_{t\to 0} \frac{s^j(t \mid \boldsymbol{x}, \boldsymbol{z})}{s^j(t \mid \boldsymbol{x}, \boldsymbol{z_0})} = \frac{\pi_j(\boldsymbol{z})}{\pi_j(\boldsymbol{z_0})} = \frac{\pi_j^*(\boldsymbol{z})}{\pi_j^*(\boldsymbol{z_0})}$. Hence, $\frac{\pi_j^*(\boldsymbol{z})}{\pi_j(\boldsymbol{z})}$ must be a constant, which we define as c_j. Using these results, differentiating $S(t \mid \boldsymbol{x}, \boldsymbol{z})$ with respect to t, (12) results.

Let $t \to 0$ on both sides of (12), then we get the well-defined $\lim_{t\to 0} \frac{h_0^{*j}(0)}{h_0^j(0)} = c_j$. When we take the derivatives on both sides of (12) with respect to x, we obtain by (A3),

$$h_0^{*j}(t) H_0^{*j}(t) \frac{\partial \phi_j^*(\boldsymbol{x})}{\partial \boldsymbol{x}} \exp\bigl(-\phi_j^*(\boldsymbol{x}) H_0^{*j}(t)\bigr) = c_j h_0^j(t) H_0^j(t) \frac{\partial \phi_j(\boldsymbol{x})}{\partial \boldsymbol{x}} \exp\bigl(-\phi_j(\boldsymbol{x}) H_0^j(t)\bigr). \tag{13}$$

Let $t \to 0$ on both sides of (13). Since $\lim_{t\to 0} \frac{H_0^j(0)}{H_0^{*j}(0)} = \frac{1}{c_j} = \lim_{t\to 0} \frac{h_0^j(0)}{h_0^{*j}(0)}$,

$$\frac{\partial \phi_j^*(\boldsymbol{x})}{\partial \boldsymbol{x}} \Big/ \frac{\partial \phi_j(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{1}{c_j}. \tag{14}$$

Integrating Equation (14) with respect to x, we get by (A3)

$$\phi_j^*(\boldsymbol{x}) = \frac{1}{c_j}\phi_j(\boldsymbol{x}) - \frac{1}{c_j} + 1. \tag{15}$$

Take $x = 0$. Because $\phi(0) = \phi^*(0) = 1$ from (A3), (12) simplifies to

$$c_j h_0^j(t) \exp\left(-H_0^j(t)\right) = h_0^{*j}(t) \exp\left(-H_0^{*j}(t)\right). \tag{16}$$

From the ratios of (12) and (16), we get that $H_0^{*1}(\boldsymbol{x})(\phi_1^*(\boldsymbol{x}) - 1) =$ $H_0^1(\boldsymbol{x})(\phi_1(\boldsymbol{x}) - 1)$. Using (15), it is then easy to show that $H_0^{*j}(t) = c_j H_0^j(t)$ and consequently $h_0^{*j}(t) = c_j h_0^j(t)$. From (16) follows that $H_0^{*1}(t) = H_0^1(t)$ and $c_1 = 1$. In addition, we obtain $\phi_j(\boldsymbol{x}) = \phi_j^*(\boldsymbol{x})$ and $\pi_j(\boldsymbol{z}) = \pi_j^*(\boldsymbol{z})$. □

A.2 Relationship between $\log \tau_{\tilde{y}|j}$ and $v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i)$.

From expression (6), the second term to be maximized is

$$\sum_{i=1}^{n} E[\log \tau_{\widetilde{Y}_i|Y_i} \mid T_i, \beta^{(r)}_{Y_i,\widetilde{Y}_i}] = \sum_{i=1}^{n} \sum_{y=1}^{J} \sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\beta_y^{(r)}; t_i, x_i) w_y(\beta^{(r)}; t_i, x_i) \log \tau_{\tilde{y}|y}.$$

Conditioning on the main group, say j, the term with $y = j$ in the above sum equals

$$\sum_{i=1}^{n} \sum_{\tilde{y}=1}^{K_j-1} v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i) \log \tau_{\tilde{y}|j} + \sum_{i=1}^{n} v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i) \log(1 - \sum_{\tilde{y}=1}^{K_j-1} \tau_{\tilde{y}|j}).$$

Setting the partial derivative with respect to $\tau_{\tilde{y}|j}$ equal to 0,

$$\frac{\partial Q^h\big((b, \beta_{Y,\widetilde{Y}})^{(r+1)} \mid (b, \beta_{Y,\widetilde{Y}})^{(r)}\big)}{\partial \tau_{\tilde{y}|j}} = \sum_{i=1}^{n} \frac{v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i)}{\tau_{\tilde{y}|j}} - \sum_{i=1}^{n} \frac{v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)}{\tau_{K_j|j}} = 0,$$

implies that the optimizer $\tau_{\tilde{y}|j}^{(r+1)}$ satisfies

$$\tau_{\tilde{y}|j}^{(r+1)} = \frac{\sum_{i=1}^{n} v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i)}{\sum_{i=1}^{n} v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)}, \tau_{K_j|j}^{(r+1)} \tag{17}$$

for $\tilde{y} = 1, \ldots, K_j - 1$. Under the constraints that for every $j = 1 \ldots, K$ the weights $v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i); \tilde{y} = 1 \ldots, K_j$ sum to 1, we obtain

$$\begin{aligned}
1 &= \sum_{\tilde{y}=1}^{K_j} \tau_{\tilde{y}|j}^{(r+1)} = \frac{\sum_{\tilde{y}=1}^{K_j} \sum_{i=1}^{n} v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i) \tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)} \\
&= \frac{\sum_{i=1}^{n} \left(\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i)\right) \tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)} \\
&= \frac{\sum_{i=1}^{n} \tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)} = \frac{n \tau_{K_j|j}^{(r+1)}}{\sum_{i=1}^{n} v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)}.
\end{aligned}$$

So, $\tau_{K_j|j}^{(r+1)} = n^{-1} \sum_{i=1}^{n} v_{K_j|j}(\beta_j^{(r)}; t_i, x_i)$. When plugging this back in (17), the same form follows for $\tilde{y} = 1, \ldots, K_j - 1$. Hence $\tau_{\tilde{y}|j}^{(r+1)} = n^{-1} \sum_{i=1}^{n} v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i)$. With the constraint that $\sum_{\tilde{y}=1}^{K_j} v_{\tilde{y}|y}(\beta_y^{(r)}; t_i, x_i) = 1$ for each i, these $v_{\tilde{y}|j}(\beta_j^{(r)}; t_i, x_i)$ are well defined, hence so is $\tau_{\tilde{y}|j}^{(r+1)}$.

References

Aitkin, M., Rubin, D.B., 1985. Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society. Series B (Methodological) 47, 67–75.

Amico, M., Van Keilegom, I., 2018. Cure models in survival analysis. Annual Review of Statistics and Its Application 5, 311–342.

Andreeva, G., 2006. European generic scoring models using survival analysis. The Journal of the Operational Research Society 57, 1180–1187.

Banasik, J., Crook, J., Thomas, L., 1999. Not if but when will borrowers default. The Journal of the Operational Research Society 50, 1185–1190.

Bellotti, T., Crook, J., 2009. Credit scoring with macroeconomic variables using survival analysis. The Journal of the Operational Research Society 60, 1699–1707.

Berrington, A., Diamond, I., 2000. Marriage or cohabitation: a competing risks analysis of first-partnership formation among the 1958 British birth cohort. Journal of the Royal Statistical Society: Series A (Statistics in Society) 163, 127–151.

Bremhorst, V., Lambert, P., 2016. Flexible estimation in cure survival models using Bayesian P-splines. Computational Statistics & Data Analysis 93, 270–284.

Burda, M., Harding, M., Hausman, J., 2015. A Bayesian semiparametric competing risk model with unobserved heterogeneity. Journal of Applied Econometrics 30, 353–376.

Cai, C., Zou, Y., Peng, Y., Zhang, J., 2012. smcure: An R-package for estimating semiparametric mixture cure models. Computer Methods and Programs in Biomedicine 108, 1255–1260.

Cai, L., 2008. SEM of another flavour: Two new applications of the supplemented EM algorithm. British Journal of Mathematical and Statistical Psychology 61, 309–329.

Cai, L., Lee, T., 2009. Covariance structure model fit testing under missing data: An application of the supplemented EM algorithm. Multivariate Behavioral Research 44, 281–304.

Cavanaugh, J.E., Shumway, R.H., 1998. An Akaike information criterion for model selection in the presence of incomplete data. Journal of statistical planning and inference 67, 45–65.

Ciochetti, D., Deng, Y., Gao, B., Yao, R., 2002. The termination of commercial mortgage contracts through prepayment and default: A proportional hazards approach with competing risks. Real Estate Economics 30, 595–633.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

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

Deng, Y., Quigley, J., Van Order, R., 2000. Mortgage terminations, heterogeneity, and the exercise of mortgage options. Econometrica 68, 275–307.

Dirick, L., Bellotti, T., Claeskens, G., Baesens, B., 2019. Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. Journal of Business and Economic Statistics 37, 40–53.

Dirick, L., Claeskens, G., Baesens, B., 2015. An Akaike information criterion for multiple event mixture cure models. European Journal of Operational Research 241, 449–457.

Dirick, L., Claeskens, G., Baesens, B., 2017. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society 68, 652–665.

Djeundje, V.B., Crook, J., 2018. Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. European Journal of Operational Research 271, 697–709.

Forcina, A., 2017. A fisher-scoring algorithm for fitting latent class models with individual covariates. Econometrics and Statistics 3, 132–140.

Gambacciani, M., Paolella, M.S., 2017. Robust normal mixtures for financial portfolio allocation. Econometrics and Statistics 3, 91–111.

Hanson, S.G., Pesaran, M.H., Schuermann, T., 2008. Firm heterogeneity and credit risk diversification. Journal of Empirical Finance 15, 583 – 612.

Hasnat, M.A., Velcin, J., Bonnevay, S., Jacques, J., 2017. Evolutionary clustering for categorical data using parametric links among multinomial mixture models. Econometrics and Statistics 3, 141–159.

Heckman, J.J., Honoré, B.E., 1989. The identifiability of the competing risks model. Biometrika 76, 325–330.

Jamshidian, M., Jennrich, R.I., 2000. Standard errors for EM estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62, 257–270.

Kuk, A., Chen, C., 1992. A mixture model combining logistic regression with proportional hazards regression. Biometrika 79, 531–541.

Lai, X., Yau, K., 2009. Multilevel mixture cure models with random effects. Biometrical Journal 51, 456–466.

Lunn, M., McNeil, D., 1995. Applying Cox regression to competing risks. Biometrics 51, 524–532.

McLachlan, G.J., Peel, D., 2000. Finite mixture models. Wiley Series in Probability and Statistics, New York.

Meng, X.L., Rubin, D.B., 1991. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association 86, 899–909.

Murray, P.M., Browne, R.P., McNicholas, P.D., 2017. A mixture of sdb skew-t factor analyzers. Econometrics and Statistics 3, 160–168.

Ng, A.K., Bernardo, M.P., Weller, E., Backstrand, K.H., Silver, B., Marcus, K.C., Tarbell, N.J., Friedberg, J., Canellos, G.P., Mauch, P.M., 2002. Long-term survival and competing causes of death in patients with early-stage Hodgkin's disease treated at age 50 or younger. Journal of Clinical Oncology 20, 2101–2108.

Patilea, V., Keilegom, I., 2020. A general approach for cure models in survival analysis. The Annals of Statistics 48, 2323–2346.

Pavlov, A., 2001. Competing risks of mortgage termination: Who refinances, who moves and who defaults. Journal of Real Estate Economics and Finance 23, 185–211.

Peng, Y., Dear, K., 2000. A nonparametric micture model for cure rate estimation. Biometrics 56, 227–236.

Peng, Y., Zhang, J., 2008. Identifiability of a mixture cure frailty model. Statistics & Probability Letters 78, 2604 – 2608.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Segal, M.R., Bacchetti, P., Jewell, N.P., 1994. Variances for maximum penalized likelihood estimates obtained via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 56, 345–352.

Stepanova, M., Thomas, L., 2002. Survival analysis methods for personal loan data. Operations Research 50, 277–289.

Sy, J., Taylor, J., 2001. Standard errors for the Cox proportional hazards cure model. Mathematical and Computer Modelling 33, 1237 – 1251.

Sy, J.P., Taylor, J.M.G., 2000. Estimation in a Cox proportional hazards cure model. Biometrics 56, 227–236.

Tawiah, R., McLachlan, G.J., Ng, S.K., 2020. Mixture cure models with time-varying and multilevel frailties for recurrent event data. Statistical Methods in Medical Research 29, 1368–1385.

Tong, E.N.C., Mues, C., Thomas, L.C., 2012. Mixture cure models in credit scoring: if and when borrowers default. European Journal of Operational Research 218, 132–139.

Tsiatis, A., 1975. A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences 72, 20–22.

Wang, W.L., Lin, T.I., 2016. Maximum likelihood inference for the multivariate t mixture model. Journal of Multivariate Analysis 149, 54–64.

Watkins, J.G.T., Vasnev, A.L., Gerlach, R., 2014. Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. Journal of Applied Econometrics 29, 627–648.

Wienke, A., Lichtenstein, P., Czene, K., Yashin, A., 2007. The role of correlated frailty models in studies of human health, ageing, and longevity., in: Auget, J., Balakrishnan, N., Mesbah, M., Molenberghs, G. (Eds.), Advances in Statistical Methods for the Health Sciences. Statistics for Industry and Technology.. Birkhäuser Boston, pp. 151–166.

Zhang, J., Thomas, L., 2012. Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. International Journal of Forecasting 18, 204–215.