Royal Institute of Technology

Master's Thesis

Estimating the term structure of default probabilities for heterogeneous credit portfolios

Author:
Felix Bogren

Supervisor:
Filip Lindskog

June 8, 2015

Abstract

The aim of this thesis is to estimate the term structure of default probabilities for heterogeneous credit portfolios. The term structure is defined as the cumulative distribution function (CDF) of the time until default. Since the CDF is the complement of the survival function, survival analysis is applied to estimate the term structures. To manage long-term survivors and plateaued survival functions, the data is assumed to follow a parametric as well as a semi-parametric mixture cure model. Due to the general intractability of the maximum likelihood of mixture models, the parameters are estimated by the EM algorithm. A simulation study is conducted to assess the accuracy of the EM algorithm applied to the parametric mixture cure model with data characterized by a low default incidence. The simulation study recognizes difficulties in estimating the parameters when the data is not gathered over a sufficiently long observational window. The estimated term structures are compared to empirical term structures, determined by the Kaplan-Meier estimator. The results indicated a good fit of the model for longer horizons when applied to each credit type separately, despite difficulties capturing the dynamics of the term structure for the first one to two years. Both models performed poorly with few defaults. The parametric model did however not seem sensitive to low default rates. In conclusion, the class of mixture cure models are indeed viable for estimating the term structure of default probabilities for heterogeneous credit portfolios.

Keywords: mixture cure model, term structure, default probabilities, heterogeneity, credit portfolios, EM algorithm

Sammanfattning

Syftet med den här uppsatsen är att estimera terminstrukturen för konkurssannolikheter i heterogena kreditportföljer. Terminstrukturen definieras som den kumulativa fördelningsfunktionen för tiden till konkurs. Eftersom den kumulativa fördelningsfunktionen är komplementet till överlevnadsfunktionen kan överlevnadsanalys appliceras för att estimera terminstrukturen. För att hantera långtidsöverlevare samt överlevnadsfunktioner som planar ut vid nivåer över noll, antar vi att observationerna kommer från en parametrisk såväl som en semiparametrisk *mixture cure model*. På grund av numeriska svårigheter att hantera maximum likelihood-funktionen för mixture modeller, så skattas parametrarna med hjälp av EM algoritmen. En simulationsstudie genomfördes för att undersöka precisionen av EM algoritmen applicerad på parametriska specifikationen av modellen, med data bestående av få antal konkurser. Simulationsstudien påvisade svårigheter att estimera parametrarna när urvalet inte tagits från en tillräckligt lång tidsperiod. En jämförelse görs med de empiriska terminstrukturerna, framtagna med Kaplan-Meier's skattning av överlevnadsfunktioner. Resultaten påvisar en bra anpassning när modellen appliceras på varje kredittyp separat, trots svårigheter att fånga dynamiken de av terminstrukturen under de första ett till två åren. Båda modellerna var otillförlitliga med få antal konkurser. Däremot var den parametriska modellen inte märkbart känslig för låga konkursfrekvenser. Sammanfattningsvis så kan klassen mixture cure modeller anses lämplig för att estimera terminstrukturen för konkurssannolikheter i heterogena kreditportföljer.

Acknowledgements

Firstly I would like to thank my supervisors at Swedbank, Fausto Molinari and Max Loxbo, for introducing me to the topic as well as for their valuable feedback and insights. Secondly I would like to thank my supervisor at KTH, Filip Lindskog, for his guidance and feedback, which has significantly improve on the overall quality of the thesis. Finally I wish to express my sincere gratitude to my family for their continuous support throughout the course of my studies.

Stockholm, June 2015

Felix Bogren

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1 Background

The increased regulatory pressure in the aftermath of recent financial crises has significantly impacted the development of credit risk models within the industry. Rather than increasing sophistication and expansion of scope, the development has primarily pursued fulfillment of regulatory standards. It has simultaneously deemed necessary for regulatory measures to recognize expected credit losses in a more timely manner (IASB, 2014). This has precipitated the works on the forthcoming reporting standards IFRS9, effective in January 2018. The new regulation is replacing its predecessor IAS39. The forward-looking impairment model of IFRS9 introduces an expected loss model for the accounting of impaired contracts, contrary to the incurred loss models of IAS39. The implication is that expected loss models for financial instruments are required to account for the entire lifespan of each instrument. This is a significant discrepancy to the IAS39 framework in which the aspect of time is limited to a one year horizon. To model expected losses for the entire lifespan of a financial instrument, a model for the term structure of the default probability is required. The term structure can either be defined as the instantaneous or the cumulative probability of default at each time t. In this thesis, the latter definition is used.

The probability of default (PD) is for many applications a level associated with a credit score. In pursuance of adequate conformity with existing regulatory framework, current industry standards of credit scoring have focused on logistic regression models (Mues, Thomas, & Tong, 2012). This facilitates a natural representation of homogeneous sub-populations of the portfolio, which is essential for any type of credit scoring. The different types of data commonly encountered in credit portfolios are however not naturally dealt with in the logistic regression. By considering the data as partial observations of the longevity of each contract, rather than defaulters or non-defaulters, the dimension of time is introduced to the model. This is necessary to efficiently estimate the term structure. In fact, for the logistic model there is no obvious representation of the term structure of default probabilities. It merely produces the cumulative default probability over a specific time horizon. A term structure can naively be obtained by extending a constant PD over the entire horizon, e.g. the one year PD. This is proved inadequate by figure 1.1, in which the cumulative default probability is determined by extending the one year PD. It is apparent that the probability of default is indeed not constant over time. As a consequence, pending the introduction of IFRS9, the need for more exhaustive models is inevitable.

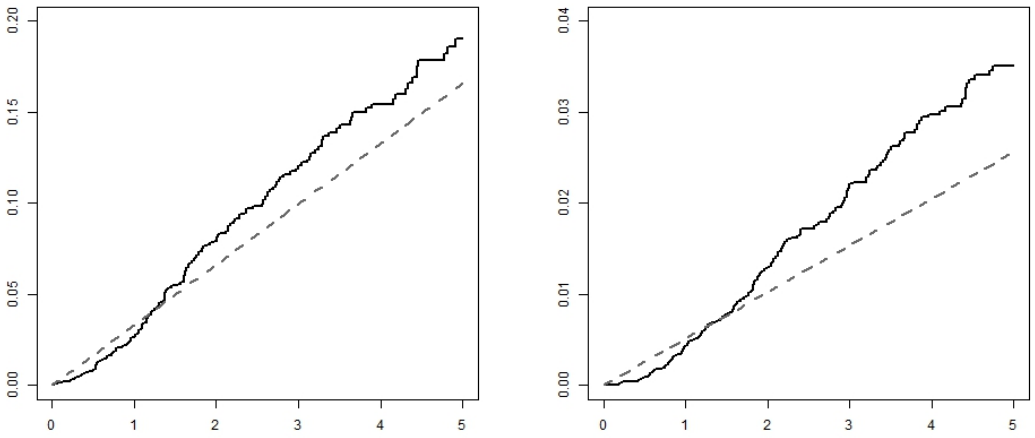
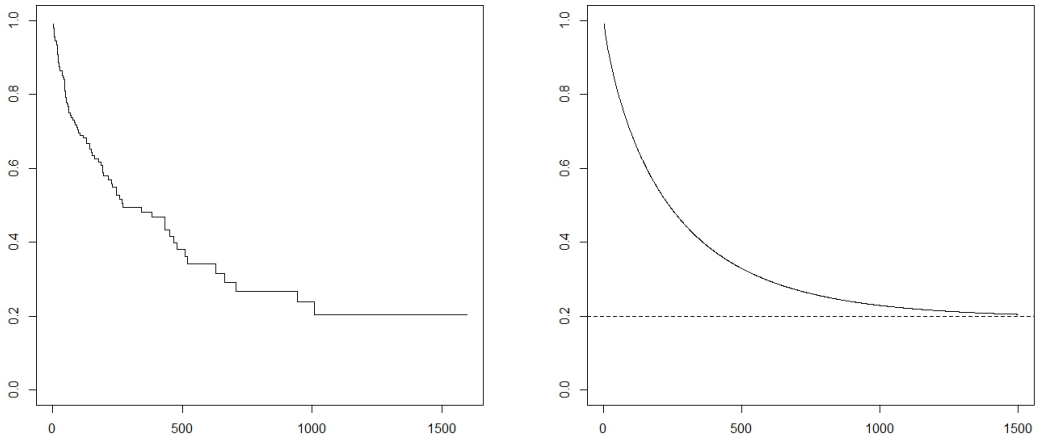


Figure 1.1: Illustration of the empirical term structure of default probabilities in comparison to extending the 1 year probability of default over the entire horizon. The term structures in the left and right figures are estimated from corporate loans with credit rating 5 and 10, respectively.

A majority of the literature on PD term structures is focused on estimating the term structure from market data. Frequently the default probabilities are implied from credit default swap or bond data. Since most credit portfolios consist of non-marketable instruments, there is no market data available for inference of the investors' subjective probability of default. Nonetheless, default probabilities implied from market data would be considered risk-neutral. Here the objective is to find the physical default probabilities. A pertinent approach is to construct credit migration matrices by utilizing the theory of Markov processes. The migration matrices are subsequently used to obtain the term structures of default probabilities. Here we will instead only consider default events. Then the data typically consist of historical default incidences in addition to characteristics of each credit. A default incidence is defined by the dichotomy of observing a default or not observing a default throughout its lifespan. Since the default incidence is observed in time, both the longevity of the loan as well as the timing of the contingent default are effectively measured, given the date the contract is established. The longevity, together with the variable indicating whether the event has occurred, constitutes the natural form of survival data. Therefore it is convenient to use survival analysis. The primary objective of survival analysis, which is an important branch within the field of biostatistics, is to model the survival function. The survival function, generally denoted $S(t)$, is a measure of the probability of surviving beyond time t. The term structure of the default probability as defined here is directly inferred from the survival function by its complement, the cumulative distribution function.

The framework of survival analysis is further justified by the fact that it has been successfully applied to credit scoring (Basanik, Crook, & Thomas, 1999; Bellotti & Crook, 2009; Malik & Thomas, 2009). Similarly as for the logistic regression, survival analysis allows each credit to be scored with respect to some characteristics deemed relevant for differentiating between the sub-populations within the portfolio. Survival analysis in its elementary form may however encounter some problems. Data with high censor rate typically result in survival functions that quickly plateaus at levels larger than zero. This occurs naturally in sub-portfolios of higher credit worthiness, with lower levels of default incidence. Farewell (1982) coined this term long-term

survivorship. This may be considered a contradiction of one of the fundamental assumptions of survival analysis, inevitable mortality, and is likely to result in biased estimates (Segovia, 2014). For data where the survival function plateaus at levels larger than zero, it is appropriate to use mixture cure models (MCM) (Sy & Taylor, 2000). Long-term survivors have been observed in credit risk modeling as well (Thomas, Tong, & Mues, 2012). The MCM postulates the existence of two types of populations, one that is susceptible to the event under study and one that is non-susceptible. As depicted by figure 1.2, the survival function of the MCM converges to the level of non-susceptible individuals. It is on the other hand unrealistic to consider some proportion of the loans to be non-susceptible to default throughout infinity. Instead, we may think of them as being at-risk or not at-risk over a sufficiently long time horizon.



(a) Empirical survival function of a MCM *(b) Survival function of a MCM*

Figure 1.2: The figures display the empirical (non-parametric) as well as the true survival function of a population with long-term survivors, i.e. a mixture cure model. As illustrated in (b), when time increases the susceptible proportion will succumb, leaving the non-susceptible individuals unaffected. For sufficiently large t, the probability of surviving is solely determined by the probability of being non-susceptible. In this example the non-susceptible proportion is 0.2, displayed by the dotted line.

The model originates from biostatistics where it was initially used to model the time until relapse of a disease, where one portion of the population was considered cured (non-susceptible to relapse) (Boag, 1949). The probability of being susceptible to default (incidence) is generally modeled by a logistic regression model, similar to current industry standards of credit scoring. The timing of the default (latency) of the susceptible sub-populations is modeled by means of ordinary survival analysis. Consequently, the MCM may be considered an integration of the logistic regression and standard survival analysis. Instead of modeling the default probability over a specific horizon as in ordinary credit scoring, the MCM models the incidence of defaults without considering the aspect of time. The timing of the default is determined by the latency.

2 Purpose

The purpose of this thesis is to investigate the suitability of the mixture cure model for determining the term structure of default probabilities for heterogeneous credit portfolios. Given the available data of observed lifetimes, we wish to model the term structure of the default probability of each member of the portfolio as well as future applicants. Interestingly, studies on survival analysis applied to credit scoring are almost exclusively focused on the retail segment of credit portfolios, and particularly retail loans. This study will instead target corporate credit, including loans as well as other forms of debt. This will shed light on whether survival analysis is an appropriate theoretical framework for other types of credit than loans.

Since the type of credit and the rating will be considered the sole determinants for the overall risk classification, these will also be the only covariates considered through out the thesis. Thus, in contrast to credit scoring, the prerequisites of the proposed model is that each credit is already scored.

3 Delimitations

Naturally, the study is limited to the observational window over which the lifetimes are recorded. The time period over which the data is recorded will influence the shape as well as the overall level of the term structure. Also, the longest conceivable observed lifetime is limited by the length of the observational window. The data used for this study consist of observed lifetimes of corporate credit over the time period June 2007 to March 2015, covering a time interval of close to eight years. Therefore, the term structure can at most be effectively estimated over an eight year horizon.

Moreover, generally the characteristics of a loan will change over time. Here, the credit rating and the product type are considered the sole determinants of the shape of the term structure. In practice, the credit rating is likely to change as the worthiness deteriorates or is strengthened. Since the model do not utilize time-dependent variables, rating migrations are not be considered directly. Instead the rating is only measured at inception of each contract. Nonetheless, there is still an underlying migration process that influence the default probability of the credit, although unobservable in this model. Consequently, for a given term structure of some credit rating and product type, potential rating migrations will only be accounted for indirectly. Since covariates are assume fixed, the impact of the macroeconomic climate on default events are not considered.

Moreover, some counterparties have multiple contracts that will potentially default closely in time. This type of dependence has not been considered. Since many of the contracts will be of different credit type, the dependence will primarily impact the inference of the models where all data is considered simultaneously. Whereas if the data of each credit type is considered separately, the potential effects are negated.

Chapter 2

Theoretical Framework

1 Survival analysis

Survival analysis is a statistical branch within biology. The field is analogous with reliability analysis in engineering and event history analysis in sociology, although there are some discrepancies in the terminology. All these fields are primarily focused on modeling the time until a specific events. In survival analysis, the event is often death or infection of a disease, whereas in reliability analysis the event is commonly the failure of a machine. The most central concept within survival theory is the survival function.

Definition 1.1. *Let T be a non-negative continuous random variable associated with the cumulative density function $F(t)$ on $[0, \infty)$, then the survival function is determined by*

$$S(t) = P(T > t) = 1 - F(t) \tag{2.1}$$

Remark 1. Although the survival function is generally denoted $S(t)$, we will in subsequent sections sometimes denote the survival function with a bar on top of the corresponding distribution function, e.g. $\bar{F}(t)$ is the survival function together with the distribution function $F(t)$ and the probability density function $f(t)$.

By the definition of the survival function, it is clear that $S(t)$ represents the probability of surviving beyond t. Alternatively, it is the probability of experiencing an event subsequent to time t. Moreover, due to the nature of how events occur the survival function is generally assumed to be restricted in the following manner

i) $S(0) = 1$

ii) $S(t) \to 0$ as $t \to \infty$.

The first restriction rejects the possibility of immediate death, whereas the second restriction postulates the idea of inevitable mortality. Although this may seem reasonable when the survival function actually represents the lifetime of an organism, for other event time models it is not necessarily so. As will be seen in subsequent sections, deviating from these conditions will prove useful. In fact, imposing these restrictions may in some circumstances result in a model which directly contradicts reality.

Another important concept in survival analysis is the hazard rate. It is defined as the instantaneous probability of experiencing an event at time t.

Definition 1.2. *The hazard function $\lambda(t)$ of the continuous random variable T is defined by*

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h \mid T \geq t)}{h} \tag{2.2}$$

The shape of the hazard function is useful for determining an appropriate failure. The different general traits of the hazard function includes constant, increase, decreasing, bathtub shaped or hump shaped distribution of T (Klein & Moeschberger, 1997). The shapes are commonly associated with different types of distributions. Further let $f(t)$ and $S(t)$ be the density and survival functions of T, respectively. By Bayes' rule it follows that the hazard function in (2.2) can be written as

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)}{dt}\frac{1}{S(t)} \tag{2.3}$$

where the last step follows directly from differentiation of the survival function. Integrating both sides in (2.3) yields

$$S(t) = e^{-\int_0^t \lambda(s)ds} = e^{-H(t)}. \tag{2.4}$$

Here $H(t)$ is used to denote the cumulative hazard function. which is simply found by integrating the hazard function $\lambda(t)$.

1.1 Survival distributions

In survival analysis we are not necessarily limited to any parametric families of distributions. It is on the other hand convenient to use distributions with sole support on $\mathbb{R}^+$. The most frequently encountered distributions; the Exponential, Weibull, Log-normal, Log-logistic and the Gamma distribution, satisfy the assumption of non-negativity.

The probability distribution function of an exponentially distributed random variable X is given by

$$F(x) = \begin{cases} 1 - e^{-x/\theta} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \tag{2.5}$$

The variable θ is referred to as the scale parameter. Occasionally the distribution is instead defined in terms of the rate parameter, given by the inverse of the scale parameter, $\lambda = 1/\theta$. The exponential distribution has the convenient property of being memory-less. This implies that, for any non-negative numbers t and s, the following property is fulfilled

$$P(X > t + s | X > t) = P(X > s). \tag{2.6}$$

This is also characterized by a constant hazard rate, i.e. the probability of experiencing an event at time t conditional on surviving until then is independent of t. This is not by any means a valid assumption in all situations. Sometimes it may prove necessary to use any of the generalizations

of the Exponential distribution; the Weibull distribution or the Gamma distribution.

Contrary to the exponential distribution, the Weibull distribution is a two-parameter family of distributions. In addition to the scale parameter, the Weibull distribution also includes a shape parameter k. The distribution function is defined as follows

$$F(x) = \begin{cases} 1 - e^{-(x/\theta)^k} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \tag{2.7}$$

It is easily seen that for $k = 1$, the distribution function equals that of the Exponential distribution.

The log-logistic function, also defined in terms of the scale and shape parameter, has the following distribution function

$$F(x) = \begin{cases} \frac{1}{1+(x/\alpha)^{-\beta}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases} \tag{2.8}$$

where α and β are the scale and shape parameters, respectively.

Finally, the log-normal distribution is defined in terms of the exponential of a normally distributed r.v. with mean μ and standard deviation σ. The distribution function of the log-normal distribution is given by

$$F(x) = \begin{cases} \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right) & \text{if } x > 0 \\ 0 & \text{if } x \leq 0, \end{cases} \tag{2.9}$$

where Φ is the CDF of the normal distribution.

1.2 Censored and truncated data

Time to event data naturally includes *censored* and *truncated* observations, as a result from the method of which data is generally gathered. Censoring occurs when the timing of the event of interest is only partially observed, contrary to uncensored observations for which the exact timing of an event is recorded. The most general type of censoring is interval censoring. As the name suggests, interval censoring occurs when the timing of an event is known to lie within a specific time span. This typically occurs in medical follow-up studies where the state of the subject only can be recorded at a finite number of points in time. If the left or right end points of the censoring interval are infinite, the observation is instead assumed to be right or left censored, respectively. The terms left and right refers to the timing of an event relative to the censoring time, for instance a right censored observations is only known to occur after (to the right on a horizontal time scale) the censoring time.

Example 1.1. *Consider a study in which one wish to model the time until an event. The study involves consecutive screenings of each subject at different points in times. Assuming continuity, the event can only occur in between two screenings. Let the event of a subject occur between two screenings at time t_i and t_{i+1}, then the exact timing of the event is only known to lie within the interval (t_i, t_{i+1}). This is referred to as interval censoring. If instead there has been no event*

recorded at the final screening t_n. Then the event is only known to occur after t_n, i.e. in the interval (t_n, ∞). This is referred to as right censoring. Finally, if the event has occurred prior to the first screening t_1, then the observation is regarded as left censored.

Because of how survival studies are often constructed, right censoring is the most frequently encountered type of censoring. Therefore, one further distinguishes between different types of right censoring schemes. Random (Independent) censoring occurs if the censoring time follows a continuous random variable, independent of the time the to event. Let T and C denote the random life time and censoring time, respectively. Denote the survival functions of T and C by $\bar{F}(t)$ and $\bar{H}(t)$, and the corresponding density functions $f(t)$ and $h(t)$. Further let time t be the observed life time and δ be the indicator if an event is observed or if the observation is right censored. The probability of observation (t, δ) is now determined by

$$P(t, \delta) = (f(t)\bar{H}(t))^{\delta}(h(t)\bar{F}(t))^{1-\delta}. \tag{2.10}$$

Type I censoring arises if the censoring mechanism is fixed in time, i.e. only events that occurs before this point in time are recorded. This is typically encountered in studies conducted over a specified period of time, where the end date of the study period acts as the censoring mechanism. If the subjects enter at different times throughout the study, each observation will effectively be associated with an individual censoring time. This is instead labeled generalized Type I censoring (Klein & Moeschberger, 1997). Finally, Type II censoring occurs when only the first r events are observed, i.e. the r smallest event times.

Truncation is somewhat similar and sometimes confused with censoring. A truncated sample is limited to an observational window over which observations can be recorded. From a statistical perspective, the probability of the observed outcome t of the random time T is conditional on T lying within the interval $[l, r]$, where l and r are the left and right truncation thresholds, respectively. Similarly as for interval censoring, either bound being infinite will result in left or right truncation. Again, left truncation implies that all observations below (to the left) of the threshold are unobservable or truncated. Contrary to censored data, truncated data results in biased samples. Nonetheless, both types data must be dealt with appropriately in order to avoid biased estimations of distribution parameters in statistical inference.

Example 1.2. *Consider now that the study of the previous examples wants to model the longevity of the subjects. Assume that for practical reasons the subjects are only included in the study after a certain age l. Then, all deaths prior to l are systematically excluded from the data set. This is regarded as a left truncated sample. In statistical terms, the probability of observing a lifetime t of the random variable T of the longevity for any subject of the study is conditional on having survived for l years*

$$P(T = t \mid T \geq l). \tag{2.11}$$

Although right truncation is not naturally encountered in survival analysis, it does appear in other settings. When a sample is both left and right truncated with thresholds l and r, respectively, then the likelihood of the observation is given by

$$P(T = t \mid l \leq T \leq r). \tag{2.12}$$

1.3 Proportional hazards

The (Cox) proportional hazards model is a class of survival models proposed by Cox (1972). The model postulates a simplified association between the hazard function and some vector of covariates $\mathbf{z} = (z_1, \ldots, z_k)^T$. More specifically, the model assumes the following representation of the hazard function

$$\lambda(t) = \lambda_0(t) \exp(\beta^T \mathbf{z}) \tag{2.13}$$

for some non-negative joint baseline hazard function $\lambda_0(t)$ and a vector of coefficients β. The covariates $\mathbf{z}$ are important as they allow for the representation of heterogeneity within the population. The term proportional refers to the proportionality of the hazard rate between two subjects. Since the baseline hazard is equal for both subjects, the hazard ratio between the subjects will only depend on their respective covariates. This is further illustrated in example 1.3.

Example 1.3. *Assuming the hazard function follows the proportional hazards assumption, with the baseline hazard function $\lambda_0(t)$ and the single fixed covariate z with corresponding coefficient β, then the hazard ratio between two subjects characterized by covariates z_1 and z_2 is given by*

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t) \exp(z_1\beta)}{\lambda_0(t) \exp(z_2\beta)} = \exp((z_1 - z_2)\beta) \tag{2.14}$$

After inserting the hazard function into (2.4), we find that the survival function of the proportional hazards model is represented by

$$S(t \mid \mathbf{z}) = S_0(t)^{\exp(\beta^T \mathbf{z})} \tag{2.15}$$

where $S_0(t)$ is the baseline survival function. The baseline survival function is the survival function associated with the baseline hazard function, e.g. if $\lambda_0(t)$ is the Weibull hazard function then $S_0(t)$ is the Weibull survival function. For a non-parametric baseline survival function, the model is generally referred to as semi-parametric.

2 Maximum-likelihood estimation

Maximum-likelihood estimation (MLE) is a method for estimating the parameters of a statistical model given a set of data. The general idea of the MLE is to find the set of model parameters that maximizes the likelihood of the observed data. Typically the data is assumed to be realized from a parametric distribution. In reality, the parametric family is of course never observed, but is chosen as see fit. Sometimes it is convenient to utilize non-parametric ML estimators, as they do no require any assumptions of the underlying distribution.

Definition 2.1. *Assuming a statistical model parametrized by a fixed and unknown parameter vector θ, then the likelihood function $\mathcal{L}(\theta; \mathbf{y})$ is the probability of the observed data $\mathbf{y}$ as a function of the parameters θ.*

For continuous random variables, the probability of the observed data $\mathbf{y}$ is given by the probabil-

ity density function. Generally, this is the multivariate probability density function of $\mathbf{y}$,

$$\mathcal{L}(\theta; \mathbf{y}) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta). \tag{2.16}$$

In most applications, the observed data is assumed to be a realization of i.i.d. random variables, for which the likelihood function is determined by the product of the likelihood of each observation. For discrete random variables, the likelihood function is instead given by the probability function of the observed data. Once again, if the observations are assumed independent, the likelihood is simply given by the product of the probability of each observation.

Example 2.1. *Let the data $\mathbf{y} = y_1, \ldots, y_n$ be the observed random sample of n independent and identically distributed continuous random variables associated with the probability density function $f_0(\cdot|\theta_0)$. It is hypothesized that f_0 belongs to a family of distributions $\{f_\theta \mid \theta \in \mathbf{\Theta}\}$, for some vector of parameters θ. Due to independence, the likelihood function of θ may be written as the product of the likelihood of each observation*

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta). \tag{2.17}$$

Further let the log-likelihood function be the logarithm of $\mathcal{L}$. Then the log-likelihood to (2.17) is given by.

$$l(\theta; \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i; \theta) \tag{2.18}$$

Definition 2.2. *Given the likelihood function $\mathcal{L}(\theta)$ of the parameter vector θ, and the corresponding log-likelihood $l(\theta) = \log \mathcal{L}(\theta)$, the score function is the gradient of $l(\theta)$ with respect to the parameter vector θ.*

$$S(\theta) = \frac{\partial l(\theta; \mathbf{y})}{\partial \theta}. \tag{2.19}$$

Considering the logarithm is a strictly monotonically increasing function, the results of the ML is indifferent to whether we maximize the likelihood function or the log-likelihood function. The latter is however practically easier to deal with. Consequently, the MLE $\widehat{\theta}$ is the point estimate of the parameters that maximizes the log-likelihood, i.e.

$$\widehat{\theta} = \underset{\theta \in \mathbf{\Theta}}{\operatorname{argmax}}\, l(\theta; \mathbf{y}) \tag{2.20}$$

The estimate is generally obtained by finding the point at which the score function is $\mathbf{0}$. For some analytically tedious likelihoods, the estimate may be found numerically by any type of optimization algorithm.

If the data is right censored and left truncated, the likelihood (2.17) must be altered accordingly. Let the data consist of the sets $\mathcal{C}$ and $\mathcal{U}$, corresponding to the censored and uncensored observations, respectively. Further let the censored set $\mathcal{C}$ be associated with the left and right censoring times y_i^l and y_i^r and the uncensored set $\mathcal{U}$ with the observed outcomes y_i. Additionally, let all observations have the left and right truncation points l_i and u_i. In this setting, the likelihood is

given by

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i \in \mathcal{U}} \frac{f(y_i; \theta)}{F(u_i) - F(l_i)} \prod_{i \in \mathcal{C}} \frac{F(y_i^l) - F(y_i^r)}{F(u_i) - F(l_i)}. \tag{2.21}$$

Extensions of the maximum-likelihood

Occasionally, it is possible to divide the likelihood into several factors each isolating a different subset of the parameters. This is particularly useful in the presence of nuisance parameters, i.e. parameters insignificant for the analysis that are only included as part of the specification of the model.

Definition 2.3 (Orthogonal parameters)**.** *If the likelihood function $\mathcal{L}(\theta, \eta; \mathbf{y})$ could be factorized according to*

$$\mathcal{L}(\theta, \eta; \mathbf{y}) = \mathcal{L}_1(\theta; \mathbf{y}) \mathcal{L}_2(\eta; \mathbf{y}). \tag{2.22}$$

then the parameters θ and η are said to be orthogonal.

Remark 2. The likelihoods $\mathcal{L}_1(\theta; \mathbf{y})$ and $\mathcal{L}_2(\eta; \mathbf{y})$ will be referred to as the partial likelihoods of θ and η given $\mathbf{y}$.

Assuming θ is the set of parameters of interest and η is the set of nuisance parameters, then only the partial likelihood $\mathcal{L}_1$ needs to be maximized to find the MLE of θ. Factorization into partial likelihoods is also advantageous if maximization of the full likelihood is computationally expensive.

Example 2.2. *Suppose we would like to find the parameters θ_f of the density function $f(y; \theta_f)$ associated with the life times $t_1, \ldots, t_n$. Due to independent random right censoring c_i, the observable life time is given by $y_i = \min(t_i, c_i)$. Let the variable δ_i indicate right censoring if $\delta_i = 0$. Further let $h(y; \theta_h)$ be the density function of the censoring time. Denote the survival functions by capital letters with a bar, i.e. $\bar{F}(y; \theta_f)$ and $\bar{H}(y; \theta_h)$. Then the likelihood of θ_f and θ_h given data $\mathbf{y}$ and δ may be written as follows*

$$\begin{aligned}
\mathcal{L}(\theta_f, \theta_h; \mathbf{y}, \delta) &= \prod_{i=1}^{n} (f(y_i; \theta_f) \bar{H}(y_i; \theta_h))^{\delta_i} (h(y_i; \theta_h) \bar{F}(y_i; \theta_f))^{1-\delta_i} \\
&= \prod_{i=1}^{n} f(y_i; \theta_f)^{1-\delta_i} \bar{F}(y_i; \theta_f))^{\delta_i} \prod_{i=1}^{n} h(y_i; \theta_h)^{\delta_i} \bar{H}(y_i; \theta_h))^{1-\delta_i} \\
&= \mathcal{L}(\theta_f; \mathbf{y}, \delta) \mathcal{L}(\theta_h; \mathbf{y}, \delta).
\end{aligned} \tag{2.23}$$

Since θ_h is generally considered a nuisance parameter, it is straight forward to maximize $\mathcal{L}(\theta_f; \mathbf{y}, \delta)$ rather than the full likelihood.

2.1 Observed information

The observed information matrix is an important concept in maximum likelihood theory. It is used as an approximation of the inverted asymptotic covariance matrix of the maximum likelihood estimate, from which the standard errors of the MLE can be obtained. Thus, the observed

information matrix $I(\widehat{\theta}, \mathbf{y})$ for data $\mathbf{y}$, yields an estimate of the reliability of the estimated parameters. The standard error of the r:th parameter θ_r is given by the square root of the r:th diagonal element of the inverted information matrix (McLachlan & Peel, 2000)

$$\mathrm{SE}(\theta_r) \approx (I^{-1}(\widehat{\theta}, y))_{rr}^{1/2}. \tag{2.24}$$

The observed information is the Hessian matrix of the negative log-likelihood. Therefore, a numerical estimation of the observed information matrix is obtained by approximating the Hessian of the negative log-likelihood function at the MLE.

2.2 Kaplan-Meier

The non-parametric maximum likelihood estimator of the survival function $S(t)$ is given by the product-limit estimator, proposed by Kaplan and Meier (1958) (therefore also labeled the Kaplan-Meier estimator). By introducing the ordered set of survival times, $t_1 \leq t_2 \leq \cdots \leq t_n$, the product limit estimator is defined as the right-continuous step function

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \Pi_{t_i \leq t}(1 - \frac{d_i}{Y_i}) & \text{if } t_1 \leq t \end{cases} \tag{2.25}$$

where d_i is the number of defaults at time t_i and Y_i is the number of subjects at risk over time period $[t_{i-1}, t_i)$. The advantage of the Kaplan-Meier estimator in comparison to other estimators of the empirical distribution function is that censoring is rigorously dealt with. Without censoring, the estimator is equivalent to the empirical survival function $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}[t_i > t]$. Moreover, although defaults can in theory never occur simultaneously following from the continuity of the random survival times T_i, this is not true in practice due to the discretization of time t. Thus d_i may be larger than 1. For survival times greater than t_n the function is ill defined.

2.3 Nelson-Aalen

The Nelson-Aalen estimator is a non-parametric estimator of the cumulative hazard function $H(t)$. Although $\hat{H}(t) = -\ln(\hat{S}(t))$ serves as reasonable estimate of $H(t)$, evident by (2.4), it is conventional to use the Nelson-Aalen estimator as it has better performance on smaller sample sizes (Klein & Moeschberger, 1997, p. 107). The estimator is defined as follows

$$\hat{H}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & \text{if } t_1 \leq t \end{cases} \tag{2.26}$$

where, as previously, d_i is the number of defaults at time t_i and Y_i is the number of subjects at risk. Censoring is dealt with similarly as in the Kaplan-Meier estimator.

2.4 Proportional hazards

An important contributing factor to the widespread use of the proportional hazards model is that the effect of covariates can be estimated without having to specify any baseline hazard rate.

Also, a non-parametric survival function is easily incorporated. Consequently, the model is not necessarily restricted to a parametric family of distributions. As discussed previously, it does on the other hand impose the restriction of a proportional effect of the covariates on the hazard rate.

To estimate the parameters β of the covariates $\mathbf{z}$, first assume there are no ties amongst the event times, $t_1 \leq \cdots \leq t_m$. Further let R_i denote the risk set just prior to time t_i, defined by the set individuals still under study at the event time t_i. Now the likelihood, as proposed by (Cox, 1975), constitute the probabilities of the events at times t_i, conditional on observing exactly one event at each time. More specifically, the contribution to the likelihood of each event is given by

$$\frac{\lambda_0(t_i)\exp(\beta^T\mathbf{z}_i)}{\sum_{j\in R_i}\lambda_0(t_i)\exp(\beta^T\mathbf{z}_j)} = \frac{\exp(\beta^T\mathbf{z}_i)}{\sum_{j\in R_i}\exp(\beta^T\mathbf{z}_j)} \tag{2.27}$$

The full likelihood is determined by multiplication of the contributing likelihoods of each observations

$$\mathcal{L}(\beta) = \prod_{i=1}^{m}\frac{\exp(\beta^T\mathbf{z}_i)}{\sum_{j\in R_i}\exp(\beta^T\mathbf{z}_j)}. \tag{2.28}$$

Taking the logarithm of the likelihood yields

$$l(\beta) = \sum_{i=1}^{m}\sum_{l=1}^{k}\beta_l z_{i,l} - \sum_{i=1}^{m}\ln\left[\sum_{j\in R_j}\exp\Big(\sum_{l=1}^{k}\beta_k z_{j,k}\Big)\right], \tag{2.29}$$

where $z_{i,l}$ is the l:th covariate of the i:th observation. The likelihood is treated in usual manner, thus the score function is obtained by differentiating the logarithm of likelihood w.r.t. the parameters β_l. Let $U_l(\beta)$ denote the derivative of $l(\beta)$ w.r.t. the l:th coefficient β_l,

$$U_l(\beta) = \sum_{i=1}^{m} z_{i,l} - \sum_{i=1}^{m}\frac{\sum_{j\in R_i} z_{j,l}\exp\left(\sum_{l=1}^{k}\beta_k z_{j,k}\right)}{\sum_{j\in R_i}\exp\left(\sum_{l=1}^{k}\beta_k z_{j,k}\right)}. \tag{2.30}$$

The non-linear set of equations $U_l(\beta) = 0$, $l = 1,\ldots,k$ can be solved numerically by some iterative optimization method (Klein & Moeschberger, 1997).

In practice tied events are commonly encountered. Particularly since the measurement of time is often limited to a discrete time scale. Assuming d_i events at time t_i, the exact probability is cumbersome to compute. Although the numerator is easily managed, the denominator is more tedious considering one has to account for d_i permutations of R_i, which grows quickly as d_i increases. Instead, several approximations has been proposed. The Berslow approximation is given by raising the denominator to d_i, which is reasonably good when d_i is small relative to the size of the risk set R_i (Grambsch & Therneau, 2000). A more accurate approximation was proposed by Efron (1977). He assumed the events to have occurred consecutively, the denominator is then approximated by multiplying the average risk sets at each default.

Example 2.3. *Let $d_i = 2$ be the number of tied events at time t_i and let R_i be the risk set at time t_i. Further let $r_j = \exp(\beta^T z_j)$, where $j = 1, 2, 3, 4$ correspond to the subjects at risk. Assume subjects $j = 1, 2$ have experienced the event. The approximative contribution to the likelihood is given as follows*

- *Berslow approximation*

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)^2} \tag{2.31}$$

- *Efron approximation*

$$\frac{r_1 r_2}{(r_1 + r_2 + r_3 + r_4)(0.5r_1 + 0.5r_2 + r_3 + r_4)}. \tag{2.32}$$

3 Quantile-quantile plots with censored data

Quantile-quantile plots are used to graphically assess the goodness-of-fit of data against a hypothesized parametric distribution, with corresponding distribution function $F(t;\theta)$ for some values of the parameters θ. Let $\hat{F}(t)$ be the empirical distribution associated with the data $t_1, \ldots, t_n$. In the case of censored data, Turnbull and Waller (1992) proposes that the Nelson-Aalen estimator or the Kaplan-Meier estimator is used to estimate the empirical distribution function. In the case where the distribution function is strictly increasing, the inverse $F_0^{-1}(t;\theta)$ is uniquely defined on the image of F. Otherwise, the inverse may be defined as $F^{-1}(t) = \inf\{y \in \mathbb{R} : F(y) \geq t\}$. Thus, plotting the data t_i against $F^{-1}(\hat{F}(t_i);\hat{\theta})$, for some optimized parameters $\hat{\theta}$, will approximate a straight line with intercept 0 and slope 1 (Turnbull & Waller, 1992). Alternatively, one may transform the distribution function such that the plot approximates a straight line with intercept and slope corresponding to functions of θ. Thereby the goodness of fit can be assessed without specifying the parameters of the distribution.

The cumulative distribution function $F(t)$ of the Weibull distribution is given by (2.7). After rearranging $F(t)$,

$$-\log(-\log(1 - F(t))) = k \log t - k \log \theta, \tag{2.33}$$

it can be noted that plotting $\ln(-\ln(1-\hat{F}(t)))$ against $\ln(t)$ yields a straight line with slope k and intercept $-k\ln(\lambda)$ (assuming a good fit of the distribution). Here k and θ are the corresponding parameter values of the Weibull distribution. Similarly, by transformation of the log-logistic distribution function (2.8) we obtain

$$\log\left(\frac{1 - F(t)}{F(t)}\right) = -\beta(\log t - \log \alpha). \tag{2.34}$$

Once again, plotting $\log(t)$ against the left-hand side should approximate a straight line with intercept $\beta \log(\alpha)$ and slope $-\beta$, assuming a good fit. Although these are not truly Q-Q plots, the idea is still the same and the plots are just as illustrative in order to assess the goodness-of-fit.

Finally, for the log-normal distribution we may plot $\log(t)$ against $\Phi^{-1}(\hat{F}(t))$, where Φ is the

CDF of the normal distribution. This is justified by the following rearrangement of the quantile function function

$$\Phi^{-1}(F(t)) = \sigma \log t + \mu. \tag{2.35}$$

4 Finite mixture models

Finite mixture models are generally used as a relatively simple tool to model complex probabilistic distributions. It is represented by the convex combination of a finite number of probability density functions. Therefore, finite mixtures provide a natural representation of heterogeneity in cluster or latent class analysis. It has been successfully applied in a wide range of fields, including astronomy, psychiatry, medicine, engineering and economics (McLachlan & Peel, 2000).

In accordance with (McLachlan & Peel, 2000), let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ denote a random sample of size n, where each $\mathbf{Y}_i$ is a multivariate random variable of dimension p. Let $f(\cdot)$ be the distribution function and $\mathbf{y}_i$ the realized value associated with $\mathbf{Y}_i$. Further let $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots \mathbf{Y}_n^T)$. A random sample of $\mathbf{Y}$ is denoted by its corresponding lower case letter, $\mathbf{y} = (\mathbf{y}_1^T, \ldots \mathbf{y}_n^T)$. Note that $\mathbf{y}$ is an n-tuple of p-dimensional vectors.

Now, for some mixing proportions π_k and corresponding component densities $f_k(\mathbf{y}_i)$, we suppose the density of $\mathbf{Y}_i$ can be written as

$$f(\mathbf{y}_i) = \sum_{k=1}^{m} \pi_k f_k(\mathbf{y}_i) \tag{2.36}$$

where the mixing proportions satisfy

$$\sum_{k=1}^{m} \pi_k = 1, \ \pi_k \geq 0, \ k = 1, \ldots, m. \tag{2.37}$$

Since $f_k(\mathbf{y}_i)$ are probability density functions, it can easily been shown that $f(\mathbf{y}_i)$ defines a density. The representation of $f(\mathbf{y}_i)$ in (2.36) is the m-component finite mixture distribution.

The realization of $\mathbf{Y}_i$ may be interpreted as a realization of the pseudo-variable $\mathbf{Y}_i^k$ with the corresponding density function f_k from (2.36). Here k is determined by the realization of the random variable $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ik})$, which follows a multinomial distribution with exactly one draw from m categories with probabilities $\pi = (\pi_1, \ldots, \pi_m)$. A sampling procedure for $\mathbf{Y}_i$ follows intuitively from this interpretation:

(1) Simulate $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, where $\mathbf{Z}_i \sim \text{Multinomial}_m(1, \pi)$.

(2) For each i, draw the random variable $\mathbf{Y}_i^k$ associated with density function $\{f_k \mid z_{ik} = 1\}$.

This interpretation also forms the foundation of the rationale of the Expectation-Maximization algorithm applied to finite mixture models, explained further in section 5.1. The likelihood of $\mathbf{y}$

for the parameters $\theta = (\theta_1, \ldots, \theta_k, \pi_1, \ldots, \pi_k)$ is given by

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y_i}) = \prod_{i=1}^{n} \sum_{k=1}^{m} \pi_k f_k(\mathbf{y_i}). \tag{2.38}$$

where θ_k is the vector of parameters of the density f_k. It will later prove convenient to construct the complete (pseudo) likelihood, i.e. the likelihood of both the observed data $\mathbf{y}$ and the unobserved latent variables $\mathbf{z}$, which may be defined as follows

$$\mathcal{L}(\theta; \mathbf{y}, \mathbf{z}) = \prod_{i=1}^{n} \prod_{k=1}^{m} f_k(\mathbf{y_i})^{z_{ik}}. \tag{2.39}$$

5 Expectation-Maximization algorithm

The Expectation-Maximization (EM) algorithm is an iterative procedure to compute the maximum-likelihood estimates of statistical models in the presence of incomplete data. It was formally introduced and generalized in the seminal paper by Dempster, Laird, and Rubin (1977), after having appeared in various setting in several preceding papers. The algorithm is particularly useful when the maximum-likelihood of the incomplete data is numerically difficult to compute. The algorithm alternates between two steps; the Expectation-step (E-step) and the Maximization-step (M-step). In the E-step, a computationally more tractable pseudo-likelihood is produced by taking the expectation of the complete likelihood data given the observed data and the model parameters. This likelihood is then maximized in the M-step.

In accordance with Dempster et al. (1977) let $\mathcal{Y}$ and $\mathcal{X}$ define two sample spaces with corresponding realizations $\mathbf{y}$ and $\mathbf{x}$. Further assume the existence of a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. The sample $\mathbf{y}$ is directly observed, whereas $\mathbf{x}$ is only observed indirectly through $\mathbf{y}$. Consequently, the complete data is only known to lie within the subset $\mathcal{X}(\mathbf{y})$. The observed data $\mathbf{y}$ is now associated with the marginalized density function of $\mathbf{x}$ over the subset $\mathcal{X}(\mathbf{y})$.

$$g(\mathbf{y}; \theta) = \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x}; \theta) d\mathbf{x}. \tag{2.40}$$

for some set of parameters θ. The log likelihood of the observed data takes the form

$$\log \mathcal{L}(\theta) = \log g(\mathbf{y}; \theta) \tag{2.41}$$

which in practice often appears as an intractable sum that is both analytically and numerically difficult to compute. Instead, we introduce the conditional density of $\mathbf{x}$ given $\mathbf{y}$ and the parameters θ

$$k(\mathbf{x}|\mathbf{y}; \theta) = \frac{f(\mathbf{x}; \theta)}{g(\mathbf{y}; \theta)}. \tag{2.42}$$

The alternate representation of the likelihood (2.41) can now be written as

$$\log \mathcal{L}(\theta) = \log f(\mathbf{x}; \theta) - \log k(\mathbf{x}|\mathbf{y}; \theta). \tag{2.43}$$

Since $f(\mathbf{x}; \theta)$ is the likelihood for parameter θ and the complete data $\mathbf{x}$, commonly referred to as

the complete likelihood, this is replaced by $\mathcal{L}(\mathbf{x};\theta)$. Now, regarding $\mathbf{x}$ as unobserved (random) and taking the conditional expectation of (2.43) given $\mathbf{Y}=\mathbf{y}$ and evaluated at the parameters of the current iteration of the algorithm, $\theta^{(p)}$, yields

$$\log \mathcal{L}(\theta) = E\big[\log \mathcal{L}(\theta;\mathbf{X}) \mid \mathbf{y};\theta^{(p)}\big] - E\big[\log k(\mathbf{X}|\mathbf{y};\theta) \mid \mathbf{y};\theta^{(p)}\big]. \tag{2.44}$$

The first and second term of the right-hand side of (2.44) are fundamental for the theory of the EM algorithm, and will be denoted

$$Q(\theta;\theta^{(p)}) = E\big[\log \mathcal{L}(\theta;\mathbf{X}) \mid \mathbf{y};\theta^{(p)}\big] \tag{2.45}$$

$$H(\theta;\theta^{(p)}) = E\big[\log k(\mathbf{X}|\mathbf{y};\theta) \mid \mathbf{y};\theta^{(p)}\big]. \tag{2.46}$$

The iteration of the EM algorithm to find the next parameter value $\theta^{(p+1)}$ is now given as follows

- *E-step.* Compute $Q(\theta;\theta^{(p)})$.
- *M-step.* Find the value $\theta^{(p+1)} \in \boldsymbol{\Theta}$ which maximizes $Q(\theta;\theta^{(p)})$.

Then we have that the difference of the log-likelihood between two iterations

$$\begin{aligned}\log \mathcal{L}(\theta^{(p+1)}) - \log \mathcal{L}(\theta^{(p)}) = &\big[Q(\theta^{(p+1)};\theta^{(p)}) - Q(\theta^{(p)};\theta^{(p)})\big] \\ &- \big[H(\theta^{(p+1)};\theta^{(p)}) - H(\theta^{(p)};\theta^{(p)})\big].\end{aligned} \tag{2.47}$$

The first difference on the right-hand side of (2.47) is non-negative, which follows immediately from the M-step, i.e.

$$Q(\theta^{(p+1)};\theta^{(p)}) - Q(\theta^{(p)};\theta^{(p)}) \geq 0. \tag{2.48}$$

Also, for any choice of θ we have

$$\begin{aligned}H(\theta;\theta^{(p)}) - H(\theta^{(p)};\theta^{(p)}) &= E\Big[\log \frac{k(\mathbf{X}|\mathbf{y};\theta)}{k(\mathbf{X}|\mathbf{y};\theta^{(p)})} \mid \mathbf{y};\theta^{(p)}\Big] \\ &\leq \log E\Big[\frac{k(\mathbf{X}|\mathbf{y};\theta)}{k(\mathbf{X}|\mathbf{y};\theta^{(p)})} \mid \mathbf{y};\theta^{(p)}\Big] \\ &= \log \int_{\mathcal{X}(\mathbf{y})} k(\mathbf{x}|\mathbf{y};\theta)d\mathbf{x} \\ &= 0\end{aligned} \tag{2.49}$$

where the second step follows from Jensen's inequality. Consequently, by (2.48) and (2.49) we have established that the sequence of likelihoods $\mathcal{L}(\theta^{(p)})$ of the incomplete data is monotonically increasing

$$\mathcal{L}(\theta^{(p+1)}) \geq \mathcal{L}(\theta^{(p)}). \tag{2.50}$$

5.1 Applied to finite mixture models

For the EM algorithm applied to mixture models, slightly modified but more intuitive notations will be used instead. First, let $\mathcal{L}(\mathbf{y},\mathbf{z};\theta) = p(\mathbf{y},\mathbf{z};\theta)$ be the complete likelihood for the parameter

vector θ and the full data set $\mathbf{x} = \{\mathbf{y}, \mathbf{z}\}$, where $\mathbf{y}$ and $\mathbf{z}$ are the observed and unobserved subsets of $\mathbf{x}$, respectively. In the finite mixture model, $\mathbf{z}$ correspond to the latent variable in (2.39). The marginal likelihood is given by

$$\mathcal{L}(\mathbf{y}; \theta) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z}; \theta) \tag{2.51}$$

with the corresponding log-likelihood

$$l(\mathbf{y}; \theta) = \log \Big[\sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z}; \theta) \Big]. \tag{2.52}$$

Again, this requires maximization of the logarithm of a sum of probabilities which is computationally difficult (Verbelen, Gong, Antonio, Badescu, & Lin, 2014). Instead, in accordance with (2.45) let

$$Q(\theta; \theta^{(p)}) = E\big[\log \mathcal{L}(\theta; \mathbf{y}, \mathbf{Z}) \mid \mathbf{y}; \theta^{(p)} \big]. \tag{2.53}$$

Inserting the expression of the complete likelihood (2.39) of the finite mixture model, assuming no truncation or censoring, yields

$$\begin{aligned} Q(\theta; \theta^{(p)}) &= E\big[\sum_{i=1}^{n} \sum_{k=1}^{m} Z_{ik} \log f_k(\mathbf{y_i}) \mid \mathbf{y}; \theta^{(p)} \big] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{m} E\big[Z_{ik} \mid \mathbf{y}; \theta^{(p)} \big] \log f_k(\mathbf{y_i}). \end{aligned} \tag{2.54}$$

Let $\widehat{z}_{ik}$ be the expected value of Z_{ik} conditional on $\mathbf{y}$. Note here that only $\widehat{z}_{ik}$'s corresponding observation $\mathbf{y}_i$ will be relevant for the estimation. Also, since Z_{ik} is an indicator, the conditional expectation is reduced to the conditional probability of $Z_{ik} = 1$. This may be rewritten and computed by the use of Bayes' rule

$$\begin{aligned} \widehat{z}_{ik} = P(Z_{ik} = 1 \mid \mathbf{y}_i; \theta^{(p)}) &= \frac{P(\mathbf{y}_i \mid Z_{ik}; \theta^{(p)}) P(Z_{ik}; \theta^{(p)})}{P(\mathbf{y}_i; \theta^{(p)})} \\ &= \frac{\pi_k^{(p)} f_k(\mathbf{y}_i)}{\sum_{k=1}^{m} \pi_k^{(p)} f_k(\mathbf{y}_i)}. \end{aligned} \tag{2.55}$$

The functions f_k are evaluated at current parameter estimates $\theta^{(p)}$.

5.2 Applied to censored and truncated finite mixtures

As mentioned previously, life time data is rarely uncensored. Therefore we will assume the presence of random right censoring times c_i of each observation i, as this is the most frequently encountered type of censoring in event time studies. This can however be generalized to interval censoring (e.g. Verbelen et al., 2014; Lee & Scott, 2010). Further assume the data is a realization of the truncated sample space $\mathcal{Y}_T \subseteq \mathcal{Y}$. The truncated sample space is defined by the p-orthotope confined by the lower and upper vectors of truncation points $l = (l_1, \ldots, l_p)$ and $u = (u_1, \ldots, u_p)$, where p is the dimension of each observation. The observation $\mathbf{y}_i$ is no longer an observation of

the the finite mixture model in (2.36), but instead of its truncated counter-party

$$g(y) = \frac{f(y)}{\int_l^u f(y')dy'}. \tag{2.56}$$

This is of course a generalization of the special case of no truncation. Since if all points in l and u were $-\infty$ and ∞, respectively, the denominator would be the density function integrated over $\mathbb{R}^p$, which is 1 by definition.

The truncated distribution $g(y)$ is in fact a mixture distribution with re-weighted mixing proportions, which is evident after minor manipulations of the right-hand side

$$\begin{aligned} g(y) &= \sum_{k=1}^{m} \pi_k \frac{\bar{F}_k(u) - \bar{F}_k(l)}{\bar{F}(u) - \bar{F}(l)} \frac{f_k(y)}{\bar{F}_k(u) - \bar{F}_k(l)} \\ &= \sum_{k=1}^{m} \beta_k \frac{f_k(y)}{\bar{F}_k(u) - \bar{F}_k(l)} \\ &= \sum_{k=1}^{m} \beta_k g_k(y; l, u) \end{aligned} \tag{2.57}$$

where $\bar{F}(u) - \bar{F}(l) = \int_l^u f(y')dy'$ and $g_k(y; l, u)$ is the truncated distribution of component k. To deal with censoring, let δ_i be the indicator of an event for observation i, where $\delta_i = 0$ implies right censoring. Assume the censoring times is a random sample of n i.i.d. random variables with density function h and survival function $\bar{H}$. Then, in accordance with (2.10), the likelihood of the now truncated sample $\mathbf{y}$ in addition to the data δ and $\mathbf{z}$ is given by

$$\begin{aligned} \mathcal{L}(\theta; \mathbf{y}, \delta, \mathbf{z}) &= \prod_{i=1}^{n} (g(\mathbf{y}_i)\bar{H}(\mathbf{y}_i))^{\delta_i} (\bar{G}(\mathbf{y}_i)h(\mathbf{y}_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} g(\mathbf{y}_i)^{\delta_i} \bar{G}(\mathbf{y}_i)^{1-\delta_i} \bar{H}(\mathbf{y}_i)^{\delta_i} h(\mathbf{y}_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \prod_{k=1}^{m} \Big(\beta_k g_k(\mathbf{y}_i)^{\delta_i} \bar{G}_k(\mathbf{y}_i)^{1-\delta_i}\Big)^{z_{ik}} \bar{H}(\mathbf{y}_i)^{\delta_i} h(\mathbf{y}_i)^{1-\delta_i} \end{aligned} \tag{2.58}$$

The truncation thresholds l and u are suppressed in g_k for ease of notation. Thereafter, taking the logarithm yields,

$$\begin{aligned} \log \mathcal{L}(\theta; \mathbf{y}, \delta, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{m} z_{ik} \Big(&\log \beta_k + \delta_i \log g_k(\mathbf{y}_i) + (1 - \delta_i) \log \bar{G}_k(\mathbf{y}_i) \Big) \\ &+ \delta_i \log \bar{H}(\mathbf{y}_i) + (1 - \delta_i) \log h(\mathbf{y}_i). \end{aligned} \tag{2.59}$$

Similarly as in (2.54), the conditional expectation in $Q(\theta; \theta^{(p)})$ only needs to be evaluated over

the Z_{ik}'s.

$$\begin{aligned}\widehat{z}_{ik} = P(Z_{ik} \mid \mathbf{y}_i, \delta_i; \theta^{(p)}) &= \frac{P(\mathbf{y}_i, \delta_i \mid Z_{ik}; \theta^{(p)}) P(Z_{ik}; \theta^{(p)})}{P(\mathbf{y}_i, \delta_i; \theta^{(p)})} \\ &= \left(\frac{\beta_k^{(p)} g_k(\mathbf{y}_i) \bar{H}(\mathbf{y}_i)}{\sum_{k=1}^m \beta_k^{(p)} g_k(\mathbf{y}_i) \bar{H}(\mathbf{y}_i)} \right)^{\delta_i} \left(\frac{\beta_k^{(p)} \bar{G}_k(\mathbf{y}_i) h(\mathbf{y}_i)}{\sum_{k=1}^m \beta_k^{(p)} \bar{G}_k(\mathbf{y}_i) h(\mathbf{y}_i)} \right)^{1-\delta_i} \\ &= \left(\frac{\beta_k^{(p)} g_k(\mathbf{y}_i)}{\sum_{k=1}^m \beta_k^{(p)} g_k(\mathbf{y}_i)} \right)^{\delta_i} \left(\frac{\beta_k^{(p)} \bar{G}_k(\mathbf{y}_i)}{\sum_{k=1}^m \beta_k^{(p)} \bar{G}_k(\mathbf{y}_i)} \right)^{1-\delta_i} \end{aligned} \tag{2.60}$$

The log-likelihood in (2.59) can be divided into three partial log-likelihoods,

$$\log \mathcal{L}(\theta; \mathbf{y}, \delta, \mathbf{z}) = \log \mathcal{L}(\beta; \mathbf{y}, \delta, \mathbf{z}) + \log \mathcal{L}(\theta_g; \mathbf{y}, \delta, \mathbf{z}) + \log \mathcal{L}(\theta_h; \mathbf{y}, \delta, \mathbf{z}) \tag{2.61}$$

where θ_g and θ_h correspond to the parameters of the density functions g_k, $k = 1, \ldots, m$ and h, respectively.

5.3 Observed information

A clear disadvantage of the EM algorithm is that the standard errors are not as easily obtained as when the likelihood function is maximized directly. Generally, a numerical approximation of the Hessian matrix can be requested at the optimal solution for most statistical software. However, since the algorithm optimizes the complete, augmented data, the Hessian of the complete log-likelihood is not truly the observed information. Instead, the literature suggests several ways to account for the missing data (Louis, 1982; Meng & Rubin, 1991; Oakes, 1999). Nonetheless, these methods can be both numerically and analytically tedious. In addition, Basford, Greenway, McLachlan, and Peel (1997) found that the standard errors produced from the observed information were unstable when applied to Gaussian mixtures, unless the sample size was large. Instead they advocated the use of bootstrapping methods.

Another alternative is to approximate the Hessian matrix at the MLE of the logarithm of the original likelihood function (2.38). Thereafter, the standard errors are produced in accordance with (2.24).

6 Mixture cure model

The mixture cure model is a bivariate mixture model. The model postulates the existence of two types of sub-populations, corresponding to subjects being susceptible or non-susceptible to the event under study. For the application to credit data, they may also referred to as being either at-risk or not at-risk over a foreseeable future. In accordance with the finite mixture model as given by (2.36), let π_k be the mixing proportions and the functions f_k and S_k be the corresponding probability density functions and survival functions, respectively. Contrary to the general framework of the finite mixture models as described previously, the probability of being at risk is assumed dependent on some some vector of covariates x. Let the indicator variable z denote if the subject is at-risk, with corresponding probability $P(z = 1 \mid x) = \pi(x)$. Then $\pi(x)$ is referred to as the link function. The probability of not being at-risk is simply given by the complement $P(z = 0 \mid x) = 1 - \pi(x)$.

For $z = 0$, the credit are not at-risk of defaulting and are therefore associated with an infinite lifetime. The conditional survival function of this sub-population is given by

$$S(t \mid z = 0, x) = P(T > t \mid z = 0, x) = 1. \tag{2.62}$$

Once again x is the vector of covariates. Although not necessary, we have here assumed that the covariates x determining the latency is identical to those determining the incidence. The probability density function will effectively be 0 at all finite times t since there is no risk of defaulting. Practically, an infinite lifetime can never be observed for obvious reasons. Analogously let the survival function of the loans at-risk be determined by

$$S(t \mid z = 1, x) = P(T > t \mid z = 1, x) = S_1(t \mid x). \tag{2.63}$$

It is straightforward to show that the survival function of the population in its entirety is given by

$$S_p(t \mid x) = 1 - \pi(x) + \pi(x)S_u(t|x). \tag{2.64}$$

The subscript u is generally used to denote the the conditional survival function of the uncured (at-risk) subjects. Since the density of the non-susceptible sub-population is zero, the probability density function of the whole population is given by

$$f_p(t \mid x) = \pi(x)f(t \mid z = 1, x) = \pi(x)f_u(t \mid x). \tag{2.65}$$

For ease of notation, we will henceforth drop the indexing of the density and survival functions for the sub-population at-risk. Now, the complete likelihood is constructed as

$$\mathcal{L}(\Theta; \mathbf{y}, \mathbf{z}, \mathbf{x}, \delta) = \prod_{i=1}^{n} \Big(\big(\pi(x_i)f(y_i|x_i)\big)^{z_i} \Big)^{\delta_i} \Big(\big(1 - \pi(x_i)\big)^{1-z_i} \big(\pi(x_i)S(y_i|x_i)\big)^{z_i} \Big)^{1-\delta_i} \tag{2.66}$$

where Θ is the set of parameters of the link and distribution functions. All right censoring is assumed to be random, therefore the censoring distribution may be omitted for the MLE as shown in example 2.2. Contrary to mixture models in general, the latent variables z_i is here partially observable. In the event of a default, that particular loan must belong to the susceptible sub-population. More specifically, if $\delta_i = 1$ then $z_i = 1$. The log-likelihood is given by

$$\begin{aligned} l(\Theta; \mathbf{y}, \mathbf{z}, \mathbf{x}, \delta) = &\sum_{i=1}^{n} \delta_i z_i \Big(\ln \pi(x_i) + \ln f(y_i|x_i) \Big) \\ &+ (1 - \delta_i) \Big((1 - z_i) \ln(1 - \pi(x_i)) + z_i \big(\ln \pi(x_i) + \ln S(y_i|x_i) \big) \Big). \end{aligned} \tag{2.67}$$

Since the two different sets of parameters of the link function and the distribution function are orthogonal, the log-likelihood can be factorized into partial log-likelihoods. After some simplifications of the expressions we obtain the two partial likelihoods

$$l_1(\alpha; \mathbf{z}, \mathbf{x}) = \sum_{i=1}^{n} z_i \ln \pi(x_i) + (1 - z_i) \ln(1 - \pi(x_i)) \tag{2.68}$$

$$l_2(\theta; \mathbf{y}, \mathbf{z}, \mathbf{x}, \delta) = \sum_{i=1}^{n} z_i \Big(\delta_i \ln \lambda(y_i|x_i) + \ln S(y_i|x_i) \Big). \tag{2.69}$$

Here the identity in (2.3) is used to rewrite l_2 in terms of the hazard function rather than the density function. The expected value of the mixing proportions is obtained from (2.60) together with the relationship between z and δ.

$$\begin{aligned}\widehat{z}_i &= P(Z_i = 1 \mid \mathbf{y}_i, \delta_i, x_i; \boldsymbol{\Theta}) \\ &= \delta_i + (1 - \delta_i)\frac{\pi(x_i)S(y_i|x_i)}{1 - \pi(x_i) + \pi(x_i)S(y_i|x_i)},\end{aligned} \tag{2.70}$$

where $\pi(x_i)$ and $S(y_i|x_i)$ are evaluated at Θ.

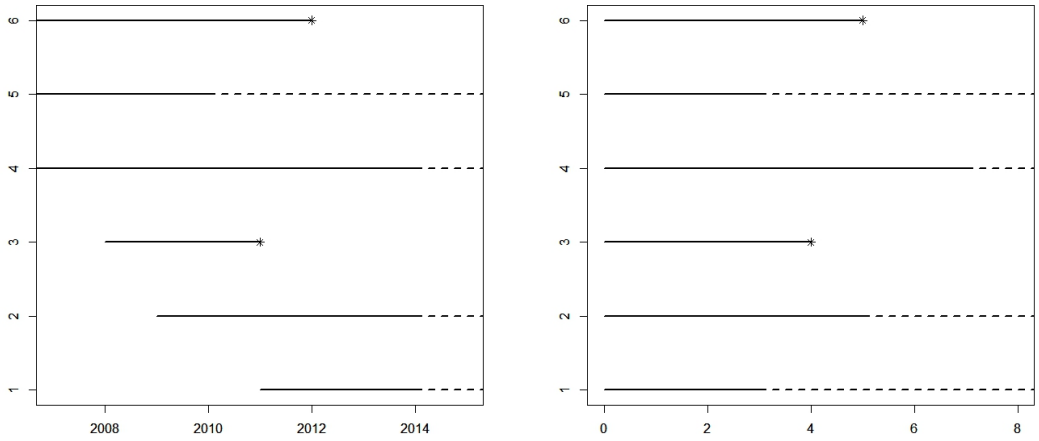
Chapter 3

Methodology

1 Data

The data consists of 314 067 contract lifetimes, extracted from the corporate loans portfolio of a large Swedish bank. This particular subset of the credit portfolio is geographically concentrated on Sweden, with close to 99 % Swedish counterparties, and comprises small- and medium sized enterprises. The contract lifetime is considered to be the time from when the contract is issued until default. In accordance with the regulatory framework of Basel, there exist multiple events after which the contract is marked by a default flag in the internal database. The default flag is consistent with the Basel definition, i.e. 90 days past due. Since a default is in many cases an elongated process, a defaulted contract will often have several different consecutive default flags. Therefore the actual time of default is generally ambiguous. For the purposes of this study, the first default event of a contract will be considered the time of default. Although some clients hold several contract that could potentially default closely in time in case of such an event, these will be treated as independent. Also, consecutive renewals of the contract will be merged into one observation, for instance assuming a 5-year loan is extended by an additional 5 years at maturity, this will be treated as one contract with an observed lifetime of 10 years. This may of course induce potential biases in the data. Assume that loans for which the credit quality deteriorates over the holding period are less likely to be renewed at maturity, relative to loans for which the credit quality remains at the same level or is strengthened. Then the loans are presumably not as likely to default subsequent to the renewal considering the qualitative assessment of the loan prior to its extension. The hazard rate at longer lifetimes is in this case likely to be underestimated. On the contrary, one could argue that loans for which the credit quality deteriorates may be renegotiated and extended in an attempt to avoid potential losses.

The data is a sample of all healthy clients as of June 2007 and all loans issued subsequently, until the end of March 2015. As a result, some contracts are started prior to the initial date of the sampling period. Since these observations are conditional on not experiencing a default event until June 2007, they are regarded as left truncated. Moreover, since only a small portion of the credit will experience default, the lifetime is merely a partial observation of the true lifetime. Hence, each observation is supplemented by the indicator variable denoting if the loan has defaulted or not. The loan will not experience a default if the contract has matured or if is still active at the end of the sampling window. These observations will be treated as random right censoring and generalized Type I censoring, respectively. Although all lifetimes are originally measured in terms of a start and end date in calendar time, the appropriate time scale for the analysis is the age of the contract. This transformation of time is illustrated in figure 3.1.



(a) Calendar date as time scale *(b) Age as time scale*

Figure 3.1: The figure displays the observed lifetime for each loan as continuous lines. The star indicates that the loan has defaulted, whereas the dashed line indicates the uncertainty of the true, unobserved lifetime. Hence the dashed line also indicates right censoring. As seen in figure (a), only loan 3,5 and 6 has been terminated, whereof loan 3 and 6 have defaulted and loan 5 is right censored after being paid back in 2010. Figure (b) displays the same observations with age as time scale.

Although the transformation is important to rigorously analyze the time until default, it is also to some extent problematic. The primary disadvantage is that the impact of events, e.g. the financial crisis of 2008, is diluted over the new time scale (age). Therefore events that could potentially cause default events to surge is not identifiable in the new time scale. This could be mitigated by introducing multiple time scales, whereof one represents calendar time. Subsequently macroeconomic variables can be used as a representation of the financial climate at the time. For the purposes of this study, neither multiple time scales nor macroeconomic variables are included into the model.

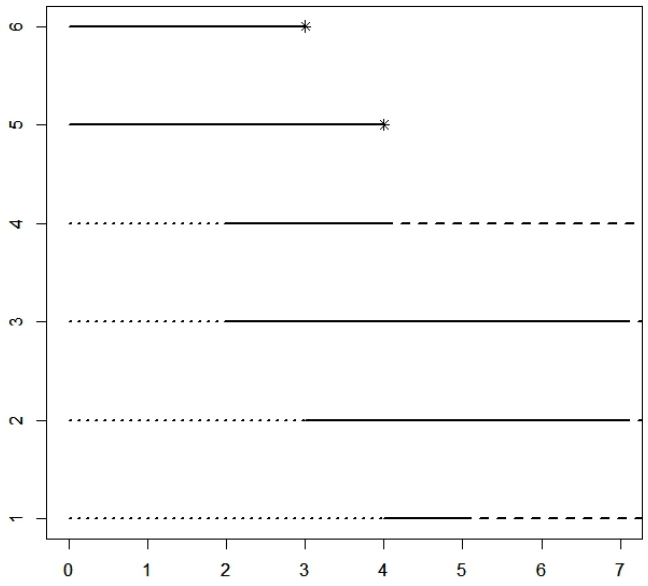


Figure 3.2: The figure displays the observed lifetimes of six loans measured with age as the time scale. Contrary to figure 3.1, here loan 1 through 4 are left truncated, indicated by the dotted lines. Thus, loan 1-4 were issued 2, 2, 3 and 4 years prior to the beginning of the sampling period, June 2007.

The credit of each counter-party is rated according to the internal classification scheme, i.e. on a 21 point scale where a higher credit rating (by number) is equivalent with higher credit worthiness. Moreover, each credit is classified according to four product classes, where each product class represent some types of credit:

- *Corporate lending.* Any loans issued to corporate entities, with the purpose of financing businesses and their investments. The principal amount varies greatly within the product class and includes minor investments as well as considerable real estate purchases.
- *Credit guarantees.* It is the commitment to reimburse another creditor if the debtor fails to repay his loan, in exchange for a fee.
- *Credit facilities.* Includes company line of credits. The obligor can at their own discretion utilize their credit. Debt is created when there is an overdraft of the current account.
- *Other.* This product class is a pooling of types of credit to which few observations are classified.

The distribution of product types within each credit class is depicted in table 3.1. Since the statistical inference is likely to be sensitive to the number of defaults, the corresponding distribution of default within each group is included in the table. Due to difficulties of specifying the EM algorithm with randomly truncated data, the subset of truncated observations have been excluded from the data set. Also, some loans with negligible exposure are considered insubstantial and may default per agreement between the bank and the counterparty. For this reason, all loans with exposure less than SEK 1000 at default will be excluded from the data set. After the reduction of the full data set, 178 811 observations remain.

With the exception of the exposure at default, all other covariates are measured at inception of its respective loan. Although most covariates are expected to vary over time, this is not taken into consideration. Instead, the term structure of default probabilities are calibrated with respect to the characteristics of each credit at initial recognition.

	Number of Observations / Defaults				
Credit Rating	Corporate Lending	Credit Guarantees	Credit Facilities	Other	Total
1	707 / 159	167 / 29	167 / 21	109 / 3	1 150 / 212
2	468 / 77	83 / 5	240 / 24	78 / 4	869 / 110
3	802 /91	152 / 15	623 / 122	98 / 5	1 675 / 233
4	1 519 / 132	262 / 26	469 / 44	222 / 7	2 427 / 209
5	2 274 / 174	340 / 27	1 698 / 191	318 / 5	4 630 / 397
6	8 941 / 404	878 / 55	3 038 / 241	589 / 8	13 446 / 708
7	8 966 / 329	1 229 / 50	3 128 / 218	638 / 9	13 961 / 606
8	11 792 / 361	1 333 / 63	6 251 / 261	853 / 3	20 229 / 688
9	15 732 / 359	1 520 / 43	1 998 / 80	1130 / 2	20 380 / 484
10	15 184 / 226	1 726 / 33	2 439 / 96	905 / 3	20 254 / 358
11	19 080 / 255	1 806 / 37	2 994 / 152	1 331 / 43	25 211 / 487
12	15 299 / 118	2 093 / 25	1 727 / 37	875 / 2	19 994 / 182
13	9 796 / 53	1 136 / 13	1 014 / 20	491 / 1	12 437 / 87
14	13 235 / 39	666 / 4	710 / 4	298 / 1	14 909 / 48
15	2 108 / 10	576 / 3	333 / 2	85 / 0	3 102 / 15
16	721 / 1	69 / 1	91 / 3	30 / 0	911 / 5
17	971 / 1	40 / 1	45 / 0	51 / 0	1 107 / 2
18	811 / 1	6 / 0	26 / 0	17 / 0	860 / 1
19	470 / 2	2 / 0	8 / 0	10 / 0	490 / 2
20	264 / 0	2 / 0	8 / 0	6 / 0	280 / 0
21	431 / 0	0 / 0	8 / 0	5 / 0	444 / 0
Total	129 571 / 2 792	14 086 / 430	27 015 / 1506	8 139 / 96	178 811 / 4 834

Table 3.1: The distribution of observations and defaults within each credit rating and product class, excluding the left-truncated data. The default is defined in accordance with the regulatory framework of Basel, thus this can be considered a conservative measure of the true number of defaults.

2 Estimation of the mixture cure model

Due to the general intractability of the log-likelihood for mixture models, the EM algorithm is used to estimate the parameters of the mixture cure model. In this study, both a parametric as well as a semi-parametric model will be used to fit the data. If the data set was large and included an abundant number of defaults within each cohort, one could consider calculating the empirical survival curves with the Kaplan-Meier estimator for each credit rating and product type. Most often this is not possible. Instead, for the semi-parametric model one has to impose some assumption on the characteristics of the hazard rate in relations to the covariates. Many studies on credit scoring assume the proportional hazard model (e.g. Basanik et al., 1999; Bellotti & Crook, 2009; Carling, Jacobson, Linde, & Roszbach, 2007). Alternatively one can apply accelerated time to failure models, although they do not seem to be encountered in literature as often. Nonetheless, these are both strong assumptions and may not necessarily be appropriate representations of reality, and should therefore be used with caution.

Example 2.1. *To illustrate the difficulties of estimating the parametric mixture cure model by the EM algorithm, 50 sets of mixture cure data is simulated. Each data set consist of 100 observations for each of the artificially constructed subsets, i.e. a full sample size of 500. The parameters are estimated by the EM algorithm and by numerical minimization of the true negative log-likelihood.*

	EM algorithm			Direct minim.		
	Avg.	Bias	S.E	Avg.	Bias	S.E
Incidence						
$\alpha_1 = -1.5$	-1.380	0.120	0.769	-1.532	-0.032	0.879
$\alpha_2 = -1.0$	-0.914	0.086	0.634	-0.839	0.161	1.098
$\alpha_3 = 0.0$	0.276	0.276	1.257	0.688	0.688	2.572
$\alpha_4 = 0.5$	0.603	0.103	0.798	1.338	0.838	2.728
$\alpha_5 = 1.0$	1.381	0.381	1.681	1.944	0.944	2.773
Latency						
$\beta_1 = 4.5$	4.434	-0.066	0.551	4.315	-0.185	0.388
$\beta_2 = 5.0$	5.078	0.078	0.448	5.052	0.052	0.447
$\beta_3 = 5.5$	5.511	0.011	0.269	5.519	0.019	0.289
$\beta_4 = 6.0$	5.992	-0.008	0.246	6.006	0.006	0.271
$\beta_5 = 6.6$	6.447	-0.053	0.204	6.455	-0.045	0.215
Shape						
$\gamma = 0.5$	0.542	0.042	0.095	0.544	0.044	0.094

Table 3.2: The table displays the results of estimating the parameters of a mixture model by either the EM algorithm or by minimizing the true negative log-likelihood directly.

Although the model here is quite simple, table 3.2 shows that with only five variables for the incidence as well as for the latency, minimization of the proper negative log-likelihood yields much larger standard errors for the incidence level. The other estimates seem to have quite similar behavior. As the problem increases in size, the EM algorithm is expected to be successively favorable.

In the parametric model, the distribution of the latency for the loans at-risk will be assumed to belong to any of the parametric families mentioned in section 1.1. The appropriateness of each distribution will be assessed graphically, and the model will be determined accordingly. To allow for a set of covariates to alter the shape of the survival function, the model parameters are defined by some regression model.

The link function $\pi(z)$, which determines the proportion of susceptible and non-susceptible loans, will be important for the risk assessment of each loan. The identity link $\pi(z) = \alpha^T z$ facilitates the interpretation of the covariates' effect on the probability of being susceptible. It may however induce problems for small or large levels of incidence (Lambert, Thompson, Weston, & Dickman, 2007). For the log-log link $\log(-\log(1 - \pi(z))) = \alpha^T z$, the covariate effects are interpreted as log excess hazard ratios, under the assumption of proportional hazards. Nonetheless, the logistic link function $\log(\pi(z)/(1 - \pi(z))) = \alpha^T z$ is used for both the parametric as well as the semi-parametric model to keep the connection with existing credit scoring models. In the logistic link function, the covariates' effect is proportional to the logarithm of the odds-ratio (Lambert et al., 2007).

2.1 Semi-parametric

The semi-parametric mixture cure model is the most frequently encountered specification of the model within the literature. The model is obtained when the susceptible sub-population conform to the proportional hazards assumption, with a non-parametric baseline survival function. Consequently, $S(y_i|x_i)$ in (2.69) is assumed to have the representation $S_{u0}(t)^{\exp(\beta^T x_i)}$, where $S_{u0}(t)$ is the baseline survival function of the susceptible population. In accordance with (Peng & Dear, 2000; Sy & Taylor, 2000), the coefficients β can in each iteration of the EM algorithm be estimated by the means of ordinary cox regression with the minor modification of including $\hat{z}_j$ as a scaling factor in the denominator

$$\mathcal{L}(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta^T \mathbf{x}_i)}{\left\{ \sum_{j \in R_i} \hat{z}_j \exp(\beta^T \mathbf{x}_j) \right\}^{d_i}}. \tag{3.1}$$

Here the Berslow approximation is used to handle tied events. Since the number of defaults at each event time t_i must be very small, considering the low default rate in the portfolio, the Berslow approximation is assumed to be of sufficient accuracy. Although the Efron approximation has greater precision, the Berslow approximation is favorable due to its computational advantage. Especially considering the EM algorithm itself is computationally expensive. The log-likelihood is maximized numerically by solving the set of equations in (2.30). Furthermore, the partial likelihood l_1 is also maximized numerically by any type of iterative optimization algorithm.

Although the parameters β are estimated without having to specify the baseline hazard, it is necessary to estimate the survival function in each iteration to calculate the $\widehat{z}_i$'s (2.70) in the E-step. The survival function is estimated by a Nelson-Aalen type estimator (2.26) with minor modifications (Peng & Dear, 2000; Sy & Taylor, 2000), the estimator takes the following form

$$\widehat{S}_{u0}(t) = \exp\left(- \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j \in R_i} \hat{z}_j \exp(\beta^T \mathbf{x}_j)} \right) \tag{3.2}$$

where $t_1, \ldots, t_m$ are the ordered default times and d_i is the number of defaults at time t_i. The Expectation and Maximization steps of the EM algorithm now involves estimating the conditional expectations $\widehat{z}_i$ w.r.t. the parameters and the baseline survival function $\widehat{S}_{u0}(t)$ estimated in the M-step of the preceding iteration. A more detailed pseudo-code is presented in algorithm 1. We choose the absolute error tolerance $\epsilon = 10^{-7}$. The algorithm is implemented using the R-package `smcure` (Cai, Zou, Peng, & Zhang, 2012).

The primary model of the covariates that is investigated includes the credit rating and the product class. Each level of the two categories is coded by a dummy variable. The corresponding dummy variables for the credit ratings are denoted C_i for each level $i = 1, \ldots, 21$. For the product classes, the dummy variables are denoted P_{CL}, P_{CG}, P_{CF} and P_O for *Corporate Lending*, *Credit Guarantees*, *Credit Facilities* and *Other*, respectively. To avoid multicolinearity, C_1 and P_{CL} are removed from the regression in favor of an intercept. The credit with rating 1 and product type Corporate Lending is considered the base case to which all other cohorts are compared, e.g. for credit rating 10 and product type Credit Guarantees the estimate is given by $\beta_0 + \beta_{10} + \beta_{CG}$. The model will be applied for the Cox regression as well as the logistic regression. It is noticeable

Algorithm 1: EM Algorithm: Semi-Parametric Mixture Cure Model

Data: $\mathbf{x}$ is a set of covariates, $\mathbf{y}$ is the observed lifetimes and δ the censoring indicators and $\mathbf{z}$ the latent variables.

Result: ML estimate of $\theta = (\alpha, \beta)$.

$p \leftarrow 0$;

$\theta^{(p)} \leftarrow \text{initialize}(\theta) \in \Theta$;

while $||\beta^{(p)} - \beta^{(p-1)}|| > \epsilon$ **do**

  /* E-step: */

  for $i = 1$ **to** n **do**

    $S(y_i|x_i) \leftarrow S_{u0}(y_i|x_i)^{\exp(\beta^T x)}$

    $z_i \leftarrow \delta_i + (1-\delta_i)\pi(x_i)S(y_i|x_i)/\big(1-\pi(x_i)+\pi(x_i)S(y_i|x_i)\big)|_{\beta^{(p)}}$;

  end

  /* M-step: */

  $p \leftarrow p + 1$;

  $\alpha^{(p)} \leftarrow \text{argmax}_\alpha \, l_1(\alpha; \mathbf{z}, \mathbf{x})$; /* l_1 as (2.68) */

  $\beta^{(p)} \leftarrow \text{argmax}_\beta \, l_2(\beta; \mathbf{y}, \mathbf{z}, \mathbf{x}, \delta)$; /* l_2 as (3.1) */

  Update $\widehat{S}_{u0}(t)$; /* by (3.2) */

  $\theta^{(p)} \leftarrow (\alpha^p, \beta^p)$;

end

from the score function (2.30) that if there exist no defaults for any of the coviarates, then the corresponding coefficient cannot be estimated. Since Credit rating 20 and 21 are not associated with any defaults, they are excluded from the analysis. The regression takes the form

$$\beta_0 + C_2\beta_2 + \cdots + C_{19}\beta_{19} + P_{CG}\beta_{CG} + P_{CF}\beta_{CF} + P_O\beta_O \tag{3.3}$$

In Cox's model, the intercept is indirectly included in the baseline survival function (the baseline hazard), and for that reason the estimate of β_0 is not directly available. Instead, only coefficients β_2 through β_{19} together with β_{CG}, β_{CF} and β_O are presented.

The observed information matrix for the parameters α can be estimated by the methods of Louis (1982). The standard errors of β are on the contrary not as easily accessible (Peng & Dear, 2000). This is partly a consequence of the non-parametric form of the baseline survival function. It is further complicated when the partial likelihood (3.1) is maximized instead of the true likelihood (2.69). A few methods have been suggested to find the standard errors of the semi-parametric cure model, e.g. Sy and Taylor (2000) defines the full, semi-parametric log-likelihood and derives the second derivatives w.r.t. to each covariate analytically to compute the Hessian matrix. Chen and Kuk (1992) uses a Monte-Carlo approach. Since the proposed methods are analytically and computationally tedious, we will instead use a bootstrap method as suggested in (Cai et al., 2012). The bootstrap procedure is defined as follows:

(1) Let $\mathbf{y}_c$ and $\mathbf{y}_u$ be the censored and uncensored subsets of the data $\mathbf{y}$, with corresponding sizes n_c and n_u.

(2) Draw n_c and n_u random samples with replacement from $\mathbf{y}_c$ and $\mathbf{y}_u$, respectively. Merge the two samples and denote this $\mathbf{y}^j$.

(3) For sample $\mathbf{y}^j$, estimate the parameter vectors α^j and β^j by algorithm 1.

(4) Repeat step (2)-(3) for $j = 1, \ldots, N$. Then estimate the sample standard deviation of each parameter α_k and β_k.

Since the EM algorithm is computationally expensive and converge slowly, we settle for a moderately high absolute error tolerance $\epsilon = 10^{-4}$. Yet, it is chosen sufficiently small such that the standard deviation is still reliable.

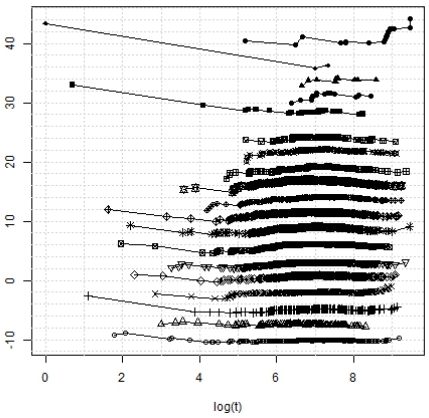
2.2 Parametric

In the parametric specification of the mixture cure model, the probability density function of the lifetime of each individual of the susceptible sub-population is assumed to belong to some parametric family of distributions. The appropriateness of each distribution is assessed graphically. The distributions defined in section 1.1, Weibull (including Exponential) distribution, Log-logistic distribution or Log-normal distribution, are considered as potential candidate distributions. We assume that the form of the survival curve is primarily explained in terms of the credit rating and the product type. Therefore, $\hat{F}(t)$ is estimated for each cohort by the Kaplan-Meier estimator in (2.25). Thereafter, $\log(t)$ is plotted against $\log(-\log(1-\hat{F}(t)))$, $\log((1-\hat{F}(t))/\hat{F}(t))$ and $\Phi^{-1}(\hat{F}(t))$ to assess the goodness of fit with the Weibull, Log-Logistic and Log-normal distributions, respectively. Since the estimator only jumps at defaults it is only necessary to plot the actual default times.

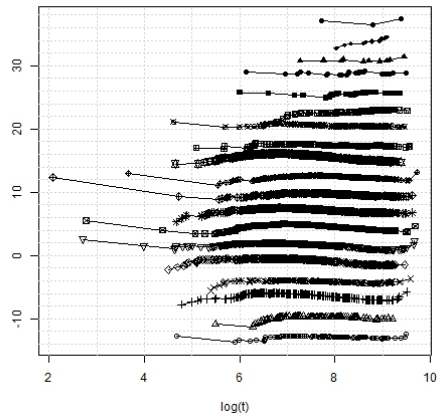
To reduce the amount of figures produced, the plots are aggregated by product type. Each figure consists of the plots for every credit rating within that particular product type. Since the scale parameter (or mean for the log-normal distribution) is expected to be increasing with lower credit rating, the produced QQ-plot of each credit rating is separated vertically from the other. To keep each cohort further distinguishable in the plot, an increasing sequence of constants for each credit rating j is added to the corresponding y-values y_j. Also, the set of points corresponding to the QQ-plot of each credit rating is transformed such that the slope relative to the other cohorts can more easily be assessed. More specifically, for the first credit rating within a product type, the slope and intercept is found by the least-squares estimated of the model $y_1 = kx_1 + m$. Thereafter, $\widehat{k}x_j$ is subtracted from y_j for each $j = 1, \ldots, 21$. Recall here that the slope of each QQ-plot is given by either the shape or the standard deviation, depending on the distribution. If all plots are approximately horizontal, then it is reasonable to assume a common shape parameter or standard deviation with the product type.

One should keep in mind that the full model is indeed not represented solely by the Weibull, Log-logistic or the log-normal distributions, but instead by a mixture cure model of the aforementioned distributions as the candidate component distributions. Thus, substituting $\hat{F}(t)$ by $\hat{F}(t)/p$, where p is the probability of being susceptible, would be more appropriate[1]. It does however not exist any transformation of this model for which the result approximates a straight line, without having to specify the parameter p. Nonetheless, for lower credit rating the susceptible proportion is expected to be high, possibly close to one. If the plots indicate a reasonable fit, the model is expected to be well approximated by the Weibull distribution, i.e. with the level of susceptible individuals $p = 1$. The QQ-plots do not necessarily need to indicate a close fit for every credit rating, since minor deviations are presumed dealt with by the additional flexibility of the mixture cure model.

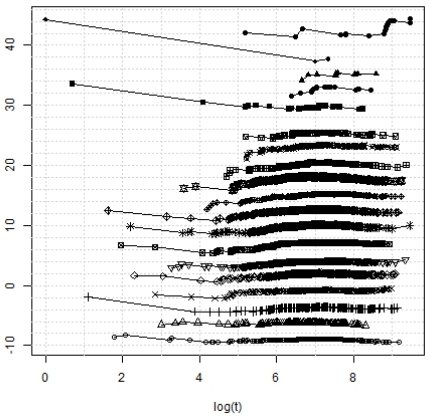
[1] The full CDF of the MCM is given by $F(t) = pF_p(t)$. It is assumed that $\hat{F}(t)$ is an empirical estimation of $pF_p(t)$.



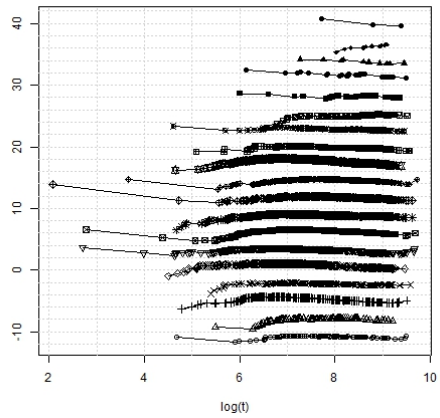
(a) Log-Logistic: Corporate Lending



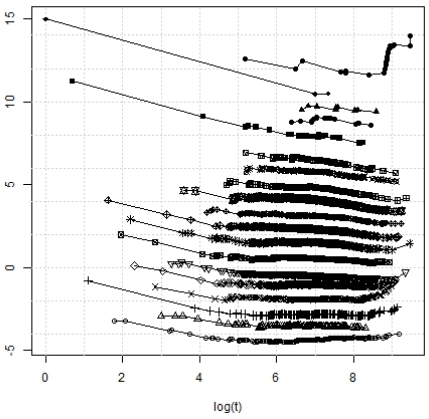
(b) Log-Logistic: Credit Facilities



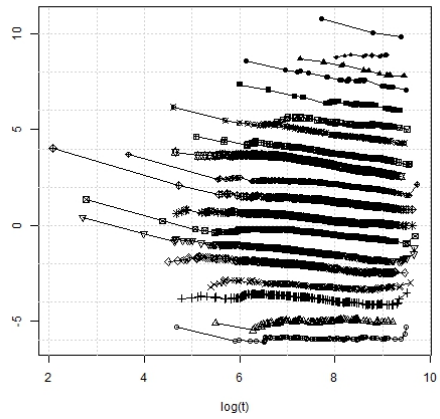
(c) Weibull: Corporate Lending



(d) Weibull: Credit Facilities



(e) Log-Normal: Corporate Lending



(f) Log-Normal: Credit Facilities

Figure 3.3: Each subfigure shows transformed QQ-plots according to (2.33)-(2.35) *for all credit ratings 1 (bottom) through 21 (top) within each product type. The slopes are transformed as discussed in section 2.2 to yield approximately vertical lines if the shape (or standard deviation) is close to equal across all ratings.*

The QQ-plots for Corporate Lending and Credit Facilities for all distributions are shown in figure 3.3. The remaining QQ-plots, for Credit Guarantees and Other, are displayed in figure (ref figure) in the Appendix. For the Log-normal distribution, the lines tend to be more curved than for the Log-logistic or the Weibull distributions, indicating an inferior fit. Also, the plots for the Log-logistic and the Weibull distributions are close to identical. Since the Weibull distribution is more frequently encountered in literature as well as in practice, it will be chosen for the parametric specification of the mixture cure model.

The parameters of the Weibull distribution are specified in terms of some set of covariates. The choice of covariates for each parameter is generally deduced from some preconceived idea on how different variables will alter the form of the survival curve. One may also try different models and assess the appropriateness in retrospect. In the parametric MCM we will refrain from averaging the effect of the credit rating across all product types. Instead, we have one model for each product type. By figure 3.3c and 3.3d it seems reasonable to assume that the shape of the Weibull distribution is determined solely by the product. Since the shape must be strictly positive, we use the following parametrization $k = \exp(\gamma)$. For the same reason, the scale is parametrized according to $\theta = \exp(\beta^T x)$, where x is the covariate vector. The covariate vector consist of the dummy variables for each credit rating, i.e.

$$\theta = \exp(\beta_0 + \beta_2 C_2 + \cdots + \beta_{21} C_{21}). \tag{3.4}$$

Thereafter, it is straightforward to apply the EM algorithm for statistical inference of the model. Since the algorithm does not ensure the convergence to the global minimum, different initializing procedures has been suggested as improvements. Commonly, it is suggested to use random starts to initialize the algorithm (McLachlan & Peel, 2000). Similarly to Gormley, O'Hagan, and Murphy (2012), a burn-in scheme will be used for finding a suitable vector of parameters to initialize the algorithm:

(1) Generate the set of 2^j candidate parameter vectors $\theta_i \in \mathbf{\Theta}$, $i = 1, \ldots, 2^j$.

(2) For each parameter vector θ_i in the set, conduct 2 iterations of EM algorithm. Concurrently update the parameters θ_i and associate it with the observed log-likelihood l_i evaluated at θ_i.

(3) Rank the parameter vectors θ_i by their corresponding log-likelihood l_i in descending order.

(4) Reduce the set of candidates to the upper-half of the ranked parameter vectors.

(5) Continue with steps (2)-(4) until only one candidate remains.

Ideally one would like to proceed until convergence for each candidate parameter vector. This is however impracticable for larger data sets when the running time of each instance of the algorithm increases. The burn-in scheme serves as a good alternative nonetheless.

Once the initial parameters are found, we proceed to compute the expected value of the latent variables conditional on the data evaluated at current parameter values, in accordance with (2.70). They are subsequently used in the minimization of the partial negative log-likelihoods in (2.68)-(2.69). This procedure is iterated until convergence. Convergence is generally defined in terms of relative changes in the parameters values or the log-likelihood. A more detailed pseudo-code for the EM algorithm applied to the parametric mixture cure model is presented in algorithm 2.

Algorithm 2: EM Algorithm: Parametric Mixture Cure Model

Data: $\mathbf{x}$ is a set of covariates, S is the survival function of the susceptible population, $\mathbf{y}$ is the vector of observed lifetimes and δ the vector of censoring indicators.

Result: ML estimate of $\theta = (\alpha, \beta, \gamma)$.

$p \leftarrow 0$;

$\theta^{(p)} \leftarrow \text{initialize}(\theta) \in \Theta$;

while $||\theta^{(p)} - \theta^{(p-1)}|| > \epsilon$ **do**

  /* E-step: */

  for $i = 1$ **to** n **do**

    $z_i \leftarrow \delta_i + (1 - \delta_i)\pi(x_i)S(y_i|x_i)/\big(1 - \pi(x_i) + \pi(x_i)S(y_i|x_i)\big)|_{\theta^{(p)}}$;

  end

  /* M-step: */

  $p \leftarrow p + 1$;

  $\alpha^{(p)} \leftarrow \text{argmax}_\alpha \, l_1(\alpha; \mathbf{z}, \mathbf{x})$; /* l_1 as (2.68) */

  $(\beta^{(p)}, \gamma^{(p)}) \leftarrow \text{argmax}_{\beta,\gamma} \, l_2(\beta, \gamma; \mathbf{y}, \mathbf{z}, \mathbf{x}, \delta)$; /* l_2 as (2.69) */

  $\theta^{(p)} \leftarrow (\alpha^p, \beta^p, \gamma^p)$;

end

2.3 Assesing the goodness-of-fit

The literature provides several ways of assessing the appropriateness of a proportional hazards model. Amongst others, the Schoenfeld residuals is one of the methods used to graphically assess the appropriateness of the proportional hazards assumption for each covariate of the model (Grambsch & Therneau, 2000). Although the method has subsequently been altered and accommodated to mixture cure models (Heitjan, Li, & Wileyto, 2012), it is not applicable to data with tied failures. Instead, QQ-plots can potentially be used to graphically assess the goodness-of-fit for both the parametric and semi-parametric models. The appropriate form of the QQ-plots is derived from the identity $pF_p(t) = \widehat{F}(t)$, which yields the QQ-plot $F_p^{-1}(\widehat{F}(t)/p)$ against t. This will induce problems as the argument $\widehat{F}(t)/p$ will occasionally surpass one, where the inverse is not defined. Therefore, the difference between the estimated survival function and the empirical survival function $S(t;\widehat{\theta}) - \widehat{S}(t)$ is adopted, similarly to (Chen, Tsay, Wu, & Horng, 2013). Contrary to the Schoenfeld residuals, other graphical methods including QQ-plots and residual plots, are only appropriate for categorical variables. Then the data can be classified according to a set of homogeneous sub-populations, for which the empirical survival function can be estimated by the Kaplan-Meier estimator. The estimator is only reliable when there are a sufficient number of defaults recorded within the sub-population. Here, the goodness-of-fit is only assessed for cohorts with more than 10 defaults. The residual are determined as follows:

(1) Retrieve the subset of observations $\mathbf{y}_{ij}$ that has credit rating i and is of product type j. If the number of defaults within $\mathbf{y}_{ij}$ does not exceed 10, we do not continue with the remaining steps.

(2) Compute the empirical survival function $\widehat{S}_{ij}(t)$ of $\mathbf{y}_{ij}$.

(3) Let the $S_{ij}(t;\widehat{\theta})$ be the survival function of some credit of product type j and credit rating i, for the MLE $\widehat{\theta}$.

(4) Let the residual be the vector of differences $R_{ij}(t) = S_{ij}(t;\widehat{\theta}) - \widehat{S}_{ij}(t)$. Since the empirical survival function $\widehat{S}_{ij}(t)$ only jumps at the defaults, the residual is only plotted for times t where a default has occurred.

It may be noted here that an alternative representation of the residual is given by

$$\begin{aligned} R_{ij}(t) &= S_{ij}(t;\widehat{\theta}) - \widehat{S}_{ij}(t) \\ &= (1 - F_{ij}(t;\widehat{\theta})) - (1 - \widehat{F}_{ij}(t)) \\ &= \widehat{F}_{ij}(t) - F_{ij}(t;\widehat{\theta}) \end{aligned} \tag{3.5}$$

which is the difference between the empirical and the estimated distribution functions. Then the residual at time t is interpreted as the difference between the predicted and the empirical cumulative default probability. Since the distribution function F is referred to as the (cumulative) term structure of default probabilities, the accuracy of the term structure can be assessed directly.

In addition to the absolute difference between the empirical and the estimated term structures, we also present the relative differences defined by

$$\left(\widehat{F}_{ij}(t) - F_{ij}(t;\widehat{\theta})\right)/F_{ij}(t;\widehat{\theta}). \tag{3.6}$$

3 Simulation study

A simulation study is conducted to assess the accuracy of the EM algorithm for the parametric model. The purpose of the simulation study is to reveal potential biases of the algorithm in terms of the estimated parameters and what type of settings or circumstances may influence the bias. Ideally the attributes of the simulated data should be in close analogy with that of the actual data. The circumstances targeted for primary investigation are

- *High censoring rate.* Since the actual data is heavily right censored, it is necessary to investigate if the censoring rate cause biased parameter estimates. The censoring rate is affected by both the mean of the true lifetime in relation to the length of the observational window, as well as the distribution of the censoring mechanism.
- *Lack of default events.* Closely related to the censoring rate is the number of defaults available for the estimation of each parameter. It is not necessarily the censoring rate that induces biases in the parameter estimations, but the number of defaults. Here we wish to establish a minimum number of defaults necessary for a reliable estimate of the parameters.

Data is simulated from the underlying cure model by the inverse probability integral transform. Knowing the distribution function F_Y of the random variable Y of the mixture cure model, Y can be simulated through the inverse of its distribution function, F_Y^{-1}, evaluated at a uniformly distribution r.v. U. The distribution function is given by $F_Y(y|x) = \pi(x)F_1(y|x)$, where F_1 is the distribution function of the sub-population at-risk, assumed to follow the Weibull distribution

(2.7). Then the data is simulated from the following identity

$$Y = F_1^{-1}\Big(\frac{U}{\pi(x)}\Big|x\Big). \tag{3.7}$$

Since the argument takes values greater than one for $\pi(x) < 1$, where the inverse is not defined, the argument is forced to take value one if the fraction is indeed larger. These will represent the credit not at-risk. A random right censoring scheme is introduced to represents the reimbursement of the credit. The random censoring times C_i are assumed independent and identically distributed. To facilitate the interpretation and understanding of the censoring scheme in relations to the true lifetimes, the random censoring times are assumed to be Weibull-distributed. Additionally, generalized Type I censoring as a result of the limited observational window is introduced by assigning each observation a random start date, assumed to be uniformly distributed over the sampling period $[T_s, T_e]$. If the length of the observed lifetime $y_i = \min(T_i, C_i)$ reaches beyond the end of the horizontal window, the observation is censored. The simulation scheme is found in detail in algorithm 3.

The outline of the simulation study begins with defining three artificial homogeneous sub-populations. The sample size of the simulated data is given by $3n$, where n is the number of observations within each sub-population. The groups are assumed to share the common shape parameter γ, but have different scale parameters. The scale of each group is determined by the exponentiated linear regression $\exp(\beta^T x)$, for the covariate vector x of dummy variables. More specifically, the scale parameter is given by

$$\theta_i = \exp(\beta^T x_i) = \exp(\beta_0 + \beta_2 x_2 + \beta_3 x_3). \tag{3.8}$$

The coefficients are chosen such that there are different levels of censored observations within the sub-populations. Each group will represent a credit rating, whereof the third group will be the credit rating extraordinary low default rate. Also, by altering the size n, the expected number of defaults can be manipulated. This allows us to differentiate between the effect of low default rate and the lack of default events. Finally, the EM algorithm is applied to the simulated data in accordance with algorithm 2.

Algorithm 3: Simulating censored random variables of the Weibull cure model

Data: n is the sample size, $\mathbf{x}$ is the vectors of covariates, α, β, γ are parameter vectors of the cure model, T_s and T_e are the start and end date of the sampling window and c_1 and c_2 are scaling constants.

Result: Vector $\mathbf{y} = (y_1, \ldots, y_n)$ of observed lifetimes.

for $i = 1$ **to** n **do**

- $T_s^i \leftarrow \text{Uniform}(T_s, T_e)$; `/* Start date of observation i */`
- $c \leftarrow \text{Weibull}(c_1\beta, c_2\gamma; x_i)$; `/* Censoring time */`
- $U \leftarrow \text{Uniform}(0, 1)$;
- $p \leftarrow \pi(x_i; \alpha)$;
- $\tilde{U} \leftarrow \min(U/p, 1)$;
- $t \leftarrow F_1^{-1}(\tilde{U}|x_i; \beta, \gamma)$; `/* Actual lifetime */`
- $y_i \leftarrow \min(t, c, T_e - T_s^i)$; `/* Observed lifetime */`

end

Chapter 4

Results

1 Semi-Parametric: proportional hazards

Table 4.1 presents the estimated parameters of the semi-parametric mixture cure model. It is reasonable to expect the proportion of susceptible credit (incidence) to be decreasing with higher credit worthiness (higher credit rating by number). Analogously, the latency is likely to be increasing with higher credit worthiness. This is represented by decreasing coefficients α_j or β_j for higher credit rating j. In accordance with the expectation, there is a clear tendency for both the incidence and the latency to be decreasing with higher credit rating, with some deviancy. The coefficient corresponding to the latency of credit rating 19 (β_{19}) is the only estimate that clearly diverge from the overall trend. It is in fact positive when all other coefficients, except β_2, are negative. The lower latency is on the other hand compensated by a significantly larger proportion of non-susceptible credit. Nonetheless, the possibility of the abnormal estimates being a result of the lack of default events should not be rejected. The estimated coefficients of the latency seem to be particularly sensitive to the number of default events. For credit rating 16 through 19, where there are only a few default events the standard errors are much larger than for the other estimates. The corresponding estimates for the incidence level does not vary as much. Instead, the standard error tends to be increasing with the frequency at which default events occur (although not explicitly presented, the default frequency can be inferred from table 3.1, and have as expected a decreasing tendency with higher credit rating).

A few examples of the estimated term structure is presented in figure 4.1. Evident by both the term structures as well as the estimated parameters, all credit ratings for all product types have a proportion of non-susceptible that is larger than zero. Thus strengthening the validity of the mixture cure model. Yet the most conspicuous element of figure 4.1 is the poor fit of credit rating 1 for both Corporate Lending as well as Credit Facilities. This is further illustrated by the residuals presented in figure 4.3 as well as in figure A.1 and A.2 in appendix A. Figure 4.3 shows that for Corporate Lending, except for credit rating 1, 2 and 15, the residuals indicate a relatively good fit over the first couple of years at the least. Whereas towards the end of the time period, the empirical and the estimated term structures seem to diverge for most credit ratings. The Kaplan-Meier estimate of the survival function tend to be unstable in the tails with high censoring rate (Klein & Moeschberger, 1997), thus potentially augmenting the magnitude of the residual. Therefore the estimated residual is not as reliable for large times t. The estimate is likely to be unreliable for lower credit ratings as well, since the risk set Y_i of the Kaplan-Meier

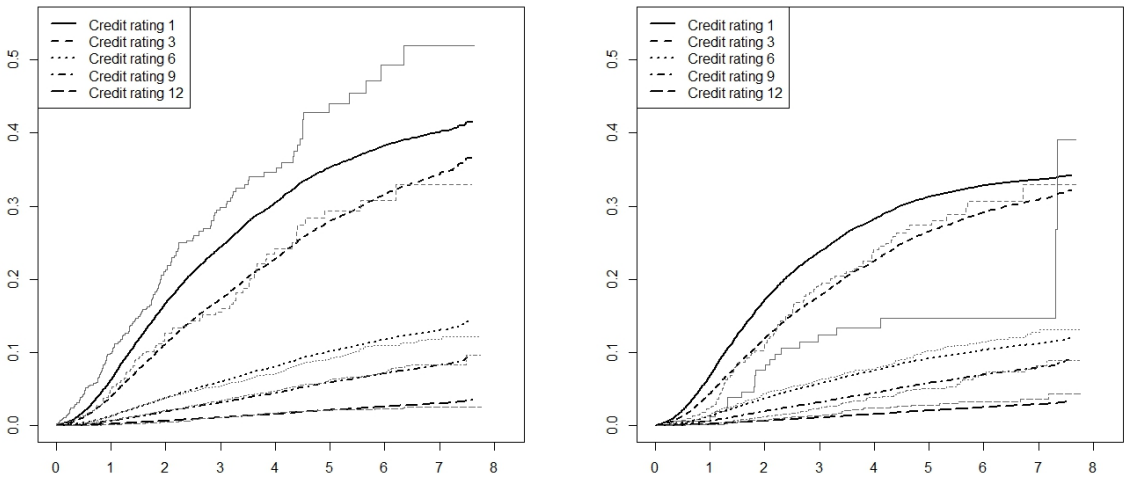
	Incidence (α)		Latency (β)	
	$\widehat{\alpha}_j$	SE($\widehat{\alpha}_j$)	$\widehat{\beta}_j$	SE($\widehat{\beta}_j$)
Intercept	-0.018	0.354		
Credit Rating				
1				
2	-0.562	0.507	0.034	0.357
3	-0.056	0.458	-0.396	0.312
4	-0.861	0.709	-0.112	0.497
5	-0.752	0.585	-0.435	0.389
6	-1.198	0.640	-0.796	0.495
7	-0.982	0.642	-1.074	0.415
8	-1.638	0.789	-0.694	0.585
9	-0.900	0.851	-1.743	0.583
10	-0.817	1.012	-2.204	0.744
11	-0.945	1.110	-2.046	0.763
12	-1.647	1.050	-2.217	0.833
13	-2.907	1.109	-1.151	0.871
14	-3.541	1.174	-1.628	0.976
15	-1.195	1.180	-3.046	0.889
16	-2.027	1.033	-2.089	1.734
17	-2.913	1.112	-2.382	3.681
18	-1.697	1.239	-3.426	5.831
19	-4.407	1.207	1.017	4.140
Product Class				
Corporate Lending				
Credit Guarantees	-0.058	0.633	0.578	0.340
Credit Facilities	-0.581	0.504	0.500	0.345
Other	-0.998	0.878	0.572	0.628

Table 4.1: The table presents the estimated parameters of the semi-parametric mixture cure model. The parameters are estimated by algorithm 1. The standard errors are computed by the bootstrap procedure. In accordance with the linear model in (3.3)*, the coefficients of Corporate Lending and Credit Rating 1 are not estimated. Also, since the intercept of the latency is included as a scaling factor of the baseline survival function, the estimate is not directly available and is therefore omitted. Credit rating 20 and 21 are omitted since they have no defaults.*

estimator (2.25) is exhausted at large t when the censoring rate is low, resulting in larger jumps when a default event occurs. This can be see in figure 4.3a where the residual has a lot of variation between time points towards the end of the time horizon.

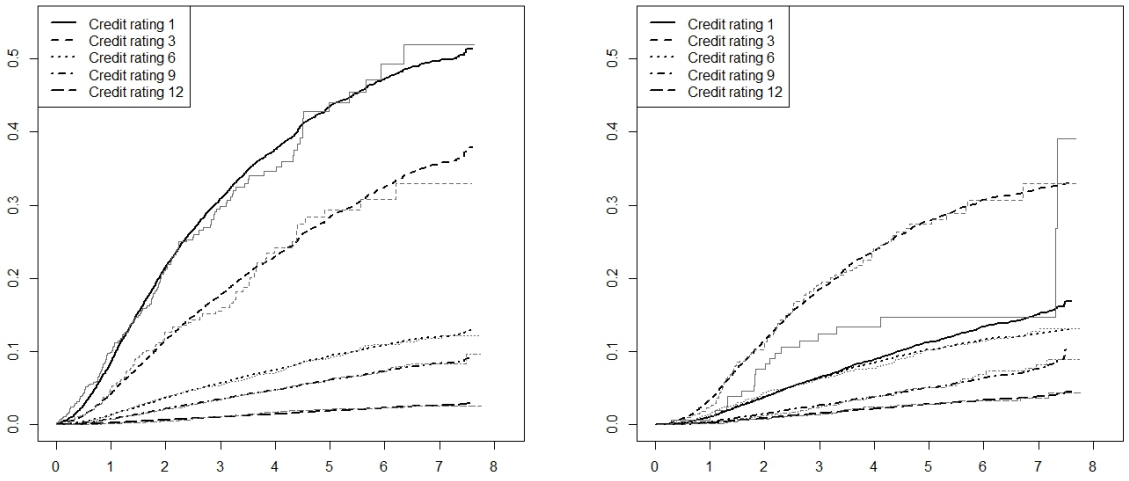
Furthermore, the overall fit for the Corporate Lending data tend to be superior than for Credit Facilities, as well as for Credit Guarantees and Other. This is illustrated in figure A.1 and A.2 in appendix A, in which the residuals for most credit ratings within the other product types tend to diverge from 0. The poor fit is either the result of a violation of the proportional hazards assumption or that the effect of the credit rating on either the incidence or the latency is not homogeneous across product types. The latter is effectively mitigated by extending the model and introducing dummy variables for the interaction between credit rating and product type. The model could also be estimated for the data of each product type separately. Nonetheless, the fit for the Corporate Lending data is superior to the other product types because it constitutes a larger proportion of the complete data, as depicted in table 3.1. Since the effect of credit rating is averaged across product types, the observations of type Corporate Lending will collectively

have greater influence on the maximum likelihood than the other products.



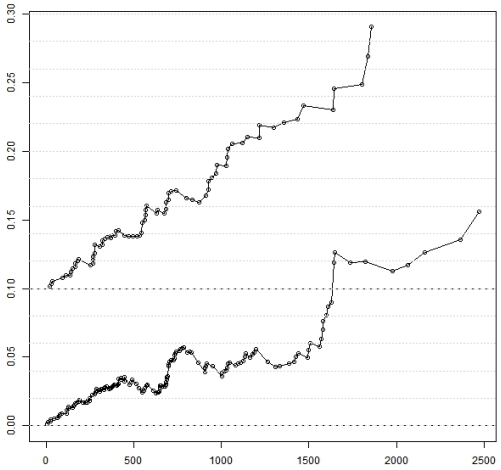
(a) Corporate Lending *(b) Credit Facilities*

Figure 4.1: The figure displays examples of the estimated term structures (in black) as well as the corresponding empirical term structures (in dim gray), for the semi-parametric mixture cure model applied to the full data set.

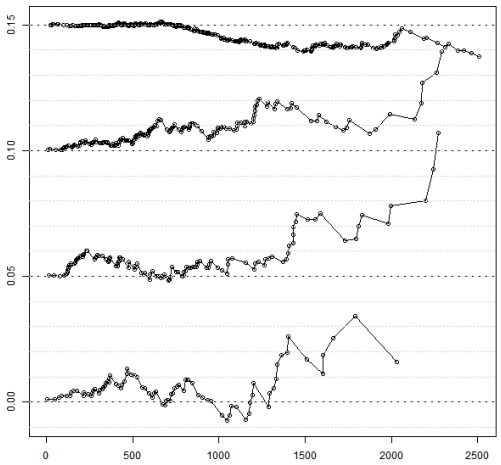


(a) Corporate Lending *(b) Credit Facilities*

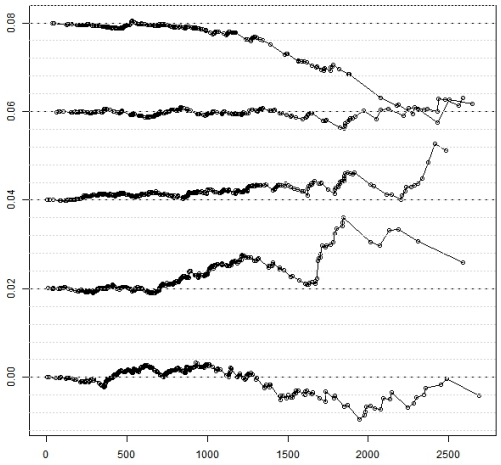
Figure 4.2: The figure displays examples of the estimated term structures (in black) as well as the corresponding empirical term structures (in dim gray), for the semi-parametric mixture cure model applied to the Corporate Lending data (left) and the Credit Facilities data (right) separately.



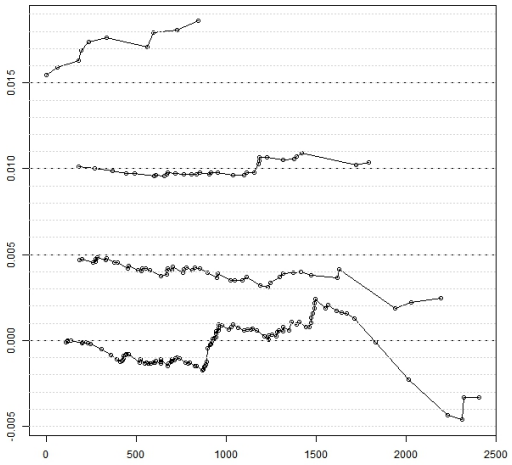
(a) Rating 1 (bottom) and 2 (top).



(b) Rating 3 (bottom), 4, 5, 6 (top).

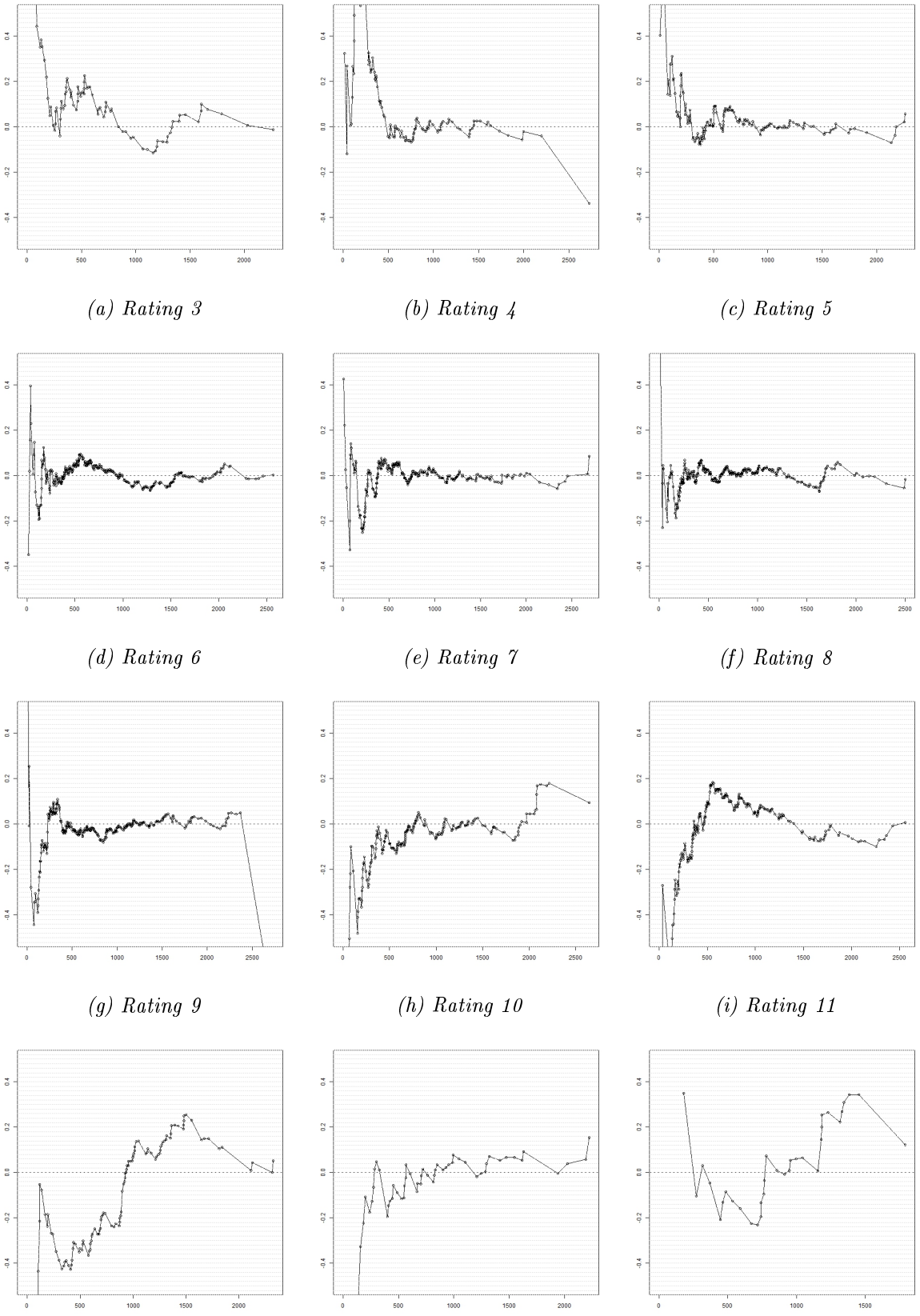


(c) Rating 7 (bottom), 8, 9, 10, 11 (top).



(d) Rating 12 (bottom), 13, 14, 15 (top).

Figure 4.3: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings of the Corporate Lending data for the semi-parametric model. The residual may also be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. The dashed line represents the shifted mean of the residuals of each rating.



(a) Rating 3 *(b) Rating 4* *(c) Rating 5*

(d) Rating 6 *(e) Rating 7* *(f) Rating 8*

(g) Rating 9 *(h) Rating 10* *(i) Rating 11*

(j) Rating 12 *(k) Rating 13* *(l) Rating 14*

Figure 4.4: Relative difference between estimated and empirical term structure for the semi-parametric model applied to the Corporate Lending data.

Applying the model to the data of each product type separately (excluding the dummy variable for product type), similar to the parametric model, yields the estimated term structures in figure 4.2. Contrary to term structures in figure 4.1, the fit of the first credit rating within both product types is remarkably improved. By the residual plots for all product types, found in figure A.3 together with figure A.4 and A.5 in appendix A, one can see that the residual fluctuates near zero for most cohorts. As anticipated, the results indicate a superior fit. Figure 4.4 displays the relative difference between the estimated and the empirical term structure for the Corporate Lending data. Although there is a considerable relative difference between the estimated and the empirical term structure for the first one or two year for all credit ratings, the residual plots indicate that the absolute difference for most credit ratings is negligible. On the contrary, residual plot A.3a illustrate a significant absolute difference for credit rating 1, 2 and 4. For these credit ratings, the term structure seem to be underestimated for small t. This is especially problematic when one wants to estimate the default probability of new applicants, e.g. the default probability over the first year. Nevertheless, subsequent to the first one or two years, the relative difference is for most credit ratings stable and varies with only a few percent. For a few credit ratings there exist spikes in the relative difference at some point in time, e.g. for credit rating 11, 12 and 14 the differences peak at approximately 20 percent.

2 Simulation study

In the simulation study, we simulated 400 independent data sets of observations from a Weibull mixture cure model in accordance with algorithm 3. First we let each data set be of size 600 (200 observations per artificial group). Then we doubled the size to 1200 (400 observations per artificial group), with the same parameter values. Table 4.2 presents the average, bias and standard error of the estimated parameters for the two sample sizes. The scale parameter was modeled by the exponentiated linear regression, analogous to (3.4), and the shape parameter was assumed constant for all groups. The proportion of susceptible individuals (incidence) was modeled by a logistic regression. The parameter values was chosen such that the default rate was decreasing, with the first group having the highest default rate and the third group having the lowest default rate. The third group represented the credit rating with extremely low default incidence. For size 600, the third group had no defaults in 35 of the 400 simulated data sets, whereas for size 1200 the corresponding figure was 1 out of 400. The lack of default events seemed to have great affect on the bias of the corresponding scale parameter, β_3. This is evident by the average and bias of the estimates when the number of defaults within group 3, d_3, is equal to 0, less than or equal to 3 or larger than 3. For the 35 simulated data sets without defaults in group 3, the bias of β_3 was 7.443. There was no clear additional bias for the estimates of the other parameters, except for the shape parameter which seemed to be slightly biased. For more than 3 defaults in the low default rate group, the bias of the shape parameter had basically reduced to zero.

Presumably the most important reason for the significant biases of α and β is that the incidence and the latency counteracts each other to some extent. The incidence determines the level of the plateau, by scaling the term structure. This means that if the observational window is not long enough to reach the point where the term structure plateaus, a higher incidence level may be compensated by a longer latency. Here, the latency is in first hand determined by the scale parameter. For the first group with scale parameter $\theta = \exp(7)$, the conditional

(on being susceptible) mean time until default is approximately 1000 days. This is relatively small in comparison to the observational window which is close to 3000 days. The bias of all parameters for the first group is also small, particularly for the scale parameter. For the second and third group, the corresponding means are approximately 3000 and 5000, respectively. For these groups, we record large biases in both the incidence as well as the scale parameters.

Despite significant biases, we do not reject that the estimated term structure fits closely with the theoretical over the time horizon for which we have observed the times until default. However, beyond the horizon of the observational window the fit of the estimated term structure (or the survival function) will be poor since it plateaus at an erroneous level of non-susceptible individuals. Consequently, the length of the term structure can only be effectively measured for a time horizon as long as the observational window.

	Incidence (α)			Scale (β)			Shape (γ)
	$\alpha_0 = 0$	$\alpha_2 = -1$	$\alpha_3 = -2$	$\beta_0 = 7$	$\beta_2 = 1$	$\beta_3 = 1.5$	$\gamma = 0.5$
$n = 200$							
Avg.	0.099	0.084	-0.147	7.004	1.328	3.008	0.522
Bias	0.099	1.084	1.853	0.004	0.328	1.508	0.022
S.E.	0.635	1.503	1.741	0.261	0.599	5.160	0.143
$d_3^j = 0$							
Avg.	0.138	-0.135	-0.234	6.986	1.307	8.943	0.561
Bias	0.138	0.865	1.766	-0.014	0.307	7.443	0.061
$d_3^j \leq 3$							
Avg.	0.088	0.062	-0.147	6.992	1.326	3.338	0.529
Bias	0.088	1.062	1.853	-0.008	0.326	1.838	0.029
$d_3^j > 3$							
Avg.	0.176	-0.062	-0.188	7.067	1.190	1.985	0.504
Bias	0.176	0.938	1.812	0.067	0.190	0.485	0.004
$n = 400$							
Avg.	0.066	0.152	-0.039	7.015	1.355	2.310	0.500
Bias	0.066	1.152	1.961	0.015	0.355	0.810	0.000
S.E.	0.483	1.442	1.669	0.201	0.523	0.813	0.091

Table 4.2: The table presents the results of the simulation study for different sample sizes. The average, bias and standard error of the estimated parameters is computed from 400 simulated data sets, each of size $3n$, where n is the number of observations within each of the three subpopulations. The table also shows the average and bias of all estimates where the number of failures d_3^j within group 3 of the jth simulated data set is equal to 0, less than or equal to 3 or more then 3 (for $n = 200$). The ratio of simulated data sets in which there are no failures in group 3 is 35/400 for $n = 200$ and 1/400 for $n = 400$.

3 Parametric: Weibull

The parametric specification of the model is applied on each product type separately. In this section, we will primarily focus on the Corporate Lending data since it constitutes the largest proportion of the data. The parameter estimates as well the residual for Corporate Lending are presented in table 4.3 and figure 4.6. The corresponding tables and figures for Credit Guarantees, Credit Facilities and Other are presented in appendix A. Additionally, an illustration of the estimated term structures for Corporate Lending and Credit Facilities is found in figure 4.5. By comparing them to the estimated term structure of the semi-parametric model applied to Corporate Lending and Credit Facilities separately, displayed in figure 4.2, it is noticeable that the term structures are similar in shape.

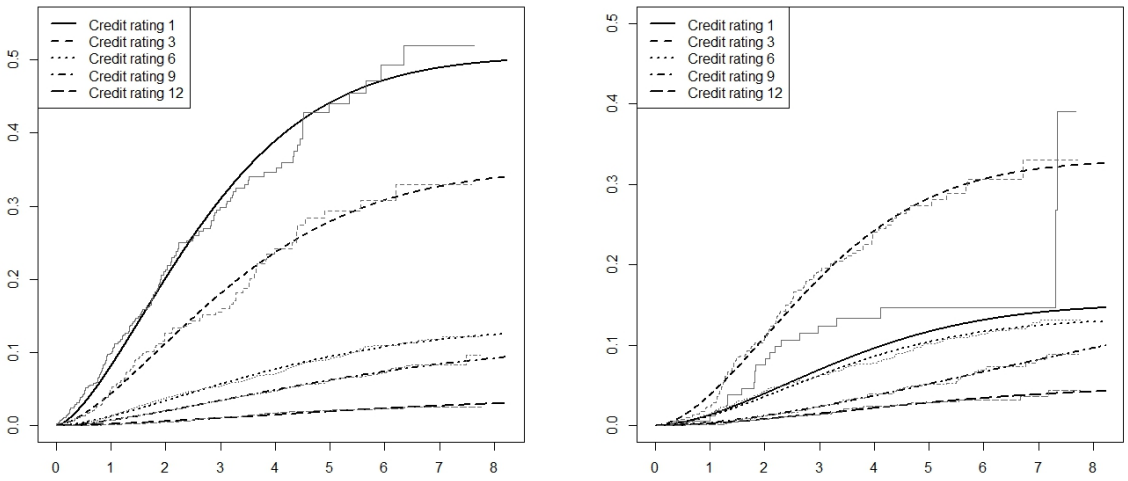
	Incidence (α)		Scale (β)	
	$\widehat{\alpha}_j$	SE($\widehat{\alpha}_j$)	$\widehat{\beta}_j$	SE($\widehat{\beta}_j$)
Intercept	0.019	0.190	7.033	0.100
Credit Rating				
1				
2	-0.283	0.312	-0,069	0.172
3	-0.628	0.305	0.182	0.179
4	-0.908	0.278	0.233	0.169
5	-1.078	0.260	0.259	0.159
6	-1.860	0.214	0.375	0.129
7	-1.771	0.245	0.481	0.154
8	-1.909	0.245	0.487	0.155
9	-2.044	0.264	0.653	0.172
10	-2.011	0.483	1.057	0.319
11	-2.903	0.244	0.487	0.163
12	-3.134	0.405	0.803	0.300
13	-3.519	0.527	0.744	0.407
14	-4.850	0.350	0.344	0.290
15	-4.652	0.423	-0.497	0.344
16	1.436	-	3.650	-
17	-0.041	-	3.669	-
18	-0.292	-	3.094	-
19	-4.385	1.088	-0.332	0.878

Shape (γ)	$\widehat{\gamma}$	SE($\widehat{\gamma}$)
	0.427	0.019

Table 4.3: The table presents the results of the parametric model applied to the Corporate Lending data. The parameters are estimated by algorithm 2. The standard errors are computed from the observed information matrix. Some diagonal elements of the inverted observation matrix are negative, yielding complex values standard errors. These values have been omitted. In accordance with the specification of the model, the coefficient for Credit Rating 1 is not estimated. Credit rating 20 and 21 are omitted since they have no defaults.

Similarly as for the semi-parametric model, the estimated parameters in table 4.3 indicates a clear tendency for the incidence to be decreasing with higher credit rating for product type Corporate Lending. Unexpectedly, the scale parameter tends to increase until it peaks at credit rating 10, from where it declines. Due to the relatively high standard error, this is likely to be a spurious deviancy.

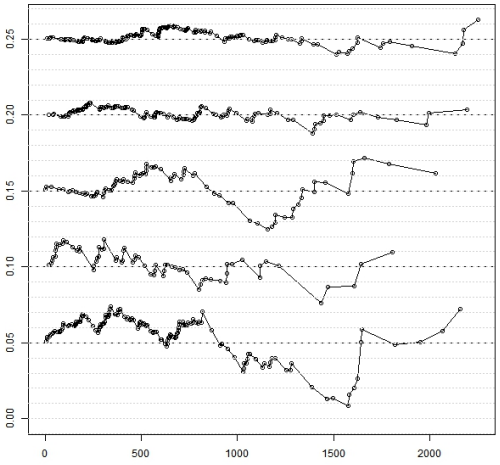
The standard errors are estimated by the observed information matrix. For some estimates of the parameters of a credit rating with few defaults, the corresponding diagonal element of the inverted observation information matrix was negative. Since the variance cannot be negative, this result must be erroneous. Instead, we believe the result to be a consequence of the high unreliability of the estimate. Furthermore, figure A.1 and A.3 reveals that the standard errors of the incidence parameter tend to soar when the estimate is large. This result is on the other hand comprehensible. The explanation is that for large α, the derivative of the incidence (determined by the derivative of the logistic function) is small, i.e. for a sizable increase in α the increased probability of being susceptible is negligible. In fact, the standard deviation of the parameters of the incidence level is relatively high. In addition to the results of the simulation study, the finding is that the incidence is much more difficult to estimate than the latency. Table 4.3, and table A.2, A.1 and A.3 in appendix A, reveals no clear tendency for the standard error to be increasing with the default rate. The increase in the standard error for the incidence parameters of higher credit rating is likely to partially be the result of the decreasing derivative of the logistic function as the argument increases, as discussed previously. The considerable increase of the standard errors for credit ratings where the number of defaults is low indicates that the reliability of the estimates is sensitive to the number of defaults rather than the default rate. This result is important, since it allows one to effectively estimate the term structure of default probabilities for portfolios with low default rate as long as there are sufficiently many default events (or equivalently, if the size of the portfolio is sufficiently large).



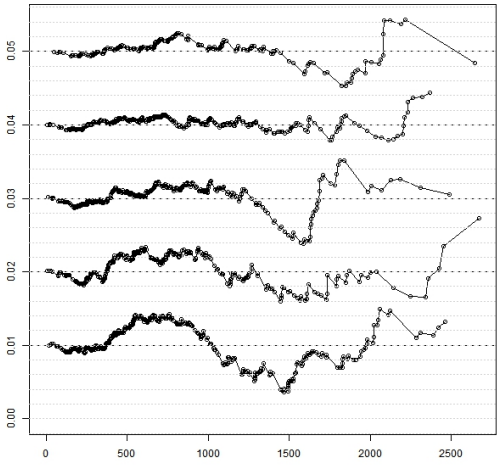
(a) Corporate Lending *(b) Credit Facilities*

Figure 4.5: The figure displays examples of the estimated term structures (in black) as well as the corresponding empirical term structures (in dim gray), for the parametric mixture cure model applied to the Corporate Lending data (left) and the Credit Facilities data (right) separately.

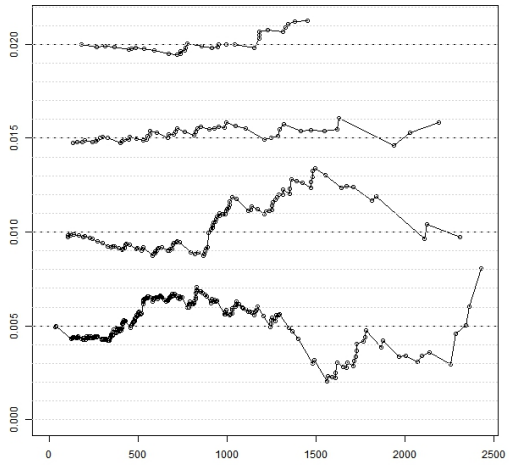
The similar shape of the estimated parametric and semi-parametric (when applied to each product type separately) term structure for the Corporate Lending data is further illustrated by the residuals in figure 4.6 together with the relative differences in figure 4.7. Similarly as for the semi-parametric model, the parametric model tend to underestimate the probabilities at the early stages for low credit ratings, and overestimate the probabilities for higher credit ratings. We do however emphasize that, once again with the exception of credit rating 1, 2 and 4, the difference in absolute terms is in most cases negligible. In accordance with the residuals in figure 4.6, the term structure of credit rating 1 and 2 is underestimated by approximately 2 percentage units at the most over the first two years. The corresponding number for credit rating 4 is close to 1 percentage unit. Despite the significant underestimation of term structure relative to the empirical term structure over the first two years, this could potentially be a result of a surge in default events during the financial crisis in 2008. Considering the observational window over which the data is sampled begins in 2007, the abundance of defaults can only appear in approximately the first two years. Although the events will influence the empirical term structure beyond the two year horizon as well, the parametric model is not as flexible for small t, resulting in a poor fit.



(a) Rating 1 (bottom), 2, 3, 4, 5 (top).

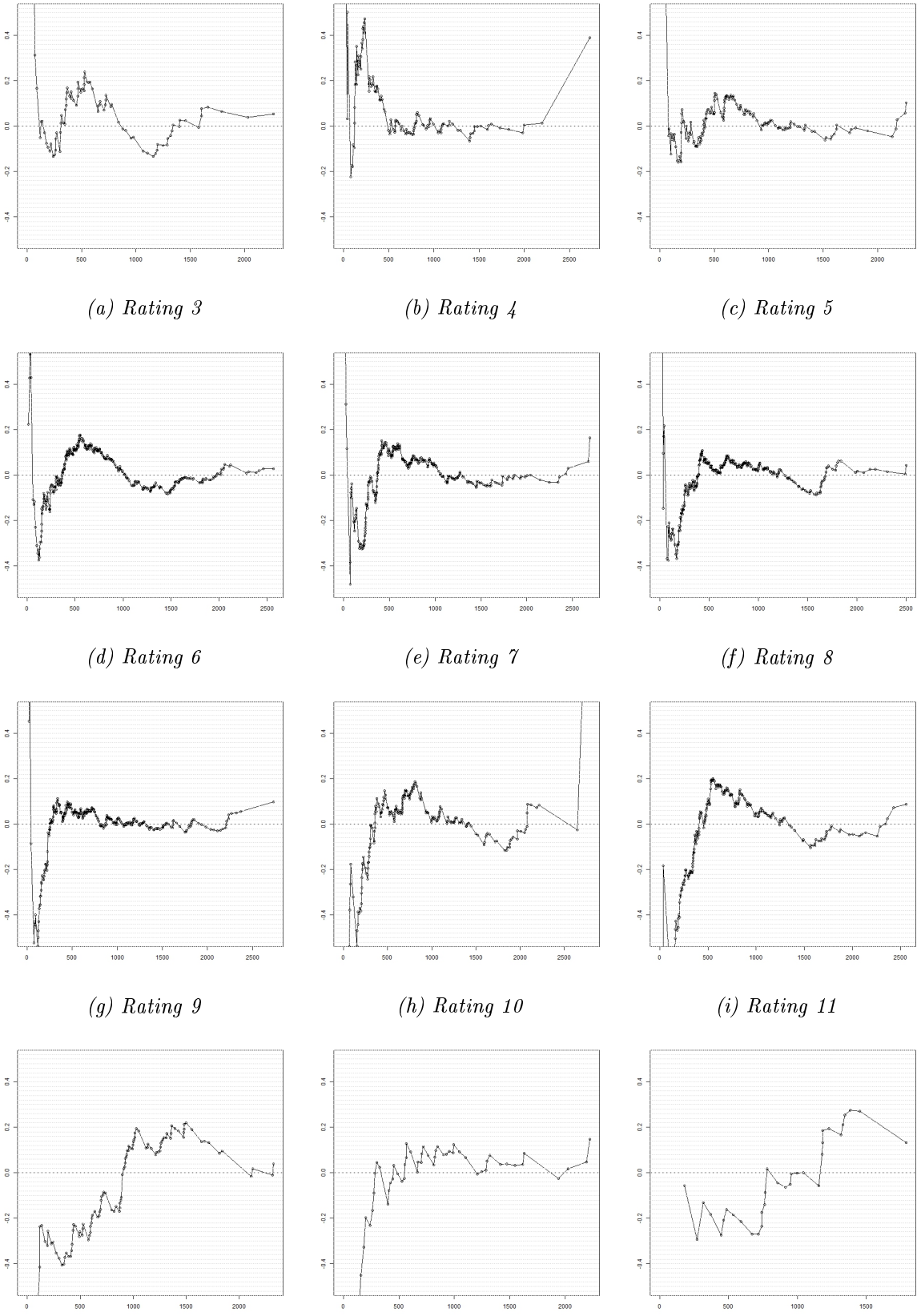


(b) Rating 6 (bottom), 7, 8, 9, 10 (top).



(c) Rating 11 (bottom), 12, 13, 14 (top).

Figure 4.6: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings for the Corporate Lending data. Equivalently, it may be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. Residuals for credit ratings with less than 10 defaults are omitted. The dashed line represents the shifted mean of the residuals of each rating.



(a) Rating 3 *(b) Rating 4* *(c) Rating 5*

(d) Rating 6 *(e) Rating 7* *(f) Rating 8*

(g) Rating 9 *(h) Rating 10* *(i) Rating 11*

(j) Rating 12 *(k) Rating 13* *(l) Rating 14*

Figure 4.7: Relative difference between estimated and empirical term structure for the parametric model applied to the Corporate Lending data.

Chapter 5

Conclusions

In accordance with the forthcoming regulatory changes with the introduction of IFRS 9, financial institutions will be required to estimate expected credit losses of financial contracts over their entire lifetime. Thus necessitating a term structure of default probabilities. In this thesis we estimated the term structure for heterogeneous credit portfolios. The heterogeneity was assumed to be fully captured by the credit rating and the credit type. We applied a mixture cure model to manage long-term survivors and plateaued survival functions of low default portfolios. Both a parametric as well as semi-parametric mixture cure model was investigated. In the parametric specification of the model, the latency was modeled by a Weibull distribution, whereas in the semi-parametric mixture cure model we utilized the proportional hazards model with a non-parametric baseline survival function. For both models, the incidence was modeled by a logistic link function. Due to the general intractability of the maximum likelihood of mixture models, the parameters were estimated by the EM algorithm. Additionally, a simulation study was conducted to asses the accuracy of the EM algorithm applied to the parametric mixture model with data characterized by a low default incidence.

The semi-parametric mixture cure model was initially applied to the full data set, with dummy variables for both the credit rating and the product type as covariates. The residual plots in figure 4.3, A.1 and A.2 indicated that the effect of the credit was not homogeneous across product types, resulting in a poor fit with the empirical term structures. Instead the model was applied separately to each data subset constituting entirely of Corporate Lending, Credit Facilities or Credit Guarantees. The updated residuals in A.3, A.4 and A.5 as well as the plot of relative differences in figure 4.4 demonstrated a remarkable improvement of the overall fit. The similarities of the estimated residuals as well as the relative differences for the parametric and semi-parametric mixture cure models indicated a strong resemblance between the shape of the estimated term structures.

Although the simulation study revealed that the estimated parameters of the parametric model experienced heavy bias for few default events (particularly for no defaults), it did not seem particularly sensitive to the censoring rate. This result is important, since it justifies the application of the model on credit data with low default rates as long as the sample is sufficiently large, i.e. containing sufficiently many defaults. The simulation study did also find difficulties in correctly estimating the incidence as well as the latency when the observational window was not sufficiently long relative to the conditional (on being susceptible) mean longevity. This was presumed not to affect the fit of term structure over the observational window. Therefore, the term

structure can only be effectively measured over a horizon with the length of the observational window at the most.

Irregardless of what model we applied, we found difficulties in capturing the dynamics of the default probability over the first one to two years. By the residual plots of the parametric and semi-parametric model applied on the Corporate Lending data, in figure 4.6 and A.3, the absolute difference between the estimated and empirical term structure surpasses 2 percentage units within a one year horizon for some credit ratings. This is of course problematic, especially considering that the one year probability of default is commonly used within the context of risk analysis. In accordance with the plots of the relative differences for both models, displayed in figure 4.7 and 4.4, the estimated term structures differs for most credit ratings by only a few percent to the empirical term structures at longer horizons. This indicate a more reliable estimate at longer horizons.

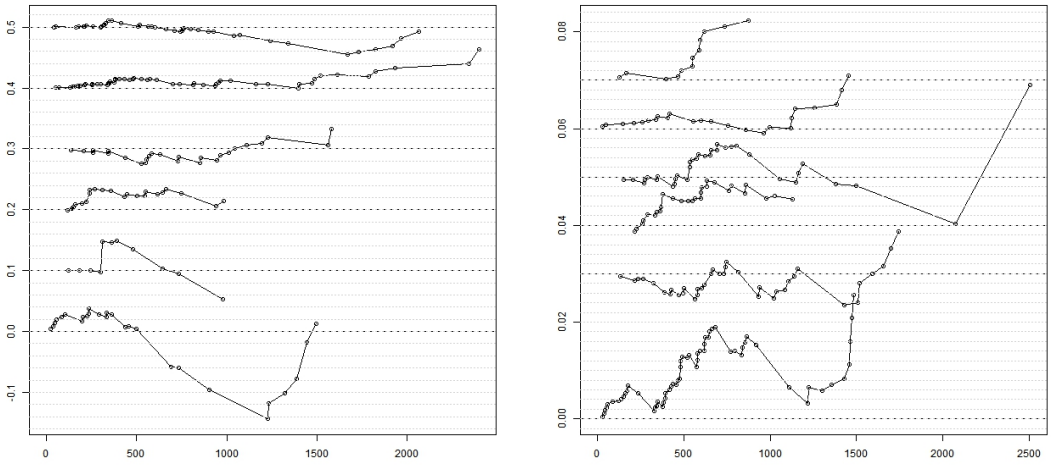
In conclusion, since the effect of the credit rating is not homogeneous across product types, the mixture cure model is favorably applied to each product type separately. Then the estimated term structures of the parametric and the semi-parametric models had strong resemblance. The model had difficulties capturing the dynamics of the term structure over the first one to two years. Beyond that the model fit well to the empirical term structures. Moreover, without being able to rigorously include truncated data, the term structure can at best be effectively estimated over a time horizon with the length of the observational window. Thus requiring data to be gathered over a sufficiently long time period. Additionally, neither model produce unbiased estimates of the term structure for credit ratings without default events. The parametric model was however not sensitive to the censor rate, therefore the term structure can be estimated as long as the data include sufficiently many default events. Moreover, the model seem particularly applicable to corporate loan data, although it had decent performance on other credit types as well.

The lack of default events can be mitigated if one does not consider predefined credit ratings, and instead use characteristics of each credit as the explanatory variables. If some covariate is not associated with a default, it can be regarded as if it affects neither the latency nor the incidence. This would result in an individual term structure of each credit. It would on the other hand be problematic to validate the model. Generally, the fit of the mixture cure model is assessed graphically by comparison of the empirical estimates of the survival function. For continuous variables or cohorts of few defaults this is rendered impossible. Therefore, methods to efficiently validate the model is needed. This could potentially form the basis of future research. Furthermore, this study has not considered events in calendar time. It is reasonable to assume that credit events surge during financial crises, e.g. the financial crisis in 2008. Events in calendar time are diluted over the definition of time in our model. Instead it would be relevant for future research to consider multiple time scales, whereof one represents calendar time. Subsequently the macroeconomic climate can be represented by either frailty variables or time dependent covariates. Frailty variables can also be used to introduce a dependence structure for the timing of default between credits. Investigating other types of dependence structures would also serve as a valid topic for future research.

Appendix A

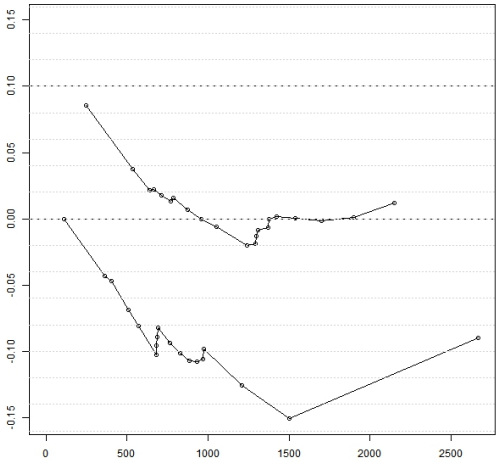
Figures and tables

1 Semi-parametric

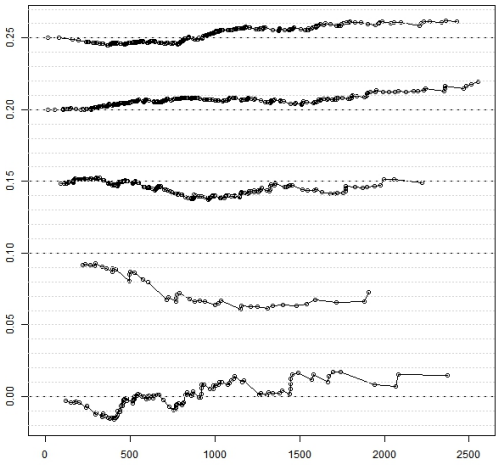


(a) Rating 1 (bottom), 3, 4, 5, 6, 7 (top). *(b) Rating 8 (bottom), 9, 10, 11, 12, 13 (top).*

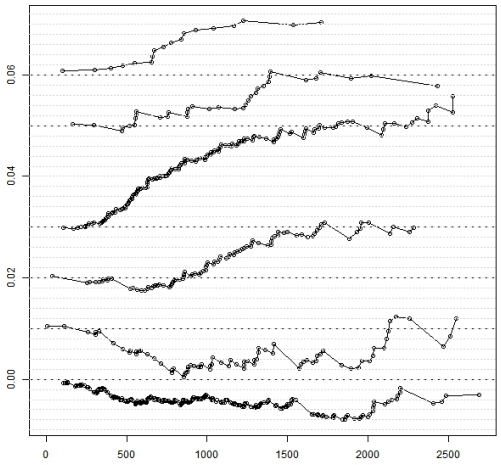
Figure A.1: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings of the Credit Guarantees data. The residual may also be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. The dashed line represents the shifted mean of the residuals of each rating.



(a) Rating 3 (bottom), 4, 5, 6, 7 (top).



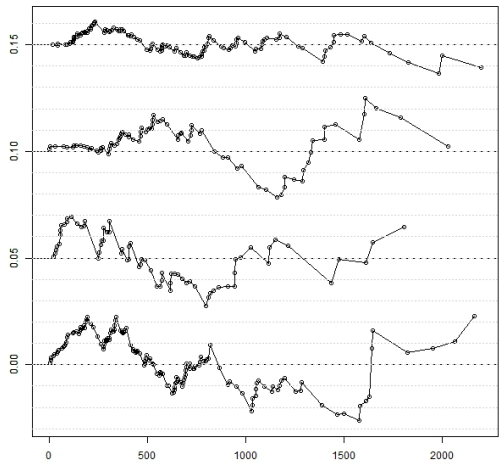
(b) Rating 3 (bottom), 4, 5, 6, 7 (top).



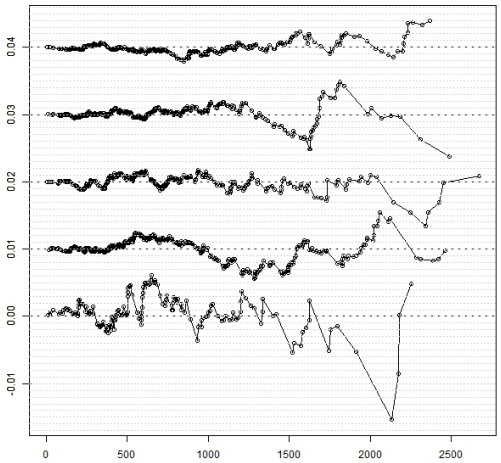
(c) Rating 8 (bottom), 9, 10, 11, 12, 13 (top).

Figure A.2: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings of the Credit Facilities data. The residual may also be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. The dashed line represents the shifted mean of the residuals of each rating.

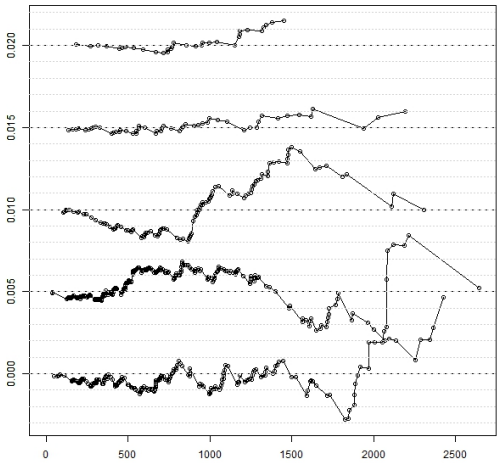
Corporate Lending



(a) Rating 1 (bottom), 2, 3, 4 (top).



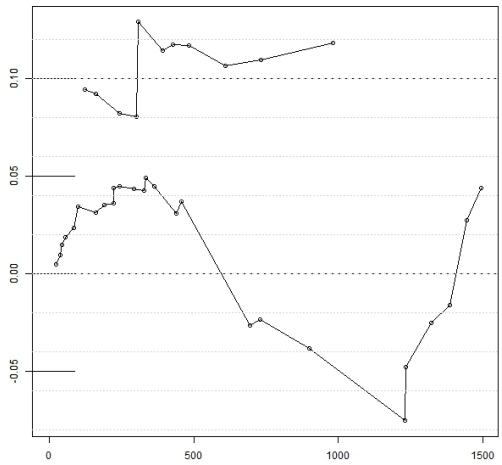
(b) Rating 5 (bottom), 6, 7, 8, 9 (top).



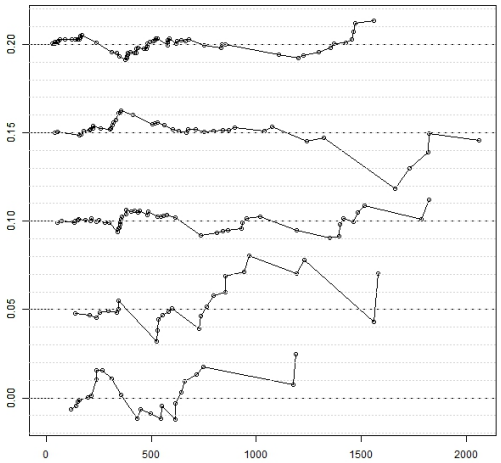
(c) Rating 10 (bottom), 11, 12, 13, 14 (top).

Figure A.3: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings of the semi-parametric model applied exclusively on the Corporate Lending data. The residual may also be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. The dashed line represents the shifted mean of the residuals of each rating.

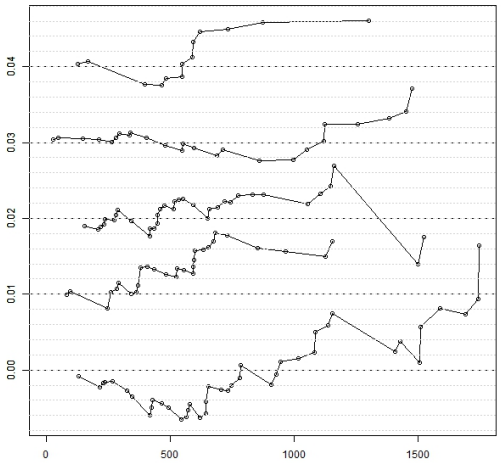
Credit Guarantees



(a) Rating 1 (bottom) and 3 (top).



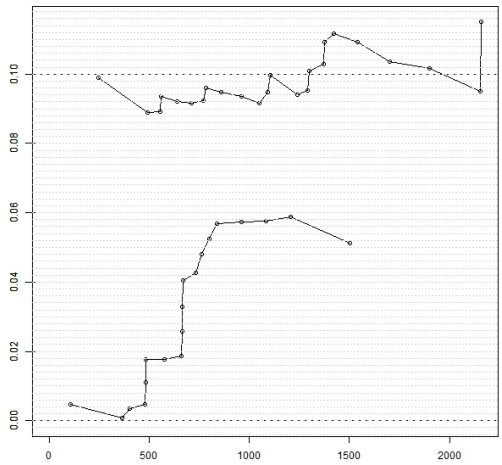
(b) Rating 4 (bottom), 5, 6, 7, 8 (top).



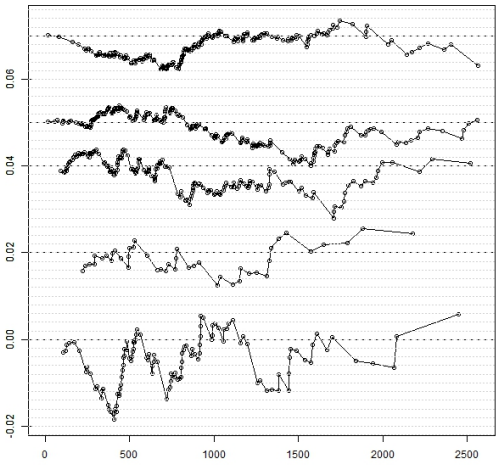
(c) Rating 9 (bottom), 10, 11, 12, 13 (top).

Figure A.4: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings of the semi-parametric model applied exclusively on the Credit Guarantees data. The residual may also be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. The dashed line represents the shifted mean of the residuals of each rating.

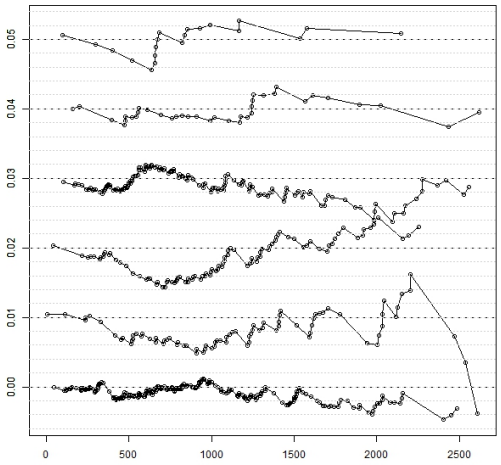
Credit Facilities



(a) Rating 1 (bottom) and 2 (top).



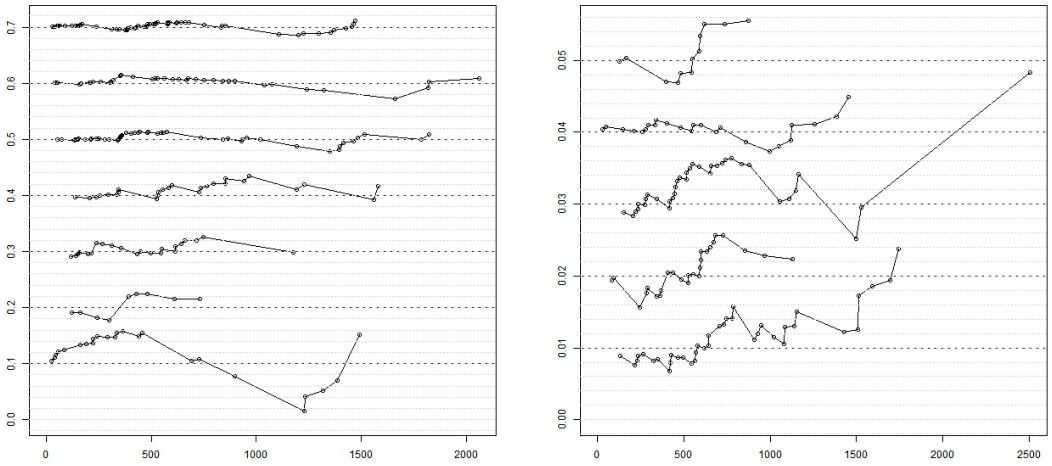
(b) Rating 3 (bottom), 4, 5, 6, 7 (top).



(c) Rating 8 (bottom), 9, 10, 11, 12, 13 (top).

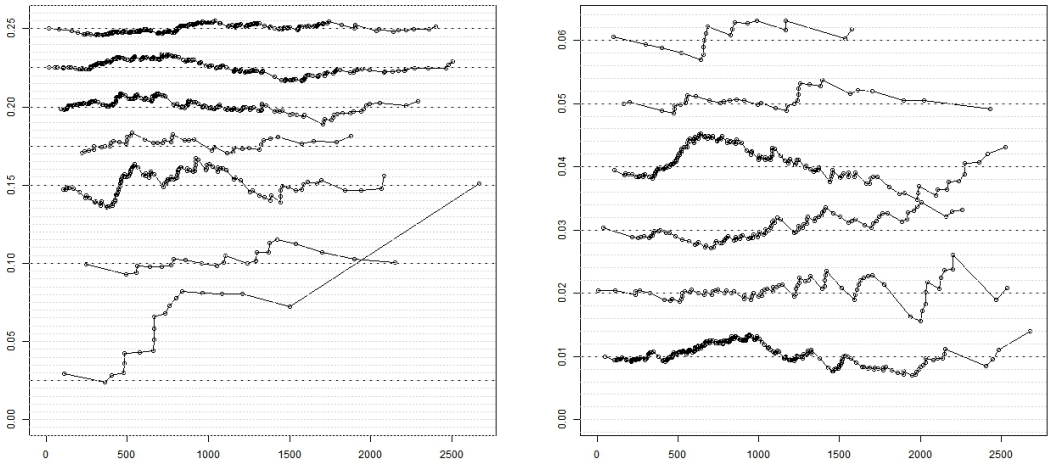
Figure A.5: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings of the semi-parametric model applied exclusively on the Credit Facilities data. The residual may also be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. The dashed line represents the shifted mean of the residuals of each rating.

2 Parametric



(a) Rating 1 (bottom), 3, 4, 5, 6, 7, 8 (top). *(b) Rating 9 (bottom), 10, 11, 12, 13 (top).*

Figure A.6: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings for the Credit Guarantees data. Equivalently, it may be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. Residuals for credit ratings with less than 10 defaults are omitted. The dashed line represents the shifted mean of the residuals of each rating.



(a) Rating 1 (bottom), 2, 3, 4, 5, 6, 7 (top). *(b) Rating 8 (bottom), 9, 10, 11, 12, 13 (top).*

Figure A.7: The figures displays the difference between the estimated and the empirical survival functions for different credit ratings for the Credit Facilities data. Equivalently, it may be interpreted as the difference between the empirical and estimated distribution functions. The residual is only plotted at times t (days, vertical axis) where a default has occurred. Residuals for credit ratings with less than 10 defaults are omitted. The dashed line represents the shifted mean of the residuals of each rating.

	Incidence (α)		Scale (β)	
	$\widehat{\alpha}_j$	SE($\widehat{\alpha}_j$)	$\widehat{\beta}_j$	SE($\widehat{\beta}_j$)
Intercept	0.346	0.942	7.348	0.359
Credit Rating				
1	-	-	-	-
2	-2.195	1.101	-0.620	0.554
3	-1.850	0.993	-1.041	0.419
4	-1.674	0.976	-0.843	0.399
5	-0.500	1.469	0.304	0.612
6	-1.674	0.975	-0.136	0.401
7	-1.817	1.025	0.113	0.455
8	-2.365	0.958	-0.452	0.383
9	-1.908	1.153	0.398	0.571
10	-3.454	0.966	-0.688	0.400
11	-2.323	1.151	0.323	0.587
12	-3.429	1.028	-0.070	0.497
13	-3.622	1.059	-0.323	0.536
14	1.591	68.097	2.572	5.516
15	-4.385	1.387	-0.184	0.955
16	5.586	101.683	2.107	0.746
17	5.255	74.346	1.411	0.747

Shape (γ)	$\widehat{\gamma}$	SE($\widehat{\gamma}$)
	0.454	0.044

Table A.1: The table presents the results of the parametric model applied to the Credit Guarantees data. The parameters are estimated by algorithm 2. The standard errors are computed from the observed information matrix. In accordance with the specification of the model, the coefficient for Credit Rating 1 is not estimated. Credit rating 18, 19, 20 and 21 are omitted since they have no defaults.

	Incidence (α)		Scale (β)	
	$\widehat{\alpha}_j$	SE($\widehat{\alpha}_j$)	$\widehat{\beta}_j$	SE($\widehat{\beta}_j$)
Intercept	-1.724	0.319	7.281	0.241
Credit Rating				
1	-	-	-	-
2	0.743	0.985	0.558	0.584
3	0.122	0.353	-0.162	0.258
4	0.018	0.383	-0.097	0.288
5	0.272	0.336	-0.133	0.252
6	-0.145	0.332	-0.013	0.251
7	-0.151	0.338	0.058	0.255
8	-0.710	0.333	0.046	0.252
9	0.244	0.839	0.878	0.508
10	-0.614	0.384	0.305	0.293
11	-0.500	0.346	0.122	0.263
12	-1.220	0.474	0.326	0.370
13	-1.801	0.404	-0.225	0.307
14	7.017	-	3.204	-
15	-3.008	0.844	-0.237	0.659
16	3.246	-	1.930	-

Shape (γ)	$\widehat{\gamma}$	SE($\widehat{\gamma}$)
	0.540	0.026

Table A.2: The table presents the results of the parametric model applied to the Credit Facilities data. The parameters are estimated by algorithm 2. The standard errors are computed from the observed information matrix. Some diagonal elements of the inverted observation matrix are negative, yielding complex values standard errors. These values have been omitted. In accordance with the specification of the model, the coefficient for Credit Rating 1 is not estimated. Credit rating 17, 18, 19, 20 and 21 are omitted since they have no defaults.

	Incidence (α)		Scale (β)	
	$\widehat{\alpha}_j$	SE($\widehat{\alpha}_j$)	$\widehat{\beta}_j$	SE($\widehat{\beta}_j$)
Intercept	-4.346	0.597	5.190	0.397
Credit Rating				
1	-	-	-	-
2	3.161	0.916	1.522	0.549
3	4.591	3.166	2.586	1.115
4	2.432	0.801	1.230	0.510
5	1.750	0.824	1.613	0.549
6	1.278	0.721	1.546	0.480
7	2.160	0.834	2.069	0.564
8	0.278	1.003	1.832	0.717
9	-0.880	0.989	1.547	0.689
10	7.801	58.353	4.700	1.171
11	12.393	71.317	3.680	0.418
12	1.564	10.172	3.371	5.945
13	6.984	38.503	5.247	1.652
14	8.406	106.534	5.039	1.269

Shape (γ)	$\widehat{\gamma}$	SE($\widehat{\gamma}$)
	0.554	0.075

Table A.3: The table presents the results of the parametric model applied to the data of product class Other. The parameters are estimated by algorithm 2. The standard errors are computed from the observed information matrix. In accordance with the specification of the model, the coefficient for Credit Rating 1 is not estimated. Credit rating 15 through 21 are omitted since they have no defaults.

Bibliography

Basanik, J., Crook, J., & Thomas, L. (1999). Not If But When Will Borrowers Default. *Journal of the Operational Research Society*, *50*, 1185–1190.

Basford, K. E., Greenway, D. R., McLachlan, G. J., & Peel, D. (1997). Standard Errors of Fitted Means under Normal Mixture Models. *Computational Statistics*, *12*, 1–17.

Bellotti, T., & Crook, J. (2009). Credit Scoring with Macroeconomic Variables Using Survival Analysis. *Journal of the Operational Research Society*, *60*, 1699–1707.

Boag, J. W. (1949). Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society*, *11*, 15–53.

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.

Carling, K., Jacobson, T., Linde, J., & Roszbach, K. (2007). Corporate Credit Risk and the Macroeconomy. *Journal of Banking and Finance*, *31*, 845–868.

Chen, C. H., & Kuk, A. Y. C. (1992). A Mixture Model Combining Logistic Regression with Proportional Hazards Regression. *Biometrika*, *79*, 531–541.

Chen, C. H., Tsay, Y. C., Wu, Y. C., & Horng, C. F. (2013). Logistic-AFT Location-Scale Mixture Regression Mdoels with Nonsusceptibility for Left-truncated and General Interval-Censored Data. *Statistics in Medicine*, *32*, 4285–4205.

Cox, D. (1972). Regression Models and Life-Tables (with discussion). *Journal of the Royal Statistical Society*, *34*, 187–220.

Cox, D. (1975). Partial Likelihood. *Biometrika*, *62*, 269–276.

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*, *39*, 1–38.

Efron, B. (1977). Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, *72*, 557–565.

Farewell, V. T. (1982). The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors. *Biometrics*, *38*, 1041–1046.

Gormley, I. C., O'Hagan, A., & Murphy, T. M. (2012). Computational Aspects of Fitting Mixture Models via the Expectation-Maximization Alogrithm. *Computational Statistics and Data Analysis*, *56*, 3843–3864.

Grambsch, P. M., & Therneau, T. M. (2000). *Modeling Survival Data: Extending the Cox Model* (1st ed.). Springer-Verlag.

Heitjan, D. F., Li, Y., & Wileyto, E. P. (2012). Assessing the Fit of Parametric Cure Models. *Biostatistics*, *14*, 340–350.

IASB. (2014, July). *Project Summary: IFRS 9 Financial Instruments.*

Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *Journal of the American Statistical Association*, *53*, 457–481.

Klein, J., & Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). Springer-Verlag.

Lambert, P. C., Thompson, J. R., Weston, C. L., & Dickman, P. W. (2007). Estimating and Modeling the Cure Fraction in Population-Based Cancer Survival Analysis. *Biostatistics*, *8*, 576–594.

Lee, G., & Scott, C. (2010, September). *EM Algorithms for Multivariate Gaussian Mixture Models with Truncated and Censored Data.*

Louis, T. A. (1982). Finding the Observed Information Matrix when using the EM Algorithm. *Journal of Royal Statistical Society*, *44*, 226–233.

Malik, M., & Thomas, L. (2009). Modelling Credit Risk of Portolio of Consumer Loans. *Journal of the Operational Research Society*, *61*, 411–420.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models.* John Wiley and Sons, Inc.

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

Mues, C., Thomas, L. C., & Tong, E. N. C. (2012). Mixture Cure Models in Credit Scoring: If and When Borrowers Default. *European Journal of Operational Research*, *218*, 132–139.

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM. *Journal of the Royal Statistical Society*, *61*, 479–482.

Peng, Y., & Dear, K. B. G. (2000). A Nonparametric Mixture Model for Cure Rate Estimation. *Biometrics*, *56*, 237–243.

Segovia, L. (2014). *Survival data analysis with heavy-censoring and long-term survivors* (Unpublished doctoral dissertation). Univesitat Autonoma de Barcelona.

Sy, J. P., & Taylor, J. M. G. (2000). Estimation in a Cox Proportional Hazards Cure Model. *Biometrics*, *56*, 227–236.

Thomas, L. C., Tong, E. N. C., & Mues, C. (2012). Mixture Cure Models in Credit Scoring: If and when Borrowers Default. *European Journal of Operational Research*, *218*, 132–139.

Turnbull, B. W., & Waller, L. A. (1992). Probability Plots with Censored Data. *The American Statistician*, *46*, 5–12.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., & Lin, S. (2014, July). *Fitting Mixtures of Erlangs to Censored and Truncated Data Using the EM Algorithm.*