Multi-Period Credit Default Prediction - A Survival Analysis Approach

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2012

vorgelegt
von

Walter Orth

aus

Krefeld

Referent:	Prof. Dr. Karl Mosler
Korreferent:	Prof. Dr. Friedrich Schmid
Tag der Promotion:	14.12.2012

Acknowledgements

In July 2008, I began to work at the Chair for Statistics and Econometrics of the University of Cologne. Looking back, I am very happy that I was given the opportunity and decided to do so. The scientific work in the context of writing this thesis but also other aspects like the teaching of various statistics and econometrics courses really helped me learning a lot. At the same time, I had the luck to enjoy my work over all these years.

My first thank goes to Prof. Dr. Karl Mosler, my supervisor. I am very thankful to him for both supervising my doctoral thesis and for employing me on a full-time basis. I greatly appreciated that he was always there for me to discuss and to inspire my research. Importantly, he also provided me with the freedom to develop my own ideas. I also want to thank Prof. Dr. Friedrich Schmid for acting as the second supervisor. I am happy that this was possible even after his retirement in the beginning of 2012. Finally, I want to express my thanks for Prof. Dr. Thomas Hartmann-Wendels for joining the board of examiners and for his helpful comments during my thesis defense.

I also want to say a big "thank you" to my colleagues at the Chair for Statistics and Econometrics, Pavlo Bazovkin, Nana Dyckerhoff, Dr. Rainer Dyckerhoff, Prof. Dr. Gabriel Frahm, Andoni Ioannidis, Dominik Liebl, Christina Loley, Jun.-Prof. Dr. Hans Manner, Pavlo Mozharovskyi, Daniel Nowak, Yulia Polyakova, Dr. Christoph Scheicher, Dr. Frowin Schulz, Tobias Wickern and Dr. Christof Wiechers. I do not remember a single occasion where the atmosphere was not cooperative and friendly. Being in such an enjoyable working environment helped me a lot to go through the ups and

downs of my research.

Last but not least, I want to thank my wife Christina, my parents, Elisabeth Orth and Prof. Dr. Walter Orth, and my sister, Dr. Lisa Bechtold, for their continuous support. Christina, the wonderful time with you was and continues to be the perfect balance to the often abstract and technical work as a statistician. You and my family always encouraged me in my ambitions to write a doctoral thesis and I am very grateful to have you all on my side.

Walter Orth

Cologne, October 2012

Contents

1 Introduction **1**

2 Measuring predictive accuracy **7**

2.1 Discriminatory power 8

2.1.1 Accuracy Ratio and related measures 8

2.1.2 Harrell's C 14

2.2 Calibration 18

2.2.1 Nonparametric calibration analysis 18

2.2.2 Parametric calibration analysis 22

2.3 Validation techniques 24

2.4 Statistical inference 28

2.4.1 Single-cohort statistical inference 29

2.4.2 Multiple-cohort statistical inference 30

2.5 Empirical illustration 36

3 Default prediction with time-varying covariates **43**

3.1 Approaches with covariate forecasting models 45

3.2 An alternative approach 48

3.2.1 The models 48

3.2.2 Estimation 53

3.2.3 Extensions to mixture models 58
3.3 Empirical analysis 60
3.3.1 Data description and model specification 60
3.3.2 Estimation results 64
3.3.3 Evaluation of discriminatory power 67
3.3.4 Calibration analysis 73

4 Default prediction with given rating grades 79
4.1 The standard estimator 82
4.2 Confidence bound approaches 86
4.3 An empirical Bayes approach 88
4.4 Application to sovereign bonds 93
4.5 Simulation study 100
4.A Variance of the standard estimator 111
4.B Consistency of the empirical Bayes estimator 113
4.C R code for the empirical Bayes estimator 116

Bibliography 117

Chapter 1

Introduction

In recent years, the subprime crisis and the euro-zone crisis have highlighted the importance of credit risk assessments. In both cases, misperceptions of credit risk have arguably led to a severe financial and economic crisis. Not least due to these recent experiences, the development of methods to measure credit risk accurately is of considerable economic importance. This thesis contributes to the corresponding academic literature by dealing with the problem to estimate credit default probabilities under a flexible multi-period prediction horizon. Among other things, credit default probabilities often serve as the basis for credit rating assignments. Obligors are commonly classified on the basis of their default probabilities while sometimes the recovery rate - the estimated portion of the debt that the lender will recover if a default event occurs - is taken into account as well.[1] Consequently, the methods presented in this work are also helpful for accurate rating assignments.

Default probabilities and ratings have several important applications in the financial industry. Obvious applications concern the decisions whether and to which conditions (loan pricing, required collateral, maturity) an obligor might borrow from a lender. Further applications include loan loss reserve

[1] For instance, Standard & Poor's ratings are based only on the likelihood of default whereas Moody's classifies obligors according to expected losses which incorporates the recovery rate as well.

analysis, portfolio monitoring, internal capital allocation, profitability analysis and frequency of loan review (Basel Committee on Banking Supervision, 2000).[2] In the investment community, restrictions on investments below a certain rating grade are commonplace and rating distributions are usually reported to investors (Cantor et al., 2007). From the regulatory side, the Basel II Accord (and its successor named Basel III) have contributed to a considerably increased interest in default probability estimation. Within the so-called Internal Ratings-Based approach, banks assign one-year default probabilities to their internal rating grades which are then plugged into the regulatory formula for the capital requirements of a bank (Basel Committee on Banking Supervision, 2006, Part 2, Ch. III). Due to the novel regulatory intiative for insurance companies called Solvency II, default risk is likely to play an increasingly important role for insurance companies as well. Under Solvency II, regulatory capital requirements for insurers will be based, among other factors, on the default risks that arise from their investments and reinsurance treaties (Committee of European Insurance and Occupational Pensions Supervisors, 2009).

What distinguishes this work from the major part of the related literature is the multi-period prediction horizon that will be considered throughout. There is an obvious point for a multi-period approach since most loans have a maturity of multiple periods so that the lender faces a multi-period risk which should be adequately modeled. For instance, only 11.99% of loans of German banks at the end of 2011 are classified as short-term whereas the rest is classified as medium- and long-term lending (Deutsche Bundesbank, 2012, S. 30*). Numerous other examples like the standard 5-year maturity of Credit Default Swaps are possible. The popularity of single-period models seems thus not to be based on economic reasoning. Rather, it is arguably a convenient simplification that has become common practice. It is indeed true that predicting over multiple periods entails certain challenges that do not arise within a single-period view. Among the main contributions of this work to the literature is to show that there are relatively simple solutions to

[2]The calculation of expected returns can be seen as one kind of profitability or performance analysis. See Altman (1989) and section 4.4.

these challenges available.

From a regulatory point of view, the fact that under Basel II one-year default probabilities are the inputs to the formula for the capital requirements of banks might have contributed to the prevalent one-year (and often one-period) view. However, even within Basel II regulators state that "banks are expected to use a longer time horizon [than one year] in assigning ratings" (Basel Committee on Banking Supervision, 2006, § 414). Similarly, within the Standardised Approach of Basel II regulators assign risk weights to externally rated obligors based on the rating-grade specific three-year cumulative default rate (Basel Committee on Banking Supervision, 2006, Annex 2). More recently, plans have been developed by the International Accounting Standards Board and the Basel Committee on Banking Supervision to base loss provisions upon the expected loss over the whole life of the credit portfolio (Basel Committee on Banking Supervision, 2009). This would necessitate the estimation of multi-year default probabilities. More forward-looking provisions based on expected loss are part of regulatory efforts to reduce the procyclicality of capital requirements. Since multi-period predictions are less volatile and less sensitive to the business cycle than short-term predictions a multi-period approach would generally contribute to the reduction of procyclicality.

The methodological approach used throughout this thesis is survival analysis. The central variable within the survival analysis framework is the lifetime or time until default of an obligor. These lifetimes are often right-censored, i.e. one does not observe the end of each lifetime. As we will illustrate in detail in the upcoming chapters, the censoring problem gets more important as the prediction horizon grows. It is thus no coincidence that we select a survival analysis approach - where censoring is easily taken into account - for our multi-period prediction problem. Alternative approaches that only consider binary variables for default events neglect the timing of default events on the one hand and censored data on the other hand and thus do not use all relevant information. Further, survival analysis methods are naturally suitable for a flexible prediction horizon and thereby allow the estimation of a complete term structure of default probabilities. Such term structures are

not directly available in a simplified binary choice framework. The benefits from a complete term structure of default predictions are closely connected to the advantages of multi-period predictions just mentioned above. Banasik et al. (1999) give an extensive discussion in this respect and mention, among other things, the opportunities to conduct sophisticated analyses with respect to the profitability and the appropriate maturity of loans.

It is important to distinguish the topics of this work from other areas of credit risk research. First, we do not deal with the dependence of default events which is an important constituent of credit portfolio models. However, the results of this thesis are insofar relevant for credit portfolio models as default probabilities are usually important inputs to these models. Second, we do not cover the area of mathematical finance that uses risk-neutral probability measures to valuate and hedge credit-risk sensitive financial instruments. The models used in this strand of the literature share many similiarities (including the adoption of survival analysis methods) with the models considered in this work. A major difference is, however, that models based on the theory of risk-neutral valuation are calibrated to market prices whereas we use actual default events for the purpose of estimation. Note that models solely based on market data are only able to deliver so-called risk-neutral default probabilities which refer to the theoretical construct of a risk-neutral world. In contrast, this work is about real-world default probabilities. Finally, we do not aim at the development of models for the migration of obligors over rating grades. Although the analysis of rating migrations will affect our analysis at certain stages our focus is clearly on the transition to the default state.[3] Nevertheless, many of the ideas presented in this paper have a natural extension in the prediction of general rating transitions so that we see considerable potential for further research in this respect.

The thesis is structured as follows. In the next chapter, we deal with the question how predictive accuracy should be measured. We will divide the analysis into two parts, namely discriminatory power - concerning the ranking of obligors according to their default risk - and calibration, which is about

[3]We will discuss at the beginning of chapter 4 under which circumstances rating migration models can be helpful for default probability estimation.

the level of default probabilities. In the part about discrimination, we contribute twofold to the existing literature. First, we introduce a measure from biostatistics called Harrell's C to the credit risk area and propose a modified version for limited prediction horizons. Second, we derive methods to conduct statistical inference for Harrell's C (and other measures) under a sampling scheme that involves overlapping lifetimes. With respect to calibration, we present a new validation technique which we name circular rolling-window validation and discuss, among other things, how this technique can be applied to analyze the shrinkage effect. In the empirical part of chapter 2, one main finding is that traditional measures of discriminative power tend to overstate the long-run predictive ability of rating systems.

After having developed methods for the evaluation of default prediction models in chapter 2, we deal with default predictions themselves in chapter 3. More precisely, we analyze the situation that a panel dataset with time-varying covariates is at hand. This situation is rather common and simpler situations with cross-sectional data are obvious special cases. The main contribution in this chapter is the development of a new approach that allows multi-period predictions without the need to specify and estimate a model that predicts the covariates as it was proposed in the related literature. An application of our methods to a dataset of North American public firms shows that it delivers high out-of-sample predictive accuracy both in absolute terms and relative to other studies that use similar datasets.

In the final chapter of this thesis, we consider the situation that - differently to chapter 3 - a ranking of the obligors according to their default risk is already given in the form of ratings but that default probabilities need to be assigned to rating grades. This situation often occurs in practice, for instance if ratings stem from an external institution or if the rating system is at least partly based on qualitative elements. With respect to the two dimensions introduced in chapter 2, chapter 4 is solely about calibration. The specific problem that is analyzed in detail in chapter 4 is the situation where the sample size and/or the true default probabilities are small. Then, there is a high probability to underestimate the true default probability by standard methods. An important case that suffers from this problem is the

estimation of default probabilities for sovereign bonds. As a potential solution, we present an empirical Bayes estimator that allows more conservative and potentially also more precise estimation of default probabilities under realistic scenarios. The latter is shown by a novel kind of simulation study where we evaluate both the standard estimator and the empirical Bayes estimator. One important economic finding of our simulation study is that Basel II capital requirements of banks for their sovereign exposure tend to be underestimated considerably by standard methods as opposed to our empirical Bayes approach.

Chapter 2

Measuring predictive accuracy

For any prediction problem an essential part of the analysis is to measure the accuracy of the predictions. Usually, and this work is no exception in this respect, the predictions are aimed to be as precise as possible *out-of-sample*, i.e. we are analyzing *ex ante* predictions, which must be reflected by the evaluation method. We will distinguish between two dimensions of predictive accuracy, discrimination and calibration, thereby following common practice for credit default predictions (Basel Committee on Banking Supervision, 2005b). Discrimination here refers to the accuracy of the ranking of obligors according to their default risk while calibration concerns the accuracy of the levels of default probabilities. For instance, a model has high discriminatory power if the obligors with the highest estimated default probabilities actually default but may at the same time be badly calibrated if the default probabilities are much different from the observed default rates. On the other hand, a prediction that assigns the same default probability to each obligor may be well calibrated if this default probability is close to the true underlying average default probability but has virtually no discriminatory power. It is also possible to measure the overall accuracy of default probability estimates, i.e. a combination of discriminatory power and calibration accuracy. However, the distinction between discrimination and calibration is useful since it corresponds to the two common parts of the modeling process. First, the focus is to develop a model with high discriminatory power. Then,

the model is checked with respect to its calibration and can eventually be re-calibrated without the need to specify a new model. In contrast, it is not possible to improve the discriminatory power of a model afterwards.

The chapter is structured as follows. In section 2.1, we will deal with the evaluation of discriminatory power which is followed by a discussion of calibration in section 2.2. Section 2.3 will be about validation techniques which concerns the different ways how out-of-sample predictions can be made. We will then turn to the problem of statistical inference for measures of predictive accuracy in section 2.4. Finally, we will empirically illustrate the presented methods in section 2.5. The contents of this chapter stem largely from Orth (2012).

2.1 Discriminatory power

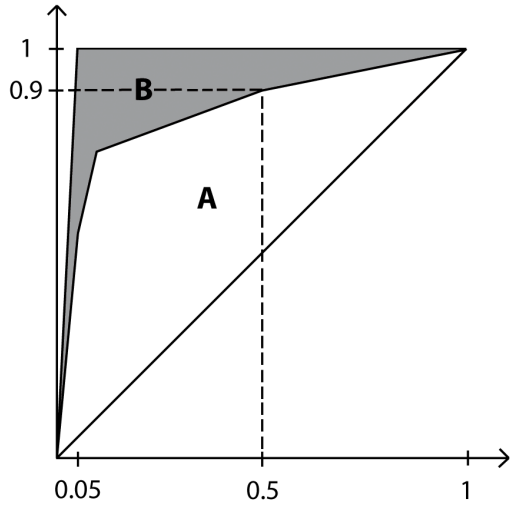
As far as discriminatory power is concerned, only the ordinal part of a default probability estimate is evaluated. More precisely, we can think of default probabilities to lead to a risk ranking of obligors which corresponds to the usual practice of assigning ratings to obligors. In this section, we will refer to the default predictions as ratings to reflect the fact that only ordinal predictions and no actual default probabilities are needed for measuring discriminatory power.

The following section covers the most common approaches but is not meant to be exhaustive. A more complete list including measures motivated from information theory can be found in Basel Committee on Banking Supervision (2005b). Further measures, especially some that are not based on a separation of discrimination and calibration like the Brier score (Brier, 1950) are documented and applied to credit default data in Krämer & Güttler (2008).

2.1.1 Accuracy Ratio and related measures

To measure discriminatory power, the approach most popular among banks, rating agencies and academics is based either on the Cumulative Accuracy

Figure 2.1: Cumulative Accuracy Profile



Profile (CAP) and its summary statistic, the Accuracy Ratio (AR), or the Receiver Operating Characteristic (ROC) curve and its summary index, the area under the ROC curve (AUROC).[1] Let us start with the explanation of the CAP which is exemplified by Figure 2.1. The CAP plots the share of defaulting obligors (ordinate) that is included in the $p \cdot 100\%$ worst rated obligors (abscissa). For instance, the point $(0.5, 0.9)$ in the graph means that the worse rated half of the obligors account for 90% of all default events in the sample. A perfect rating system would assign the worst ratings exactly to those who default (5% in our example), a situation which is visualized by the upper curve in Figure 2.1. On the contrary, a naive rating system where all obligors have the same rating corresponds to the main diagonal in the graph.[2] Thus, the better the predictive power of the rating system the closer is the realized CAP to the perfect CAP which motivates the Accuracy Ratio, $AR = A/(A+B)$, as a summary statistic with A being the area between the

[1] Sometimes, the Accuracy Ratio is also referred to as the Gini coefficient.

[2] Note that the CAP curve is monotonically increasing but neither necessarily concave nor above the main diagonal.

CAP and the main diagonal and B being the grey shaded area. Obviously, the CAP is closely related to the well-known Lorenz curve and is in fact sometimes simply referred to as Lorenz curve or, alternatively, power curve.[3] In some studies (Shumway, 2001; Roszbach, 2004), instead of the Accuracy Ratio some point of the CAP is simply chosen as the measure of predictive accuracy, for instance, the share of defaulters included in the lowest rated decile.

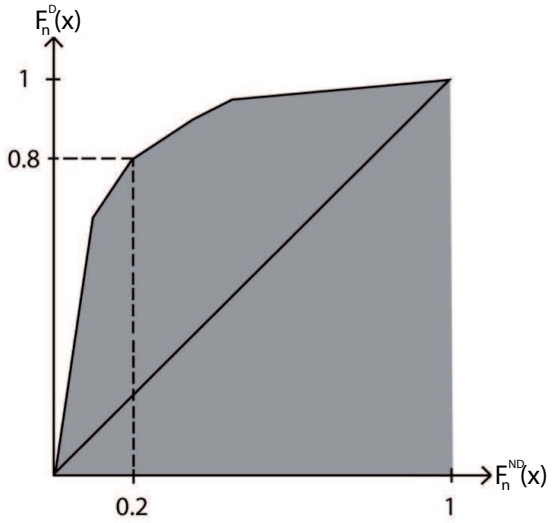
We now turn to the ROC curve which is similar but not identical to the CAP. For the ROC curve, the empirical cumulative distribution functions of the ratings of the defaulting and the non-defaulting obligors $(F_n^D(x), F_n^{ND}(x))$ are plotted against each other. An example for a ROC curve is given in Figure 2.2. For instance, the point $(0.2, 0.8)$ in the graph has the interpretation that there is a rating where 20% of the non-defaulting obligors and 80% of the defaulting obligors had that or a lower rating. The perfect ROC curve would be constant at the level one since in this case there is a rating where none of the non-defaulting obligors but all of the defaulting obligors have this or a worse rating. The AUROC is then simply the grey shaded area in Figure 2.2 which is obviously the greater the closer the ROC curve is to the perfect ROC curve. We can interpret each point of the ROC curve as representing a potential cut-off rating meaning that, for instance, a bank would only lend to obligors being rated better than the cut-off rating. Due to such an interpretation, the ordinate of the ROC curve is sometimes labeled as the hit rate, i.e. the share of the defaulting obligors that did not receive a loan, whereas the abscissa is called the false alarm rate, i.e. the share of non-defaulting obligors that did not get a loan.[4]

Interestingly, the maximum horizontal distance of the ROC curve to the main diagonal equals the classic Kolmogorov-Smirnov statistic for the one-

[3] One may argue that the CAP is more closely related to a concentration curve than to a Lorenz curve.

[4] A specific cut-off also leads to the construction of a so-called confusion matrix which contains the hit rate and the false alarm rate on the main diagonal. See Thomas et al. (2002, Ch. 7.2) for details. In another terminology, the terms sensitivity and specificity are used where sensitivity equals the hit rate and 1−specificity corresponds to the false alarm rate.

Figure 2.2: ROC curve



sided test that the rating distribution of the defaulting firms is stochastically larger than the rating distribution of the non-defaulting firms.[5] In a few studies, the Kolmogorov-Smirnov statistic is used instead of the AUROC as a measure of discriminatory power.

Besides their graphical derivations, there is a simple algebraic method to calculate the Accuracy Ratio and AUROC that provides a good intuition about what both indices measure. We will first focus on the Accuracy Ratio and then provide the link to the AUROC. Denote the rating (high values indicate low risk) of the ith defaulting obligor and the jth non-defaulting obligor by X_i^D and X_j^{ND}, respectively. The number of defaulting and non-defaulting obligors in the sample are referred to as n_1 and n_2. Define

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i^D < X_j^{ND}\,, \\ -1 & \text{if} \quad X_i^D > X_j^{ND}\,, \\ 0 & \text{if} \quad X_i^D = X_j^{ND}\,. \end{cases} \tag{2.1}$$

[5] To be specific, $H_0 : F^D(x) \leq F^{ND}(x)$ for all x against H_1: not H_0. Other variants of the Kolmogorov-Smirnov test exist (Mosler, 1995).

Then, the Accuracy Ratio is given by

$$AR = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij} \,. \tag{2.2}$$

We will call c_{ij} the concordance score of the pair of the ith defaulting and the jth non-defaulting obligor. Concordance is given if the rating of the defaulting obligor was worse than the rating of the non-defaulting obligor, while we have discordance in the opposite case. The case of identical ratings is captured by a concordance score of zero. The concordance score is evaluated for every pair of a defaulting and a non-defaulting obligor and then averaged over all pairs. It can be shown that the AUROC can also be calculated using Formulas (2.1) and (2.2) by simply replacing the concordance scores of 1,0 and -1 by $1, \frac{1}{2}$ and 0. Consequently, the following simple linear relation between the two measures holds (Engelmann et al., 2003):

$$AR = 2 \cdot AUROC - 1 \tag{2.3}$$

Obviously, the Accuracy Ratio and AUROC contain the same information and are only scaled differently. Note that the Accuracy Ratio and AUROC are closely related to the Mann-Whitney statistic (Mann & Whitney, 1947) for the test that the rating distributions of defaulters and non-defaulters are equal against the alternative of first-order stochastic dominance.[6] From now on, we will concentrate on the Accuracy Ratio but of course all arguments apply to the AUROC as well.

Formula (2.2) shows that the Accuracy Ratio is the fraction of pairs where the rating was concordant with the outcome minus the fraction of discordant pairs. In line with this interpretation, the corresponding population value is

$$P(X_i^D < X_j^{ND}) - P(X_i^D > X_j^{ND}) \,, \tag{2.4}$$

for a randomly selected pair i, j of the population (DeLong et al., 1988).

The Accuracy Ratio is a special case of a more generally defined measure of predictive accuracy called Somers' D (Somers, 1962). The connection is

[6] The Mann-Whitney statistic for our case is simply the number of pairs where $X_i^D < X_j^{ND}$ plus half of the number of tied pairs.

important in our context since the index which will be introduced in the next section, Harrell's C, is also based on Somers' D. Consider predicting a variable Y with a predictor X. The sample size is denoted by n. For ease of exposition, sort the values of Y in ascending order so that $Y_i \leq Y_j$ for $i < j$.[7] Let

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i < X_j \,,\, Y_i < Y_j \,, \\ -1 & \text{if} \quad X_i > X_j \,,\, Y_i < Y_j \,, \\ 0 & \text{else} \,. \end{cases} \tag{2.5}$$

Then Somers' D is defined as follows:

$$D_{XY} = \frac{1}{n_u} \sum_{i=1}^{n} \sum_{j>i} c_{ij} \tag{2.6}$$

$$n_u = \sum_{i=1}^{n} \sum_{j>i} 1_{[Y_i \neq Y_j]} \tag{2.7}$$

The denominator of D_{XY} (the number of usable pairs n_u) excludes any ties in Y since in these cases it is not possible to assess a "correct" or "incorrect" order of the predictors. In contrast, ties on X represent a case of mediocre prediction and are subsumed under "else". The Accuracy Ratio is simply the special case with Y being a binary variable (coded as 0 in the case of default and 1 otherwise).

Interestingly, there is a close relation between Somers' D and the well-known dependence measure Kendall's τ which will be important for the purpose of statistical inference in section 2.4. There are different versions of Kendall's τ, especially τ_a and τ_b (Kendall & Gibbons, 1990, Ch. 3), which all count the number of concordant pairs minus the number of discordant pairs in the numerator as does Somers' D. For τ_a, the denominator is just the number of all pairs, $n(n-1)$, so that Somers' D can be expressed as

$$D_{XY} = \frac{\tau_{a,XY}}{\tau_{a,YY}} \,, \tag{2.8}$$

where $\tau_{a,YY}$ is just the fraction of pairs not tied on Y so that $\tau_{a,YY} = n_u/n(n-1)$. Formula (2.8) will be helpful in section 2.4. With respect to τ_b, where

[7] It does not matter how ties on Y are ordered since pairs with equal values of Y are not "usable" for Somers' D. See Equation (2.7).

there is a symmetric adjustment in the denominator for ties, it can be shown that $D_{XY}D_{YX} = \tau_b^2$ (Somers, 1962).

For the Accuracy Ratio, we have seen that we need to classify the obligors into defaulters and non-defaulters to construct the corresponding binary variable Y. However, this means that any information about the timing of default events and certain censored observations are disregarded. How this loss of information can be avoided will be the topic of the next section.

2.1.2 Harrell's C

Consider first the following motivating example. At time t, two firms have ratings AA and B, respectively. When the prediction horizon is five years and the AA rated firm defaults prior to $t + 1$, while the B rated firm is censored at $t + 4$,[8] this pair is dropped for the calculation of the Accuracy Ratio although for this pair ratings and outcomes are clearly discordant. In fact, the firm that was rated B at time t has to be dropped completely in the case of the Accuracy Ratio since it can not be classified in either the defaulting or the non-defaulting group. In contrast, we will see that Harrell's C (Harrell et al., 1996) uses every observation. In the example given above the corresponding pair would - in line with intuition - receive a concordance score of -1 (with the analogous meaning as above).

We will now give the formal definition of Harrell's C and then discuss the various individual cases. Again, X_i is the rating (high values correspond to low risk) of obligor i, $i = 1, \ldots, n$. After being rated X_i, obligor i is observed not to default for a time denoted by Y_i. We will refer to Y_i as the observed lifetime of obligor i. If the observation is then ended by a default event, the censoring indicator variable C_i is set to zero. If obligor i is no longer observed due to right censoring, the value of C_i is one. Again, it is convenient to sort the lifetimes in ascending order so that $Y_i \leq Y_j$ for $i < j$. As a natural extension of Somers' D to censored data, we then define the concordance

[8] This means that the B firm does not default until period $t + 4$, but is no longer observed thereafter.

score as[9]

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i < X_j\,,\, Y_i < Y_j\,,\, C_i = 0\,, \\ -1 & \text{if} \quad X_i > X_j\,,\, Y_i < Y_j\,,\, C_i = 0\,, \\ 0 & \text{else}\,. \end{cases} \tag{2.9}$$

Then Harrell's C is given by:

$$C = \frac{1}{n_u} \sum_{i=1}^{n} \sum_{j>i} c_{ij} \tag{2.10}$$

$$n_u = \sum_{i=1}^{n} \sum_{j>i} 1_{[Y_i \neq Y_j\,,\, C_i = 0]} \tag{2.11}$$

n_u is the number of usable pairs. In words, a pair of observations is usable if (a) the obligors' observed lifetimes are not equal and (b) the obligor with the shorter observed lifetime experiences a default event, i.e. the lifetime is not censored. These conditions ensure that for every usable pair one obligor has indeed "outlived" the other obligor thereby enabling a sensible comparison of both. Given a usable pair, we can distinguish two cases. The first one consists of two obligors, both defaulting but after different time spans. Concordance is achieved if the rating of the obligor with the earlier default event was worse than the rating of the obligor defaulting later, while discordance is given in the opposite case and a concordance score of zero is assigned in the case of equal ratings. In the second case, one obligor defaults after a certain time span and the other obligor's lifetime is censored at a later point in time. For concordance, we require that the defaulting obligor was lower rated. Accordingly, we assign a concordance score of -1 in the opposite case and a score of zero for equal ratings.

Since Harrell's C is based on Somers's D it can also be written as a ratio of censored versions of Kendall's τ: $C = \tau_{a,XY,cens}/\tau_{a,YY,cens}$, where $\tau_{a,XY,cens}$ and $\tau_{a,YY,cens}$ are the estimators proposed by Oakes (2008) for censored data. Again, the relation is useful in the context of statistical inference as we will see in section 2.4.

[9] Harrell et al. (1996) actually normalize the measure between zero and one by assigning concordance scores of 1, $\frac{1}{2}$ and 0 instead of 1,0 and -1 as we do. We stick to the latter version to ensure comparibility with the Accuracy Ratio.

Similar to Pencina & D'Agostino (2004) we define the population value of Harrell's C as

$$P(X_i < X_j | Y_i < Y_j, C_i = 0) - P(X_i > X_j | Y_i < Y_j, C_i = 0)\,, \tag{2.12}$$

for two randomly selected individuals i and j from the population. That is, given a pair is found to be usable, Harrell's C estimates the probability of concordance minus the probability of discordance.

A potential source of criticism may be the fact that Harrell's C does not cover a specific prediction horizon since only the sample length provides a limit. This may not be suitable in credit risk applications since the maturity of most credits is limited and risk managers usually have a certain planning horizon. For this reason, we propose the following modification of Harrell's C. Denote the maximum prediction horizon that is of practical interest as H. Let

$$c_{ij} = \begin{cases} 1 & \text{if} \quad X_i < X_j\,,\, Y_i < Y_j\,,\, C_i = 0\,,\, Y_i < H\,, \\ -1 & \text{if} \quad X_i > X_j\,,\, Y_i < Y_j\,,\, C_i = 0\,,\, Y_i < H\,, \\ 0 & \text{else}\,. \end{cases} \tag{2.13}$$

The adjusted index is then calculated analogously to before:

$$C_{adj} = \frac{1}{n_u} \sum_{i=1}^{n} \sum_{j>i} c_{ij} \tag{2.14}$$

$$n_u = \sum_{i=1}^{n} \sum_{j>i} 1_{[Y_i \neq Y_j\,,\, C_i = 0\,,\, Y_i < H]} \tag{2.15}$$

The rationale of the adjustment is simple. Everything what happens after H is ignored. For instance, with H equal to 3 years, pairs of observations that do not include a default within the first 3 years are now not usable. This corresponds to the fact that we now require for a usable pair that the shorter observed lifetime is ended by a default event that occurs before H. The modification is easy to implement by simply conducting an artificial censoring at H for lifetimes that last longer than H. To be specific, values of Y equal to or larger than H are set to H and their censoring indicator is set to one.

While the number of unusable pairs grows with this adjustment it is important to note that still no observations have been completely removed. Thus, the amount of information – as measured by the number of usable pairs – included in C_{adj} is still distinctively higher than in the case of the Accuracy Ratio. We can distinguish two kind of pairs that are used for C_{adj} but not for the Accuracy Ratio. The first type covers obligors defaulting at different points in time before H. The second type refers to cases with one obligor defaulting at a certain point in time and another obligor whose lifetime is censored at a later point in time but before H. Harrell's C (in both its original and its adjusted version) has thus the advantages of using (a) the timing of the default events and (b) more information from censored observations.

In contrast to the Accuracy Ratio, for a reasonable use of Harrell's C one has to assume that the (conditional) survival functions of two obligors do not cross (before H). With respect to ratings, this means that there should be no difference in the ranking of obligors in terms of short-term and long-term ratings, respectively. If this was the case, the ordering of the obligors according to their default risk would change over time and thus the assessment of concordance of ratings and lifetimes would have to reflect these changes. Extensions of Harrell's C that are capable of crossing survival functions seem to be possible but are not covered in this work. Note that for the models we propose in chapter 3 the assumption of non-crossing survival functions is implied by the model structure.

The interpretation of C_{adj} is still in line with Harrell's C and the Accuracy Ratio. All these measures are bounded between -1 and 1 and yield the proportion of concordant pairs minus the proportion of discordant pairs among all usable pairs.[10] Further, Harrell's C has been implemented in various software packages. For instance, it is available in STATA through the user-written *somersd* program by Roger B. Newson and in R it is part of the *Hmisc* package (function *rcorr.cens*).

[10] If a naive rating system that assigns the same rating to each obligor is considered as the lower bound, the measures are bounded between 0 and 1.

2.2 Calibration

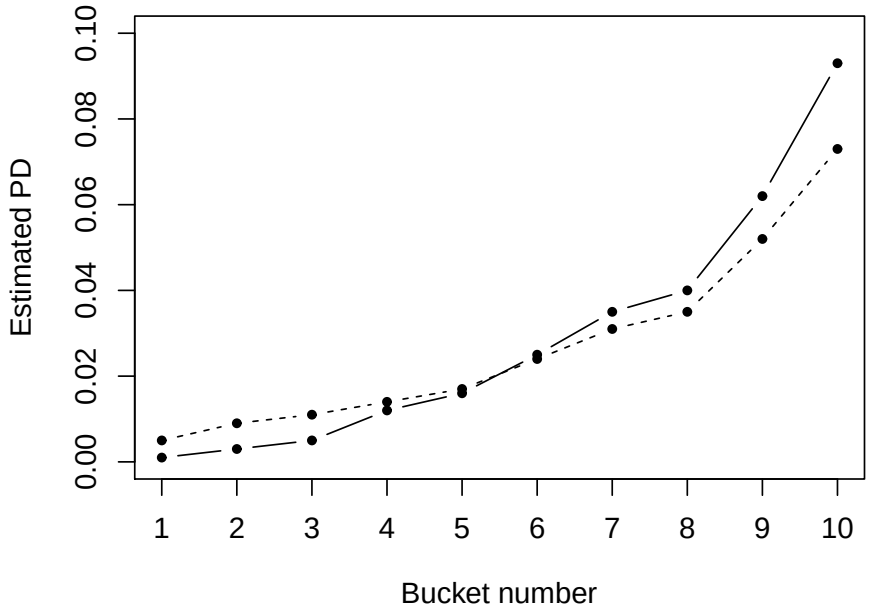
To analyze calibration, we need estimated default probabilities instead of only ordinal predictions like ratings which were sufficient in the context of discrimination. Let Y^* be the possibly unobserved lifetime or time until default of an obligor so that $Y = Y^*$ if the lifetime is not censored and $Y \leq Y^*$ otherwise. Further, let $PD_H = P(Y^* \leq H|\widehat{PD}_H)$ be the probability to default within a time horizon H conditional on some estimate $\widehat{PD}_H$.[11] We define predictions to be perfectly calibrated if $PD_H = \widehat{PD}_H$. We will investigate two ways to analyze departures from perfect calibration. The first is based on grouping the obligors into buckets based on their estimated default probabilities while the second approach is based on a calibration model. Due to their different statistical nature, we call these two approaches nonparametric and parametric calibration analysis, respectively.

2.2.1 Nonparametric calibration analysis

Since out-of-sample predictive accuracy is our central objective, any calibration analysis should be based on a *validation sample* which was not used in estimating the model that generated the default probabilities. For the moment, assume that such a validation sample exists. We will analyze different validation schemes in detail in section 2.3. To generate out-of-sample predictions, the model fitted to the *training sample* is applied to the validation sample. In the nonparametric calibration approach, the resulting out-of-sample default probabilities are grouped into J (approximately) equally sized buckets where the first bucket consists of the lowest default probability estimates and so on. A suitable procedure is then to compare the average out-of-sample default probability estimate in each bucket, $\overline{\widehat{PD}}^{OS}_{H,j}$, with an estimate of the default probability in the validation sample, $\widehat{PD}^{VS}_{H,j}$. In the

[11] The conditioning on $\widehat{PD}_H$ is necessary since unconditionally all obligors have the same default probability. Note also that the higher the discriminatory power of a model the more are the conditional default probablities apart from the unconditional default probability.

Figure 2.3: Calibration plot



Hypothetical estimated out-of-sample default probability predictions (solid line) and observed default rates (dashed line) in a validation sample

simplest case, $\widehat{PD}_{H,j}^{VS}$ is simply the observed default frequency in the validation sample for the members of bucket j. However, in the presence of censored data, default frequencies are not directly available and one should use a suitable survival analysis extension like the life-table or the Kaplan-Meier estimator (see chapter 4.1) instead.

A suitable way to visualize how the out-of-sample predictions fit to the observed default behaviour in the validation sample is a so-called calibration plot which plots $\overline{\widehat{PD}_{H,j}^{OS}}$ and $\widehat{PD}_{H,j}^{VS}$ against the bucket numbers. Figure 2.3 gives a hypothetical but typical example for such a plot.

The main observation is that the line connecting the out-of-sample default probabilities is steeper than the line connecting the default rates observed in the validation sample. This phenomenon is called *shrinkage* or *regression to*

the mean (Copas, 1983; Harrell, 2001) and is based on the fact that the fit of a model to new data is usually worse than the fit to the training sample. Note that the steeper the line of the observed default rates in a calibration plot the higher is the discriminatory power of the model. In this sense, the shrinkage effect corresponds to the fact that discriminatory power is usually lower out-of-sample than in-sample. Shrinkage is closely related to the notion of overfitting and is, everything else equal, more pronounced if the (effective) size of the estimation sample is small and if the number of parameters is high.[12] It indicates that (unadjusted) out-of-sample predictions tend to be too "extreme" and that properly adjusted predictions would be an improvement. Thus, the shrinkage effect is an important argument to recalibrate a model.

Besides the shrinkage effect, a calibration plot can also indicate that the functional form of a default prediction model is inappropriate. For instance, if the line connecting the out-of-sample default probabilities would be concave as opposed to the obviously convex line connecting the default rates, this would point to a misspecification of the default probability model.

Of course, it is desirable to analyze if the departures of $\widehat{PD}_{H,j}^{VS}$ from $\overline{\widehat{PD}}_{H,j}^{OS}$ are statistically significant. If $\widehat{PD}_{H,j}^{VS}$ is asymptotically normally distributed and a consistent variance estimator is at hand, we can use the fact that under the joint null hypothesis that $\overline{\widehat{PD}}_{H,j}^{OS} = PD_{H,j}, j = 1, \ldots, J$,

$$Q = \sum_{j=1}^{J} \frac{\left(\widehat{PD}_{H,j}^{VS} - \overline{\widehat{PD}}_{H,j}^{OS}\right)^2}{\widehat{V}(\widehat{PD}_{H,j}^{VS})} \overset{\text{asy.}}{\sim} \chi_J^2 \quad . \tag{2.16}$$

Note that we have to assume independence of $\widehat{PD}_{H,j}^{VS}$ and $\widehat{PD}_{H,j'}^{VS}$ for $j \neq j'$. The test statistic Q has an intuitive interpretation as a measure of fit between $\overline{\widehat{PD}}_{H,j}^{OS}$ and $\widehat{PD}_{H,j}^{VS}$ as it sums up the squares of their standardized differences. It can thus be interpreted as a measure of calibration accuracy. In particular, it is much more sensible to use Q than any unstandardized

[12] For instance, the effective sample size of a dataset that exhibits dependencies is smaller than an independent sample of the same size.

measure like $\sum_{j=1}^{J}\left(\widehat{PD}_{H,j}^{VS}-\overline{\widehat{PD}}_{H,j}^{OS}\right)^2$ or $\sum_{j=1}^{J}\left|\widehat{PD}_{H,j}^{VS}-\overline{\widehat{PD}}_{H,j}^{OS}\right|$ since these will be dominated by the accuracy in estimating the default probability in the highest-risk buckets.[13]

In the case that a simple default frequency is used for $\widehat{PD}_{H,j}^{VS}$ the test from Equation (2.16) is very similar to the Hosmer-Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1980) for logistic regression models with the only difference being that the denominator is there equal to $\overline{\widehat{PD}}_{H,j}^{OS}(1-\overline{\widehat{PD}}_{H,j}^{OS})/n_j$ (with n_j as the number of observations in bucket j). Given the different standardization, the Hosmer-Lemeshow test is also referred to as a Score test whereas our test can be interpreted as a Wald test. The Hosmer-Lemeshow test is quite popular among regulators for validating default probability estimates of banks (Basel Committee on Banking Supervision, 2005b). However, it is not straightforwardly extended to the survival data setting which we focus on.[14]

In a survival analysis context, $\widehat{PD}_{H,j}^{VS}$ can be estimated by the aforementioned life-table estimators.[15] The life-table estimator and an estimator for its variance are presented in sections 4.2 and 4.A, respectively. It remains the question how to choose the number of buckets J. As an extreme case, if $J=1$, one merely tests what is sometimes called global calibration (as opposed to local calibration). Analyzing global calibration does not reveal anything about possible shrinkage effects but can nonetheless be useful. From a methodological point of view, the case $J=1$ has the advantage that the sometimes questionable assumption of independence between the different groups is no longer necessary. Moreover, the smaller J, the better the asymptotic approximation of both the variance estimator for $\widehat{PD}_{H,j}^{VS}$ and of the χ_J^2 distribution will be. However, as J decreases our test loses its ability to detect patterns of miscalibration that are hidden by the aggregation into a few

[13] Such unstandardized measures are commonly used in the evaluation of probability predictions from time series. See, for instance, Diebold & Rudebusch (1989) for details.

[14] To see this, note that the sample size in bucket j is not so easily defined in a survival analysis context because usually not every obligor will be observed until H because of censoring.

[15] Alternatively, one may fit a parametric survival distribution to the lifetimes of each bucket in the validation sample instead. This is done, for instance, by Dwyer et al. (2004).

buckets.

Testing for potential miscalibration and improving calibration are closely connected tasks. An obvious way to correct for the shrinkage effect and possible problems with the functional form is to use the default rate in each bucket, $\widehat{PD}_{H,j}^{VS}$, as a recalibrated default probability estimate. Sometimes, the evolution of the default rates over the buckets is additionally smoothed by a nonparametric or parametric regression (Hartmann-Wendels et al., 2010, Ch. I1.2.6).

2.2.2 Parametric calibration analysis

As an alternative to the nonparametric analysis presented above a model-based approach to calibration is also possible. Our main concern here is to account for the shrinkage effect. There are estimators available like the Lasso (Tibshirani, 1996) or Penalized Maximum Likelihood (Harrell et al., 1996, Ch. 9.10) that directly incorporate shrinkage in the estimation process. Instead, we consider the simpler solution to use a calibration model for evaluating and potentially revising the original, i.e. unshrunk, default probability estimates (Van Houwelingen & Le Cessie, 1990; Medema et al., 2009). In a calibration model, the outcomes in the validation sample are regressed on the predictions made with the parameter estimates from the training sample.[16] Besides simplicity, such an approach has the advantage that calibration can be treated separately from the task to derive a model with high discriminatory power. This is because using a calibration model to recalibrate the original estimates will not change the ranking of the default predictions. In contrast, using the Lasso or Penalized Maximum Likelihood will usually introduce at least some change in the ranking.

Consider as an example a Logit calibration model. Using the log odds trans-

[16] This principle was already proposed by Mincer & Zarnowitz (1969) in a time series context.

formation for the out-of-sample default probabilities the model is given by

$$P(Y_i^* \leq H|\widehat{PD}_i^{OS}) = \left(1 + \exp\left[-\gamma_0 - \gamma_1 \log\left(\frac{\widehat{PD}_i^{OS}}{1 - \widehat{PD}_i^{OS}}\right)\right]\right)^{-1} . \tag{2.17}$$

If the out-of-sample default probabilities were derived from a Logit model as well (which must not necessarily be the case)[17] $\log\left(\frac{\widehat{PD}_i^{OS}}{1-\widehat{PD}_i^{OS}}\right)$ is equal to the linear predictor $\widehat{\beta}' x_i$ where $\widehat{\beta}$ is the estimated parameter from the training sample and x_i refers to the vector of predictor variables of observation i in the validation sample. Note that the right-hand side of Equation (2.17) simplifies to $\widehat{PD}_i^{OS}$ if $\gamma_0 = 0$ and $\gamma_1 = 1$. Thus, a natural way to test calibration is to estimate the model given by Equation (2.17) with the observations of the validation sample and to check if the estimates $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ differ significantly from 0 and 1, respectively. Given the shrinkage effect, we will usually expect $\widehat{\gamma}_1 < 1$ and $\widehat{\gamma}_0 > 0$ which would mean that less extreme predictions give a better fit in the validation sample. As a measure of calibration accuracy, we can use, for instance, the Wald statistic from testing the joint null hypothesis that $(\gamma_0, \gamma_1) = (0, 1)$.

The Logit model is just one simple example for a calibration model. It suffers from the fact that it is not able to take the timing of default events and censored data appropriately into account. If the lifetimes censored before H are simply omitted, the Logit model systematically ignores "positive" information, i.e. information about obligors not defaulting for a certain time, and default probability estimates are thus upward biased. Hazard models, which will be introduced in chapter 3, do not suffer from this drawback and can also be used as calibration models. A detailed presentation of these kinds of calibration models together with an application will be given in section 3.3.4.

As in the nonparametric approach, a parametric calibration analysis directly offers an opportunity to recalibrate and thus to possibly improve the model.

[17] The most natural approach is, of course, to use the same model structure (Logit in this case) for estimating the original model and for calibration. However, sometimes the calibration analysis is conducted from "outsiders" (regulators, for instance) so that a different approach may be used.

In the Logit example, estimates of γ_0 and γ_1 can be simply plugged into Equation (2.17) to get recalibrated default probability estimates. We see thus that measuring and optimizing calibration can be seen as more or less the same thing. This is an important difference to the measurement of discriminatory power where a potential correction for low power is not directly available. Of course, a recalibration is only sensible if the parameters of the calibration model (γ_0, γ_1) are estimated reliably which in turn heavily depends on the validation scheme. Consequently, we will analyze different validation techniques in detail in the following section.

2.3 Validation techniques

In the preceding sections, we have taken the existence of a training and a validation sample as given. It turns out, however, that the precise validation scheme is an issue that deserves more attention. If the training sample and the validation sample are indeed separate samples the scheme is typically referred to as a sample split. The drawback of a sample split is that either the model fitted to the training sample is estimated inefficiently (if the validation sample is a large part of the whole sample) or that the validation exercise suffers from limited data (if the validation sample is small).

A solution to this problem is the use of cross-validation where each observation is used for model estimation *and* validation. For instance, if we use classical leave-one-out cross-validation to estimate a calibration model (Van Houwelingen & Le Cessie, 1990), we estimate the model without the ith observation to imitate an out-of-sample prediction for the default probability of observation i, $\widehat{PD}_i^{OS}$, $i = 1, \ldots, n$. Then, the calibration model is fitted to the whole sample using only $\widehat{PD}_i^{OS}$ or its linear part (see above) as predictors. Similarly, with respect to nonparametric calibration, buckets can be built according to $\widehat{PD}_i^{OS}$, and the whole sample can be used to estimate the default probability for each bucket and to compare it with the bucket averages of $\widehat{PD}_i^{OS}$.

Leave-one-out cross-validation requires the estimation of $n + 1$ models and

is thus computationally quite intensive. The computational burden can be considerably decreased by using K-fold cross-validation where the sample is split into K approximately equally sized parts and in each step one of these subsamples is held out to simulate the corresponding out-of-sample predictions.

An important assumption underlying ordinary cross-validation is the independence of the observations.[18] In many credit risk applications, including our empirical study in chapter 3, this assumption will not be met. For the case of stationary dependent data, Burman et al. (1994) introduced block cross-validation. Within this approach, for the prediction of the observation in some period t one estimates the model omitting the observations $t - B, \ldots, t, \ldots, t + B$. B is selected such that approximate independence between the training and the validation sample is achieved. Importantly, the rationale is here to simulate predictions for an observation of a "process that has the same distribution as [the original process] but is independent of it" (Burman et al., 1994, p. 351). This is arguably not the most relevant situation at least with respect to credit default predictions. Rather, in practice one usually uses the information up to some period t to make predictions for the subsequent periods *of the same process*. Note that this may well mean that there is some dependence between the training sample (which includes all observations up to period t) and the outcomes in periods $t + 1, \ldots$.

An alternative approach that takes the latter argument into account is the application of recursive or rolling-window estimation schemes. Within the recursive approach, one estimates the model with all the information available up to period t to make out-of-sample predictions for the upcoming periods and then increases t step-by-step to generate a series of predictions. The size of the estimation window thus increases by one period in each step. In contrast, under a rolling-window approach the size of the estimation window is fixed and in each step one period is added in the end and one period is omitted in the beginning. The recursive approach is also known as forward validation (Hjorth, 1982) or prequential analysis (Dawid, 1984). In the credit

[18] While being generally invalid, Burman & Nolan (1992) show that in certain cases ordinary cross-validation can still be saved even when the data exhibit dependencies.

risk context, Stein (2004) argues in favor of the recursive scheme that it is closest to the actual application of default prediction models in practice.

As a tool to analyze possible shrinkage effects, recursive and rolling-window validation schemes have important drawbacks which are similar (albeit less pronounced) to the problems of a sample split. On the one hand, when one starts building validation samples at an early point in time, the first models are estimated on a rather small dataset. This will usually result in an overestimation of the shrinkage effect since the amount of shrinkage decreases with the sample size. On the other hand, when the validation period starts late, only a rather small amount of data can be used for validation purposes. If the validation samples are used to estimate shrinkage parameters, this will result in inefficient estimation.

To overcome the problems of both block cross-validation and recursive or rolling-window validation we propose a new kind of validation scheme which we call circular rolling-window (CRW) validation. The precise procedure is as follows:

1. Choose a block length B so that it is reasonable to assume that observations in period t and period $t + B$ are approximately independent. Choose B such that $B \geq H$, where H denotes the prediction horizon.

2. For calendar period t, estimate the model after omitting all information from periods $t+1, \ldots, t+B$. This includes a possible adjustment of the lifetimes and censoring indicators for the observations in period t and before as these may contain information about the omitted periods.

3. Use the model estimated in step 2 to make out-of-sample predictions from period t (with a horizon of H).

4. Let t run from the first period to period T where T is the last calendar period in the sample.

The CRW method differs from block cross-validation only in the fact that one omits a block only on the right-hand side (the future) and not on both sides of period t (for reasons that were discussed above). It is also very

closely related to a recursive or rolling-window estimation scheme. The differences are here that for the CRW approach the periods $t+B+1,\ldots,T$ are additionally attached to each training sample (thereby motivating the name "circular") and that the validation period already starts with the first period, i.e. there are more validation periods. An application of the CRW procedure to calibration analyses is straightforward. The CRW method produces out-of-sample default probabilities for each observation in the sample which can be used for a nonparametric or parametric calibration analysis as described in the preceding sections. We will apply the CRW approach in section 3.3.4, which also involves some further discussion from a practical point of view. Of course, the CRW scheme is also an interesting option for the evaluation of out-of-sample discriminatory power. However, it is more important in the context of calibration for two reasons. First, compared to the recursive scheme, the size of the training samples is much closer to the full sample size, which is important as too small samples would lead to an overestimation of the shrinkage effect. Such kind of systematic bias is usually not present in the context of discrimination. Second, a calibration analysis amounts not only to test predictive accuracy but also to potentially recalibrate the model. Since the final estimates may thus depend on the validation exercise, it is important to validate as efficiently as possible. This is achieved by the CRW method as *every* period in the sample is used as a validation period.

Finally, it is important to note that methods like block cross-validation were originally designed for time series data. Typical credit default datasets, including the ones that are used in this work, have a panel structure so that the question of transferability arises. More precisely, while block cross-validation or CRW validation clearly simulate predictions for *new periods*, predictions for *new obligors* are of interest as well. In many relevant datasets, however, such predictions for new obligors are automatically done by these methods since new obligors enter the dataset over time so that a prediction for a new period will usually also involve predictions for new obligors. This is also likely to be the relevant case in practice where in a new period some obligors are already known to a lender and some additional obligors appear. Thus, as it is closely related to the actual prediction processes in practice, the application of CRW to panel data seems to be appropriate.

2.4 Statistical inference

Measures of predictive accuracy are, of course, subject to sampling variability. Analyzing this variation is useful not least for confidence intervals and hypothesis tests. For instance, it is often interesting to test if one model has significantly more predictive accuracy than another model. We structure our analysis of statistical inference into two parts. In section 2.4.1 we cover single-cohort statistical inference which can be thought of as the standard case where the sample is assumed to consist of independent observations.[19] For single-cohort settings, suitable methods are already available in the literature and we will present these methods briefly in section 2.4.1 to provide an appropriate background. In section 2.4.2, we will then derive methods for the multiple-cohort case, where the analysis is complicated by strong dependencies in the data. With multiple cohorts the dataset has a panel structure. We will explain the multiple-cohort sampling scheme in detail below and we will argue that it uses the maximum amount of information that is contained in the data.

The main part of our work in this section concerns the Accuracy Ratio and (adjusted) Harrell's C. However, the results from section 2.4.2 generalize to any asymptotically normal index of predictive accuracy and, as far as bootstrap methods are concerned, the requirement of asymptotic normality might even be not necessary. Since our measures of calibration accuracy are test statistics we have already discussed statistical inference in their context so that there is nothing to add in the standard single-cohort case. At the end of section 2.4.2, we will give some remarks on measures of calibration in a multiple-cohort setting.

[19] The assumption of independence (within a single cohort) is standard in the literature so far but may not be fulfilled in the presence of common market shocks or clustering within industry sectors. If dependence is sizeable and ignored, one has to be aware that standard errors are likely to be too low.

2.4.1 Single-cohort statistical inference

As we have seen in section 2.1, both the Accuracy Ratio and (adjusted) Harrell's C can be represented as ratios of Kendall's τ_a coefficients. Using the same notation as in section 2.1 we can apply the following representation:

$$\tau_{a,XY} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \,, \tag{2.18}$$

where $c_{ij} = c((X_i, Y_i, C_i), (X_j, Y_j, C_j))$ for Harrell's C and $c_{ij} = c((X_i, Y_i), (X_j, Y_j))$ for the Accuracy Ratio are the functions that assign the concordance score to its (vector-valued) arguments. Given (2.18) and the fact that the function $c(\cdot)$ is invariant to permutation of its arguments it follows that $\tau_{a,XY}$ is a so-called U-statistic. Note that the same holds for $\tau_{a,YY}$, the denominator in the corresponding representation of the Accuracy Ratio and Harrell's C, where $c(\cdot)$ is simply a function that is equal to one if the pair is usable and zero otherwise. U-statistics have been shown to be asymptotically normally distributed (Hoeffding, 1948). Further, the corresponding asymptotic covariance matrix can be consistently estimated by the jackknife method (Arvesen, 1969), which we will now briefly explain. The jackknife is based on leaving out the ith observation, $i = 1, \ldots, n$, and calculating the corresponding statistic with the remaining $n-1$ observations. For our case, denote this statistic by $\tau_{a,XY,-i}$. Then the jackknife pseudo-values are defined as

$$\widetilde{\tau}_{a,XY,i} = \tau_{a,XY} + (n-1)(\tau_{a,XY} - \tau_{a,XY,-i}) \,. \tag{2.19}$$

Tukey (1958) introduced these pseudo-values and argued that they could be treated as approximately independently and identically distributed random variables. We now assume that we have d competing predictors, $X_{(1)}, \ldots, X_{(d)}$, and define the vector $\tau_{a,d} = (\tau_{a,YY}, \tau_{a,X_{(1)}Y}, \ldots, \tau_{a,X_{(d)}Y})'$ (and analogously $\widetilde{\tau}_{a,d,i}$ for the vector of pseudo-values). The jackknife covariance matrix is then given by

$$\widehat{Cov}_{jack}(\tau_{a,d}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\widetilde{\tau}_{a,d,i} - \tau_{a,d})(\widetilde{\tau}_{a,d,i} - \tau_{a,d})' \,. \tag{2.20}$$

The jackknife - similar to other resampling methods - can be computationally quite intensive. However, for our case Newson (2006b) has developed an al-

gorithm that allows computation of the jackknife covariance matrix in a time of order $n \log(n)$. In our empirical analysis we have found the computational effort of the jackknife to be relatively low and to be considerably less than for the bootstrap.

Since we are ultimately interested in inference for the Accuracy Ratio and Harrell's C and not for τ_a one more step is needed. Since both indices have a representation as a ratio of τ_a coefficients they are asymptotically normal as well and we can apply the multivariate Delta method to obtain an appropriate covariance matrix (Newson, 2006a). Denote the d indices referring to our d predictors as $I_{(1)}, \ldots, I_{(d)}$. The application of the multivariate Delta method then yields the following result:

$$\widehat{Cov}(I_{(1)}, \ldots, I_{(d)}) = \Gamma \widehat{Cov}_{jack}(\tau_{a,d}) \Gamma' \,, \tag{2.21}$$

$$\Gamma_{i1} = \frac{\partial I_{(i)}}{\partial \tau_{a,YY}} = -\frac{\tau_{a,X_{(i)}Y}}{(\tau_{a,YY})^2} \,, \tag{2.22}$$

$$\Gamma_{ij\,(j>1)} = \frac{\partial I_{(i)}}{\partial \tau_{a,X_{(j-1)}Y}} = \begin{cases} \frac{1}{\tau_{a,YY}} & \text{if} \quad i = j-1 \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{2.23}$$

For the Accuracy Ratio, there also exist closed-form formulas for the covariance matrix (Bamber, 1975; DeLong et al., 1988). In the empirical analysis, we use the jackknife approach, however, to enhance comparibility of the Accuracy Ratio and Harrell's C.

2.4.2 Multiple-cohort statistical inference

To the best of our knowledge, there is no study so far that deals with the problem of statistical inference in a multiple cohort setting (as defined below) and takes into account the dependence structure of such datasets. The multiple cohort case is relevant because it allows to extract the maximum amount of information out of the dataset. To see this, let us first clarify what is meant by a multiple cohort sampling scheme.[20]

[20]Sometimes the term "static pool" instead of "cohort" is used.

A cohort consists of all obligors that have a rating at a given point in time t.[21] For the members of the cohort, the rating at t and the lifetimes beginning at t (together with the censoring indicators) are recorded. As an example, consider a firm that was rated, say, BB at the beginning of 2009 and defaulted in October 2010. The firm thus enters the cohort that was built at the start of 2009 with its BB rating and a lifetime of 21 months (and a censoring indicator of 0). Now assume that in the beginning of 2010 the same firm was rated CCC. In the cohort built at the start of 2010 the same firm is included again with its CCC rating and a lifetime of 9 months (and again a censoring indicator of 0). The reason why the same firm is included in both cohorts is that we want to evaluate both the performance of the BB rating in the beginning of 2009 and of the CCC rating in the beginning of 2010. Note also that the firm would be included in the cohort of 2010 even if the rating would not have changed. Obviously, if we build an aggregate or pooled cohort out of all the individual cohorts the pooled observations are dependent because we have a panel dataset of partially overlapping lifetimes. In our example, the overlapping period consists of the 9 months in 2010. The overlapping sample problem gets more pronounced if we build cohorts at a higher frequency and if we have longer lifetimes.[22] For instance, in the empirical section we will build cohorts on a monthly basis to use as much information as possible at the same time leading to even larger overlappings than in our example.[23] Due to the strong dependencies in the pooled cohort methods for statistical inference designed for approximately independent samples as the ones mentioned in section 2.4.1 are not directly applicable in our setting. While dependence does not introduce any bias to the indices themselves standard errors that ignore dependencies are usually downward biased so that confidence intervals and hypothesis tests are not reliable.

Returning to a more general setup, let us assume that there is a sequence

[21] Since we refer to measures of discriminatory power in the following, we refer to our default predictions as ratings. Of course, the cohort terminology generalizes to default probabilities as predictions.

[22] If we censor our lifetimes at the prediction horizon, the amount of overlapping increases with the prediction horizon.

[23] Moody's is also building cohorts each month in its calculation of Accuracy Ratios (Cantor & Mann, 2003).

of points in time $t, t = 1, \ldots, T$, and a cohort is built at each t with I_t denoting the chosen index of predictive accuracy (e.g., the Accuracy Ratio or Harrell's C) for the cohort built at time t. Given a prediction horizon H, this would correspond to a sample length of $T + H$.[24] As a first issue, one has to decide how to combine the indices $I_t, t = 1, \ldots, T$, to one single measure of predictive accuracy. Cantor & Mann (2003) propose either using some type of weighted mean of I_t or simply calculating the index for the pooled cohort. As weighting schemes, the authors consider equal weights, the number of observations and the number of defaults while finally using the second alternative in their empirical part. We first analyze the weighted mean approach. Consider the following general weighted mean:

$$I = \sum_{t=1}^{T} w_t I_t \tag{2.24}$$

The weights are normalized to sum up to one. Due to the "overlapping lifetimes problem" sketched above we expect strong autocorrelation of the time series I_t.[25] We assume that $\mathrm{Corr}(I_t, I_{t+j}) = \rho_j$ depends only on j but not on t (assumption 1). This assumption seems reasonable since the main source of dependence between I_t and I_{t+j} is the overlapping fraction of the underlying lifetimes which is equal to $\max(0, 1 - j/H)$, regardless of t. In contrast, the variance of I_t, denoted by σ_t^2, is allowed to vary with t so that we do not assume stationarity. Further, we assume that the correlation of the indices vanishes if the time between the cohort building dates is equal to or larger than the prediction horizon, i.e. $\rho_j = 0$ for $j \geq H$ (assumption 2). In these cases, overlapping lifetimes do not occur anymore. Under these assumptions, the variance of I can be expressed as

$$V(I) = \sum_{t=1}^{T} w_t^2 \sigma_t^2 + 2 \sum_{j=1}^{H-1} \rho_j \sum_{t=1}^{T-j} w_t \sigma_t w_{t+j} \sigma_{t+j} \,. \tag{2.25}$$

For the derivation of this formula, we have also used the additional assumption that the weights are deterministic. Strictly speaking, from the

[24] Since we do not build cohorts in periods $T+1, \ldots, T+H$ and consider only measures with a limited prediction horizon we avoid any boundary problems.

[25] This is confirmed by the empirical analysis in section 2.5 with empirical first-order autocorrelations ranging from 0.539 to 0.946.

three types of weights mentioned above, only equal weights are deterministic. However, we have conducted bootstrap experiments with fixed and varying weights that show that this source of variation is negligible. Estimators for σ_t are available for every t by the procedures presented in section 2.4.1. For ρ_j, a natural choice are the empirical autocorrelations, which are consistent estimators of the true autocorrelations and do not require the construction of a time series model. The formula used in the empirical section is

$$\widehat{\rho}_j = \max\left(0, \frac{1/(T-j)\sum_{t=1}^{T-j}(I_t - \overline{I})(I_{t+j} - \overline{I})}{1/T\sum_{t=1}^{T}(I_t - \overline{I})^2}\right) \tag{2.26}$$

which cancels out the effect of occasionally occurring negative autocorrelation estimates. $\overline{I}$ refers to the simple mean of the time series of indices.

We now turn to the sampling distribution of I. For both the Accuracy Ratio and Harrell's C, asymptotic normality of I_t follows from the arguments given in section 2.4.1. We then assume that the weighted average I converges to some value $\mu(I)$ as the cohort sizes and the number of periods approach infinity (assumption 3). This excludes any trending behaviour. Under these assumptions, we can apply Slutsky's theorem and the Central Limit Theorem for M-dependent random variables[26] to derive the following formula, which can be used for confidence intervals and hypothesis tests:

$$\frac{I - \mu(I)}{\sqrt{\widehat{V}(I)}} \xrightarrow{d} N(0,1) \tag{2.27}$$

Note that the asymptotics of Formula (2.27) require both the cohort sizes and the number of cohorts approaching infinity. As was already indicated above, Formulas (2.25) and (2.27) are applicable not only for the Accuracy Ratio and Harrell's C but for any asymptotically normal index where an estimator for σ_t is available. In order to perform hypothesis tests regarding

[26]Our time series of indices is M-dependent in the sense that we assume that indices separated by more than M periods are assumed to be independent, i.e. in our case $M = H - 1$. Since independence implies uncorrelatedness but not vice versa this assumption is stronger than assumption 2 – which suffices for the variance formula – and might thus be referred to as assumption 2*. For details about this kind of Central Limit Theorem see, for instance, Shumway & Stoffer (2006), appendix A.

the difference in the predictive accuracy of two different rating systems, say A and B, we only have to substitute I_t by $(I_{t,(A)} - I_{t,(B)})$, σ_t^2 by $\sigma_{t,(A-B)}^2 = \sigma_{t,(A)}^2 - 2 \cdot Cov(I_{t,(A)}, I_{t,(B)}) + \sigma_{t,(B)}^2$ and ρ_t by $\rho_{t,(A-B)}$, the autocorrelation of the time series $(I_{t,(A)} - I_{t,(B)})$. The necessary covariances, $Cov(I_{t,(A)}, I_{t,(B)})$, can be computed with the methods of section 2.4.1. Asymptotic normality of $(I_{t,(A)} - I_{t,(B)})$ follows from the joint asymptotic normality of $I_{t,(A)}$ and $I_{t,(B)}$.

Alternatively, resampling methods can be used for inference. They are an especially important alternative for datasets with just a few number of cohorts where it is not possible to estimate the autocorrelations for the time series of indices reliably. Jackknife and bootstrap approaches can be applied to both the weighted average and the pooled version of the indices. Clearly, the resampling procedures have to take the dependence structure of the data into account as well. If we interpret all the observations of an individual obligor as one cluster and assume independence between clusters, i.e. between different obligors, we can apply the cluster versions of the jackknife and the bootstrap which we will briefly outline in the following.[27]

For the bootstrap, this amounts to resampling with replacement from the set of obligors instead of the set of all observations which contains several observations per obligor.[28] The indices are calculated for each bootstrap sample and are used for inference in the usual way. More precisely, if I_b^* denotes the index for the bth bootstrap sample and if we draw B bootstrap replications, the bootstrap estimate of the standard error of I is

$$\widehat{\sigma}(I) = \left(\frac{1}{B-1} \sum_{b=1}^{B} (I_b^* - \overline{I}^*)^2 \right)^{1/2} \quad , \quad \overline{I}^* = \frac{1}{B} \sum_{b=1}^{B} I_b^* \quad . \tag{2.28}$$

Bootstrap confidence intervals can be obtained by simply looking at the corresponding percentiles of the bootstrap distribution (Efron & Tibshirani,

[27] A more detailed discussion is given in Field & Welsh (2007) and Cameron et al. (2008) for the cluster bootstrap and in Wolter (2007, Ch. 4.6) for the cluster jackknife. Besides clustering, another non-standard feature of our data is the censoring of the lifetimes. Efron (1981) shows that the bootstrap works well even when the data are censored.

[28] Hanson & Schuermann (2006) and Cantor et al. (2008) use this kind of bootstrap to calculate confidence intervals for default probabilities.

1993, Ch. 13.3).

With respect to bootstrap hypothesis tests, we are especially interested in testing $H_0 : \Delta\mu(I) = \mu(I_{(A)}) - \mu(I_{(B)}) = 0$ (again for two rating systems A and B) against the two-sided alternative. To do so, we would ideally generate bootstrap samples under the null distribution in order to estimate $P^*_{H_0}(|\Delta I^*| \geq |\Delta I|)$ (where the probability refers to the bootstrap distribution under the null hypothesis) giving the p value of the test. Under the nonparametric bootstrap approach which we consider here the bootstrap null distribution is usually approximated by the translation approach (Efron & Tibshirani, 1993, Ch. 13.3), i.e. the empirical distribution is used to draw bootstrap samples and ΔI^* is centered. This leads to $P^*(|\Delta I^* - \Delta I| \geq |\Delta I|)$ as an approximation for the bootstrap p value. This quantity may then be estimated by (Davison & Hinkley, 1997, Ch. 4.4)

$$p = \frac{1 + \#(|\Delta I^* - \Delta I| \geq |\Delta I|)}{1 + B} \quad , \tag{2.29}$$

which is the formula that we will use in all the empirical parts of this work.

The extension of the jackknife to clustered data is also rather straightforward. Formulas (2.19)-(2.23) can be directly applied to the pooled index where the subscript i, $i = 1, \ldots, n$, refers to an individual cluster (obligor).[29] Instead of using the subsamples where a single observation is omitted now the entire cluster of observations (all the observations of one obligor) has to be removed for the calculation of jackknife pseudo-values to account for the dependencies within clusters. The pooled indices are ratios of U statistics and therefore the jackknife can be applied along the lines of the single-cohort case. In contrast, the weighted average indices are weighted averages of dependent ratios of U statistics and the Delta method can only be applied if we estimate the autocorrelations of the time series of U statistics. This is a major drawback since we wanted to avoid estimating autocorrelations in the context of resampling methods. Of course, the jackknife can be directly

[29] In Formula (2.18), the double sum is then over all observations of the panel their number being $\sum_{i=1}^{n} T_i$ (T_i denoting the number of observations of obligor i.)

applied to the weighted average index,[30] but since this is not a U statistic the validity of the jackknife is not clear in this case. In the upcoming empirical illustration, we will therefore use the jackknife and the bootstrap for the pooled indices whereas we use the asymptotic Formulas (2.25)-(2.27) and the bootstrap for the weighted average indices. For the jackknife, as opposed to the bootstrap, we exploit the asymptotic normality of our indices for the purpose of confidence intervals and hypothesis tests.

We close this section with some remarks on measures of calibration. In section 2.2, we have introduced a χ^2 test statistic (Equation (2.16)) and the test statistic of a test for perfect calibration in a calibration model (Equation (2.17)) as measures of calibration accuracy. Both measures are not intended to be calculated as a weighted average over different cohorts (although this is possible as well). Rather, similarly to the pooled Accuracy Ratio and Harrell's C, they are based on the pooled dataset. Again, this means that the dependencies arising from overlapping lifetimes have to be taken into account. The cluster bootstrap is one option to do so and can be applied as explained just above. For the χ^2 statistic from Equation (2.16), the bootstrap can be avoided by using a cluster-robust estimator for the variance of the default rates in the validation sample ($\widehat{PD}^{VS}_{H,j}$). If the life-table or the Kaplan-Meier estimator is used to calculate default rates, such a cluster-robust variance estimator exists. It will be presented in detail in section 4.A. Some further discussion about statistical inference for measures of calibration is given in the context of our empirical calibration analysis in section 3.3.4.

2.5 Empirical illustration

For our empirical illustration in this section, we restrict ourselves to measures of discriminatory power. Calibration will be treated in chapter 3 since it requires a model for the default probability which will then be introduced. In the following, we will deal with the evaluation of credit ratings which - as

[30] For each subsample with the ith obligor omitted, $i = 1, \ldots, n$, the weighted average index is calculated and Formulas (2.19) and (2.20) are used with the weighted average index replacing Kendall's τ.

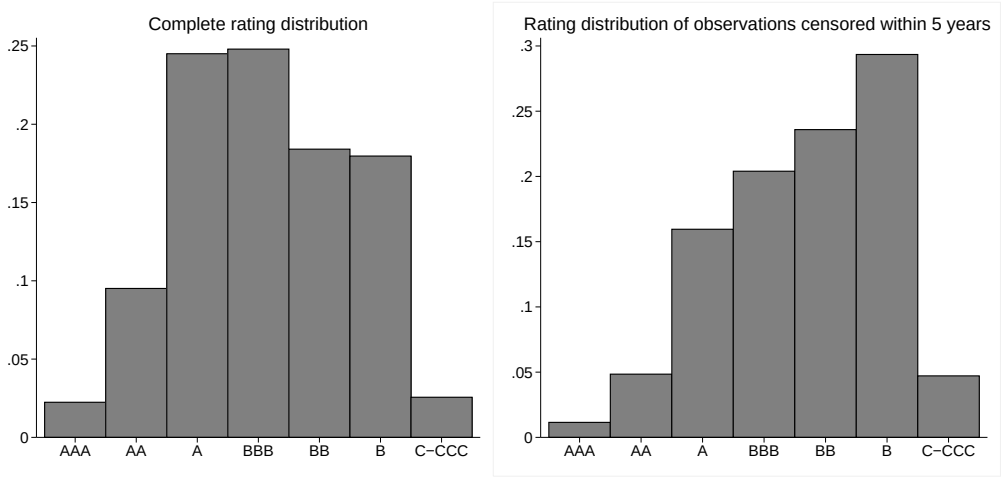
they are pure ordinal predictions - suffices as far as discriminatory power is concerned.

Our dataset consists of monthly Standard & Poor's (S&P) long term issuer credit ratings for North American public firms provided by Compustat. Long term ratings are particularly suitable in our context since the benefits from Harrell's C compared to the Accuracy Ratio get most visible in the evaluation of long term predictions. Note that the ratings used here do not refer to any specific security but rather are assessments of the overall solvability of a firm. Since S&P rating assignments are done "in real time" and are not adjusted "ex post", they are naturally out-of-sample predictions. We consider prediction horizons from six months up to five years which is the maximum time horizon of S&P's long term ratings (Standard & Poor's, 2010). After excluding missing observations we have 512 685 firm-months of 5151 firms in the period from December 1985 to June 2009, including 609 defaults. Cohort building is performed on a monthly basis starting in December 1985 until June 2004. Thus, our time series of indices consists of 223 periods.

Figure 2.4 shows the rating distribution of our sample on the left-hand side. To investigate the censoring scheme in our data, Figure 2.4 also shows on the right-hand side the rating distribution of the observations censored within five years. Clearly, lower rated firms have higher censoring rates so that the subsample of firms which were not censored within the first five years tends to contain primarily highly rated and defaulting firms. Recall that the five-year Accuracy Ratio (in contrast to Harrell's C) uses only this uncensored subsample which obviously has to some degree different characteristics than the whole sample. Apart from this finding, the censoring problem leads to a substantial loss of information as 30.99% of all observations are omitted for the five-year Accuracy Ratio.

We can see the consequences of these problems among other things in Table 2.1 which gives Accuracy Ratios and adjusted Harrell's C's together with their standard errors and 95% confidence intervals. Looking at the levels of the indices, we see that for both the weighted average and the pooled versions the Accuracy Ratio declines distinctively less with the prediction horizon than Harrell's C. In particular the five-year Accuracy Ratio is almost

Figure 2.4: Rating distribution of the full and censored sample



as high as the three-year Accuracy Ratio. This is most likely due to the aforementioned fact that, as the prediction horizon grows, the subsample relevant for the Accuracy Ratio tends to consist of "very good" and "very bad" firms making discrimination obviously easier. We can interpret this finding as a missing data problem. The subsample used for the Accuracy Ratio is the part of the sample without missing data (the complete cases in missing data terminology) and is – as Figure 2.4 shows – very likely not a random subsample of the whole sample. In such cases, a bias typically arises (Little & Rubin, 2002, Ch. 3.2). In our application, the Accuracy Ratios at long horizons are evidently upward biased and indicate a prognostic power of the rating system that is not really existent. Thus, investors and risk managers relying on the Accuracy Ratio are endangered to be too optimistic about the long-run predictive accuracy of ratings.

While Harrell's C declines more as the prediction horizon increases it is also lower than the Accuracy Ratio for short horizons where the missing data problem is rather unimportant. This can be explained by the fact that Harrell's C is more demanding in the sense that a perfect Harrell's C requires the correct prediction of the temporal order of defaults and not only the correct prediction of which firms will default within a certain time horizon.

Table 2.1: Indices of predictive accuracy, their standard errors and 95% confidence intervals

Panel A: Weighted average indices (number of firms per cohort as weights)

	Adjusted Harrell's C				Accuracy Ratio			
Prediction horizon (months)	6	12	36	60	6	12	36	60
Index	.8686	.8340	.7475	.7168	.8725	.8422	.7768	.7679
Formulas (2.25)-(2.27)								
Standard error	.0087	.0114	.0133	.0144	.0086	.0112	.0130	.0163
CI lower bound	.8515	.8116	.7214	.6886	.8557	.8202	.7514	.7360
CI upper bound	.8856	.8563	.7736	.7450	.8893	.8643	.8022	.7999
Cluster bootstrap								
Standard error	.0105	.0116	.0175	.0204	.0103	.0114	.0173	.0204
CI lower bound	.8436	.8074	.7111	.6715	.8477	.8165	.7406	.7237
CI upper bound	.8831	.8512	.7790	.7509	.8869	.8594	.8077	.8025

Panel B: Pooled indices

	Adjusted Harrell's C				Accuracy Ratio			
Prediction horizon (months)	6	12	36	60	6	12	36	60
Index	.8562	.8116	.7368	.7135	.8599	.8200	.7682	.7660
Cluster jackknife								
Standard error	.0106	.0115	.0141	.0155	.0106	.0114	.0141	.0157
CI lower bound	.8354	.7891	.7091	.6831	.8391	.7977	.7406	.7353
CI upper bound	.8769	.8340	.7645	.7440	.8806	.8424	.7959	.7967
Cluster bootstrap								
Standard error	.0111	.0114	.0144	.0152	.0107	.0113	.0140	.0157
CI lower bound	.8340	.7886	.7077	.6837	.8386	.7971	.7397	.7332
CI upper bound	.8773	.8325	.7640	.7424	.8798	.8413	.7958	.7985

The number of bootstrap replications is 1000. Bootstrap confidence intervals are calculated via the percentile method. Jackknife confidence intervals are calculated using jackknife standard errors and assuming normality.

Further, we see that the weighted average measures are generally higher than the pooled measures. This does not surprise as the weighted average indices aggregate measures for predictions made at certain points in time and do not compare ratings from different points of the business cycle as in the pooled cohort approach.

We now turn to the analysis of standard errors and confidence intervals. Regarding the weighted average indices, the asymptotic formulas derived in section 2.4.2 tend to be more liberal than the bootstrap which is a common

finding in such comparisons (Horowitz, 2001). While the suggested finite-sample bias of the asymptotic formulas seems to be moderate, the computational effort of the bootstrap might be worthwhile for more precise inference. For the pooled indices, the differences between the cluster jackknife and the cluster bootstrap are very small. Since the jackknife is computationally more efficient in this situation, we recommend its use for the pooled indices. Finally, looking at the standard errors and the length of the confidence intervals over time, it is obvious that the uncertainty about rating accuracy grows with the prediction horizon. Note that for a single cohort, the standard errors do not rise with the prediction horizon. However, for the aggregate indices they do, since the overlapping lifetimes problem is more pronounced in this case leading to higher dependencies in the data for longer horizons.

In section 2.4.2, we have argued that inference based on multiple cohorts including overlapping lifetimes extracts the maximum amount of information out of the dataset. From a statistical point of view, this leads to smaller standard errors, narrower confidence intervals and more powerful tests. We now demonstrate these improvements by example. The following test is motivated by the observation that the information that a firm reached its rating by a downgrade may be useful in predicting future defaults (Lando & Skodeberg, 2002; Guettler & Raupach, 2010). Thus, we created a rating scale that includes new additional grades for downgraded firms. For instance, we classify a firm that reached a BBB+ rating by a downgrade between the BBB+ firms that did not reach their rating by a downgrade and the firms which are one grade lower, in this case BBB. The null hypothesis of the test presented in Table 2.2 is that this augmented rating scale has the same predictive power as the original rating scale which is tested against the two-sided alternative. We use Harrell's C in the weighted average version as our measure of predictive accuracy. On the one hand, in the first four columns of Table 2.2, we perform the test using again monthly cohort building. On the other hand, we do the same test using only cohorts where the time between the cohort building dates is $H - 1$ months (H being the prediction

Table 2.2: Tests for significant differences in adjusted Harrell's C

	Overlapping lifetimes included				Overlapping lifetimes excluded			
Pred. horizon (months)	6	12	36	60	6	12	36	60
C_{adj}	.8686	.8340	.7475	.7168	.8678	.8318	.7466	.6780
$C_{adj}+$	.8695	.8350	.7482	.7173	.8686	.8322	.7467	.6781
Difference	9.92e-4	1.05e-3	7.52e-4	5.59e-4	8.24e-4	4.14e-4	1.04e-4	9.82e-5
Form. (2.25)-(2.27)								
St.err. diff.	1.71e-4	1.79e-4	1.29e-4	1.27e-4	3.39e-4	2.38e-4	2.54e-4	2.91e-4
p value	6.06e-9	4.09e-9	5.64e-9	1.06e-5	.0152	.0808	.6834	.7355
Cluster bootstrap								
St.err. diff.	1.81e-4	1.83e-4	1.12e-4	7.87e-05	4.03e-4	2.59e-4	2.73e-4	3.20e-4
p value	.001	.001	.001	.001	.038	.100	.668	.752

The columns under "Overlapping lifetimes included" refer to monthly cohort building. "Overlapping lifetimes excluded" columns use only data from cohorts which are separated by $H-1$ months where H is the prediction horizon. C_{adj} refers to adjusted Harrell's C in the weighted average version for the S&P fine-grained rating scale as in Table 2.1. $C_{adj}+$ augments the rating scale by an additional grade for firms who reached their rating grade by a downgrade. The equality of the population indices is tested against the two-sided alternative. The number of bootstrap replications is $B = 999$. Bootstrap p values are calculated according to Formula (2.29).

horizon) so that no overlapping lifetimes occur.[31] The latter case refers to inference which avoids to deal with the dependence induced by the overlapping lifetimes problem. We apply both the asymptotic formulas as described in section 2.4.2 and the cluster bootstrap to perform the tests. In the case of no overlapping lifetimes, Formula (2.25) (more precisely its extension to differences) reduces to its elementary part that does not include the terms which involve autocorrelations.

The results show that we can reject the null hypothesis at any horizon and at any conventional significance level if we include overlapping lifetimes. We conclude that the consideration of the downgrade effect indeed yields incremental predictive accuracy. However, such a conclusion is hardly possible without the use of overlapping lifetimes. In this case, we observe only marginally significant improvements at short horizons and no significant dif-

[31] For instance, for five-year Harrell's C, we use the cohorts build in June of 2004, 1999, 1994 and 1989.

ferences for longer horizons. One reason for this caused by chance is that the point estimates for the difference of the indices are lower throughout all horizons. The other and systematic reason is that the standard errors of the differences are considerably higher reflecting the higher variability that is caused by the reduction of the dataset. The results of the asymptotic formulas and the cluster bootstrap are quite similar. The test decisions are essentially the same for both approaches while the bootstrap again tends to be somewhat more conservative. To conclude, we see that there are realistic examples where the greater power of tests based on overlapping lifetimes results in different decisions.

Chapter 3

Default prediction with time-varying covariates

In this chapter, we deal with the common situation that the dataset has a panel structure and consists of the default histories for a set of obligors together with time-varying covariates. Of course, the cases of cross-sectional datasets or time-constant covariates are then just special cases. A typical example for time-varying covariates is given by firms, where balance sheet variables are updated over time. In the area of consumer credit, time-varying covariates are often gathered as well although sometimes only the data reported at the time of credit application are used in practice (Crook & Bellotti, 2010). For sovereign obligors, the covariates, typically macroeconomic and fiscal variables, are also time-varying, although default prediction models are not as common in this area as for other types of obligors due to sparse default data.

As was explained in the introduction, we will use a survival analysis approach for our default prediction problem. Over the past 15 years hazard models have emerged to become the state of the art in the credit risk literature.[1] Hazard models are formulated in terms of the hazard rate which is defined

[1] Early contributions in this respect are Lee & Urrutia (1996) and Banasik et al. (1999).

in continuous-time as

$$\lambda^c(y) = \lim_{\Delta y \to 0} \frac{P(y \leq Y^* < y + \Delta y \,|\, Y^* \geq y)}{\Delta y}\,, \tag{3.1}$$

where Y^* denotes again the possibly unobservable lifetime or the time until default of an obligor.[2] In discrete-time, the hazard rate is defined to be

$$\lambda^d(y) = P(Y^* = y | Y^* \geq y)\,. \tag{3.2}$$

In both cases, the hazard rate is a measure of instantaneous default risk which may be linked to a set of covariates to arrive at a hazard regression model. Interestingly, standard discrete-time hazard models can be shown to be equivalent to panel data models with a binary dependent variable (Sueyoshi, 1995; Jenkins, 1995).[3] Before the introduction of hazard models to the credit risk literature, most studies considered a simple cross-sectional setting with a binary dependent variable representing default.[4] In such a framework, only a small part of the information is used since i) only one cross-section of the panel data set is utilized and ii) the exact default times as well as iii) the information from censored lifetimes are not incorporated into the analysis.

In recent years, it has become a standard approach in the literature to estimate a discrete-time hazard model with yearly data directly yielding one-year default probabilities (Shumway, 2001; Chava & Jarrow, 2004; Hillegeist et al., 2004; Beaver et al., 2005; Hamerle et al., 2006; Cheng et al., 2010). While such a framework has the aforementioned benefits compared to cross-sectional analyses it still suffers from the fact that default predictions for time horizons of more than one year are not directly available since the future evolution of the covariates is unknown. Further, these models still do not use all information if data are available at shorter time intervals, say quarterly or monthly. Therefore, even for one-year horizons, models that allow for multi-period predictions are useful.

[2] As in chapter 2 the observed lifetime is denoted by Y. If the lifetime is uncensored it holds that $Y^* = Y$ and $Y^* \geq Y$ otherwise.

[3] Intuitively, a discretely measured lifetime is the result of a sequence of binary variables which are the default indicators for a sequence of periods.

[4] Methods used in such a context include discriminant analysis (Altman, 1968) and cross-sectional logistic regression (Ohlson, 1980).

The models considered in this chapter are sometimes referred to as reduced-form models as opposed to structural models. Structural models are based on some theory *why* an obligor defaults. For instance, in the popular structural model of Merton (1974),[5] the assets of a firm are assumed to follow a certain stochastic process and a default occurs if the asset values drop below the face value of the firm's debt. The main drawback of this and other structural models is that they require very strong assumptions, especially about the functioning of capital markets. Most likely because of this reason, recent studies that compared structural and reduced-form models in terms of their predictive accuracy concluded that appropriate reduced-form approaches yield superior predictions (Bharath & Shumway, 2008; Campbell et al., 2008).

In this chapter, we analyze and develop reduced-form approaches that allow for multi-period predictions in a panel data framework. In the next section, we will present approaches in the literature that overcome the problem of unknown future covariates by developing a model to forecast these covariates. Then, we will introduce a new alternative approach that delivers multi-period predictions within a parsimonious setting that does not need a covariate forecasting model. In section 3.3, we apply our estimators to a large sample of North American public firms, evaluate their predictive accuracy and show how the original estimates can be eventually recalibrated. The contents of the upcoming sections stem to a large extent from Orth (2011b).

3.1 Approaches with covariate forecasting models

We first introduce the study of Duffie et al. (2007). In an application to U.S. public firms, the authors use the following specification for the continuous-

[5] Further developments of Merton's model have led to the approach from Moody's KMV (Crosbie & Bohn, 2003) which has received considerable attention in the financial industry.

time hazard rate of firm i in period t, given a covariate vector x_{it}:[6]

$$\lambda^c(t|x_{it}) = \exp(\beta' x_{it}) \tag{3.3}$$

The model is a proportional hazard model with a constant baseline hazard and time-varying covariates. The authors use four covariates (taken from balance sheet and market data) which are modelled with Gaussian panel vector autoregressions, partly assuming independence among the covariates. Using this forecast model, the probability to default within a time horizon H can be calculated as follows:

$$P(Y_{it}^* \leq H|x_{it}) = 1 - E\left[\exp\left(-\int_t^{t+H} \lambda^c(t+s|x_{i,t+s})\, ds\right)\right], \tag{3.4}$$

Y_{it}^* denotes the lifetime of firm i starting at t and the expectation is with respect to the path of the vector of covariates from time t to $t+H$. Duffie et al. (2007) state that they evaluate this expression by numerical methods. To enable a better understanding of the necessary calculations let us do some rearrangements of Formula (3.4). Since the covariates are observed at discrete points in time the hazard rates are piecewise constant and the default probability can be rewritten as

$$P(Y_{it}^* \leq H|x_{it}) = 1 - E\left[\exp\left(-\sum_{s=1}^{H} \exp(\beta' x_{i,t+s})\right)\right] \tag{3.5}$$

$$= 1 - E\left[\prod_{s=1}^{H} \exp\left(-\exp(\beta' x_{i,t+s})\right)\right] \tag{3.6}$$

It is clear from above that the model for the covariate processes and the necessary numerical calculations get more and more involved as the number of covariates rises. Duffie et al. (2007) use only four covariates which is a relatively low number compared to other studies. An interesting approach to overcome this dimensionality problem is given in the study of Hamerle et al. (2007). In an application to German firms from the manufacturing industry, they choose a discrete-time hazard model of the following form:

$$\lambda^d(t|x_{it}) = \Phi(\beta' x_{it} + \varepsilon_t) \tag{3.7}$$

[6] In the following, we disregard the fact that Duffie et al. (2007) consider not only the exit of a firm due to default but also other exits like, for instance, mergers. We do so to make comparisons with other approaches easier.

$\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and ε_t is a normally distributed random variable that represents common unobserved shocks. To simplify covariate forecasting, Hamerle et al. (2007) partition the covariate vector in a firm-specific part, $\beta_1' x_{it,1}$, and a macroeconomic part, $\beta_2' x_{it,2}$, ($\beta' x_{it} = \beta_1' x_{it,1} + \beta_2' x_{it,2}$), which may be interpreted as a credit score and a macroeconomic default risk index, respectively. Instead of modeling all covariates, Hamerle et al. (2007) only specify a model for the credit score and the macroeconomic index thereby reducing the complexity of the problem considerably. To be specific, Hamerle et al. (2007) use univariate autoregressive panel and time series models assuming independence between $\beta_1' x_{it,1}$ and $\beta_2' x_{it,2}$. Similarly to Duffie et al. (2007), the default probability for a time horizon H is calculated by taking the expectation over the possible paths of the covariate processes:

$$P(Y_{it}^* \leq H | x_{it}) = 1 - E\left[\prod_{s=1}^{H} \left(1 - \lambda^d(t+s | x_{i,t+s})\right)\right] \tag{3.8}$$

Hamerle et al. (2007) approximate this expression by performing Monte Carlo simulations of their covariate processes.

Approaches that involve covariate forecasting models have some drawbacks that make it worthwile to look for alternatives. First, there is a considerable additional burden in model building, programming and computing time. To see this, note that panel vector autoregressions like the one used by Duffie et al. (2007) are usually not implemented in statistical software packages. Further, the evaluation of the expectations in Equations (3.4) and (3.8) requires multidimensional numerical integration which is often computationally demanding. Second, and maybe more importantly, there are purely statistical disadvantages since the econometrician is left with the choice between a large covariate forecasting model containing many parameters – which may lead to low out-of-sample predictive power – and quite restrictive assumptions to reduce dimensionality. As examples for the latter note that the model of Duffie et al. (2007) contains only four covariates and that in the approach of Hamerle et al. (2007) equal credit scores lead to equal credit score forecasts regardless of the composition of the covariate vector. More generally, since any errors in the forecasting model for the covariates will

impact the overall model's ability to predict defaults an additional source of model and parameter uncertainty arises.

3.2 An alternative approach

3.2.1 The models

As a background to the subsequently presented approach we briefly address the study of Campbell et al. (2008) which also deals with North American public firms. In their work, the authors estimate discrete-time hazard models (using a Logit specification) lagging their time-varying covariates by s months, $s = 6, 12, 24, 36$:

$$\lambda^d(t+s|x_{it}) = [1 + \exp(-\beta_s' x_{it})]^{-1} \tag{3.9}$$

The authors point out that this approach can be extended by letting s run from 1 to H (H denoting the prediction horizon) meaning a stepwise increase of the lag index in the hazard regressions. Then, multi-period default probabilities can be calculated in closed form since the hazard rate in period $t+s$ is directly given as a function of the covariates in period t:

$$P(Y_{it}^* \leq H|x_{it}) = 1 - \prod_{s=1}^{H} \left(1 - \lambda^d(t+s|x_{it})\right) \tag{3.10}$$

However, estimation gets a bit cumbersome since one has to estimate H different parameter vectors which also increases the numbers of parameters substantially and thereby raises questions about out-of-sample predictive power. While Campbell et al. (2008) do not perform and validate such an extended approach,[7] it nevertheless provides an interesting means to overcome the burden to specify a covariate forecasting model. Therefore, we will consider this approach in the empirical analysis. Before we do so, we will now introduce a framework that does not need a covariate forecasting model as well and thereby involves the estimation of just one parameter vector.

[7] In a very recent study, Duan et al. (2012) employ such kind of sequential lagging procedure using a complementary log-log instead of a Logit specification.

Let us first introduce the basic notation. We observe obligor i, $i = 1, \ldots, n$, for T_i periods thereby recording his default history and a vector of time-varying covariates x_{it}. Importantly, we define Y_{it} to be the observed lifetime of obligor i starting in period t, for each period t, $t = t_{i1}, \ldots, t_{i1} + T_i - 1$, so that we have T_i partially overlapping lifetimes for each obligor. In real datasets we will not observe the end of every lifetime, so that we have to define additionally the corresponding censoring indicator variable C_{it} which is zero in the case of no censoring, i.e. the lifetime ends with a default event, and one for censored lifetimes. Further, Y_{it}^* again denotes the uncensored and sometimes unobservable lifetime, i.e. $Y_{it} = Y_{it}^*$ if $C_{it} = 0$ and $Y_{it} \leq Y_{it}^*$ if $C_{it} = 1$. We will specify our models in terms of the continuous-time hazard rate. We choose the continuous-time specification since it is more common in the survival analysis literature and gives us a greater variety of models to choose from. Additionally, software packages usually offer more implementations for continuous-time hazard models.[8]

The idea behind the models we propose is as follows. Given the information, i.e. the covariates, available at some point in time t, we want to predict the probability to default within a time horizon of H. A simple solution is to specify the hazard rate at point in time $t + s$ as a function of the covariates in period t, x_{it}, and the "forecast time" s. To do so we define, in contrast to all of the aforementioned studies, our hazard rate in terms of the lifetimes starting at the variable point in time t:[9]

$$\lambda^c(s|x_{it}) = \lim_{\Delta s \to 0} \frac{P(s \leq Y_{it}^* < s + \Delta s | Y_{it}^* \geq s, x_{it})}{\Delta s} \tag{3.11}$$

We may call $\lambda^c(s|x_{it})$ the time-t conditional hazard rate at time $t + s$ or the time-s ahead hazard rate given the information available at t. Under this definition, we can use, for instance, the proportional hazard (PH) framework to specify our model. Then,

$$\lambda^c(s|x_{it}) = \lambda_0(s) \exp(\beta' x_{it}) \,. \tag{3.12}$$

[8]Standard discrete-time hazard models can be estimated by routines for binary panel models. However, this is not possible for our kind of models since we deal with a panel of lifetimes and not with a panel of binary variables.

[9]The studies cited before consider only one lifetime per obligor which is implicitly defined to start at the beginning of the observation period.

$\lambda_0(s)$ is called the baseline hazard and captures the variation of the hazard rate over the forecast time which may also be interpreted as a kind of duration dependence. The model given by Equation (3.12) is essentially an ordinary hazard model. The innovative part is simply the specification of the hazard rate in terms of the forecast time s (instead of the calendar time t) which allows us to use ordinary hazard models to express the evolution of default risk over the forecast time. Note the difference to the usual PH specification as, for instance, used in Duffie et al. (2007). There, the hazard rate in period $t+s$ is a function of the covariates in period $t+s$ leaving those models with the problem that the covariates are not known in $t+s$.[10] Further, notice that the forecast time s is the analogon to the lag length in the approach of Campbell et al. (2008) outlined in the beginning of this section. There, the hazard rate freely fluctuates for different s due to the repeated estimation of the model. In contrast, we impose a structure on the evolution of the hazard rate over the forecast time by integrating s as an argument into the functional form of the model.

Importantly, the default probabilities are easily calculated in closed form:

$$P(Y_{it}^* \leq H|x_{it}) = 1 - \exp\left(-\int_0^H \lambda^c(s|x_{it})\, ds\right) \tag{3.13}$$

For instance, if we choose the PH model we have

$$P(Y_{it}^* \leq H|x_{it}) = 1 - \exp\big(-\exp(\beta' x_{it})\Lambda_0(H)\big)\,, \tag{3.14}$$

where $\Lambda_0(H) = \int_0^H \lambda_0(s)ds$ is the so-called cumulative baseline hazard.

Within our approach we neither claim that only x_{it} (and not $x_{i,t+s}$) is relevant for the hazard rate in period $t+s$ nor do we think that the vector of covariates is not forecastable to some degree. Rather, we argue that the analysis can be simplified by tailoring the model directly for its purpose, namely to deliver multi-period predictions.[11] As we will see in the empirical section, this simplicity does not come at the cost of low predictive accuracy.

[10] Under our definition of the hazard rate, the model of Duffie et al. (2007) would read as $\lambda^c(0|x_{it}) = \exp(\beta' x_{it})$.

[11] Our approach has some similarities with an idea from the time series literature called direct multi-step estimation. See Clements & Hendry (1998, Ch. 11) for details.

PH models have received great popularity not least because it is possible to estimate β without specifying the baseline hazard. This approach, developed by Cox (1972), can be followed by a nonparametric estimation of the baseline hazard and is thus often called semiparametric. Alternatively, one may use a fully parametric PH model like, for instance, the Weibull model for which $\lambda_0(s) = \gamma s^{\gamma-1}$ with parameter γ. In any case, the PH model in our version implies that the hazard ratios for two obligors i and j, $\lambda^c(s|x_{it})/\lambda^c(s|x_{jt})$, are constant with respect to the forecast time s. There is evidence in the literature that this assumption is not realistic at least in the area of corporate credit. Fons (1994) finds that marginal default rates (which are estimates of dicrete-time hazard rates) tend to rise with forecast time for low-risk investment-grade firms whereas marginal default rates tend to decrease for high-risk speculative-grade firms. The empirical evidence given in Figure 3.1 creates a similar picture. There, we have plotted nonparametric estimates of the hazard ratios for firms having CCC-C and A ratings on the left hand side and firms having B and BBB ratings on the right hand side.[12] Obviously, hazard ratios are declining and are not constant over the forecast time.

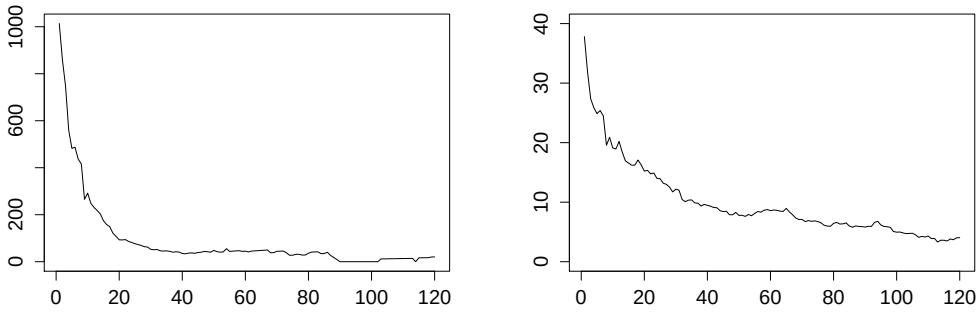
An intuitive interpretation for this phenomenon is that the importance of the information in period t decays with the forecast time s. Fortunately, there is a class of hazard models that covers the case of converging hazard rates. Proportional odds (PO) models generally imply that the hazard ratios converge monotonically towards one (Bennett, 1983) where the convergence is with respect to the forecast time s in our setting. In PO models the survival odds and not the hazard rates of two firms are constant multiples of each other:

$$\frac{P(Y_{it}^* \leq H|x_{it})}{1 - P(Y_{it} \leq H|x_{it})} \propto \frac{P(Y_{jt} \leq H|x_{it})}{1 - P(Y_{jt} \leq H|x_{it})} \tag{3.15}$$

The most common PO specification is the log-logistic model where the condi-

[12]For these calculations, we have used the dataset which will be introduced in section 3.3. The hazard rates which underlie the hazard ratios are calculated by the nonparametric life-table estimator which will be presented in section 4.1. Consequently, the finding of declining hazard ratios is not caused by any parametric assumptions.

Figure 3.1: Hazard ratios for different rating grades



Hazard ratios for firms having CCC-C and A ratings on the left hand side and firms having B and BBB ratings on the right hand side. Calculations are based on monthly discrete-time hazard rates estimated by the life-table method presented in section 4.1. The firms included in these calculations are North American public firms (see section 3.3).

tional distribution of Y_{it} is assumed to be log-logistic.[13] Then, in our framework the hazard rate is given by

$$\lambda^c(s|x_{it}) = \frac{\alpha \, \exp(\beta' x_{it})^\alpha s^{\alpha-1}}{1 + [\exp(\beta' x_{it})s]^\alpha} \, . \tag{3.16}$$

α is a shape parameter whereas $\exp(\beta' x_{it})$ determines the scale of the distribution with $\exp(-\beta' x_{it})$ being the median lifetime. The cumulative distribution function evaluated at H (yielding the default probabilities) under this model is

$$P(Y_{it} \leq H|x_{it}) = 1 - \left(1 + [\exp(\beta' x_{it})H]^\alpha\right)^{-1} \, . \tag{3.17}$$

Note that the log-logistic model belongs to the class of accelerated failure time models which have the interpretation that lifetimes are stretched or contracted by some constant acceleration factor. While the model is fully parametric, there also exist semiparametric specifications for the PO model. For instance, Royston & Parmar (2002) use cubic splines thereby achieving

[13] Very similar to the log-logistic distribution is, of course, the log-normal distribution which, however, does not have the proportional odds property.

similar flexibility for the trajectory of the hazard function as in the Cox model. In our empirical analysis, we also experimented with this approach. However, this did not lead to improved predictive accuracy.[14] Thus, we do not document it further here.

3.2.2 Estimation

For the models we propose, the observed lifetimes starting at t, Y_{it}, are simply connected to the covariates in period t, x_{it}. Clearly, the multiple lifetimes of an individual obligor are not conditionally independent, i.e. Y_{it} is not conditionally independent from Y_{it^*}, $t \neq t^*$. To see this, note that for instance Y_{it} already covers the lifetime $Y_{i,t+1}$ plus one additional period so that we have a sample of partially overlapping lifetimes. The reason why $Y_{i,t+1}$ is included although it is completely covered by Y_{it} is that the covariates vary from period t to $t+1$ and provide additional information. For the purpose of point estimation, it is possible to ignore the dependencies due to our overlapping sample and still to consistently estimate the parameters. This is a result from multivariate survival analysis (Lawless, 2003, Ch. 11) where the asymptotics only require that the lifetimes of different obligors are conditionally independent and that the number of obligors (n) approaches infinity. No assumptions are made about the dependence structure within the lifetimes of an individual obligor. With respect to the censoring mechanism, we need the assumption that censoring events are conditionally independent from default events.[15] Our pseudo log likelihood function is

$$\log L = \sum_{i=1}^{n} \sum_{t=t_{i1}}^{t_{i1}+T_i-1} (1 - C_{it}) \cdot \log(\lambda^c(Y_{it}|x_{it})) + \log(S(Y_{it}|x_{it}))\,, \qquad (3.18)$$

[14] The reason for this is arguably the fact that we measure predictive accuracy primarily in terms of an accurate risk ordering of the obligors which is usually invariant to changes in the shape of the hazard function.

[15] For instance, if, conditionally on the covariates, smaller firms would have higher default risk and earlier censoring times, this would only be a violation of our assumption if firm size is not included as a covariate. More formally, we require that the distribution of Y_{it}^*, conditionally on x_{it}, is not changed if one additionally conditions on C_{it}. See Wooldridge (2002, Ch. 20.3.2) and Lawless (2003, Ch. 2.2.2) for further discussion.

where $S(\cdot) = 1 - F(\cdot)$ is the so-called survival function referring to the cumulative distribution function $F(\cdot)$ of Y_{it}^*. In many applications the assumption of conditional independence of the lifetimes of different obligors may at best be approximately true because of common shocks which jointly affect the obligors over the forecast time and which are not reflected in the covariates at the start of the lifetime. However, results on Maximum Likelihood estimation under multi-way clustering show that an additional clustering (dependence) within the time dimension does not lead to inconsistency of our estimator (Cameron et al., 2011).[16] The question remains whether an approach that explicitly models the different sources of dependencies would be favorable. While such an approach is theoretically more efficient, efficiency gains are often found to be small (Joe, 1997, Ch. 10.1.2) and the computational burden may rise considerably. Further, misspecified dependence models can lead to inconsistent estimates so that our "independence working" approach is more robust in this sense (He & Lawless, 2003; Sullivan Pepe & Anderson, 1994). Nevertheless, we partly investigated this issue empirically by introducing dummy variables for each year which should capture common shocks to a large extent. We found – similar to Campbell et al. (2008) – no important effects on our results. Finally, the high out-of-sample predictive power of our models (the central objective of our analysis) which will be reported in the upcoming section provides further support for our approach.

For the estimation of the log-logistic model, we simply substitute the definitions of the hazard rate and the survival function as given in the preceding section into Equation (3.18). For the semiparametric Cox model, the likelihood of Equation (3.18) is not applicable. Instead, the approach is as follows. Suppose that we have r distinct values of uncensored lifetimes in our sample, $Y_{(1)}, \ldots, Y_{(r)}$ and, for the moment, assume that there are no ties, i.e. there is only one lifetime ending at $Y_{(j)}$. Further, denote by $R(Y_{(j)})$ the set of observations with a lifetime of at least $Y_{(j)}$ (those "at risk" at $Y_{(j)}$). Then, following Cox (1972, p. 191), the probability that the particular failure at

[16] The asymptotics in this case, however, require that the time dimension approaches infinity as well. In our empirical analysis the sample length is 352 months.

$Y_{(j)}$ is as observed, conditional on the composition of the risk set, is

$$\frac{\lambda^c(Y_{(j)}|x_{(j)})}{\sum_{l\in R(Y_{(j)})}\lambda^c(Y_{(j)}|x_l)} = \frac{\exp(\beta' x_{(j)})}{\sum_{l\in R(Y_{(j)})}\exp(\beta' x_l)}\,. \tag{3.19}$$

Note that we use the single index l to abbreviate the notation although we are still in a panel data setting so that $l \in \{1,\ldots,\sum_{i=1}^n T_i\}$. Taking logarithms the expression gives the log likelihood contribution

$$\beta' x_{(j)} - \log\Big(\sum_{l\in R(Y_{(j)})}\exp(\beta' x_l)\Big)\,. \tag{3.20}$$

Now, we allow for ties and denote their number by $d_{(j)}$, i.e. $d_{(j)} = \sum_{i=1}^n \sum_{t=t_{i1}}^{t_{i1}+T_i-1} 1_{[Y_{it}=Y_{(j)},C_{it}=0]}$. Under the Breslow approximation (Breslow, 1974), we simply sum up the log likelihood contributions of all observations ending at $Y_{(j)}$ and then take the sum over all r distinct failure times to arrive at our pseudo log partial likelihood:

$$\log L^p = \sum_{j=1}^r \beta' z_{(j)} - d_{(j)}\log\Big(\sum_{l\in R(Y_{(j)})}\exp(\beta' x_l)\Big) \tag{3.21}$$

Here, $z_{(j)}$ is the sum of the covariate vectors of all observations that ended with a default at $Y_{(j)}$, i.e. $z_{(j)} = \sum_{i=1}^n \sum_{t=t_{i1}}^{t_{i1}+T_i-1} 1_{[Y_{it}=Y_{(j)},C_{it}=0]} x_{it}$. Within the Breslow approach, the original partial likelihood, developed by Cox (1972) for the case without ties, is simply left unchanged meaning that there is no adjustment to the risk set in the presence of tied lifetimes.[17]

Maximizing the likelihood (3.21) gives an estimate of β which suffices to determine the relative risk of different obligors. If default probabilities are desired as well, an estimate of the baseline survivor function is needed. From (3.14) it follows that for the PH model the default probabilities are given by

$$P(Y_{it}^* \le H|x_{it}) = 1 - S_0(H)^{\exp(\beta' x_{it})}\,, \tag{3.22}$$

[17]Besides the Breslow approximation, there also exist other methods to handle ties, especially Efron's approximation (Efron, 1977). In our empirical analysis, we used both methods and found that Spearman's rank correlation coefficient for the predictions derived from both approaches is equal to 0.99999991 in our final model. Therefore, all of our reported results are based on the computationally faster Breslow approximation.

where $S_0(H) = \exp(-\int_0^H \lambda_0(s)ds)$ is the baseline survivor function. Kalbfleisch & Prentice (1973) propose to specify the baseline survivor function as $S_0(Y_{(k)}) = \prod_{j=1}^{k} \alpha_j$ at the observed failure times and being constant in between. $\alpha_j, j = 1, \ldots, r$ are parameters to be estimated with $1 - \alpha_j$ being interpretable as a discrete-time hazard rate for firms with $x_{it} = 0$. Using this specification and the estimate of β already obtained, the log likelihood function can be written as

$$\log L = \sum_{j=1}^{r} \Big(\sum_{i,t:Y_{it}=Y_{(j)},C_{it}=0} \log\left(1 - \alpha_j^{\exp(\widehat{\beta}' x_{it})}\right) + \sum_{i,t:Y_{it}>Y_{(j)}} \log\left(\alpha_j^{\exp(\widehat{\beta}' x_{it})}\right) \Big) \tag{3.23}$$

The first sum of the term in brackets is over all observations which ended with a default at $Y_{(j)}$ while the second sum is over those observations which were at risk before the jth period but did not default in that period.[18] See Kalbfleisch & Prentice (2002, Ch. 4.3) for a detailed derivation of this likelihood. Obviously, each α_j can be estimated separately. The estimator is a nonparametric Maximum Likelihood estimator and can be interpreted as a generalization of the Kaplan-Meier estimator (Kaplan & Meier, 1958) since it reduces to the latter if no covariates are included.[19]

We now turn to the issue of finding appropriate covariance matrices for our estimators. While we can consistently estimate our models under the working independence approach, the dependencies due to overlapping lifetimes must not be ignored for covariance matrix estimation. In particular, unadjusted standard errors would be much too low. Instead, if we view all the lifetimes of an individual obligor as one cluster, we can apply cluster-robust covariance matrix estimation. Let $\widehat{V}_H = \left(-\frac{\partial^2 \log L}{\partial\beta\partial\beta'}|_{\widehat{\beta}}\right)^{-1}$ be the conventional covariance matrix estimator based on the Hessian of the log likelihood function. Further, denote by $s_i(\widehat{\beta})$ the contribution of obligor i to the score vector.[20] Then, the

[18] In constrast, $R(Y_{(j)})$ was defined to be the entire risk set also including those observations that ended with a default in the jth period.

[19] See section 4.1 for more details about the Kaplan-Meier estimator.

[20] $s_i(\widehat{\beta}) = \sum_{t=t_{i1}}^{t_{i1}+T_i-1} s_{it}(\widehat{\beta}) = \sum_{t=t_{i1}}^{t_{i1}+T_i-1} \partial \log L_{it}/\partial\beta|_{\widehat{\beta}}$, where (see (3.18)) $\log L_{it} = (1 - C_{it}) \cdot \log(\lambda^c(Y_{it}|x_{it})) + \log(S(Y_{it}|x_{it}))$ for fully parametric models like the log-logistic model.

cluster-robust covariance matrix estimator is

$$\widehat{V}(\widehat{\beta}) = \widehat{V}_H \left(\sum_{i=1}^{n} s_i(\widehat{\beta}) s_i(\widehat{\beta})' \right) \widehat{V}_H \,. \tag{3.24}$$

Again, the estimator is consistent for $n \to \infty$. While the calculation of score contributions is straightforward for fully parametric models like the log-logistic model, the score contributions are not immediately available for the Cox model. However, cluster-robust covariance matrix estimation is possible for the Cox model as well as was shown by Wei et al. (1989). The basis of the calculations is an additive decomposition of the partial log likelihood function, $\partial \log L^p / \partial \beta|_{\widehat{\beta}} = \sum_{i=1}^{n} \sum_{t=t_{i1}}^{t_{i1}+T_i-1} W_{it}(\widehat{\beta})$, where

$$W_{it}(\widehat{\beta}) = (1 - C_{it})(x_{it} - \overline{x}_{it}) - \exp(\widehat{\beta}' x_{it}) \sum_{k:Y_k \leq Y_{it}} \frac{(1 - C_k)(x_{it} - \overline{x}_k)}{\sum_{l \in R(Y_k)} \exp(\widehat{\beta}' x_l)} \tag{3.25}$$

with

$$\overline{x}_{it} = \frac{\sum_{l \in R(Y_{it})} \exp(\widehat{\beta}' x_l) x_l}{\sum_{l \in R(Y_{it})} \exp(\widehat{\beta}' x_l)} \tag{3.26}$$

and $\overline{x}_k$ analogously defined. The W_{it} terms are sometimes called score residuals (Therneau et al., 1990) and are uncorrelated across different clusters which is not true for other decompositions (Therneau & Grambsch, 2000, Ch. 4.5). A consistent cluster-robust covariance matrix is given by

$$\widehat{V}(\widehat{\beta}) = \widehat{V}_H \left(\sum_{i=1}^{n} W_i(\widehat{\beta}) W_i(\widehat{\beta})' \right) \widehat{V}_H \quad , \quad W_i(\widehat{\beta}) = \sum_{t=t_{i1}}^{t_{i1}+T_i-1} W_{it}(\widehat{\beta}) \,. \tag{3.27}$$

The implementation of our models is easy. If for every observation of the panel dataset the lifetime Y_{it} and the corresponding censoring indicator C_{it} is calculated, standard survival analysis routines for time-constant covariates can be employed.[21] An option for cluster-robust standard errors is also available in many software packages. There is a final point to note about the

[21] Although we have time-varying covariates, the covariates can be regarded as pseudo-constant over the forecast time since we explicitly decide not to update the information on the covariates. Doing so would result in the problem of unknown covariates as far as forecasting is concerned. Nevertheless, the full panel of covariates is used in the estimation process since they are linked to a panel of lifetimes.

definition of the lifetimes Y_{it}. Given that we usually assume a limited prediction horizon, H, it may be sensible to conduct (again) an artificial censoring of the lifetimes at H thereby omitting possibly irrelevant information about what happened after H. For instance, with H equal to 60 months, we would set - along the lines of chapter 2 - a value of 60 to all lifetimes larger than 60 together with a change in the censoring indicator if the lifetime ended with a default event before. Empirical tests show that while the differences are rather small it is indeed preferable to conduct such an additional censoring.

3.2.3 Extensions to mixture models

The models presented above are relatively simple so that extensions are easily possible. One such extension is the specification of mixture models which have received considerable popularity in the survival analysis literature. Following Mosler (2003), reasons for mixture models would be especially (a) unobserved heterogeneity among obligors possibly caused by unobservable covariates and (b) the possibility to specify flexible parametric models by using mixtures. Clearly, it is reasonable to assume in credit risk applications that unobserved heterogeneity is present even after conditioning on a set of covariates. Further, flexible parametric models may also be helpful. A popular way to incorporate unobserved heterogeneity is to specify an additional random variable V (often referred to as frailty) that enters the hazard rate multiplicatively, which in our setting gives

$$\lambda^c_{SS}(s|v_{it}, x_{it}) = v_{it}\lambda^c(s|x_{it})\,. \tag{3.28}$$

We use the index SS in Equation (3.28) to clarify that the hazard rate should be interpreted as a so-called subject-specific hazard rate, i.e. conditional on the unobservable v_{it}. In contrast, if we integrate out the frailty variable V_{it} we arrive at the so-called population-averaged or marginal model,

$$\lambda^c_{PA}(s|x_{it}) = -\frac{S'_{PA}(s|x_{it})}{S_{PA}(s|x_{it})}\,, \tag{3.29}$$

$$S_{PA}(s|x_{it}) = \int_0^\infty \exp\left(-\int_0^s \lambda^c_{SS}(u|x_{it})du\right) dG(v)\,. \tag{3.30}$$

$G(\cdot)$ denotes the cumulative distribution function of V_{it}. Importantly, the subject-specific and the population-averaged model have different interpretations. While the subject-specific model has the aforementioned conditional interpretation, "the marginal survival function under heterogeneity is the expected survival function of a randomly drawn individual from a heterogeneous population." (Xue & Brookmeyer, 1997, p. 1987). With respect to our application, these marginal survival functions yield the default probabilities and their accurate estimation is the central objective of our analysis. Therefore, as opposed to other applications, we have no direct interest in a subject-specific model. The question remains if it is preferable to start with a subject-specific model and to integrate out the frailty variable afterwards or if one should only specify a population-averaged model. Investigating this issue, Xue & Brookmeyer (1997) show in a discrete-time framework that it suffices to specify an appropriate population-averaged model in the way that it yields the same results as estimating the subject-specific model and integrating out the frailty variable afterwards.[22] In our application, it might be easier to specify an appropriate population-averaged model since, for instance, the empirical findings about declining hazard ratios that led us to the proportional odds model related to the default behaviour of the "population average" of possibly heterogeneous firms in certain rating categories. Nevertheless, mixture models may still be interesting for the purpose of credit default prediction due to their ability to provide flexible survival distributions. For instance, in the classical case that $E[V_{it}] = 1$ for identifiability and that $V[V_{it}] = \theta$ is the only free parameter of $G(\cdot)$, the population-averaged hazard function contains one additional parameter (θ) and is thus more flexible than comparable approaches without frailty.

It follows from the discussion given above that it is largely an empirical question if a mixture model leads to more predictive accuracy. In the credit risk literature, unobserved heterogeneity across firms is so far very rarely incor-

[22] Similarly, Nicoletti & Rondinelli (2010) show in a Monte Carlo study that when unobserved heterogeneity is neglected in a discrete-time survival model there is no major bias as far as the marginal survival probabilities are concerned.

porated into the analysis.[23] Exceptions are the studies of De Leonardis & Rocci (2008) and Crook et al. (2011) where the results are mixed as the first study finds improvements in predictive accuracy whereas the second study does not. In our empirical analysis, we also experimented with adding a Gamma distributed multiplicative frailty term to the log-logistic model. We considered the cases that the frailty variable is allowed to vary across all firm-months and the special case of a firm-specific constant (shared) frailty, i.e. $V_{it} = V_i$. While the predictive accuracy in the first case was comparable to not modeling heterogeneity,[24] the predictive accuracy was somewhat lower under the second specification. Given these findings and the additional computational burden caused by mixture models, there seem to be no major benefits from the mixture approach in our application. Thus, the results that we will present in the upcoming section are based only on models without frailty.

3.3 Empirical analysis

3.3.1 Data description and model specification

To construct our dataset for the empirical analysis, we have merged three different datasets all of them referring to North American public firms. First, we collect monthly Standard & Poor's (S&P) rating and default data from Compustat. The dataset contains three types of ratings: Long term issuer credit ratings, short term issuer credit ratings and subordinated debt ratings with most data of the first type. We define default in our study to be a

[23] Frailty specifications have received some popularity for modeling the dependence of default events. In these models, $V_{it} = V_t$, so that the heterogeneity across periods is addressed only. In Duffie et al. (2009) and Koopman et al. (2008) V_t follows an autoregressive process and can be interpreted as an unobservable common risk factor.

[24] Especially, the ranking of the firms according to their default risk changed very little as Spearman's rank correlation coefficient for the predictions from the log-logistic model with and without frailty was equal to 0.9987.

default rating (D or SD) by S&P in any of the three rating types.[25] Consequently, a firm is defined not to be in the default status in a given period if it does not have a default rating of any type and has a non-default rating of at least one type. We then merge the default histories with quarterly balance sheet data from Compustat and monthly stock market data from Compustat and the Center for Research in Security Prices (CRSP). The balance sheet variables are taken to be constant over the months between financial statements so that the final dataset has monthly time intervals. Since there are on average two months between the end of the corresponding fiscal period and the reporting date we lag the balance sheet variables by two months so that the values of the variables should have been indeed available in each month. Further, following common practice we exclude financial firms (Standard Industrial Classification (SIC) codes 6000-6799) since these are assumed to be structurally different. In a study concerning a very similar dataset as ours, Chava & Jarrow (2004) find that predictive accuracy is higher when financial firms are omitted from the sample.

In the data preparation process, we had to deal with both missing data and outliers. With respect to missing data, we imputed missing values for some variables based on regressions on their leads and lags.[26] The main criterion was that the goodness-of-fit of such regressions is very high. For instance, the variable total assets can be very accurately predicted from past and future values whereas returns are known to be hard to predict. Consequently, we used imputations only for "stable" variables like total assets and did not impute any values for variables like returns or net income.[27] Using such imputations will usually result in more efficient estimation. However, standard errors can be expected to be too low due to the reduced variability of imputed values (Harrell, 2001, Ch. 3.6). Since the share of missing values is rather low (no variable of our final model is missing in more than 10% of all cases)

[25] If a firm defaults on all its securities it receives a D (Default) rating while it is rated SD (Selective Default) if the default event applies only to selected securities.

[26] This method is often called single conditional mean imputation (Harrell, 2001, Ch. 3).

[27] The variables where imputations were used are total assets, cash and short-term investments, market value, interest expenses, retained earnings and total liabilities. The cases where missing variables remained had to be dropped from the subsequent analysis.

and since our focus is on prediction rather than on inference, efficiency gains should be more important.

To eliminate the effect of outliers, we winsorized all variables at the 5th and 95th percentile. An inspection of the data showed that implausible values ("wrong signs") occasionally occur pointing to a need for winsorization. We further fitted our models to the data before and after winsorization and observed a remarkably better goodness-of-fit for the winsorized dataset. By winsorizing the data we follow the related literature where the use of this procedure is very common. The final dataset consists of 339 222 firm-months from 3575 firms in the period from December 1980 until March 2010. We observe 498 different default events, but note that our definition of Y_{it} leads to 18 914 partially overlapping lifetimes in our sample that end with a default event.

For the selection of our covariates, we used the experience from studies based on similar datasets (Shumway, 2001; Chava & Jarrow, 2004; Duffie et al., 2007; Campbell et al., 2008; Löffler & Maurer, 2011) to choose candidate variables. Table 3.1 is a list of the covariates considered together with descriptive statistics. The final specification of our models was derived by a backward selection approach that entailed the sequential reduction of the model containing all candidate variables.[28] As the main criteria in the model selection process we used the Wald statistics and the associated p values of the covariates since we have to be careful with likelihood ratio tests and information criteria in a pseudo likelihood setting. The liquidity variable (CATA) as well as retained earnings (RETA) were found to be insignificant (with p values larger than 0.5) so that we did not include them in the final model although the signs of the coefficients were as theoretically expected. Interest coverage (NII) was found to be significant but is strongly correlated with profitability (NITA). Due to this finding and the fact that the share of missing values was considerably higher for NII (16.6% vs. 3%) we dropped NII. For the covariates of the final model, all correlations are below 0.5 (see

[28] When deciding between forward and backward selection one must weigh up potential biases arising from starting with a very simple model against potential data mining problems when a very large model is the starting point (Greene, 2008, Ch. 7.2.4). Since the set of candidate variables is moderate in our analysis, we decided to use backward selection.

Table 3.1: Summary statistics for covariates

Name	Description	Mean	St.dev.	Min	Max
Selected for final model					
NITA	Net income over previous year / Total assets	.007	.020	-.155	.079
TLTA	Total liabilities / Total assets	.636	.168	.115	1
GRO	Dummy for extreme growth of total assets	.5	.5	0	1
RET	Excess one-year log stock return over S&P 500	-.029	.367	-1.317	1.220
VOLA	St. dev. of monthly log returns in previous year	.110	.063	.039	.298
SIZE	Log(market value / S&P 500 total market value)	-8.99	1.72	-13.27	-6.34
Not selected for final model					
CATA	Cash and short-term investments / Total assets	.071	.085	.001	.344
RETA	Retained earnings / Total assets	.134	.255	-.582	.537
NII	Net income / Interest expenses over previous year	2.797	5.513	-3.993	25.295

Table 3.2: Correlations of covariates

	NITA	TLTA	GRO	RET	VOLA	SIZE
NITA	1.000					
TLTA	-0.343	1.000				
GRO	-0.221	0.107	1.000			
RET	0.255	-0.103	-0.061	1.000		
VOLA	-0.438	0.201	0.261	-0.266	1.000	
SIZE	0.380	-0.278	-0.171	0.280	-0.431	1.000

Table 3.2) so that multicollinearity should not pose a problem. Further, we looked for possible non-monotone effects of the variables on the hazard rate by grouping the covariates into quartiles and including the corresponding dummy variables into our model. We found strongly non-monotone effects for growth of total assets. Both high and low (highly negative) growth rates are associated with higher default risk. Therefore, our final model contains a dummy variable which is one if annual growth of total assets is in the upper or lower quartile and zero otherwise. The other covariates are quite standard and are used in this way or very similarly in the aforementioned studies.

3.3.2 Estimation results

We now turn to our estimation results. Table 3.3 shows the parameter estimates for the Cox model and the log-logistic model.[29] The results refer to lifetimes which have been (additionally) censored at 60 months as described at the end of section 3.2.2. All coefficients have the expected sign and are highly significant. Higher profitability (NITA), higher stock market returns (RET) and larger firm sizes (SIZE) are associated with lower default risk. On the opposite side, higher debt levels (TLTA), extreme growth (GRO) and more volatile returns (VOLA) correspond to higher hazard rates. As we estimate the Cox model with the partial likelihood approach, we do not estimate parameters for the baseline hazard (and thus no intercept) here. The results from the Cox model and the log-logistic model turn out to be quite similar. The goodness-of-fit can not be directly compared due to the different estimation procedures but if we compare the log likelihood values of the log-logistic and the Weibull model (-68249.16 vs. -69558.03) we find that the PO approach has a better fit than the PH approach.

Since the absolute parameter values are hard to interpret in nonlinear models we calculated the marginal effects of the covariates in Table 3.4. The marginal effects are evaluated at the means of the covariates and refer to the ceteris paribus effect of a one-unit increase of a covariate on the 5-year default probability. For instance, in the log-logistic model increasing profitability (NITA) by 1% is estimated to lower the 5-year default probability by 0.2782% and a 10% increase in leverage (TLTA) is estimated to raise the 5-year default probability by 0.945%. The results for the log-logistic model and the Cox model are relatively close to each other while the marginal effects are somewhat lower for the Cox model.

While the better goodness-of-fit of the log-logistic model gives us first evidence on its appropriateness we will now do some further analyses. Figure 3.2 shows on the left hand side the course of the median firm's hazard rate

[29] Using the Weibull instead of the Cox model makes almost no difference. The coefficients are very close and Spearman's rank correlation for the predictions from the Weibull and the Cox model is as high as 0.99999883.

Table 3.3: Results from hazard regressions

	Cox model (PH)		Log-logistic model (PO)	
	Coef.	Std. error	Coef.	Std. error
NITA	-5.598	(1.358)	-6.804	(1.271)
TLTA	2.426	(0.296)	2.311	(0.254)
GRO	0.212	(0.054)	0.184	(0.053)
RET	-0.826	(0.056)	-0.813	(0.053)
VOLA	6.142	(0.526)	6.062	(0.461)
SIZE	-0.374	(0.031)	-0.336	(0.027)
const.			-11.992	(0.278)
α			1.255	(0.023)
firm-months	339 222		339 222	
$\log L$	-214084.69		-68249.16	
Wald χ^2	2351.88		2415.62	

Table 3.4: Marginal effects on 5-year default probability

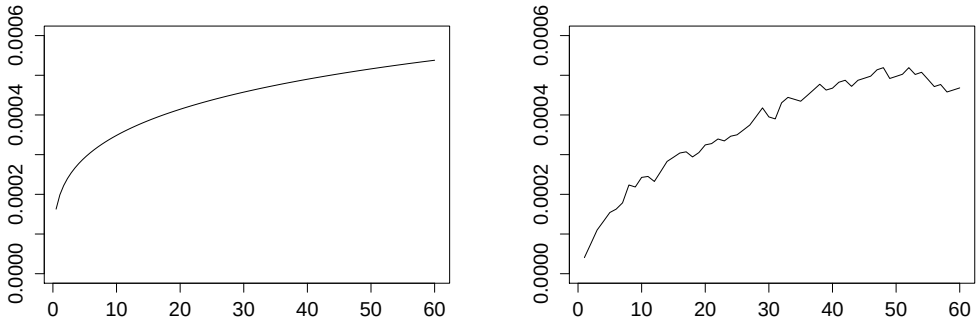
	Log-logistic model (PO)	Cox model (PH)
NITA	-0.2782	-0.1971
TLTA	0.0945	0.0854
GRO	0.0075	0.0075
RET	-0.0332	-0.0291
VOLA	0.2478	0.2161
SIZE	-0.0137	-0.0132

over the forecast time according to the log-logistic model.[30] The hazard rate is monotonically increasing in the first 60 months which makes sense since the median firm is not supposed to be close to default in the beginning.[31] The right-hand side of Figure 3.2 provides some support to our conjecture

[30]The median firm refers to the median of the predictions from the log-logistic model, i.e. it refers to the observation where 50% of all firm-months are estimated to be less risky.

[31]Note that the log-logistic model implies, given that the shape parameter α is larger than 1, that the hazard rate rises until $\frac{(\alpha-1)^{1/\alpha}}{\exp(\beta' x_{it})}$ and decreases thereafter (Kalbfleisch & Prentice, 2002, Ch. 2.2.6).

Figure 3.2: Hazard rate curves



The plot shows the continuous-time hazard rate of the median firm according to the log-logistic model (left hand side) and the nonparametrically estimated discrete-time hazard rate of a BBB firm (right hand side) plotted against forecast time (in months). Continuous-time and discrete-time hazard rates are comparable if hazard rates are small and periods are short since $\lambda^d(s) = 1 - \exp\left(-\int_{s-1}^{s} \lambda^c(u)du\right) = 1 - \exp(-\overline{\lambda}^c(s)) \approx \overline{\lambda}^c(s)$.

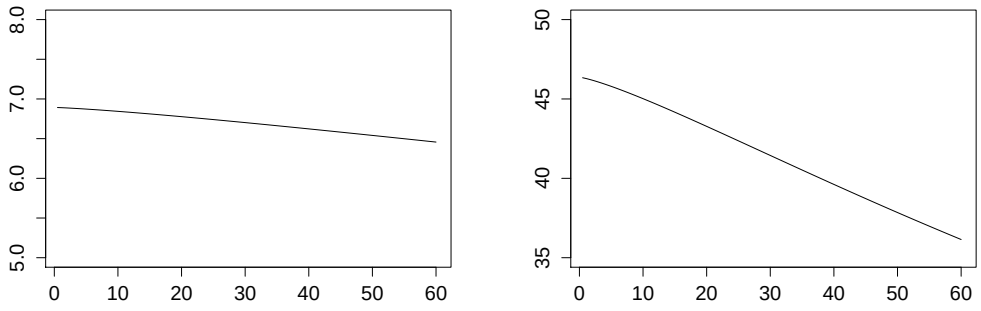
that the hazard curve of the log-logistic model is a realistic one. We see that the hazard curve of a BBB rated firm,[32] estimated completely nonparametrically with the life-table estimator (see section 4.1), exhibits a similar pattern and does not seem to have any important characteristics that are smoothed away by the parametric structure of the log-logistic model.

Our primary motivation to estimate the log-logistic model was its property of declining hazard ratios. To study this aspect we plotted in Figure 3.3 the evolution of the hazard ratios of selected pairs of firms over the forecast time. The left hand side shows the hazard ratios for the upper and lower quartile firm while the right hand side refers to the upper and lower decile firm.[33] The decline of the hazard ratios is evident but happens at a moderate pace in our model. This does not surprise as we do not expect that the hazard rates of high-risk and low-risk firms approach each other very quickly. By comparing both curves we further see that more extreme hazard ratios decline more quickly. Note that in the Cox model, the hazard ratios for the same

[32] In our sample, the median S&P rating is BBB.

[33] The quantiles are defined analogously to the median firm.

Figure 3.3: Evolution of hazard ratios



Hazard ratios of the upper and lower quartile firm (left-hand side) and upper and lower decile firm (right-hand side) derived from the log-logistic model. The abscissa represents forecast time (s) in months.

quantiles are constant at 5.30 and 24.28, respectively.

3.3.3 Evaluation of discriminatory power

While the analysis of hazard curves and hazard ratios provides relevant insights we will now evaluate the predictive power of our models which is the central criterion in our context. We will first focus on the ability of our models to provide an accurate rank order of the firms according to their default risk. The second dimension of predictive accuracy, calibration, will be investigated subsequently in section 3.3.4. To measure predictive accuracy, we will use the Accuracy Ratio and Harrell's C, which were both presented in detail in chapter 2. We use our measures in basically the same way as in section 2.5. For a given sample calendar month t, we calculate the Accuracy Ratio and Harrell's C for the predictions made in period t (and the corresponding lifetimes starting at t). We do this in monthly steps for a range of values for t and then take a weighted average of our indices with the number of firms observed in each period as weights. We measure both in-sample and, more importantly, out-of-sample predictive power. In the in-sample part, t is ranging from December 1985 to March 2005 which covers

all periods where the indices can be calculated.[34] In the out-of-sample part, t is ranging from December 1995 to March 2005. There, each month the models are re-estimated using only the information available until period t, a procedure known as a recursive estimation scheme (see section 2.3). We start in December 1995 to ensure that we estimate our models with at least 10 years of data (twice the maximum prediction horizon). Stein (2004) calls the recursive approach alternatively a walk-forward approach and argues that it is closest to the practical use of default prediction models. While other studies often use only a single sample split our recursive scheme removes the problem that the results may not be robust to a different choice of the split period.

Besides the Cox model and the log-logistic model we consider as competitors the stepwise lagging procedure (SLP) as outlined in the beginning of section 3.2.1 (using a logit specification for the discrete-time hazard rate as in Campbell et al., 2008) and S&P long term issuer credit ratings. As prediction horizons we choose one, three and five years. The results are shown in Table 3.5. We observe high predictive accuracy for all our models. While comparisons with other studies have to be taken with care, note that Duffie et al. (2007) report out-of-sample Accuracy Ratios of 87% (one year) and 70% (five years) using a similar dataset thereby achieving less accuracy than our models.[35] Our finding is supported by a recent study by Duan et al. (2012) where the authors use their dataset to estimate both a version of the SLP approach (with a complementary log-log-specification) and the model of Duffie et al. (2007). In line with our findings, Duan et al. (2012) find lower predictive accuracy for the approach of Duffie et al. (2007).

Comparing our different specifications, we see that the log-logistic model performs best in every category. The Cox model is second-best in the out-of-sample part and similar to the SLP procedure in-sample. This difference

[34] Prior to December 1985 the dataset is relatively sparse and does not contain a lifetime that ends with a default event. After March 2005, there are less than 5 years left in our sample so that the 5-year Accuracy Ratio, which requires some firms surviving the whole 5 years, can not be calculated anymore.

[35] Duffie et al. (2007) use a covariate forecasting approach as described in section 3.1 and state that their model is an improvement over available alternatives.

Table 3.5: Model performance statistics

Panel A: In-sample predictive accuracy

	Harrell's C			Accuracy Ratio		
Prediction horizon (months)	12	36	60	12	36	60
log-logistic	.9011	.8071	.7593	.9086	.8283	.7931
Cox	.9003	.8061	.7580	.9077	.8274	.7917
SLP	.9004	.8065	.7571	.9078	.8279	.7910
S&P	.8264	.7616	.7284	.8353	.7929	.7784

Panel B: Out-of-sample predictive accuracy

	Harrell's C			Accuracy Ratio		
Prediction horizon (months)	12	36	60	12	36	60
log-logistic	.8862	.7672	.7104	.8939	.7864	.7436
Cox	.8840	.7628	.7059	.8917	.7819	.7389
SLP	.8829	.7586	.6993	.8906	.7785	.7338
S&P	.8149	.7338	.6943	.8234	.7625	.7417

is most likely due to the fact that the SLP approach is more highly parameterized and thus suffers more from out-of-sample instability than the other models.[36] S&P ratings throughout have the lowest predictive power with the exception of the out-of-sample five-year Accuracy Ratio. The gains from our models as compared to S&P are highest for the shorter horizons. This is similar to findings in related studies (Löffler, 2007; Hilscher & Wilson, 2009) and is also in line with the common perception that rating agencies are not making the most efficient use of short-term relevant information.[37]

[36]Note that even in-sample the SLP approach suffers from quite implausible developments over the different lag lenghts. For instance, the marginal effect of net income (NITA) is -.00295 in the model with covariates lagged by 34 months, more than halves to be -.00123 at lag 36 only to decrease to -.00408 for lag 39. The marginal effects are evaluated again at the means of the covariates.

[37]The finding that rating agencies may react relatively slowly to new information can be explained by the objective of rating stability which is - besides rating accuracy - explicitly stated at least by Moody's (Cantor & Mann, 2003).

Table 3.6: Bootstrap hypothesis tests for out-of-sample predictive accuracy

Prediction horizon of 12 months

	Harrell's C				Accuracy Ratio			
	log-l.	Cox	SLP	S&P	log-l.	Cox	SLP	S&P
log-l.	.	.002	.002	.001	.	.001	.001	.001
Cox		.	.009	.001		.	.008	.001
SLP			.	.001			.	.001
S&P				.				.

Prediction horizon of 36 months

	Harrell's C				Accuracy Ratio			
	log-l.	Cox	SLP	S&P	log-l.	Cox	SLP	S&P
log-l.	.	.001	.001	.012	.	.001	.001	.068
Cox		.	.009	.022		.	.029	.135
SLP			.	.056			.	.217
S&P				.				.

Prediction horizon of 60 months

	Harrell's C				Accuracy Ratio			
	log-l.	Cox	SLP	S&P	log-l.	Cox	SLP	S&P
log-l.	.	.001	.001	.223	.	.001	.001	.871
Cox		.	.001	.383		.	.009	.816
SLP			.	.700			.	.575
S&P				.				.

The table contains p values for the null hypothesis that the population values of the indices (Accuracy Ratio or Harrell's C) for two predictors are equal which is tested against the two-sided alternative. The test refers to the results of Table 3.5, Panel B. The number of bootstrap replications is $B = 999$. Bootstrap p values are calculated by Formula (2.29).

We now go on to analyze if the differences in out-of-sample predictive power between our competing predictors are statistically significant. We choose the cluster bootstrap as described in section 2.4 for this purpose, i.e. we resample from the set of firms instead of the set of firm-months again interpreting all observations of a firm as one cluster. By resampling from our out-of-sample predictors and the associated lifetimes we can perform bootstrap hypothesis tests for the null that two models have the same predictive power.

The results of Table 3.6 show that the log-logistic model is a significant improvement over all alternatives with the exception of S&P ratings at the

five-year time horizon.[38] Further, the SLP approach performs significantly worse than the more parsimonious Cox and log-logistic models. This result holds regardless of the prediction horizon and the accuracy measure used. Our findings give rise to the following two main interpretations. On the one hand, we see that it pays off to choose a parsimonious model with relatively few parameters. This is of course a common finding especially in the forecasting literature. On the other hand, we observe that it is worthwhile to thoroughly analyze the structure imposed by the functional form. Here, the more realistic assumption of converging hazard rates of the log-logistic model as opposed to the constant hazard ratio assumption of the Cox model leads to a significantly higher predictive accuracy.

Before we turn to the calibration of our models, we will briefly investigate the impact of the prediction horizon on out-of-sample predictive accuracy. Since we argue that multi-period models are useful there should be a non-negligible difference in the long-run predictive accuracy between models that have a long-run horizon and models that have a shorter horizon. We tested this issue by calculating Harrell's C and the Accuracy Ratio with a prediction horizon of three and five years (thereby measuring long-run accuracy) for the Cox model and the log-logistic model under i) a corresponding three- or five-year prediction horizon and ii) under a shorter prediction horizon of one year.[39] Table 3.7 shows that we can clearly reject the null hypothesis that models with a short-term horizon do their job as well as models with a long-term horizon in terms of long-run predictive accuracy. We observe differences of about one percentage point in three-year accuracy if a one-year model is used instead of a three-year model and differences of up to about 1.6% for a one-year vs. a five-year model. Bootstrap tests, conducted as in the calculations for Table 3.6, reveal that the differences are statistically

[38]The reason that the improvements of our models to S&P ratings are sometimes not statistically significant although the differences in the indices are higher than between our models is that the S&P predictions are less correlated with the predictions from our models than the predictions from our different models are with each other.

[39]The specific horizon for estimating the models is accounted for by censoring the lifetimes artficially after H months, with H being the prediction horizon. Note that our models do not imply a change in the risk ordering of firms with varying prediction horizon so that this kind of additional censoring is the only source of differences.

Table 3.7: Sensitivity of model performance to the prediction horizon

	Cox model		Log-logistic model	
	Harrell's C	Accuracy Ratio	Harrell's C	Accuracy Ratio
$H = 12$	.7535	.7706	.7577	.7750
$H = 36$	.7628	.7819	.7672	.7864
p value	.006	.001	.004	.001
$H = 12$	.6939	.7227	.6984	.7274
$H = 60$	.7059	.7389	.7104	.7436
p value	.003	.001	.001	.001

The upper half of the table refers to the three-year versions of Harrell's C and the Accuracy Ratio whereas the bottom half contains five-year indices. p values refer to differences in the indices for models estimated with different prediction horizons H (in months), i.e. $H = 12$ vs. $H = 36$ and $H = 12$ vs. $H = 60$, and were calculated using the bootstrap analogously to Table 3.6.

significant at any conventional significance level. An important question is whether the differences are economically significant as well. A deeper investigation of this issue is beyond the scope of this work since it would require a complete model for the loan market including assumptions about how credit decisions are made and so on. However, we note that in such an extended framework Blöchlinger & Leippold (2006) find that relatively small Accuracy Ratio differences of the order we find here may already have a sizeable economic impact in a competitive environment. The main reason for the findings of Blöchlinger & Leippold (2006) is adverse selection: If, for instance, Bank A lends to certain obligors in contrast to Bank B because it has omitted a certain important risk factor in its rating model, Bank A will attract a high share of the obligors that are exposed to this disregarded risk factor and is thus likely to realize relatively large unexpected losses.

3.3.4 Calibration analysis

As was pointed out in section 2.2, a model with high discriminative power may be improved with respect to its calibration because of the shrinkage effect and also because of possible questionable restrictions induced by its parametric structure. Although our sample is relatively large, which indicates that at least the shrinkage effect should not be too pronounced, we will now investigate the calibration of our best model, the log-logistic specification. As in section 2.2 we will divide our analysis into a nonparametric and a parametric calibration analysis. We will use the circular rolling-window (CRW) validation scheme which was introduced in section 2.3. The reason for switching from the recursive estimation scheme applied in section 3.3.3 to the CRW scheme is that the CRW method is particularly helpful for calibration analyses (see section 2.3). As the block length for the CRW approach we choose $B = H$. For $H = 60$, this should be enough since dependencies induced by common shocks like recessions should have largely been disappeared after five years. For $H = 36$ or $H = 12$ we found almost no sensitivity of the results if B was increased to 60 so that we uniformly selected the forecast horizon to be the block length.

We start with the nonparametric calibration analysis. The first step is to generate the out-of-sample default probabilities ($\widehat{PD}^{OS}$) for each sample period except the last one (where no predictions can be evaluated) by using the CRW method which amounts to estimating the model 351 times. The observations were then grouped into buckets according to the deciles of the distribution of the out-of-sample default probabilities. For each bucket, we applied the nonparametric life-table estimator (see section 4.1) giving $\widehat{PD}^{VS}$ which can then be compared to the average out-of-sample default probability $\left(\overline{\widehat{PD}}^{OS}\right)$ in each bucket. The results are displayed in Table 3.8.

We observe some discrepancies between $\overline{\widehat{PD}}^{OS}$ and $\widehat{PD}^{VS}$ which can be attributed to different reasons. First, we see some general differences in the evolution of the default probabilities over the deciles. For instance, the model-based default probabilities ($\overline{\widehat{PD}}^{OS}$) are smaller in the 8th decile for all horizons and are then increasing more sharply than their nonparametric

Table 3.8: Nonparametric calibration analysis of log-logistic model

Prediction horizon of 60 months

Decile	1	2	3	4	5	6	7	8	9	10
$\overline{\widehat{PD}}^{OS}$	0.43	0.78	1.14	1.58	2.18	3.11	4.68	7.76	15.70	50.05
$\widehat{PD}^{VS}$	0.54	0.67	0.85	1.16	1.61	2.95	5.52	9.80	16.33	38.75
$\widehat{\sigma}(\widehat{PD}^{VS})$	0.21	0.19	0.16	0.19	0.24	0.33	0.51	0.73	1.02	1.83
$Q = 63.002$, $p = 9.76e-10$										

Prediction horizon of 36 months

Decile	1	2	3	4	5	6	7	8	9	10
$\overline{\widehat{PD}}^{OS}$	0.20	0.36	0.52	0.73	1.02	1.47	2.26	3.85	8.29	36.31
$\widehat{PD}^{VS}$	0.28	0.29	0.32	0.43	0.59	1.27	2.45	5.22	10.33	30.32
$\widehat{\sigma}(\widehat{PD}^{VS})$	0.10	0.08	0.08	0.09	0.10	0.17	0.25	0.42	0.65	1.38
$Q = 80.160$, $p = 4.67e-13$										

Prediction horizon of 12 months

Decile	1	2	3	4	5	6	7	8	9	10
$\overline{\widehat{PD}}^{OS}$	0.02	0.04	0.06	0.09	0.13	0.19	0.31	0.58	1.46	14.68
$\widehat{PD}^{VS}$	0.01	0.03	0.04	0.06	0.06	0.13	0.16	0.61	1.59	14.30
$\widehat{\sigma}(\widehat{PD}^{VS})$	0.01	0.01	0.02	0.03	0.02	0.04	0.03	0.09	0.15	0.61
$Q = 37.636$, $p = 4.39e-05$										

$\overline{\widehat{PD}}^{OS}$ is calculated by building buckets according to the sorted out-of-sample default probabilities generated by the CRW method and then taking the average of the out-of-sample default probabilities for each bucket. $\widehat{PD}^{VS}$ is based on the same buckets but applies the life-table estimator to the observations of each bucket. $\widehat{\sigma}(\widehat{PD}^{VS})$ are standard errors for $\widehat{PD}^{VS}$ calculated by the cluster-robust extension to the Greenwood formula (see section 4.A). Q gives the test statistics for the null hypothesis of correct calibration of $\overline{\widehat{PD}}^{OS}$ (see section 2.2.1), with p being the associated p value. Numbers (except test statistics and p values) are in percentage points.

counterparts in the 9th and 10th decile. Similarly, $\overline{\widehat{PD}}^{OS}$ grows at a faster pace in the first three deciles than $\widehat{PD}^{VS}$. In contrast, we observe the oppo-

site behaviour in the middle part of the deciles. Most likely, these differences are based on the fact that $\widehat{\overline{PD}}^{OS}$ is derived from the log-logistic model and is thus based on a specific functional form which restricts the evolution of the default probabilities over the deciles. In contrast, no such restrictions appear in the nonparametric approach used to calculate $\widehat{PD}^{VS}$.

Another reason for the differences between $\widehat{\overline{PD}}^{OS}$ and $\widehat{PD}^{VS}$ is the existence of the shrinkage effect. Looking at the most extreme deciles we see that the unshrunk estimates ($\widehat{\overline{PD}}^{OS}$) tend to be too extreme as expected from theory. An exception is the first decile at the 1-year horizon. To understand this finding note that with few default events the life-table estimator has a tendency to underestimate the true default probability (see chapter 4). As the log-logistic model does not only use the default events in one particular decile the difficulties arising from few default events are less pronounced.

Of course, it is of interest to analyze the departures of $\widehat{PD}^{VS}$ from $\widehat{\overline{PD}}^{OS}$ with respect to their statistical significance. The standard errors of $\widehat{PD}^{VS}$ reveal that there are even individual deciles where the differences are statistically significant. Using the χ^2 test introduced in section 2.2.1, which aggregates the standardized differences over the deciles, we see by looking at the test statistics (denoted by Q) and their p values that we can clearly reject the null hypothesis of correct calibration for all horizons.[40]

While Table 3.8 provides some evidence for the shrinkage effect its overall amount is quite hard to disentangle as we also see other effects that cause differences between $\widehat{\overline{PD}}^{OS}$ and $\widehat{PD}^{VS}$. The upcoming parametric calibration analysis will shed some more light on the size of the shrinkage effect. We will use two different calibration models. The first is a straightforward extension of the Logit example given in section 2.2 using the log-logistic specification

[40] As we introduced the test in section 2.2.1 we have mentioned that one has to assume independence of $\widehat{PD}^{VS}$ for different buckets. Here, this assumption is likely not to be literally true as the same obligor may in some cases enter different buckets. As the assumption underlying the χ^2_{10} null distribution is thus not perfectly met, one may use the cluster bootstrap to approximate the true null distribution. However, in our case the test results are very clear-cut so that this does not seem to be necessary.

instead:

$$S(s|\widehat{\beta}_t' x_{it}, \widehat{\alpha}) = \left(1 + \left[\exp(\gamma_0 + \gamma_1 \widehat{\beta}_t' x_{it}) s\right]^{\widehat{\alpha}}\right)^{-1} \tag{3.31}$$

$\widehat{\beta}_t$ is the parameter estimate from the corresponding training sample which means in the case of the CRW procedure that $\widehat{\beta}_t$ was estimated from the subsample involving the information from periods $1, \ldots, t, t+B+1, \ldots, T$ (T being the number of sample periods). $\widehat{\alpha}$ is the estimate of the shape parameter from the full sample and is fixed for the CRW estimations and in the calibration model. The reason for doing so is that we want the linear predictors, $\widehat{\beta}_t' x_{it}$, to be comparable for different t since they are pooled together for the estimation of the calibration model. If $\widehat{\alpha}$ would vary with t as well, ordering according to the linear predictors would not be exactly the same as ordering according to the corresponding default probabilities. Thus, we have chosen to fix $\widehat{\alpha}$ although the results are similar if $\widehat{\alpha}$ is allowed to fluctuate. The log-logistic calibration model can simply be fitted by doing a log-logistic regression with the whole sample on $\widehat{\beta}_t' x_{it}$ with a restricted shape parameter. If the model is correctly calibrated γ_0 and γ_1 should not be significantly different from zero and one, respectively. It should be noted that unlike in the Logit case of section 2.2 this calibration regression does only make sense if the original model is also log-logistic. However, any hazard model containing a linear part can be calibrated analogously.

An alternative calibration model (Van Houwelingen, 2000) is given by

$$\Lambda(s|\widehat{\beta}_t' x_{it}, \widehat{\alpha}_t) = \gamma_0 \widehat{\Lambda}_t(s|x_{it})^{\gamma_1} . \tag{3.32}$$

The idea of the model is based on the fact that if $\Lambda(\cdot)$ is the true cumulative hazard of a lifetime Y^* it holds that $\Lambda(Y^*) \sim Exp(1)$.[41] Now, note that the cumulative hazard of a Weibull distribution can be written as $\Lambda(y) = \gamma_0 y^{\gamma_1}$ which reduces to the cumulative hazard of an $Exp(1)$ distributed random variable if $\gamma_0 = \gamma_1 = 1$. This means we can fit a Weibull model to the transformed lifetimes, $\widehat{\Lambda}_t(Y_{it}|x_{it})$, to check calibration. The better the calibration the closer the values of $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ will be to one. However, in the presence of the shrinkage effect, we expect both $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ to be below one.

[41] This can be easily seen as $P(\Lambda(Y^*) \leq y) = 1 - S(\Lambda^{-1}(y)) = 1 - \exp(-\Lambda(\Lambda^{-1}(y))) = 1 - \exp(-y)$.

Table 3.9: Parametric calibration analysis of log-logistic model

Panel A: Log-logistic calibration model

	$\widehat{\gamma_0}$	$\widehat{\sigma}(\widehat{\gamma_0})$	$\widehat{\gamma_1}$	$\widehat{\sigma}(\widehat{\gamma_1})$	W	p value
$H = 12$	0.0682	(0.0175)	0.9830	(0.0038)	25.112	3.52e-06
$H = 36$	0.1876	(0.0428)	0.9634	(0.0078)	22.197	1.51e-05
$H = 60$	0.2599	(0.0619)	0.9574	(0.0100)	18.126	1.16e-04

Panel B: Weibull calibration model

	$\widehat{\gamma_0}$	$\widehat{\sigma}(\widehat{\gamma_0})$	$\widehat{\gamma_1}$	$\widehat{\sigma}(\widehat{\gamma_1})$	W	p value
$H = 12$	0.9210	(0.0107)	0.9750	(0.0034)	56.817	4.60e-13
$H = 36$	0.8467	(0.0186)	0.9449	(0.0068)	72.620	2.22e-16
$H = 60$	0.7860	(0.0279)	0.9262	(0.0097)	62.619	2.53e-14

Standard errors in parentheses are calculated via the cluster bootstrap using 100 replications. W denotes the Wald statistics (using bootstrap standard errors) for the joint tests that $(\gamma_0, \gamma_1) = (0, 1)$ (log-logistic calibration model) and $(\gamma_0, \gamma_1) = (1, 1)$ (Weibull calibration model), respectively. The p values to the Wald statistics are based on a χ^2 distribution with 2 degrees of freedom. H is the prediction horizon in months.

Note that $\widehat{\Lambda}_t(\cdot)$ is the cumulative hazard from the log-logistic model estimated again by using the information from periods $1, \ldots, t, t+B+1, \ldots, T$, i.e. $\widehat{\Lambda}_t(Y_{it}|x_{it}) = \log(1 + [\exp(\widehat{\beta}_t' x_{it})Y_{it}]^{\widehat{\alpha_t}})$. The Weibull calibration model has the advantage over the log-logistic calibration model that we do not have to fix the shape parameter. Further, the Weibull calibration model is completely general, i.e. it can be used regardless of the specification of the original model. It follows that it may detect misspecification more generally. To see this note that in our first approach the log-logistic model is imposed for the original estimation and the calibration. This is a valid approach to detect shrinkage effects but it will not reveal any problems with the log-logistic specification.

The results from applying both calibration models to our data are given in Table 3.9. For all horizons and for both calibration models, the estimates reveal that shrunk estimates give a better fit out-of-sample. The size of the shrinkage effect increases with the prediction horizon. This makes sense since

the shrinkage effect typically grows as predictability decreases.[42] To explore the statistical significance of our results we utilized again the cluster bootstrap as explained in section 2.4. Since each bootstrap replication involves a new application of the CRW method, i.e. estimating the model 351 times, we restricted ourselves to 100 replications. As the results are very clear the limited number of bootstrap replications should not pose a problem. The standard errors and the results from the Wald tests for the hypotheses that $(\gamma_0, \gamma_1) = (0, 1)$ (log-logistic calibration model) and $(\gamma_0, \gamma_1) = (1, 1)$ (Weibull calibration model), respectively, show that our findings are statistically significant. The statistical evidence is strongest for the Weibull model. We conclude that despite our relatively large sample the shrinkage effect is non-negligible and that the original estimates can be improved by a recalibration. In the parametric approach, such a recalibration can be done by simply plugging the estimates of γ_0 and γ_1 into the calibration models. The recalibrated default probabilities are then given by $1 - \left(1 + [\exp(\widehat{\gamma}_0 + \widehat{\gamma}_1 \widehat{\beta}' x_{it}) H]^{\widehat{\alpha}}\right)^{-1}$ for the log-logistic calibration model and by $1 - \exp(-\widehat{\gamma}_0 \widehat{\Lambda}(H)^{\widehat{\gamma}_1})$ for the Weibull calibration model.[43]

As an alternative to a parametric recalibration, we can use the life-table estimates from Table 3.8 as recalibrated default probability estimates. This is simply done by mapping default probabilities to the appropriate bucket of out-of-sample default probabilities, $\widehat{\overline{PD}}^{OS}$, and using the corresponding life-table estimate from Table 3.8 as a revised default probability. Given that Table 3.8 reveals certain problematic restrictions of the log-logistic model regarding the evolution of the default probabilities over the deciles, we recommend the nonparametric recalibration for deriving the final default probability estimates.

[42] In linear models, Copas (1983) showed that the shrinkage effect is more pronounced as the error variance increases, ceteris paribus. A high error variance can be interpreted as low predictability.

[43] We have dropped the index t from $\widehat{\beta}_t$ and $\widehat{\Lambda}_t(\cdot)$ since we consider now the recalibration of the model estimated with the full sample.

Chapter 4

Default prediction with given rating grades

In this chapter, we deal with the problem of assigning default probabilities to given rating grades. On the one hand, ratings may result from a statistical model like the one presented in chapter 3. In this case, computing default probabilities for rating grades (or buckets) can be sensible as a part of the calibration process but is not absolutely necessary since default probabilities can also be directly derived from the model. However, in many practical situations ratings are at least partly the result from qualitative assessments of creditworthiness (Grunert et al., 2005; Treacy & Carey, 2000; Standard & Poor's, 2009). For instance, a bank will usually judge the management quality of a firm it lends to and this judgement will often influence the firm's rating in a non-statistical way. Especially when default data are sparse the relative importance of models for ratings is reduced due to the relatively low prognostic power of model-based predictions (Standard & Poor's, 2007). In this work, we will not deal with possible non-statistical elements of the rating process. However, we have to recognize that ratings will often not be based simply on a default probability model and that thus the assignment of default probabilities to given rating grades is a relevant situation encountered in practice.

With respect to the two dimensions of discrimination and calibration this

chapter is solely about calibration since our analysis is conditional on the rank order of the default predictions as given by the rating system. While we have already dealt with calibration in the preceding chapters we will now extend our coverage of the issue of calibration in several ways. First, we will introduce what we will call the standard estimator in the next section. This is the estimator which was already applied in the nonparametric calibration analysis of section 3.3.4 and which is among the most commonly applied methods to assign default probabilities for given rating grades. The standard estimator will serve as a benchmark for alternative methods. We will then focus on the problems that arise when sample sizes and/or default probabilities are rather small resulting in only few, if any, default events for a given sample. Such samples - sometimes labeled low-default portfolios - are not only interesting from a theoretical perspective but are also highly relevant in practice since for many important classes of obligors a very limited default history exists, especially in the higher rating grades. Important examples for such sparse datasets are samples of sovereigns and financial institutions.

Standard approaches applied to low-default portfolios have serious drawbacks. Besides the obvious effect that estimation uncertainty is high the skewness of the sampling distribution leads to a high probability of underestimating the true default probability. For example, given small true default probabilities and small sample sizes, it is quite likely not to observe any default event in a particular sample leading to a default probability estimate of zero under standard approaches. More generally, Kurbat & Korablev (2002) show in a simple binomial framework that the likelihood of underestimating the true default probability rises as i) the true default probability decreases, ii) the sample size decreases and iii) the correlation of default events increases.[1] Given these properties, it would be desirable to improve upon the standard estimator by applying a more efficient and more conservative estimator in small samples. The latter is especially important since a conservative approach may be a general guideline for prudent risk management and is also demanded from the regulatory side (Basel Committee on Banking

[1] Our simulation study in section 4.5 supports these findings in an extended framework and provides numerical evidence on the probability to underestimate the true default probability under various scenarios.

Supervision, 2006, §416,451).

The problem of low-default portfolios has already received some attention in the literature. One approach is to employ the idea of confidence intervals and to use an appropriate upper confidence bound as a conservative default probability estimator (Pluto & Tasche, 2006; Benjamin et al., 2006). However, the aforementioned studies consider a fixed one-year prediction horizon and do not use potentially available within-year information. We will deal with these approaches and possible generalizations to our more flexible multi-period setup in section 4.2. Then, in section 4.3, we will deal with Bayesian approaches to default probability estimation. For single-period predictions, Bayesian methods using priors specified by expert eliciation (Kiefer, 2009) and non-informative priors (Tasche, 2011) have been proposed in the literature. The main contributions of this chapter are the introduction of an *empirical* Bayes estimator for *multi*-period predictions (section 4.3), the application of this estimator to a comprehensive sovereign bond dataset (section 4.4) and its evaluation by means of a novel kind of simulation study (section 4.5). In the application and simulation sections, we also consider the standard approach and compare it to the empirical Bayes estimator. The analysis of this chapter is based to a large extent on Orth (2011a).

We do not consider models for rating migrations although these could also be used for default probability estimation. For instance, the use of Markovian rating migration models is quite standard and has the benefit that it usually removes default probability estimates of zero if the time intervals are chosen small enough (Lando & Skodeberg, 2002). However, there is strong evidence against the Markovian assumption as migration probabilities have been found to depend on the direction of the prior rating action for corporates (Lando & Skodeberg, 2002) and for sovereigns as well (Fuertes & Kalotychou, 2007). In particular, downgrades are more often followed by subsequent downgrades than implied by a Markovian model so that default probabilities derived from Markovian migration models tend to be downward biased (Hanson &

Schuermann, 2006).[2] More sophisticated models like Hidden Markov Models (Christensen et al., 2004) have been proposed in the literature but come at the cost of considerably more complexity not least in terms of a higher number of parameters. In small samples, such an augmented parameterization is likely to cause instability, i.e. high variance of the parameter estimates and is thus rather not suitable in our context.

4.1 The standard estimator

The estimator that we will present in this section is the approach used by the major rating agencies in their calculation of cumulative default rates (Hamilton & Cantor, 2006).[3] Cumulative default rates are estimates of default probabilities that are constructed by marginal default rates (see below). Let us first introduce the notation. All the obligors that have the rating r at time t, $t = 1, \ldots, T$, form a cohort. We denote by $N^r_{t,1}$ the number of obligors that comprise the cohort at its beginning (t) and we denote by $N^r_{t,s}$ those members of the cohort that are still at risk before period $t+s$. Being at risk means that an obligor has not defaulted or is not censored in the first $s-1$ periods after the cohort building date t. Out of the $N^r_{t,s}$ obligors entering period $t+s$, let $D^r_{t,s}$ be the number of those that default in period $t+s$ and let $L^r_{t,s}$ be the number of those which are lost, i.e. which are censored, in period $t+s$. Further, let λ^r_s be the discrete-time hazard rate which is the probability that an obligor rated r at a certain point in time will default s periods later conditional on surviving the first $s-1$ periods.[4] If we define Y^*_{it} to be the discretely measured lifetime (the time until default) of obligor i

[2]Since the cited empirical evidence concerns certain agency ratings it is of course possible that for any other (internal) rating system a Markovian approach is appropriate. Still, one has to be aware that any violation of the Markovian assumption can lead to seriously biased estimates.

[3]It is also largely equal to the approach of Altman (1989). However, in that study the analysis is restricted to the cumulative default rates of newly issued bonds so that no overlapping lifetimes occur.

[4]In the notation of chapter 3, the discrete-time hazard rate would be written as $\lambda^d(s|r)$. We skip the index d for discrete and use λ^r_s instead to save space in the following derivations.

that starts in period t and define R_{it} to be the corresponding rating, we can write this probability formally as

$$\lambda_s^r = P(Y_{it}^* = s | Y_{it}^* > s-1, R_{it} = r)\,. \tag{4.1}$$

Under our notation, the standard approach for the marginal default rate, which is an estimator for the discrete-time hazard rate, is

$$\widehat{\lambda}_s^r = \frac{\sum_{t=1}^T D_{t,s}^r}{\sum_{t=1}^T N_{t,s}^r - L_{t,s}^r/2}\,. \tag{4.2}$$

Usually the main interest is on the estimation of default probabilities which we will denote by PD_s^r and define as

$$PD_s^r = P(Y_{it}^* \leq s | R_{it} = r)\,. \tag{4.3}$$

Cumulative default rates, i.e. estimators for the default probabilities, are easily constructed from the marginal default rates:

$$\widehat{PD}_s^r = 1 - \prod_{j=1}^{s}(1 - \widehat{\lambda}_j^r) \tag{4.4}$$

Let us briefly interpret what the estimator actually does. The estimator starts with the calculation of marginal default rates by taking a weighted average of the marginal default rates of individual cohorts, $\frac{D_{t,s}^r}{N_{t,s}^r - L_{t,s}^r/2}$. The weights are easily seen to be the adjusted number of obligors at risk, $N_{t,s}^r - L_{t,s}^r/2$, since $\frac{\sum_{t=1}^T D_{t,s}^r}{\sum_{t=1}^T N_{t,s}^r - L_{t,s}^r/2} = \sum_{t=1}^T \frac{D_{t,s}^r}{N_{t,s}^r - L_{t,s}^r/2} \frac{N_{t,s}^r - L_{t,s}^r/2}{\sum_{t=1}^T N_{t,s}^r - L_{t,s}^r/2}$. While it would also be possible to take a weighted average of the cumulative default rates of individual cohorts, averaging marginal default rates results in more efficient estimation as was already shown by Cutler & Ederer (1958). Further, notice the adjustment in the denominator of Formula (4.2) which involves the subtraction of half of the censored observations. This correction is based on the assumption that censored obligors have still survived on average half of the corresponding period. Finally, the estimator calculates cumulative default rates from the marginal default rates (Equation (4.4)).

The presented estimator is also known as the life-table or actuarial estimator and is approximately equal to the widely-used Kaplan-Meier or Product-Limit estimator (Kaplan & Meier, 1958) as the period length becomes small.

The difference between the two estimators is that the life-table estimator is constructed for a setup where the data are interval-censored, i.e. the default and censoring times are observed to lie in a certain interval (or period) with the exact times being unknown. In contrast, the Product-Limit estimator assumes a continuous-time setting where the exact times are observed. Then, the withdrawal adjustment in the denominator of Equation (4.2) becomes unnecessary. As the period length decreases, the number of censored observations per interval decreases and the life-table estimator approaches the Product-Limit estimator. In this work, we will use the life-table estimator with monthly intervals making the difference to the Product-Limit estimator very small. Consequently, there are only minor differences in the theoretical properties of the Product-Limit estimator and the life-table estimator in our case.

With respect to the censoring scheme, we require for the consistency of the Product-Limit estimator[5] that censoring is noninformative, i.e. that default and censoring events at time s are independent, conditionally on the history of the default and censoring processes at time s (Lawless, 2003, Ch. 2.2.2). As we have seen in section 2.5, this assumption is very doubtful in certain applications if one applies the estimator to the overall sample since worse rated firms tend to default earlier and have earlier censoring times as well. However, if one partitions the sample based on rating grades and applies the estimator for each subsample the assumption becomes much more realistic. For large US corporates, the question of informative censoring is analyzed in detail by Hamilton & Cantor (2006). In particular, they analyze whether censoring rates are higher after rating downgrades which may indicate higher subsequent default risk. The authors do not find such an effect and conclude their analysis that there is no evidence against the noninformative censoring assumption if the sample is split according to rating grades.

The Product-Limit estimator has been shown to be a nonparametric Maxi-

[5] As was shown by Breslow & Crowley (1974), the life-table estimator is generally inconsistent unless one assumes a very specific structure of the distribution of the censoring times (see Theorem 1 of Breslow & Crowley (1974)). However, as was shown in the same study, the asymptotic bias becomes very small if the number of intervals is sufficiently large.

mum Likelihood estimator (Johansen, 1978) but note that in our setting we rather have a pseudo Maximum Likelihood estimator since our observations are not independent. To see this, notice that the same obligor enters a new cohort every period and that default events are used several times in the estimation process. For instance, consider an obligor which is rated A in period t, stays A rated in period $t+1$ and subsequently defaults in period $t+2$. The same default event enters the calculation of $\widehat{\lambda}_1^A$ and $\widehat{\lambda}_2^A$. Put another way, our estimator is the classical life-table estimator applied to a pooled sample of partially overlapping lifetimes, $Y_{it_{i1}}, Y_{i,t_{i1}+1}, \ldots, i = 1, \ldots, n$, ($t_{i1}$ being the first calendar period where obligor i is observed) which clearly leads to dependencies. As the simulation study in section 4.5 will confirm, these dependencies (and additional dependencies through common shocks) do not introduce any relevant bias to the estimator (see Table 4.4, Panel B). Further, consistency under rather mild assumptions for the dependence structure has been established as well (Ying & Wei, 1994). Nevertheless, an estimator that incorporates the apparent dependencies, for instance a full Maximum Likelihood approach, might be more efficient. Nonparametric multivariate extensions to the Product-Limit estimator exist and have been compared by Kang & Koehler (1997) to the "independence-working" approach that we present here. They find that efficiency losses are minimal and do not offset the additional computational burden required for the more complicated multivariate estimators.

What remains to be specified for an empirical analysis is the period length. Rating agencies differ in this respect. While Moody's has switched to building cohorts on a monthly basis since 2005 (Hamilton & Cantor, 2006), Standard & Poor's uses cohorts of obligors built at the end of every calendar year (Standard & Poor's, 2011a). In our application, we will use monthly cohorts and accordingly construct our default probability estimates using monthly hazard rates in order to use as much sample information as possible. Since more than one rating change per month for the same obligor occurs only very rarely, almost no information is lost under a monthly periodicity.

4.2 Confidence bound approaches

An obvious approach to conservative default probability estimation is the use of confidence intervals which has been proposed by Pluto & Tasche (2006) and Benjamin et al. (2006). In these studies, a one-period view is employed, i.e. only $\lambda_1^r = PD_1^r$ is estimated (and no adjustment for censoring is made). Note that in this case dependencies through overlapping lifetimes do not occur. However, there may still be dependencies due to common shocks. Pluto & Tasche (2006) consider both the case that defaults are independent and the dependent case in a simple single-factor model. Under independence, defaults are binomially distributed, $D_{t,1}^r \sim Bin(N_{t,1}^r, \lambda_1^r)$, so that confidence interval calculation is straightforward. Using one-sided Clopper-Pearson intervals, the corresponding conservative default probability estimator is given as a quantile of a beta distribution:

$$\widehat{PD}_{1,\gamma}^r = Q_X(1-\gamma) \quad , \quad X \sim beta(D_{t,1}^r + 1, N_{t,1}^r - D_{t,1}^r) \tag{4.5}$$

γ is the significance level of the corresponding one-sided test from which the Clopper-Pearson interval is derived. Clopper-pearson intervals guarantee a coverage probability of at least $1-\gamma$ but often the coverage is considerably larger so that they are sometimes seen as overly conservative (Brown et al., 2001). However, other approaches are based on asymptotic approximations and deliver degenerate or overly optimistic intervals if no default event is observed, something which is particularly problematic for our situation.

Pluto & Tasche (2006) and Benjamin et al. (2006) consider different choices for γ but we note here that $\gamma = 0.5$ seems to be particularly interesting. For $\gamma = 0.5$, $\widehat{PD}_{1,0.5}^r$ approaches the standard estimator most quickly as the sample size increases since the underlying binomial distribution approaches a normal distribution. This property ensures a smooth transition to the standard estimator which is likely to be preferred in large samples. Further, $\widehat{PD}_{1,0.5}^r$ has the nice intuitive interpretation that it does not underestimate the true default probability in at least 50% of all cases under repeated sampling.

Dependence through common shocks will arise in many situations in practice and its existence will typically widen confidence intervals. Pluto &

Tasche (2006) show how to calculate confidence bounds when dependence comes from a single unobserved factor that exhibits exponentially decaying autocorrelation over the sample period.[6] Although the model is relatively simple, the calculation of confidence bounds gets quite involved and requires T-dimensional integration for T sample periods. Against this background, it is quite clear that a potential extension of the confidence bound approach to multi-period predictions which additionally needs to account for dependencies caused by overlapping lifetimes is very challenging especially since asymptotic theory is likely to be of little help for our small sample problem. A heuristic solution would be to construct confidence intervals for $\lambda_1^r, \ldots, \lambda_s^r$ and to construct a conservative estimate for PD_s^r from the conservative marginal default rates in the usual way. It turns out, however, that such an approach is overly conservative and depends heavily on the periodicity of the data. Consider, for instance, a very simple case where we assume independence, $N_{t,1}^r = 100$, $D_{t,1}^r = L_{t,1}^r = 0$ and the periodicity is one year. Then, $\widehat{PD}^r_{1\text{ year},0.5} = .0069$. If alternatively the data are exactly the same but the periodicity is monthly the estimator is $\widehat{PD}^r_{1\text{ year},0.5} = 1 - (1 - \widehat{\lambda}^r_{1,0.5})^{12} = 1 - (1 - .0069)^{12} = .0798$. Of course, this discrepancy is not sensible.

We have seen in the preceding chapters that the cluster bootstrap is a valuable method for inference under dependence structures involving overlapping lifetimes. In fact, the cluster bootstrap has been used by Cantor et al. (2008) to calculate confidence intervals for cumulative default rates (in a large sample application). Similarly, an analytical estimator for the variance of the standard estimator under clustered data exists (Williams, 1995) and was applied in section 3.3.4. We present this variance estimator in appendix 4.A. However, both the cluster bootstrap and the analytical approach break down when no default events are observed at all. In the case of the bootstrap, the problem is that no bootstrap sample will contain any default. Analogously, the analytical formula will give an implausible variance estimate of zero under no defaults. Taken together, all approaches relying on confidence intervals

[6] The model used by Pluto & Tasche (2006) is the one underlying the Basel II capital formula extended to multiple periods.

presented in the literature suffer from major disadvantages especially if no or few default events are observed and multi-period predictions are desired. How we can still find conservative (and also quite precise) default probability estimators is the topic of the next section.

4.3 An empirical Bayes approach

Bayesian parameter estimation is a potentially very useful approach especially in the case of small samples. The reason for this is that in small samples the data provide only little information about the parameter of interest so that the incorporation of prior information can be particularly helpful. To use such prior information, a Bayesian data analyst specifies a prior distribution for the parameters of interest. This can be done by different means. Possibilities that have been proposed in the context of default probability estimation include the specification of the prior by means of expert elicitation (Kiefer, 2009) and the use of uninformed priors (Tasche, 2011). While expert eliciation is potentially useful, there are also important caveats: The elicitation process requires experts that are well-trained in thinking about probabilities; it is relatively time-consuming; and the subjectivity of the approach may also be critized, not least by regulators, especially with respect to possible incentive problems that arise when the expert has benefits from being liberal with his prior guess.

Non-informative priors, in contrast, suffer from the major disadvantage that possible efficiency gains from the introduction of prior information are non-existent. In many cases, point estimates based on non-informative priors will even not be different from their corresponding sampling theory counterparts. Note, however, that this is generally not true for the Bayesian estimation of probabilities. For instance, Tasche (2011) proposes as one possibility the use of a uniform prior for the default probability leading to default probability estimates which are shrunk towards the prior mean, 0.5.[7] This approach

[7]It should be noted that there is no consensus that the uniform prior is non-informative in this situation. For example, an alternative would be to use Jeffrey's prior (Jeffreys, 1946). Also see Berger (1980, Ch. 3.2.2).

succeeds in giving non-zero estimates even when no default event is observed. However, besides not providing relevant efficiency improvements, there are some difficulties with non-informative priors as far as multi-period default probabilities are concerned. Recall that the default probability is not the parameter estimated in the first place but we rather need a prior for the discrete-time hazard rates from which multi-period default probabilities are constructed. For instance, a uniform prior for a monthly discrete-time hazard rate results in much more conservative estimates than a uniform prior for a one-year hazard rate.[8]

In this chapter, we will alternatively propose a data-driven way to specify the prior distribution by using an empirical Bayes (EB) approach.[9] To enable the estimation of the prior distribution, further datasets besides the original sample are needed. For instance, in our empirical analysis regarding sovereigns, we will further use data on firms to estimate the prior distribution. In many practical situations, such auxiliary datasets will be available. For example, a bank will typically have a variety of different portfolios and the information from all these portfolios can be used within the EB approach to estimate the default probabilities for each particular portfolio. Similarly, default histories referring to external ratings can be used as a prior for default probability estimation based on internal data. The combination of different datasets is - without any explicit proposal - also mentioned from the regulatory side as one tool for default probability estimation in the case of low-default portfolios (Basel Committee on Banking Supervision, 2005c).

We will now formally introduce our EB estimator and subsequently give some further discussion. Like in the case of the standard estimator, we will start with the estimation of discrete-time hazard rates which are then used to construct cumulative default rates. Suppose that we have G different groups or portfolios (corresponding to the different datasets at hand) where $G \geq 2$. We make the parametric assumption that for each group g, $g = 1, \ldots, G$, the

[8] This resembles very much the problems of constructing a conservative default probability estimate from conservative estimates for the hazard rate. See section 4.1.

[9] For an introduction to EB methods we recommend Casella (1985) and Carlin & Louis (2008).

hazard rates are a priori beta distributed,

$$\lambda^r_{s,g} \sim beta(\alpha^r_s, \beta^r_s)\,. \tag{4.6}$$

Note that each group has the same prior parameters. Further, we assume that the conditional distribution of the number of defaults in period s is binomial,

$$D^r_{s,g}|\lambda^r_{s,g} \sim Bin(\widetilde{N}^r_{s,g}, \lambda^r_{s,g})\,, \tag{4.7}$$

where we have now, to simplify notation, skipped the index t to indicate aggregation over the cohort building dates, i.e. $D^r_{s,g} = \sum_{t=1}^{T} D^r_{t,s,g}$ and $\widetilde{N}^r_{s,g} = \sum_{t=1}^{T}(N^r_{t,s,g} - L^r_{t,s,g}/2)$. The presented framework is known as the beta-binomial model and is quite common for the Bayesian analysis of proportions. The beta distribution is a pretty flexible distribution for parameters bounded in the interval [0,1] and has also been suggested by Kiefer (2009).[10] The crucial part of the binomial assumption is the conditional independence of default events. Note that although we aggregate over different cohort building dates we do not use the same default event more than once (and thus do not have dependence caused by overlapping lifetimes) for fixed s. However, we disregard the dependence of default events induced by common shocks to keep the analysis as simple as sensibly possible. In our simulation study of section 4.5, we will show that the estimator works well even for data generating processes that involve dependencies through common shocks.

The next step is now to estimate the prior parameters. We do so by using the Method of Moments hereby essentially following the analysis of Kleinman (1973).[11] For convenience, we reparameterize the beta distribution setting $\mu^r_s = \alpha^r_s/(\alpha^r_s + \beta^r_s)$ to be the prior mean of $\lambda^r_{s,g}$ and $\tau^r_s = 1/(1 + \alpha^r_s + \beta^r_s)$ to be a measure of prior precision.[12] We estimate μ^r_s as a weighted average of

[10] Kiefer (2010) shows how the beta distribution can be generalized for even more flexibility. Note that prior distributions with more parameters will increase the minimum number of groups.

[11] A more recent study that deals with the estimation of beta-binomial parameters is from Tamura & Young (1987). There, a stabilized estimator is introduced that should be more robust to small changes in the data. We also experimented with this approach but did not find it useful in our application.

[12] In terms of μ^r_s and τ^r_s the prior variance is given by $\tau^r_s\mu^r_s(1-\mu^r_s)$.

the group-specific standard marginal default rates:

$$\widehat{\mu}_s^r = \sum_{g=1}^{G} w_{s,g}^r \frac{D_{s,g}^r}{\widetilde{N}_{s,g}^r} = \sum_{g=1}^{G} w_{s,g}^r \widehat{\lambda}_{s,g}^r \tag{4.8}$$

The formula we use to estimate τ_s^r is

$$\widehat{\tau}_s^r = \frac{\frac{G-1}{G}\sum_{g=1}^{G} w_{s,g}^r(\widehat{\lambda}_{s,g}^r - \widehat{\mu}_s^r)^2 - \widehat{\mu}_s^r(1-\widehat{\mu}_s^r)\left(\sum_{g=1}^{G} w_{s,g}^r(1-w_{s,g}^r)/\widetilde{N}_{s,g}^r\right)}{\widehat{\mu}_s^r(1-\widehat{\mu}_s^r)\left(\sum_{g=1}^{G}(1-1/\widetilde{N}_{s,g}^r)w_{s,g}^r(1-w_{s,g}^r)\right)}\,. \tag{4.9}$$

See Kleinman (1973) for a detailed derivation. Natural choices for the weights are equal weights, i.e. $w_{s,g}^r = 1/G$, or the number of observations for each group so that $w_{s,g}^r = \widetilde{N}_{s,g}^r / \sum_{g=1}^{G} \widetilde{N}_{s,g}^r$. Kleinman (1973) shows that the optimal weights depend on the true parameters and proposes to use one iteration to refine the estimates, namely to set

$$w_{s,g}^r = \frac{\widetilde{N}_{s,g}^r}{1+\widehat{\tau}_s^r(\widetilde{N}_{s,g}^r - 1)} \Bigg/ \sum_{j=1}^{G} \frac{\widetilde{N}_{s,j}^r}{1+\widehat{\tau}_s^r(\widetilde{N}_{s,j}^r - 1)}\,, \tag{4.10}$$

using a preliminary estimate of τ_s^r to get improved weights which are subsequently employed to re-estimate the prior parameters. In our implementation, we used this one-time iteration step with starting weights $w_{s,g}^r = 1/G$. Note that we also experimented with an omission of the iteration step and found no large sensitivity of the results in this respect. Since there is no guarantee that $\widehat{\tau}_s^r$ will be in the interval $(0,1)$ which is necessary for a proper prior the estimates of $\widehat{\tau}_s^r$ should be truncated at zero and one, respectively.

With the estimated prior parameters at hand, we can apply the Bayesian theorem to arrive at the posterior distribution of our parameters. Since the beta distribution is the conjugate prior for the binomial distribution, the posterior distribution of $\lambda_{s,g}^r$ is beta as well. The mean of the posterior distribution minimizes the Bayes risk under quadratic loss functions and is the standard choice for a Bayesian point estimator. In our case, the posterior mean, i.e. our EB estimator for $\lambda_{s,g}^r$, can be written as

$$\widehat{\lambda}_{s,g,EB}^r = \frac{1-\widehat{\tau}_s^r}{1+\widehat{\tau}_s^r(\widetilde{N}_{s,g}^r - 1)}\widehat{\mu}_s^r + \frac{\widehat{\tau}_s^r \widetilde{N}_{s,g}^r}{1+\widehat{\tau}_s^r(\widetilde{N}_{s,g}^r - 1)}\widehat{\lambda}_{s,g}^r\,. \tag{4.11}$$

The EB estimator is obviously a weighted average of the prior mean (which itself is a weighted average of the group-specific standard estimates) and the standard marginal default rate for group g. The estimator can be interpreted as a shrinkage estimator since it shrinks the standard estimates, $\widehat{\lambda}^r_{s,g}$, towards the prior means which are equal for all groups. Note also that the weighting scheme is such that there is a smooth transition to the standard estimator if $\widetilde{N}^r_{s,g}$ grows. Thus, the amount of shrinkage will, in line with intuition, decline as the sample size of a specific group increases. We provide R code for Formulas (4.8)-(4.11) in appendix 4.C.

Like in section 4.1, our estimate for the default probability is constructed from the marginal default rates,

$$\widehat{PD}^r_{s,g,EB} = 1 - \prod_{j=1}^{s}(1 - \widehat{\lambda}^r_{j,g,EB}) \tag{4.12}$$

It is worth mentioning that our estimator minimizes the Bayes risk with respect to the marginal default rates instead of the cumulative default rates. Doing the latter would actually be preferable but would considerably increase the complexity of the problem in our setting since we would have to deal with the dependencies due to overlapping lifetimes. Note, however, that our estimator can be interpreted as minimizing the Bayes risk for the default probability under a working independence assumption. In this case, our estimator is equal to the one derived by Hjort (1990) for the discrete-time case except for the fact that Hjort (1990) does not consider *empirical* Bayes estimation. Our estimator may thus be seen as the corresponding extension. Similarly, we extend the (empirical Bayesian) beta-binomial framework - originally developed for binary data - to survival data.

Although derived from Bayesian theory, EB methods have been shown to be an improvement over standard Maximum Likelihood methods in many applications even by frequentist measures such as Mean Squared Error (Casella, 1985). Usually EB methods lead to a lower variance as compared to their Maximum Likelihood counterparts at the cost of introducing or magnifying some (finite-sample) bias. Consider for instance our application where we will combine sovereign and corporate datasets. The more the true default probabilities for both groups are apart the larger will be the bias introduced

by the EB approach. However, if the differences are rather small the effect of variance reduction will prevail and lead to smaller Mean Squared Errors. Especially in small samples, where the variance of the standard estimator is high, the potential gains from variance reduction can be substantial. In our simulation study in section 4.5, we will illustrate this bias-variance trade-off under realistic scenarios. Moreover, if conservativeness is by itself desirable, as it is stated at least by regulators with respect to default probability estimation, a moderate upward bias induced by EB methods may even be seen as a benefit rather than a weakness.

In appendix 4.B, we show that our EB estimator is consistent for $\widetilde{N}^r_{s,g} \to \infty, g = 1, \ldots, G$. Consistency is not trivial since we consider fixed G so that the prior parameter estimates $\widehat{\mu}^r_s$ and $\widehat{\tau}^r_s$ do not generally converge to their population counterparts. Further, the consistency of the EB estimator is an important difference to the simple weighted average, $\widehat{\mu}^r_s$, which is generally not consistent.[13] Practically, this means that if one would consider only $\widehat{\mu}^r_s$ and $\widehat{\lambda}^r_{s,g}$ (the standard estimator) one would have to make ad hoc decisions up to which sample size different portfolios should be pooled. The smooth transition of the EB estimator to the standard estimator makes such decisions unnecessary.

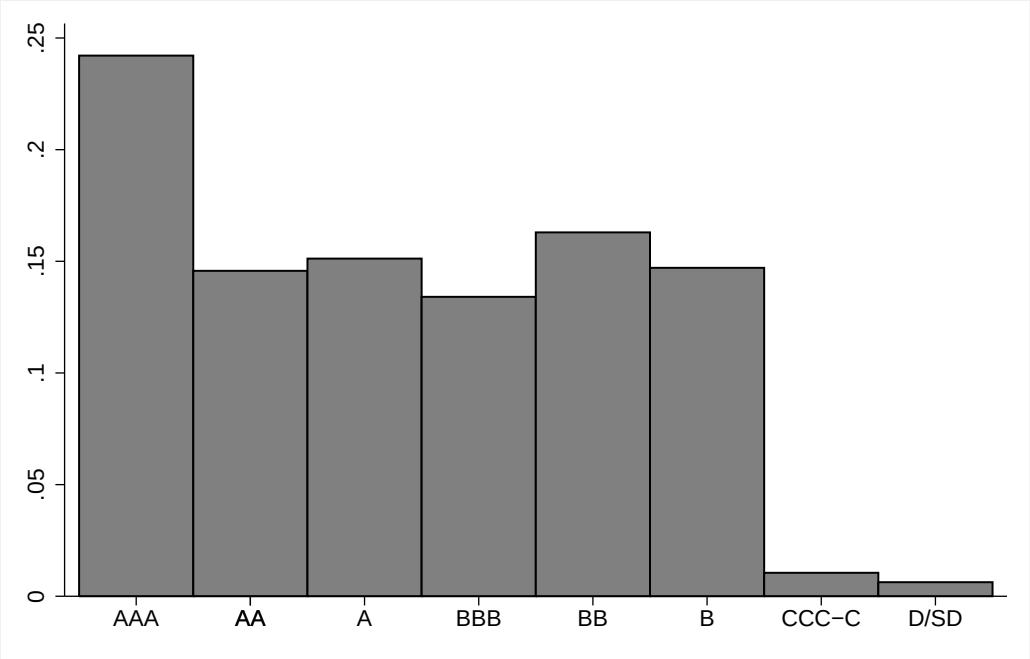
4.4 Application to sovereign bonds

Sovereign bonds provide possibly the most important application for our methods since they are among the most important asset classes and sovereign defaults are rare events. In this section, we use the complete rating and default histories of sovereigns with public ratings from Standard & Poor's (S&P) in the period from January 1975 until April 2011.[14] The data are from Standard & Poor's (2011b) and consist of 130 sovereigns observed over a total of 23014 country-months. The dataset may thus appear not that small but the sample size per rating class is of course considerably lower

[13] Instead, $\widehat{\mu}^r_s$ converges to a weighted average of the group-specific probability limits. See appendix 4.B, where these probability limits are denoted by $\overline{\lambda}_1$ and $\overline{\lambda}_2$, respectively.

[14] S&P also rates a few sovereigns on a confidential basis. See Standard & Poor's (2011a).

Figure 4.1: Rating distribution of S&P rated sovereign bonds, 1975-2011



and, importantly, we observe only 15 default events. More precisely, these default events are foreign-currency selective defaults.[15] Accordingly, we will use foreign-currency issuer credit ratings in our analysis since these have longer rating histories and are probably in most cases more relevant to investors than local-currency ratings. Figure 4.1 shows the rating distribution in our sample.

Apart from the data concerning sovereigns, we further utilize S&P rating and default histories of North American public firms from Compustat covering the period from January 1981 until April 2011. The corresponding ratings used here are S&P's long term issuer credit ratings which are on the same scale and have the same definition as their sovereign counterparts. The latter can be seen as a justification of our Bayesian assumption that there is no difference between sovereigns and firms *a priori*. This second dataset is large (5355 firms, 563809 firm-months, 755 defaults) and will be used for the EB approach as explained in the previous section.

[15] A selective default means that a sovereign entity defaults only on a part of its bonds.

Table 4.1 shows cumulative default rates for sovereigns and corporates using the EB and the standard estimator, respectively. To start with Panel B, which gives the standard estimator for sovereigns, we see that we get default probability estimates of zero throughout all time horizons for the three highest grades and under a one-year horizon even for BBB rated sovereigns. This is clearly an unpleasant feature since such estimates are anticonservative and also not in line with market perceptions given that credit default swaps are traded even for highly rated sovereigns. In contrast, the EB estimator manages to remove most of the zeros with the exception of the AAA category where we do not have any corporate default in our sample as well. Due to the relatively small size of the sovereign sample the EB estimator is dominated by the standard estimator for corporates as can be seen by comparing Panel A and Panel D. However, this closeness is varying. For instance, in the case of sovereigns rated B, where we have relatively much information in the sense that we have some defaults and not too few sovereigns rated B, we observe that the sovereign estimates are less close to the corporate estimates as they are for other grades. Overall, we see that the sovereign default probability estimates are more conservative under the EB approach while the increase in conservativeness seems to be at a reasonable degree.

The EB estimator for corporates is of less interest but reported in Panel C for completeness. As expected from theory, EB estimator and standard estimator are very close to each other due to the large sample size. Further, it is interesting to see that for grades AAA-BBB the EB cumulative default rates are the same for sovereigns and corporates. This corresponds to the fact that in these cases the EB estimator equals the pooled estimator ($\widehat{\mu}$ in the terminology of section 4.3) which is also in line with expectations as in these categories there are very few, if any, default events and a possible variance reduction from pooling dominates the weighting scheme in Formula (4.11).

We now go on to analyze the economic impact of our different estimators. The two applications we have chosen are the estimation of expected returns and the calculation of economic capital. With respect to the former, we consider sovereign bond investments with a maturity of up to 10 years. For these bonds, we consider a simple hold-to-maturity strategy and calculate

Table 4.1: Cumulative default rates (1-10 years)

Panel A: Empirical Bayes estimator for sovereigns

	1	2	3	4	5	6	7	8	9	10
AAA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	0.00	0.01	0.05	0.08	0.08	0.09	0.12	0.13	0.17	0.22
A	0.06	0.16	0.33	0.50	0.65	0.79	0.92	1.08	1.26	1.42
BBB	0.20	0.57	1.07	1.65	2.24	2.81	3.40	3.95	4.60	5.34
BB	0.76	2.42	4.24	5.70	7.56	9.40	11.27	13.04	14.74	15.95
B	3.88	7.64	11.43	15.80	19.58	23.07	26.01	28.66	31.04	33.21
CCC-C	24.38	33.66	39.87	43.89	48.02	52.72	66.77	82.27	82.46	82.88

Panel B: Standard estimator for sovereigns

	1	2	3	4	5	6	7	8	9	10
AAA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BBB	0.00	0.66	1.60	2.62	3.71	4.85	5.27	5.27	5.27	5.27
BB	0.56	1.88	3.06	3.66	5.10	6.75	8.74	10.29	11.23	11.43
B	2.60	5.16	7.59	11.01	13.12	15.48	18.30	22.17	25.44	29.02
CCC-C	32.27	44.50	51.56	55.76	63.45	72.19	83.85	91.92	91.92	91.92

Panel C: Empirical Bayes estimator for corporates

	1	2	3	4	5	6	7	8	9	10
AAA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	0.00	0.01	0.05	0.08	0.08	0.09	0.12	0.13	0.17	0.22
A	0.06	0.16	0.33	0.50	0.65	0.79	0.92	1.08	1.26	1.42
BBB	0.20	0.57	1.07	1.65	2.24	2.81	3.40	3.95	4.60	5.34
BB	0.76	2.42	4.24	6.05	7.90	9.74	11.59	13.36	15.06	16.59
B	4.40	9.85	14.72	18.96	22.59	25.95	28.78	31.33	33.63	35.71
CCC-C	24.35	33.52	39.74	43.77	47.91	51.17	53.45	54.13	54.61	55.70

Panel D: Standard estimator for corporates

	1	2	3	4	5	6	7	8	9	10
AAA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	0.00	0.01	0.05	0.08	0.09	0.10	0.13	0.14	0.18	0.24
A	0.06	0.16	0.33	0.51	0.67	0.81	0.95	1.11	1.29	1.45
BBB	0.20	0.57	1.06	1.63	2.20	2.76	3.36	3.92	4.59	5.36
BB	0.77	2.45	4.29	6.15	8.02	9.86	11.72	13.50	15.24	16.89
B	4.47	9.96	14.91	19.21	22.98	26.38	29.22	31.70	33.96	35.97
CCC-C	24.20	33.31	39.49	43.52	47.48	50.60	52.89	53.58	54.07	55.17

Numbers are in percentage points.

expected returns by replacing contractual cash flows with their expected values and computing the corresponding yield-to-maturity.[16] Besides a term structure of default probabilities, this requires an assumption for the recovery rate, i.e. the proportion of the face value of the bond that is recovered if default occurs. For our calculations, we assume a recovery rate of 0.55 which is the middle of the interval [0.5, 0.6] reported in Standard & Poor's (2011c) as the estimated historical average sovereign recovery rate. Our choice for the recovery rate also coincides with the loss given default (= 1− recovery rate) assumption of 0.45 which is prescribed in the foundation Internal Ratings Based (IRB) approach of Basel II (Basel Committee on Banking Supervision, 2006, §287). For the results shown in Table 4.2, we have selected one US Dollar denominated bond for each rating grade with the exception of the CCC-C grades since no sovereign had such a rating on May 2, 2011 (our hypothetical bond purchase date). By comparing the maximum return, i.e. the return that an investor will receive if no default occurs, with the expected returns we can see which part of the maximum return an investor can expect to lose on average by taking the risk of the corresponding bond investment. Under each estimator, the results are in line with the basic risk-return paradigm in the sense that expected returns monotonically rise as ratings worsen. Further, with the exception of the BBB class, the reward for risk is estimated to be lower under the EB approach which, of course, directly follows from its relative conservatism.

Our second application is to use our one-year default probability estimates as inputs to the Basel II capital formula (Basel Committee on Banking Supervision, 2006, §272).[17] Besides the default probability, the formula requires as an input an estimate of the loss given default which we again set to 0.45. The correlations which are also part of the formula are defined by the regulators

[16] By choosing a hold-to-maturity scenario we do not need a model for the bond price process or for rating transitions since these do not affect expected returns in this case.

[17] For an explanation of the theoretical underpinning of the Basel II capital formula see Basel Committee on Banking Supervision (2005a). Under the new regulatory initiative called Basel III the capital formula is not intended to be changed in its structure. However, the capital requirements are likely to be scaled up. See Basel Committee on Banking Supervision (2010) for details.

Table 4.2: Expected returns for selected USD denominated sovereign bonds

				Expected returns	
Country	Rating	Maturity	Max. return	Standard estimator	Empirical Bayes
USA	AAA	2/2021	3.363	3.363	3.363
Qatar	AA	1/2020	4.584	4.584	4.576
Poland	A	7/2019	4.750	4.750	4.687
Lithuania	BBB	3/2021	5.580	5.312	5.335
Egypt	BB	4/2020	6.557	5.962	5.766
Argentina	B	6/2017	8.783	7.332	6.982

Sovereign bond data are from Boerse Frankfurt. Expected returns are calculated under the assumption of a bond purchase on May 2, 2011, and a hold-to-maturity strategy.

Table 4.3: Basel II capital requirements

	ST	SE	EB	EB*	% of MP
AAA	0.00	0.00	0.00	0.00	46.6
AA	0.00	0.00	0.00	0.77	35.7
A	1.60	0.00	1.69	1.69	8.4
BBB	4.00	0.00	3.48	3.48	5.8
BB	8.00	5.88	6.66	6.66	2.8
B	8.00	9.87	11.06	11.06	0.7
CCC-C	12.00	19.85	19.67	19.67	0.0
MP	0.65	0.23	0.61	0.88	

Numbers in percentage points. ST: Basel II Standardised Approach; SE: Standard estimator; EB: Empirical Bayes estimator; EB*: Empirical Bayes estimator with one-year AA default probability calculated by scaling down the associated three-year default probability. MP refers to an approximate market portfolio (see the text for details).

as a function of the default probability. Further, under the advanced IRB approach there are potential maturity adjustments if the effective maturity differs from 2.5 years. To facilitate comparisons, we assume the standard maturity of 2.5 years. Table 4.3 shows the results for the capital requirements. The numbers can be interpreted as the capital which banks must

hold for a corresponding investment of 100 currency units. The first column contains the capital requirements under the standardised approach of Basel II where banks do not estimate default probabilities themselves and use instead fixed ratings-based risk weightings (Basel Committee on Banking Supervision, 2006, §53). The corresponding capital requirements are intended to be conservative in order to give banks an incentive to intensify their own risk analysis and to move eventually to the IRB approach. Comparing the first with the second column, we see that indeed under an IRB approach that uses our standard estimator the capital requirements are considerably lower than under the standardised approach. However, the capital ratios under the standard estimator seem to be very liberal since no capital at all is needed for BBB or better rated sovereigns including, for instance, South Africa and Peru at the end of our sample period.[18] In contrast, the EB estimates are more conservative while not implying unrealistically high levels of capital as is seen by their closeness to the standardised approach. In the column of Table 4.3 which is labelled by EB*, we have kept the original EB estimates with the exception of the AA category. To get a non-zero estimate for this class, we used a proposal from Basel Committee on Banking Supervision (2005c) to scale down multi-year default probabilities in the case of sparse data. Specifically, we used the non-zero AA three-year default probability estimated under the EB approach to calculate the one-year default probability under the assumption of constant marginal default rates, i.e. $\widehat{PD}_1^{AA,*} = 1 - (1 - \widehat{PD}_3^{AA})^{1/3} = 0.015\%$. At first sight, the new AA capital ratio of 0.77% seems to be negligibly small. However, matters change if we analyze the impact on the capital requirements of a bank which holds an approximate market portfolio. The composition of the market portfolio is given in the last column of Table 4.3 and is calculated from data of the Bank for International Settlements on the total amounts of outstanding debt

[18] Portugal and Ireland would be examples from the Euro area which are rated BBB at the end of our sample period. Note, however, that the EU capital requirements directives which refer to the implementation of the Basel II rules introduced a general zero capital charge for member states' sovereign bonds - against the intention of the Basel Committee for Banking Supervision (Hannoun, 2011). The soundness of this exemption can be seen very critical, of course.

in government securities.[19] Holding the market portfolio, a bank would increase its capital requirement to 0.88% from 0.61% under the unadjusted EB approach and from 0.23% under the standard estimator. These differences are of course substantial and show the high sensitivity of capital ratios to default probability estimates. Moreover, it is obvious that the standard default probability estimator is far from being conservative in this respect.

4.5 Simulation study

While we have seen that EB estimators have nice theoretical properties and give reasonable results in our empirical application it is clearly of interest to study the performance of the EB estimator and the standard estimator in more detail.[20] Out-of-sample tests are no appropriate option in our case since our small sample size would not allow us to draw meaningful conclusions. Instead, we will evaluate our estimators by means of a simulation study. The specification of our data generating process is as follows. With respect to the sample size, we stick to the data used in the previous section. More precisely, we drop all observations from our datasets with the exception of the observations where a firm or sovereign first entered the dataset. For instance, the United States enter our dataset in January 1975 with a AAA rating which remains constant until the end of our sample. For our simulation, we keep only the rating in January 1975 whereas the subsequent ratings are now filled

[19]The data are available under http://bis.org/statistics/secstats.htm. We aggregated the outstanding amounts of international and domestic debt securities per government (Tables 12D and 16A) as of December 2010. The corresponding S&P ratings for the same date were used to compute aggregate amounts of debt per rating class. Note that no government was rated CCC or lower at this point in time so that the CCC-C category estimates do not influence the market portfolio calculations.

[20]Even for the standard estimator, a comprehensive evaluation of its properties in the presence of overlapping lifetimes does not exist in the literature so far, to the best of our knowledge.

up in the simulation process.[21] We choose a Markovian rating migration model (again on a monthly basis) to simulate rating transitions. While we have argued that the Markovian model has serious drawbacks for default probability estimation it should nevertheless be suitable for our simulations since none of our estimators relies on the Markovian assumption. We tested several structures for the data-generating migration matrix and found the following one to lead to a realistic migration behavior as well as reasonable levels of pseudo-true default probabilities:[22]

	AAA	AA	A	BBB	BB	B	$CCC-C$	D/SD
AAA	$1-\frac{7}{4}m$	m	$m/2$	$m/4$	0	0	0	0
AA	m	$1-\frac{11}{4}m$	m	$m/2$	$m/4$	0	0	0
A	$m/2$	m	$1-\frac{13}{4}m$	m	$m/2$	$m/4$	0	0
BBB	$m/4$	$m/2$	m	$1-\frac{7}{2}m$	m	$m/2$	$m/4$	0
BB	0	$m/4$	$m/2$	m	$1-\frac{7}{2}m$	m	$m/2$	$m/4$
B	0	0	$m/4$	$m/2$	m	$1-\frac{13}{4}m$	m	$m/2$
$CCC-C$	0	0	0	m	$2m$	$4m$	$1-15m$	$8m$
D/SD	0	0	0	0	0	0	0	1

The entry in the ith row of the jth column is the probability to migrate from class i to class j over the next month. m is a parameter that refers to the basic migration rate into the neighboring classes and has to be specified. We choose $m = 0.003$ for sovereigns. Probabilities for migrations over more than one class are assumed to halve with each step giving migration probabilities of $m/2$ and $m/4$. Migrations over more than three classes within one month seem quite unrealistic so that we set the corresponding migration probabilities to zero. While the migration rates for the upper six categories follow the same pattern it was necessary to introduce higher migration rates for the CCC-C category to account for the high CCC-C default rates observed empirically. The main diagonale of the migration matrix is simply specified in the way that the rows sum up to one. Based on the empirical finding that

[21] In our corporate sample, we have firms that are not observed for some periods and then return at a later point in time. To account for these censoring events, we keep the first rating of these firms after their return for our simulations and treat them as if they were new firms.

[22] Default probabilities are taken from the exponentiated migration matrix. If M denotes the migration matrix, the last column of M^H contains the default probabilities. Their values are presented below together with the simulation results.

firms have higher migration rates (resulting in higher default probabilities) we simply rescale the migration matrix for corporates by multiplying m with some constant k. In our simulations, we will consider $k = 1, 1.25, 1.5, 1.75$.

At first sight it might seem more appropriate to simply choose a migration matrix based on historical migration rates for the data-generating process. However, we have chosen not to do so because of two reasons. First, for reasons which we have discussed in the introduction, the implied pseudo-true default probabilities would be at an unrealistically low level. For instance, in this case we would have implied pseudo-true default probabilities of 0.06% for BB rated sovereigns at a one-year horizon and 0.13% for BBB rated sovereigns at a five-year horizon. These default probabilities are considerably lower than our standard estimates which - as will be confirmed by our simulations - already have a tendency to underestimate true default probabilities in small samples. Second, we want to investigate different scenarios for the difference between sovereigns and firms (specified by different choices for k) which is more straightforward within our setting.

Sovereign and corporate default and migration rates are very likely to be affected by common shocks like, for instance, recessions. We account for this kind of dependence by applying a CreditMetrics^{TM}-type approach (Gupton et al., 1997). The procedure involves simulating observations from a multivariate normal distribution and mapping these realizations to rating changes. Consider for example a AAA rated sovereign which has a probability to remain AAA over the next month of $1 - \frac{7}{4} \cdot 0.003 = .99475$ and a probability to migrate to AA of 0.003. If the corresponding realization of the normal distribution is smaller than $\Phi^{-1}(.99475) \approx 2.5589$ the rating for the next month is set to AAA again. If instead the realization is in the interval $[\Phi^{-1}(.99475), \Phi^{-1}(.99775)] \approx [2.5589, 2.8408]$ the sovereign migrates to AA, and so on. The correlations of the corresponding multivariate normal distribution have the same meaning as the so-called asset correlations that are part of the IRB formula in the Basel II framework.[23] There, the asset corre-

[23] Asset correlations are meant to be the correlations between the asset values of obligors. The underlying theory is that once these asset values cross a certain lower threshold, a default event occurs.

Table 4.4: Evaluation of the standard estimator

	Panel A: Pseudo-true PDs (%)				Panel B: Relative bias			
	1 year	3 years	5 years	10 years	1 year	3 years	5 years	10 years
AAA	4.7e-4	0.014	0.061	0.414	-0.213	0.044	0.083	0.058
AA	0.005	0.067	0.223	1.069	0.085	0.059	0.032	0.018
A	0.018	0.203	0.609	2.414	0.070	0.022	0.011	0.003
BBB	0.132	0.981	2.245	6.101	0.005	0.016	0.007	-0.001
BB	1.082	3.905	6.958	14.124	0.001	-0.005	-0.006	-0.008
B	2.122	7.355	12.615	23.544	0.007	0.004	0.001	-0.010
CCC-C	22.786	44.379	52.797	60.227	0.038	0.014	-0.004	-0.019

	Panel C: Relative RMSE				Panel D: % $\widehat{PD} < PD$			
	1 year	3 years	5 years	10 years	1 year	3 years	5 years	10 years
AAA	15.393	5.722	3.680	2.105	99.62	94.98	88.48	74.02
AA	7.271	3.529	2.525	1.693	97.06	87.98	80.08	68.32
A	3.685	1.871	1.440	1.123	89.68	71.64	63.90	59.30
BBB	1.358	0.914	0.801	0.719	61.74	56.62	56.56	56.50
BB	0.561	0.498	0.479	0.478	53.52	53.74	53.70	54.30
B	0.453	0.415	0.407	0.428	53.22	52.44	53.14	52.72
CCC-C	0.378	0.355	0.352	0.356	49.50	50.24	50.54	51.52

Relative bias and Relative Root Mean Squared Error (RMSE) are calculated as $(\overline{\widehat{PD}} - PD)/PD$ and $RMSE/PD$, respectively. % $\widehat{PD} < PD$ is the percentage of simulations for which the estimated default probability was below the pseudo-true default probability. The number of simulations is 5000.

lations are specified as a function of the one-year default probability (Basel Committee on Banking Supervision, 2006, §272). We adopt this approach to specifiy the correlations of our multivariate normal distribution.

For the sake of illustration, we will from now on concentrate on prediction horizons of 1, 3, 5 and 10 years. All upcoming results are based on 5000 simulations. We start the presentation of our simulation results with the evaluation of the standard estimator which is given in Table 4.4. Panel A of Table 4.4 shows the pseudo-true default probabilities implied by our data-generating process which are of similar magnitude as our empirical estimates

but, importantly, are not zero even for the highest rating grades. In Panel B we see the estimated bias of the standard estimator relative to the pseudo-true values. While we know that the standard estimator has only minimal asymptotic bias[24] (see section 4.1) it is interesting to see that there is also no significant bias (with the exception of the CCC-C category) in small samples.[25]

The accuracy of the standard estimator as measured by the Root Mean Squared Error (RMSE) relative to the pseudo-true default probabilities is shown in Panel C of Table 4.4. It is clearly visible that estimation uncertainty rises in relative terms as the pseudo-true values decline. Therefore, especially in these cases there should be potential to improve upon the standard estimator. Finally, in Panel D we report the proportions of the simulations where the pseudo-true default probability has been underestimated by the standard estimator. The fact that we observe values well above 50% is caused by the highly skewed sampling distribution of the standard estimator in small samples and especially under small true default probabilities. Note that this feature can also be interpreted as a kind of bias called median bias. Following Birnbaum (1963), the median bias of a default probability estimator is given as $P(\widehat{PD} > PD|PD) - P(\widehat{PD} < PD|PD)$, and the estimator is called median-unbiased if the median bias is equal to zero. Under this concept, although being approximately mean-unbiased, the standard estimator is clearly downward median-biased which is an obvious drawback at least if conservativeness is among the criteria to evaluate an estimator.

We now turn to the evaluation of the EB estimator. Table 4.5 shows its precision as measured by the ratio of the RMSEs of the EB and the standard estimator so that values smaller than 1 indicate a superior performance of the EB estimator. We report results for three scenarios, $k = 1.25$ in Panel

[24]Serious asymptotic biases can occur if the assumptions regarding the censoring scheme (see section 4.1) are not met. In our simulation study, censoring times are fixed so that the assumption of noninformative censoring is fulfilled.

[25]We explored the significance of the bias by using Monte Carlo standard errors. At a confidence level of $\gamma = 0.05$, the bias was only significant for the CCC-C default probabilities at horizons of 1, 3 and 10 years. The special role of the CCC-C category is not too surprising given that it has the smallest sample size.

Table 4.5: Precision of the empirical Bayes estimator

Panel A: $k = 1.25$

	Pseudo-true PD ratio				RMSE ratio			
	1 year	3 years	5 years	10 years	1 year	3 years	5 years	10 years
AAA	1.94	1.90	1.86	1.79	0.841	0.737	0.708	0.735
AA	1.65	1.68	1.67	1.61	0.814	0.656	0.608	0.642
A	1.61	1.62	1.58	1.50	0.650	0.584	0.621	0.648
BBB	1.52	1.45	1.40	1.34	0.555	0.597	0.611	0.605
BB	1.29	1.29	1.27	1.23	0.627	0.672	0.662	0.616
B	1.29	1.27	1.24	1.18	0.688	0.705	0.666	0.560
CCC-C	1.18	1.09	1.05	1.03	0.559	0.480	0.487	0.580

Panel B: $k = 1.5$

	Pseudo-true PD ratio				RMSE ratio			
	1 year	3 years	5 years	10 years	1 year	3 years	5 years	10 years
AAA	3.32	3.19	3.07	2.83	0.848	0.819	0.881	1.056
AA	2.49	2.57	2.52	2.34	0.830	0.746	0.819	0.955
A	2.39	2.38	2.27	2.04	0.696	0.884	0.965	0.980
BBB	2.13	1.94	1.83	1.67	0.826	0.930	0.946	0.883
BB	1.60	1.59	1.53	1.43	0.944	1.005	0.964	0.845
B	1.58	1.54	1.47	1.34	1.082	1.081	0.975	0.732
CCC-C	1.34	1.16	1.09	1.06	0.798	0.543	0.509	0.574

Panel C: $k = 1.75$

	Pseudo-true PD ratio				RMSE ratio			
	1 year	3 years	5 years	10 years	1 year	3 years	5 years	10 years
AAA	5.22	4.92	4.67	4.13	1.023	1.190	1.318	1.532
AA	3.54	3.67	3.54	3.17	0.818	0.975	1.178	1.327
A	3.33	3.28	3.07	2.63	0.829	1.225	1.344	1.314
BBB	2.81	2.46	2.27	2.01	1.178	1.302	1.274	1.181
BB	1.91	1.88	1.79	1.62	1.284	1.330	1.258	1.050
B	1.88	1.80	1.68	1.48	1.374	1.359	1.220	0.896
CCC-C	1.48	1.21	1.12	1.09	1.006	0.611	0.555	0.604

The pseudo-true PD ratio is calculated as PD(corporate)/PD(sovereign) and the RMSE ratio is defined as $RMSE$(EB)/$RMSE$(standard estimator). The number of simulations is 5000.

A, $k = 1.5$ in Panel B and $k = 1.75$ in Panel C. To provide insight into the relative level of default probabilities implied by these specifications we show in the left parts of Table 4.5 the ratio of the pseudo-true default probabilities of corporates and sovereigns. For instance, the BBB one-year default probability is 52% higher for corporates than for sovereigns under $k = 1.25$. Still, in this case the RMSE of the EB estimator is 44.5% lower than the RMSE of the standard estimator. Overall, we observe an improvement by using the EB estimator in all cases for $k = 1.25$, in all cases with a few exceptions at $k = 1.5$ and even in some cases for $k = 1.75$. The relative strength of the EB estimator increases i) as the sample size decreases (as can be seen by the large improvements for the CCC-C category), ii) as the pseudo-true default probabilities decrease (see the robust EB performance with respect to the AAA and AA one-year default probabilities) and iii) as the distance between the corporate and the sovereign pseudo-true default probabilities decreases (Panel C up to Panel A). Case i) is expected from theory and further confirmed by additional simulations which we do not report here. In these simulations, we randomly dropped half of our sample from the simulations which is still likely to be a practically realistic sample size. Under this reduced sample the relative EB performance is even better. For instance, under $k = 1.5$, the RMSE ratio of the A one-year default probability then decreases from 0.696 to 0.520.

We see that, depending on the true data generating process, the EB estimator may or may not be more precise than the standard estimator. More clearly, we can ascribe the EB estimator to be more conservative which can be seen from the results in Table 4.6. To evaluate the conservativeness of the EB estimator, we have chosen the scenarios $k = 1$ and $k = 1.25$. For larger values of k the conservativeness of the EB estimator will obviously further rise. But even for $k = 1$, the relative frequency of underestimating the pseudo-true default probability is considerably lower than for the standard estimator which was given in Table 4.4. Since no bias is introduced in the case of $k = 1$ the reason for this finding is just the less skewed sampling distribution of the EB estimator, an effect similar to the effect of an increasing sample size. In the case of $k = 1.25$, an additional upward bias is present so that underestimation of the pseudo-true default probability happens only in less

Table 4.6: Conservativeness of the empirical Bayes estimator

	% $\widehat{PD} < PD$ for $k = 1$				% $\widehat{PD} < PD$ for $k = 1.25$			
	1 year	3 years	5 years	10 years	1 year	3 years	5 years	10 years
AAA	95.6	73.5	60.4	52.5	91.2	54.9	38.3	28.7
AA	64.7	53.9	49.3	46.1	50.2	27.6	20.2	16.1
A	53.2	46.3	47.3	46.9	27.3	13.3	11.6	13.0
BBB	47.3	51.5	53.6	53.2	10.0	12.9	19.0	21.9
BB	53.3	53.7	53.7	53.6	19.5	19.5	21.4	25.7
B	54.3	54.8	53.8	53.7	16.5	16.8	18.8	24.4
CCC-C	53.2	50.7	46.9	40.6	12.8	23.3	30.5	31.3

% $\widehat{PD} < PD$ is again the percentage of simulations for which the estimated default probability was below the pseudo-true default probability. The number of simulations is 5000.

than 50% of all cases with a few exceptions for very small pseudo-true default probabilities where the skewness effect still dominates.

The analysis of our estimators on a portfolio basis provides further insights. We again stick to our approximate market portfolio (see section 4.4) and consider the estimation of expected losses and capital requirements again assuming a recovery rate of 0.55. As was already mentioned in chapter 1, the estimation of expected losses over the whole life of the portfolio to calculate loan loss reserves may gain importance due to recent regulatory efforts (Basel Committee on Banking Supervision, 2009). Differently to the expected return calculations in section 4.4, since we now do not refer to any specific bond, we do not consider any coupon payments instead assuming only one hypothetical cash flow at the end of the prediction horizon. Panel A of Table 4.7 shows the performance of our estimators in predicting expected losses over various time horizons. Interestingly, the standard estimator now improves relative to the EB estimator. The RMSE of the EB estimator is now lower than that of the standard estimator only for the scenario with $k = 1.25$, whereas the RMSE is now higher for $k = 1.5$. This is because, when estimating expected losses for a portfolio, there is a variance reducing effect as compared to estimating expected losses for a single obligor. More specifically, note

that $\sigma(\sum_i w_i\widehat{EL_i}) \leq \sum_i w_i\sigma(\widehat{EL_i})$, i.e. the standard error of the weighted average of expected loss estimators is lower than the weighted average of their standard errors. This effect works here since the estimated portfolio-wide expected losses are weighted averages of the estimated rating-specific losses. On the other hand, $\text{Bias}(\sum_i w_i\widehat{EL_i}) = \sum_i w_i\text{Bias}(\widehat{EL_i})$, so that no such reduction effect holds for the bias of the estimators. Since the EB estimators benefits hinge on its ability to reduce variance at the cost of some bias, the EB estimators merits diminish somewhat in this case. As far as conservativeness is concerned, the standard estimator is still liberal, underestimating pseudo-true expected losses in more than 50% of all simulations. In contrast, the EB estimator tends to overestimate pseudo-true expected losses.

Similarly to section 4.4, we also analyze our estimators with respect to implied Basel II capital requirements. More precisely, we investigate how good our estimators perform in estimating pseudo-true economic capital which we define as the capital requirements which follow from our pseudo-true default probabilities and, again, the IRB formula. The results of Panel B of Table 4.7 are astonishing. Now, the standard estimator has a large downward bias as pseudo-true economic capital is underestimated by 45% on average.[26] This finding can be explained by the concavity of the IRB formula. Denote by $K(\cdot)$ the function that calculates economic capital using the default probability as an argument. Then, by a simple second-order Taylor series expansion:

$$\begin{aligned} E[K(\widehat{PD}) - K(PD)] &\approx K'(PD)E[\widehat{PD} - PD] + \frac{1}{2}K''(PD)E[(\widehat{PD} - PD)^2] \\ &\approx \frac{1}{2}K''(PD) \cdot V[\widehat{PD}] < 0 \end{aligned} \tag{4.13}$$

The equation holds for an approximately unbiased estimator. Since the IRB function is concave we have $K''(PD) < 0$ which results in a downward biased estimate for economic capital under an unbiased estimate for the default probability.[27] The bias is proportional to the variance of the estimator, $V[\widehat{PD}]$, and may thus only be negligible if estimation uncertainty is low.

[26] The bias of the standard estimator in our expected loss calculations is, in contrast, negligibly small.

[27] This type of bias has already been explored by Kiefer & Larson (2003). In that study, the authors also calculate the second derivative of the IRB economic capital function.

Table 4.7: Portfolio evaluation of estimators

Panel A: Estimation of expected losses for the market portfolio

		Relative RMSE			$\% \widehat{EL} < EL$		
	Pseudo-true EL (%)	SE	EB (k=1.25)	EB (k=1.5)	SE	EB (k=1.25)	EB (k=1.5)
1 year	0.06	0.49	0.45	0.66	55.9	12.5	4.2
3 years	0.26	0.54	0.52	0.78	58.6	10.7	2.0
5 years	0.57	0.57	0.54	0.84	59.3	11.1	1.7
10 years	1.69	0.60	0.59	0.92	57.6	12.4	2.3

Panel B: Estimation of capital requirements for the market portfolio

	Relative Bias			Relative RMSE			$\% \widehat{EC} < EC$		
Pseudo-true EC (%)	SE	EB (k=1.25)	EB (k=1.5)	SE	EB (k=1.25)	EB (k=1.5)	SE	EB (k=1.25)	EB (k=1.5)
0.79	-0.45	-0.01	0.16	0.57	0.32	0.36	94.6	57.8	32.4

The composition of the market portfolio is given in Table 4.3. Relative RMSE is defined as RMSE/Pseudo-true EL and RMSE/Pseudo-true EC, respectively, where EL means expected losses and EC means economic capital. $\% \widehat{EL} < EL$ and $\% \widehat{EC} < EC$ are the relative frequencies of simulations where pseudo-true EL/EC was underestimated. Relative bias is the estimated bias divided by pseudo-true EC. SE is the abbreviation for the standard estimator whereas EB refers to the empirical Bayes estimator. The number of simulations is 5000.

However, the results of Table 4.7 show that the opposite case is true and that the bias is substantial under a realistic scenario. In contrast, the upward bias of the EB estimator now compensates this effect and leads to nearly unbiased estimation for $k = 1.25$ and only slightly upward biased estimation for $k = 1.5$. Further, the precision of the EB estimator is now higher than for the standard estimator, even for $k = 1.5$. On top of that, the standard estimator now underestimates pseudo-true economic capital at an extremely high rate of 94.6% whereas the same figures are 57.8% and 32.4% for the EB estimator. Evidently, the EB estimator is more appropriate for economic

capital calculations in our scenario.

4.A Variance of the standard estimator

In this appendix, we derive a consistent variance formula for the life-table estimator (the standard estimator) under overlapping lifetimes. We allow for arbitrary dependence of the lifetimes of an individual obligor but we assume independence between obligors. The following derivations are based on the article of Williams (1995) which itself utilizes a general method of Woodruff (1971).[28] Williams (1995) provides formulas for the Kaplan-Meier estimator and indicates how the formulas can be extended to the life-table estimator which is what we do below. The estimator under consideration is (see section 4.1)

$$\widehat{PD}_s^r = 1 - \prod_{j=1}^{s} \left(1 - \frac{\sum_{t=1}^{T} D_{t,j}^r}{\sum_{t=1}^{T} N_{t,j}^r - L_{t,j}^r/2} \right) \quad .$$

From now on, we will omit the index r (indicating the rating grade) for convenience. Defining $\widetilde{N}_{t,j} = N_{t,j} - L_{t,j}/2$, a first-order Taylor series approximation for the variance of $\widehat{PD}_s$ is given by

$$V(\widehat{PD}_s) \approx E\left[\left(\sum_{j=1}^{s} \sum_{t=1}^{T} \frac{\partial \widehat{PD}_s}{\partial D_{t,j}} (D_{t,j} - E[D_{t,j}]) + \frac{\partial \widehat{PD}_s}{\partial \widetilde{N}_{t,j}} (\widetilde{N}_{t,j} - E[\widetilde{N}_{t,j}]) \right)^2 \right] \quad ,$$

where the partial derivatives are evaluated at expected values. As we will see we can use this Taylor approximation without expanding the squared sum into the corresponding variances and covariances. First, define

$$d_{i,t,j} = \begin{cases} 1 & \text{if the lifetime of obligor } i \text{ starting in period } t \text{ ends in period } t+j \,, \\ 0 & \text{otherwise} \,, \end{cases}$$

$$n_{i,t,j} = \begin{cases} 1 & \text{if the lifetime of obligor } i \text{ starting in period } t \text{ is at risk in period } t+j \,, \\ 0 & \text{otherwise} \,, \end{cases}$$

$$l_{i,t,j} = \begin{cases} 1 & \text{if the lifetime of obligor } i \text{ starting in period } t \text{ is censored in period } t+j \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

[28] An alternative to the variance estimator by Williams (1995) is given by Kang & Koehler (1997). In that study, however, it is assumed that the lifetimes within one cluster (of one obligor) are exchangeable. This assumption is unrealistic under the overlapping structure that we deal with in our application.

Further, let $\widetilde{n}_{i,t,j} = n_{i,t,j} - l_{i,t,j}/2$. Then, $D_{t,j} = \sum_{i=1}^{n} d_{i,t,j}$ and $\widetilde{N}_{t,j} = \sum_{i=1}^{n} \widetilde{n}_{i,t,j}$, where (by a slight abuse of notation) n denotes the total number of obligors. Using the fact that $\partial\widehat{PD}_s/\partial D_{t,j} = \partial\widehat{PD}_s/\partial d_{i,t,j}$ and $\partial\widehat{PD}_s/\partial\widetilde{N}_{t,j} = \partial\widehat{PD}_s/\partial\widetilde{n}_{i,t,j}$ and rearranging, the Taylor approximation gets

$$\begin{aligned} V(\widehat{PD}_s) \approx\ & E\Big[\Big(\sum_{j=1}^{s}\sum_{t=1}^{T}\frac{\partial\widehat{PD}_s}{\partial d_{i,t,j}}\Big(\sum_{i=1}^{n} d_{i,t,j} - E\Big[\sum_{i=1}^{n} d_{i,t,j}\Big]\Big) \\ & \qquad + \frac{\partial\widehat{PD}_s}{\partial\widetilde{n}_{i,t,j}}\Big(\sum_{i=1}^{n}\widetilde{n}_{i,t,j} - E\Big[\sum_{i=1}^{n}\widetilde{n}_{i,t,j}\Big]\Big)\Big)^2\Big] \\ = E\Big[\Big(\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j=1}^{s} & \frac{\partial\widehat{PD}_s}{\partial d_{i,t,j}}d_{i,t,j} + \frac{\partial\widehat{PD}_s}{\partial\widetilde{n}_{i,t,j}}\widetilde{n}_{i,t,j} - E\Big[\frac{\partial\widehat{PD}_s}{\partial d_{i,t,j}}d_{i,t,j} + \frac{\partial\widehat{PD}_s}{\partial\widetilde{n}_{i,t,j}}\widetilde{n}_{i,t,j}\Big]\Big)^2\Big]. \end{aligned}$$

Defining $U_{it} = \sum_{j=1}^{s} \frac{\partial\widehat{PD}_s}{\partial d_{i,t,j}} d_{i,t,j} + \frac{\partial\widehat{PD}_s}{\partial\widetilde{n}_{i,t,j}}\widetilde{n}_{i,t,j}$, which are the linearized values in the terminology of Woodruff (1971), we see that the Taylor approximation amounts to finding the variance of $\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}$. Under our assumptions regarding the dependencies, we can use the so-called between-cluster variance estimator

$$\widehat{V}\Big(\sum_{i=1}^{n}\sum_{t=1}^{T} U_{it}\Big) = \widehat{V}\Big(\sum_{i=1}^{n} U_{i\cdot}\Big) = \frac{n}{n-1}\sum_{i=1}^{n}(U_{i\cdot} - \overline{U}_{i\cdot})^2\,,$$

where $U_{i\cdot} = \sum_{t=1}^{T} U_{it}$. The between-cluster variance estimator is consistent as the number of clusters (obligors) approaches infinity. We now calculate the linearized values. To simplify notation, let $D_{j\cdot} = \sum_{t=1}^{T} D_{tj}$ and $\widetilde{N}_{j\cdot} = \sum_{t=1}^{T}\widetilde{N}_{tj}$ so that $\widehat{\lambda}_k = D_{j\cdot}/\widetilde{N}_{j\cdot}$. The partial derivatives are then

$$\begin{aligned} \frac{\partial\widehat{PD}_s}{\partial d_{i,t,j}} &= -\prod_{k\neq j}(1-\widehat{\lambda}_k)\big(-\frac{1}{\widetilde{N}_{j\cdot}}\big) = \frac{1-\widehat{PD}_s}{(1-\widehat{\lambda}_j)\widetilde{N}_{j\cdot}} = \frac{1-\widehat{PD}_s}{\widetilde{N}_{j\cdot} - D_{j\cdot}}\,, \\ \frac{\partial\widehat{PD}_s}{\partial\widetilde{n}_{i,t,j}} &= -\prod_{k\neq j}(1-\widehat{\lambda}_k)\frac{D_{j\cdot}}{\widetilde{N}_{j\cdot}^2} = -\frac{(1-\widehat{PD}_s)\widehat{\lambda}_j}{(1-\widehat{\lambda}_j)\widetilde{N}_{j\cdot}} = -\frac{(1-\widehat{PD}_s)\widehat{\lambda}_j}{\widetilde{N}_{j\cdot} - D_{j\cdot}}\,. \end{aligned}$$

Since an evaluation of the partial derivatives at expected values is not feasible, we have used the appropriate sample counterparts. Using our results for the

partial derivatives we have

$$U_{it} = \sum_{j=1}^{s} \frac{1-\widehat{PD}_s}{\widetilde{N}_{j\cdot} - D_{j\cdot}} d_{i,t,j} - \frac{(1-\widehat{PD}_s)\widehat{\lambda}_j}{\widetilde{N}_{j\cdot} - D_{j\cdot}} \widetilde{n}_{i,t,j} = (1-\widehat{PD}_s) \sum_{j=1}^{s} \frac{d_{i,t,j} - \widehat{\lambda}_j \widetilde{n}_{i,t,j}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} ,$$

$$U_{i\cdot} = \sum_{t=1}^{T} (1-\widehat{PD}_s) \sum_{j=1}^{s} \frac{d_{i,t,j} - \widehat{\lambda}_j \widetilde{n}_{i,t,j}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} = (1-\widehat{PD}_s) \sum_{j=1}^{s} \frac{d_{i,\cdot,j} - \widehat{\lambda}_j \widetilde{n}_{i,\cdot,j}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} ,$$

$$\overline{U}_{i\cdot} = \frac{1}{n} \sum_{i=1}^{n} (1-\widehat{PD}_s) \sum_{j=1}^{s} \frac{d_{i,\cdot,j} - \widehat{\lambda}_j \widetilde{n}_{i,\cdot,j}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} = \frac{1-\widehat{PD}_s}{n} \sum_{j=1}^{s} \frac{D_{j\cdot} - \widehat{\lambda}_j \widetilde{N}_{j\cdot}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} = 0 ,$$

where $d_{i,\cdot,j} = \sum_{t=1}^{T} d_{i,t,j}$ and $\widetilde{n}_{i,\cdot,j} = \sum_{t=1}^{T} \widetilde{n}_{i,t,j}$. The result for $\overline{U}_{i\cdot}$ follows from $\widehat{\lambda}_k = D_{j\cdot}/\widetilde{N}_{j\cdot}$. Plugging our results for the linearized values into the between-cluster variance formula we get

$$\begin{aligned}
\widehat{V}(\widehat{PD}_s) &= \frac{n}{n-1} \sum_{i=1}^{n} \Big((1-\widehat{PD}_s) \sum_{j=1}^{s} \frac{d_{i,\cdot,j} - \widehat{\lambda}_j \widetilde{n}_{i,\cdot,j}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} \Big)^2 \\
&= \frac{n}{n-1} (1-\widehat{PD}_s)^2 \sum_{i=1}^{n} \Big(\sum_{j=1}^{s} \frac{d_{i,\cdot,j} - \widehat{\lambda}_j \widetilde{n}_{i,\cdot,j}}{\widetilde{N}_{j\cdot} - D_{j\cdot}} \Big)^2 .
\end{aligned}$$

For the special case that we only observe one lifetime for each obligor (thus assuming an independent sample) one can show that our variance formula reduces to the well-known Greenwood formula (Greenwood, 1926) except for the factor $n/(n-1)$.

4.B Consistency of the empirical Bayes estimator

In the following, we prove the consistency of the empirical Bayes (EB) estimator which was introduced in section (4.3). We show the consistency of the estimator for the discrete-time hazard rate λ_s^r given by Equation (4.11) with the number of groups G fixed, $\widetilde{N}_{s,g}^r \to \infty$ and $\widetilde{N}_{s,g}^r / \sum_{j=1}^{G} \widetilde{N}_{s,j}^r \to c_{s,g}^r$, where $c_{s,g}^r \in (0,1)$. The consistency of the corresponding estimator for PD_s^r (Equation (4.12)) then directly follows from Slutsky's theorem. For convenience, we drop the indices r and s in the following.

From Equation (4.11),

$$\widehat{\lambda}_{g,EB} = \frac{1-\widehat{\tau}}{1+\widehat{\tau}(\widetilde{N}_g - 1)}\widehat{\mu} + \frac{\widehat{\tau}\widetilde{N}_g}{1+\widehat{\tau}(\widetilde{N}_g - 1)}\widehat{\lambda}_g \,.$$

We assume consistency of the standard estimator,[29] i.e. $\widehat{\lambda}_g \to \lambda_g$, so that it suffices to show that $\widehat{\tau}$ converges to a non-zero constant. From Equation (4.9), the estimator for τ is

$$\widehat{\tau} = \frac{\frac{G-1}{G}\sum_{g=1}^{G} w_g(\widehat{\lambda}_g - \widehat{\mu})^2 - \widehat{\mu}(1-\widehat{\mu})\left(\sum_{g=1}^{G} w_g(1-w_g)/\widetilde{N}_g\right)}{\widehat{\mu}(1-\widehat{\mu})\left(\sum_{g=1}^{G}(1-1/\widetilde{N}_g)w_g(1-w_g)\right)} \,.$$

We consider first the case that $w_g = 1/G$. Since

$$\begin{aligned}
&\text{plim}\left(\sum_{g=1}^{G}\frac{1}{G}(\widehat{\lambda}_g - \widehat{\mu})^2\right) = \frac{1}{G}\sum_{g=1}^{G}(\lambda_g - \overline{\lambda}_1)^2 \quad , \quad \overline{\lambda}_1 \equiv \frac{1}{G}\sum_{g=1}^{G}\lambda_g \quad , \\
&\text{plim}\left(\sum_{g=1}^{G}\frac{1}{G}\left(1-\frac{1}{G}\right)\frac{1}{\widetilde{N}_g}\right) = 0 \quad , \\
&\text{plim}\left(\sum_{g=1}^{G}\left(1-\frac{1}{\widetilde{N}_g}\right)\frac{1}{G}\left(1-\frac{1}{G}\right)\right) = \frac{G-1}{G} \quad ,
\end{aligned}$$

we have

$$\text{plim}(\widehat{\tau}) = \frac{\frac{G-1}{G}\frac{1}{G}\sum_{g=1}^{G}(\lambda_g - \overline{\lambda}_1)^2}{\overline{\lambda}_1(1-\overline{\lambda}_1)\frac{G-1}{G}} = \frac{1}{G}\frac{\sum_{g=1}^{G}(\lambda_g - \overline{\lambda}_1)^2}{\overline{\lambda}_1(1-\overline{\lambda}_1)} \equiv c_{\tau_1} \,.$$

The probability limit exists if $\overline{\lambda}_1 \neq 0$. If $\overline{\lambda}_1 = 0$, given our truncation of $\widehat{\tau}$ in the interval $[0,1]$, the EB estimator still exists and is consistent since $\widehat{\mu} = \widehat{\lambda}_g = 0$. Further, $\text{plim}(\widehat{\tau})$ is not equal to zero except if $\lambda_{g_1} = \lambda_{g_2}\, \forall\, g_1, g_2 \in \{1, \ldots, G\}$. Under this exception, $\text{plim}(\widehat{\lambda}_{g,EB}) = \text{plim}(\widehat{\mu})$. However, since then for all g $\widehat{\lambda}_g$ has the same probability limit, this case does not lead to inconsistency as well.

[29] As was discussed in section 4.1, the standard estimator is only consistent under quite strong assumptions. However, the asymptotic bias is very small if the periodicity is as small as it is in our application. If one is not willing to assume consistency, the proof only shows that the EB estimator has the same probability limit as the standard estimator.

We now consider $w_g = \widetilde{N}_g / \sum_{j=1}^G \widetilde{N}_j$. Then,

$$\begin{aligned}
&\text{plim}\left(\sum_{g=1}^G \frac{\widetilde{N}_g}{\sum_{j=1}^G \widetilde{N}_j}(\widehat{\lambda}_g - \widehat{\mu})^2\right) = \sum_{g=1}^G c_g(\lambda_g - \overline{\lambda}_2)^2 \quad , \quad \overline{\lambda}_2 \equiv \sum_{g=1}^G c_g \lambda_g \quad , \\
&\text{plim}\left(\sum_{g=1}^G \frac{\widetilde{N}_g}{\sum_{j=1}^G \widetilde{N}_j}\left(1 - \frac{\widetilde{N}_g}{\sum_{j=1}^G \widetilde{N}_j}\right)\frac{1}{\widetilde{N}_g}\right) = 0 \quad , \\
&\text{plim}\left(\sum_{g=1}^G \left(1 - \frac{1}{\widetilde{N}_g}\right)\frac{\widetilde{N}_g}{\sum_{j=1}^G \widetilde{N}_j}\left(1 - \frac{\widetilde{N}_g}{\sum_{j=1}^G \widetilde{N}_j}\right)\right) = \sum_{g=1}^G c_g(1-c_g) \quad ,
\end{aligned}$$

so that

$$\text{plim}(\widehat{\tau}) = \frac{\frac{G-1}{G}\sum_{g=1}^G c_g(\lambda_g - \overline{\lambda}_2)^2}{\overline{\lambda}_2(1-\overline{\lambda}_2)\sum_{g=1}^G c_g(1-c_g)} \equiv c_{\tau_2} \, .$$

For the same reasons as before, consistency can be established even if $\overline{\lambda}_2 = 0$ or $\lambda_g = \overline{\lambda}_2 \, \forall \, g \in \{1, \ldots, G\}$.

Finally, we consider the situation that the weights are refined by a one-time iteration (see Equation (4.10)):

$$w_g = \frac{\widetilde{N}_g}{1 + \widehat{\tau}(\widetilde{N}_g - 1)} \Big/ \sum_{j=1}^G \frac{\widetilde{N}_j}{1 + \widehat{\tau}(\widetilde{N}_j - 1)}$$

Since

$$\text{plim}\left(\frac{\widetilde{N}_g}{1 + \widehat{\tau}(\widetilde{N}_g - 1)}\right) = \text{plim}\left(\frac{1}{\frac{1}{\widetilde{N}_g} + \widehat{\tau}(1 - \frac{1}{\widetilde{N}_g})}\right) = \text{plim}(\widehat{\tau})^{-1}$$

we have $\text{plim}(w_g) = 1/G$ and thus $\text{plim}(\widehat{\tau}) = c_{\tau_1}$ if the iteration step is performed. □

4.C R code for the empirical Bayes estimator

The following code refers to Formulas (4.8)-(4.11). It returns the empirical Bayes estimator for the discrete-time hazard rate.

```
eb <- function(w,lambda,n,iter) {
# w: Vector of weights; sum(w) should be 1
# lambda: Vector of standard hazard rate estimators
# n: Vector of numbers at risk
# iter: TRUE or FALSE, referring to iteration step
G <- length(w)
# G is the number of groups/portfolios
mu <- sum(w*lambda)
SS <- (G-1)/G*sum(w*(lambda-mu)^2)
tau <- (SS-mu*(1-mu)*sum(w*(1-w)/n))/(mu*(1-mu)*sum((1-1/n)*w*(1-w)))
tau <- ifelse(tau < 0,0,ifelse(tau > 1,1,tau))
if (iter==TRUE) {
w <- n/(1+tau*(n-1))/sum(n/(1+tau*(n-1)))
mu <- sum(w*lambda)
SS <- (G-1)/G*sum(w*(lambda-mu)^2)
tau <- (SS-mu*(1-mu)*sum(w*(1-w)/n))/(mu*(1-mu)*sum((1-1/n)*w*(1-w)))
tau <- ifelse(tau < 0,0,ifelse(tau > 1,1,tau))
}
B <- (1-tau)/(1+tau*(n-1))
B <- ifelse(B<0,0,ifelse(B>1,1,B))
lambda.eb <- B*mu + (1-B)*lambda
if (sum(lambda)==0) lambda.eb <- lambda
return(lambda.eb)
}
# Example:
eb(c(1/2,1/2),c(0.04,0),c(1000,100),iter=TRUE)
0.03875106 0.01130409
```

Bibliography

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589–609.

Altman, E. I. (1989). Measuring corporate bond mortality and performance. *Journal of Finance*, 44(4), 909–922.

Arvesen, J. N. (1969). Jackknifing U-statistics. *Annals of Mathematical Statistics*, 40(6), 2076–2100.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387–415.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when borrowers default. *Journal of the Operational Research Society*, 50(12), 1185–1190.

Basel Committee on Banking Supervision (2000). *Range of Practice in Banks' Internal Ratings Systems.* Discussion paper, January 2000. www.bis.org/publ/bcbs66.pdf.

Basel Committee on Banking Supervision (2005a). *An Explanatory Note on the Basel II IRB Risk Weight Functions.* July 2005. www.bis.org/bcbs/irbriskweight.pdf.

Basel Committee on Banking Supervision (2005b). *Studies on the Validation of Internal Rating Systems.* BCBS Working papers No. 14. www.bis.org/publ/bcbs_wp14.pdf.

Basel Committee on Banking Supervision (2005c). *Validation of low-default portfolios in the Basel II Framework*. Basel Committee Newsletter No. 6, September 2005. www.bis.org/publ/bcbs_nl6.pdf.

Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework - Comprehensive Version*. June 2006. www.bis.org/publ/bcbs128.pdf.

Basel Committee on Banking Supervision (2009). *Guiding principles for the replacement of IAS 39*. August 2009. www.bis.org/publ/bcbs161.pdf.

Basel Committee on Banking Supervision (2010). *Group of Governors and Heads of Supervision announces higher global minimum capital standards*. Press release, September 12, 2010. www.bis.org/press/p100912.pdf.

Beaver, W. H., McNichols, M. F., & Rhie, J.-W. (2005). Have financial statements become less informative? Evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting Studies*, 10(1), 93–122.

Benjamin, N., Cathcart, A., & Ryan, K. (2006). *Low Default Portfolios: A Proposal for Conservative Estimation of Default Probabilities*. Financial Services Authority, London. Discussion Paper. www.fsa.gov.uk/pubs/international/default_probabilities.pdf.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2), 273–277.

Berger, J. O. (1980). *Statistical Decision Theory*. New York: Springer.

Bharath, S. T. & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3), 1339–1369.

Birnbaum, A. (1963). Median-unbiased estimators. *Bulletin of Mathematical Statistics*, 11, 25–34.

Blöchlinger, A. & Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, 30(3), 851–873.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1), 89–100.

Breslow, N. E. & Crowley, J. (1974). A large sample study of the Life Table and Product Limit estimates under random censorship. *Annals of Statistics*, 2(3), 437–453.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101–117.

Burman, P., Chow, E., & Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2), 351–358.

Burman, P. & Nolan, D. (1992). Data-dependent estimation of prediction functions. *Journal of Time Series Analysis*, 13(3), 189–207.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3), 414–427.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29 (2), 238–249.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance*, 63(6), 2899–2939.

Cantor, R., Ap Gwilym, O., & Thomas, S. (2007). The use of credit ratings in investment management in the U.S. and Europe. *Journal of Fixed Income*, 17(2), 13–26.

Cantor, R., Hamilton, D. T., & Tennant, J. (2008). Confidence intervals for corporate default rates. *Risk*, March 2008, 93–98.

Cantor, R. & Mann, C. (2003). *Measuring the Performance of Corporate Bond Ratings*. Moody's Investors Service. Special comment, April 2003. Available at papers.ssrn.com.

Carlin, B. P. & Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. Boca Raton: CRC Press, 3rd edition.

Casella, G. (1985). An introduction to empirical Bayes analysis. *The American Statistician*, 39(2), 83–87.

Chava, S. & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, 8(4), 537–569.

Cheng, K. F., Chu, C. K., & Hwang, R.-C. (2010). Predicting bankruptcy using the discrete-time semiparametric hazard model. *Quantitative Finance*, 10(9), 1055–1066.

Christensen, J. H., Hansen, E., & Lando, D. (2004). Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance*, 28(11), 2575–2602.

Clements, M. P. & Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Committee of European Insurance and Occupational Pensions Supervisors (2009). *CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: SCR Standard Formula - Counterparty Default Risk Module*. CEIOPS-DOC-23/09. October 2009. eiopa.europa.eu/consultations/consultation-papers/index.html.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society Series B*, 45(3), 311–354.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B*, 34(2), 187–220.

Crook, J. & Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 173(2), 283–305.

Crook, J., Leow, M., & Bellotti, T. (2011). *What seems right for all may not be right for any of us! Survival modelling with unobserved heterogeneity.*

Presented at the Credit Scoring & Credit Control XII Conference, August 2011, Edinburgh.

Crosbie, P. & Bohn, J. (2003). *Modeling Default Risk*. Moody's KMV. December 18, 2003.

Cutler, S. J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Diseases*, 8(6), 699–712.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge Unversity Press.

Dawid, A. P. (1984). Statistical theory - the prequential approach. *Journal of the Royal Statistical Society Series A*, 147(2), 278–292.

De Leonardis, D. & Rocci, R. (2008). Assessing the default risk by means of a discrete-time survival analysis approach. *Applied Stochastic Models in Business and Industry*, 24(4), 291–306.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845.

Deutsche Bundesbank (2012). *Monatsbericht - Mai 2012*. Frankfurt am Main. 64(5).

Diebold, F. X. & Rudebusch, G. D. (1989). Scoring the leading indicators. *Journal of Business*, 62(3), 369–391.

Duan, J.-C., Sun, J., & Wang, T. (2012). Multiperiod corporate default prediction - a forward intensity approach. *Journal of Econometrics*, 170(1), 191–209.

Duffie, D., Eckner, A., Horel, G., & Saita, L. (2009). Frailty correlated default. *Journal of Finance*, 64(4), 2089–2123.

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3), 635–665.

Dwyer, D. W., Kocagil, A. E., & Stein, R. M. (2004). *Moody's KMV RiskCalc v3.1 Model*. Moody's KMV white paper. April 5, 2004.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359), 557–565.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374), 312–319.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Engelmann, B., Hayden, E., & Tasche, D. (2003). Testing rating accuracy. *Risk*, January 2003, 82–86.

Field, C. A. & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B*, 69(3), 369–390.

Fons, J. S. (1994). Using default rates to model the term structure of credit risk. *Financial Analysts Journal*, 50(5), 25–32.

Fuertes, A.-M. & Kalotychou, E. (2007). On sovereign credit migration: A study of alternative estimators and rating dynamics. *Computational Statistics & Data Analysis*, 51(7), 3448–3469.

Greene, W. H. (2008). *Econometric Analysis*. Upper Saddle River, New Jersey: Prentice Hall, 6th edition.

Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33, 1–26.

Grunert, J., Norden, L., & Weber, M. (2005). The role of non-financial factors in internal credit ratings. *Journal of Banking & Finance*, 29(2), 509–531.

Guettler, A. & Raupach, P. (2010). The impact of downward rating momentum. *Journal of Financial Services Research*, 37(1), 1–23.

Gupton, G. M., Finger, C. M., & Bhatia, M. (1997). *CreditMetrics^{TM} - Technical Document*. J.P. Morgan, New York. April 2, 1997.

Hamerle, A., Jobst, R., Liebig, T., & Rösch, D. (2007). Multiyear risk of credit losses in SME portfolios. *Journal of Financial Forecasting*, 1(2), 1–29.

Hamerle, A., Liebig, T., & Scheule, H. (2006). Forecasting credit event frequency – empirical evidence for West German firms. *Journal of Risk*, 9 (1), 75–98.

Hamilton, D. T. & Cantor, R. (2006). *Measuring Corporate Default Rates*. Moody's Investors Service, Global Credit Research. Special comment, November 2006.

Hannoun, H. (2011). *Sovereign Risk in Bank Regulation and Supervision: Where Do We Stand?* Speech at the Financial Stability Institute High-Level Meeting, Abu Dhabi, UAE, 26 October 2011. www.bis.org/speeches/sp111026.htm.

Hanson, S. & Schuermann, T. (2006). Confidence intervals for probabilities of default. *Journal of Banking & Finance*, 30(8), 2281–2301.

Harrell, F. E. (2001). *Regression Modeling Strategies - With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Harrell, F. E. J., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4), 361–387.

Hartmann-Wendels, T., Pfingsten, A., & Weber, M. (2010). *Bankbetriebslehre*. Berlin: Springer, 5th edition.

He, W. & Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics*, 59(4), 837–848.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9(1), 5–34.

Hilscher, J. & Wilson, M. (2009). *Credit Ratings and Credit Risk*. Working paper. Available at papers.ssrn.com.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3), 1259–1294.

Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics*, 9(2), 95–105.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3), 293–325.

Horowitz, J. L. (2001). The Bootstrap. In J. J. Heckman & E. Leamer (Eds.), *Handbook of Econometrics*, volume 5 (pp. 3159–3228). Amsterdam: Elsevier.

Hosmer, D. W. & Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10), 1043–1069.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London Series A: Mathematical and Physical Sciences*, 186(1007), 453–461.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics*, 57(1), 129–138.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. New York: Chapman & Hall.

Johansen, S. (1978). The Product Limit estimator as Maximum Likelihood estimator. *Scandinavian Journal of Statistics*, 5(4), 195–199.

Kalbfleisch, J. D. & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2), 267–278.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, New Jersey: Wiley, 2nd edition.

Kang, S.-S. & Koehler, K. J. (1997). Modification of the Greenwood formula for correlated response times. *Biometrics*, 53(3), 885–899.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Kendall, M. & Gibbons, J. D. (1990). *Rank Correlation Methods*. New York: Oxford University Press, 5th edition.

Kiefer, N. M. (2009). Default estimation for low-default portfolios. *Journal of Empirical Finance*, 16(1), 164–173.

Kiefer, N. M. (2010). Default estimation and expert information. *Journal of Business & Economic Statistics*, 28(2), 320–328.

Kiefer, N. M. & Larson, C. E. (2003). *Biases in Default Estimation and Capital Allocations under Basel II*. Working paper. Available at www.arts.cornell.edu/econ/kiefer.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent sample. *Journal of the American Statistical Association*, 68(341), 46–54.

Koopman, S. J., Lucas, A., & Monteiro, A. (2008). The multi-state latent factor intensity model for credit rating transitions. *Journal of Econometrics*, 142(1), 399–424.

Krämer, W. & Güttler, A. (2008). On comparing the accuracy of default predictions in the rating industry. *Empirical Economics*, 34(2), 343–356.

Kurbat, M. & Korablev, I. (2002). *Methodology for Testing the Level of the EDF^{TM} Credit Measure*. Moody's KMV Technical Report No. 020729.

Lando, D. & Skodeberg, T. M. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance*, 26(2-3), 423–444.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Hoboken, New Jersey: Wiley.

Lee, S. H. & Urrutia, J. L. (1996). Analysis and prediction of insolvency in the property-liability insurance industry. *Journal of Risk and Insurance*, 63(1), 121–130.

Löffler, G. (2007). The complementary nature of ratings and market-based measures of default risk. *Journal of Fixed Income*, 17(1), 38–47.

Löffler, G. & Maurer, A. (2011). Incorporating the dynamics of leverage into default prediction. *Journal of Banking & Finance*, 35(12), 3351–3361.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley, 2nd edition.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1), 50–60.

Medema, L., Koning, R. H., & Lensink, R. (2009). A practical approach to validating a PD model. *Journal of Banking & Finance*, 33(4), 701–708.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29(2), 449–470.

Mincer, J. & Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (Ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Mosler, K. (1995). Testing whether two distributions are stochastically ordered or not. In H. Rinne, B. Rüger, & H. Strecker (Eds.), *Grundlagen der Statistik und ihre Anwendungen - Festschrift für Kurt Weichselberger*. Heidelberg: Physica.

Mosler, K. (2003). Mixture models in econometric duration analysis. *Applied Stochastic Models in Business and Industry*, 19(2), 91 – 104.

Newson, R. B. (2006a). Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal*, 6(3), 309–334.

Newson, R. B. (2006b). Efficient calculation of jackknife confidence intervals for rank statistics. *Journal of Statistical Software*, 15(1), 1–10.

Nicoletti, C. & Rondinelli, C. (2010). The (mis)specification of discrete duration models with unobserved heterogeneity: A Monte Carlo study. *Journal of Econometrics*, 159(1), 1–13.

Oakes, D. (2008). On consistency of Kendall's tau under censoring. *Biometrika*, 95(4), 997–1001.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131.

Orth, W. (2011a). *Default Probability Estimation in Small Samples - With an Application to Sovereign Bonds*. Discussion Papers in Statistics and Econometrics No. 05/11, University of Cologne.

Orth, W. (2011b). *Multi-Period Credit Default Prediction with Time-Varying Covariates*. Discussion Papers in Statistics and Econometrics No. 03/11, University of Cologne.

Orth, W. (2012). The predictive accuracy of credit ratings: Measurement and statistical inference. *International Journal of Forecasting*, 28(1), 288–296.

Pencina, M. J. & D'Agostino, R. B. (2004). Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence intervals. *Statistics in Medicine*, 23(13), 2109–2123.

Pluto, K. & Tasche, D. (2006). Estimating probabilities of default for low default portfolios. In B. Engelmann & R. Rauhmeier (Eds.), *The Basel II Risk Parameters* (pp. 79–103). Berlin: Springer.

Roszbach, K. (2004). Bank lending policy, credit scoring, and the survival of loans. *Review of Economics and Statistics*, 86(4), 946–958.

Royston, P. & Parmar, M. K. B. (2002). Flexible parametric-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15), 2175–2197.

Shumway, R. H. & Stoffer, D. S. (2006). *Time Series Analysis and Its Applications*. New York: Springer, 2nd edition.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, 74(1), 101–124.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27(6), 799–811.

Standard & Poor's (2007). *Quantitative Analysis vs. Expert Judgement: When and Why*. Commentary Report. March 23, 2007.

Standard & Poor's (2009). *On the Use of Models by Standard & Poor's Ratings Services*. RatingsDirect, Global Credit Portal: December 21, 2009.

Standard & Poor's (2010). *The Time Dimension of Standard & Poor's Credit Ratings*. RatingsDirect, Global Credit Portal: September 22, 2010.

Standard & Poor's (2011a). *Sovereign Defaults and Rating Transition Data, 2010 Update*. February 23, 2011. Available at www.standardandpoors.com.

Standard & Poor's (2011b). *Sovereign Rating and Country T&C Assessment Histories*. June 3, 2011. Available at www.standardandpoors.com.

Standard & Poor's (2011c). *Standard & Poor's Noninvestment Grade Sovereign Debt Recovery Ratings: 2011 Update*. RatingsDirect, Global Credit Portal: March 29, 2011.

Stein, R. M. (2004). *Benchmarking Default Prediction Models*. Moody's KMV technical report No. 030124. Available at www.moodyskmv.com/research/.

Sueyoshi, G. T. (1995). A class of binary response models for grouped duration data. *Journal of Applied Econometrics*, 10(4), 411–431.

Sullivan Pepe, M. & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, 23(4), 939–951.

Tamura, R. N. & Young, S. S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics*, 43(4), 813–824.

Tasche, D. (2011). *Bayesian Estimation of Probabilities of Default for Low Default Portfolios*. Working paper. Available at arxiv.org.

Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data - Extending the Cox Model*. New York: Springer.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1), 147–160.

Thomas, L. C., Edelman, D. B., & Crook, J. B. (2002). *Credit Scoring and its Applications*. Philadelphia: SIAM.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1), 267–288.

Treacy, W. F. & Carey, M. (2000). Credit risk rating systems at large US banks. *Journal of Banking & Finance*, 24(1-2), 167–201.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (Abstract). *Annals of Mathematical Statistics*, 29(2), 614.

Van Houwelingen, H. C. (2000). Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19(24), 3401–3415.

Van Houwelingen, H. C. & Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, 9(11), 1303–1325.

Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408), 1065–1073.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis*, 1(2), 171–186.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. New York: Springer.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66(334), 411–414.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambrige, Massachusetts: MIT Press.

Xue, X. & Brookmeyer, R. (1997). Regression analysis of discrete time survival data under heterogeneity. *Statistics in Medicine*, 16(17), 1983–1993.

Ying, Z. & Wei, L. J. (1994). The Kaplan-Meier estimate under dependent failure time observations. *Journal of Multivariate Analysis*, 50(1), 17–29.