2010

Multistate Intensity Model with AR-GARCH Random Effect for Corporate Credit Rating Transition Analysis

Zhi Li

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

MULTISTATE INTENSITY MODEL WITH AR-GARCH RANDOM EFFECT FOR CORPORATE CREDIT RATING TRANSITION ANALYSIS

By

ZHI LI

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2010

The members of the Committee approve the Dissertation of Zhi Li defended on October 19, 2010.

Xufeng Niu
Professor Co-Directing Dissertation

Fred Huffer
Professor Co-Directing Dissertation

Alec Kercheval
Outside Committee Member

Wei Wu
Committee Member

Approved:

Dan McGee, Chair, Department of Statistics

Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above named committee members.

This thesis is dedicated to my family.

ACKNOWLEDGEMENTS

First, I would like to acknowledge my gratitude to my major advisors Dr. Fred Huffer and Dr. Xu-Feng Niu. This research would not be possible without their generous support, ideas, and feedback. They guided me through the whole research process from choosing the topic to working out each detail. Also, their suggestions to my course study and career development have made my past five years very fruitful. It was fortunate for me to have them as my advisors.

Secondly, I would like to express my appreciation to my committee members Dr. Wei Wu and Dr. Alec Kercheval for their kind support.

Finally, I want to give special thanks to my family. This thesis would not have been possible without their support and love.

TABLE OF CONTENTS

List of Tables vi

List of Figures viii

Abstract ix

1. INTRODUCTION 1
- 1.1 Markov Models 3
- 1.2 Multi-State Models 5

2. BACKGROUND 8
- 2.1 Agency Credit Ratings and Market Implied Ratings 8
- 2.2 Survival Analysis 14
- 2.3 Time Series Analysis 26

3. PROPOSED MODEL 31
- 3.1 Model Specification 31
- 3.2 Model Estimation 35

4. SIMULATION STUDY 45
- 4.1 The Practical Issues in the Simulation Study 52
- 4.2 The Simulation Study 56

5. APPLICATION STUDY 63

6. Future Work 76
- 6.1 The Fixed Covariates 76
- 6.2 The Model Application to Agency Rating Transition Data and Comparison with the Markov Chain Model 77

APPENDICES 78

A. The Simulation Results on Individual η and α parameters 78

REFERENCES 111

BIOGRAPHICAL SKETCH 114

LIST OF TABLES

1.1 S&P Rating Categories 3

1.2 Moody's Long-Term Rating Categories 4

4.1 All possible transitions 46

4.2 The Simulation Study Results 61

5.1 Initial Intensity Parameter Values for η 70

5.2 Initial Intensity Parameter Values for α_s 71

5.3 Estimated Intensity Parameter Values for η_s 72

5.4 Estimated Intensity Parameter Values for α_s 72

5.5 Estimated Transition Matrix 75

A.1 The simulation results for η for $T = 50$, $K = 200$ 79

A.2 The simulation results for α for $T = 50$, $K = 200$ 80

A.3 The simulation results for η for $T = 50$, $K = 300$ 81

A.4 The simulation results for α for $T = 50$, $K = 300$ 82

A.5 The simulation results for η for $T = 50$, $K = 400$ 83

A.6 The simulation results for α for $T = 50$, $K = 400$ 84

A.7 The simulation results for η for $T = 50$, $K = 500$ 85

A.8 The simulation results for α for $T = 50$, $K = 500$ 86

A.9 The simulation results for η for $T = 100$, $K = 200$ 87

A.10 The simulation results for α for $T = 100$, $K = 200$ 88

A.11 The simulation results for η for $T = 100$, $K = 300$ 89

A.12 The simulation results for α for $T = 100$, $K = 300$ 90

A.13 The simulation results for η for $T = 100$, $K = 400$ 91
A.14 The simulation results for α for $T = 100$, $K = 400$ 92
A.15 The simulation results for η for $T = 100$, $K = 500$ 93
A.16 The simulation results for α for $T = 100$, $K = 500$ 94
A.17 The simulation results for η for $T = 150$, $K = 200$ 95
A.18 The simulation results for α for $T = 150$, $K = 200$ 96
A.19 The simulation results for η for $T = 150$, $K = 300$ 97
A.20 The simulation results for α for $T = 150$, $K = 300$ 98
A.21 The simulation results for η for $T = 150$, $K = 400$ 99
A.22 The simulation results for α for $T = 150$, $K = 400$ 100
A.23 The simulation results for η for $T = 150$, $K = 500$ 101
A.24 The simulation results for α for $T = 150$, $K = 500$ 102
A.25 The simulation results for η for $T = 200$, $K = 200$ 103
A.26 The simulation results for α for $T = 200$, $K = 200$ 104
A.27 The simulation results for η for $T = 200$, $K = 300$ 105
A.28 The simulation results for α for $T = 200$, $K = 300$ 106
A.29 The simulation results for η for $T = 200$, $K = 400$ 107
A.30 The simulation results for α for $T = 200$, $K = 400$ 108
A.31 The simulation results for η for $T = 200$, $K = 500$ 109
A.32 The simulation results for α for $T = 200$, $K = 500$ 110

LIST OF FIGURES

2.1 S&P Rating Process 9

2.2 Market Implied Rating - Barra 12

4.1 Agency Credit Rating: the no-transition movements are also included. 47

4.2 Bond Implied Rating: the no-transition movements are also included. 48

4.3 Simulated intensity function over 100 time points. 59

5.1 MIR: empirical transition probability time trend for rating 1 65

5.2 MIR: empirical transition probability time trend for rating 2 66

5.3 MIR: empirical transition probability time trend for rating 3 66

5.4 MIR: empirical transition probability time trend for rating 4 67

5.5 MIR: empirical transition probability time trend for rating 5 67

5.6 MIR: empirical transition probability time trend for rating 6 68

5.7 MIR: empirical transition probability time trend for rating 7 68

5.8 MIR: empirical transition probability time trend for rating 8 69

5.9 MIR: empirical transition probability time trend for rating 9 69

ABSTRACT

This thesis presents a stochastic process and time series study on corporate credit rating and market implied rating transitions. By extending an existing model, this paper incorporates the generalized autoregressive conditional heteroscedastic (GARCH) random effects to capture volatility changes in the instantaneous transition rates. The GARCH model is a crucial part in financial research since its ability to model volatility changes gives the market practitioners flexibility to build more accurate models on high frequency financial data. The corporate rating transition modeling was historically dealing with low frequency data which did not have the need to specify the volatility. However, the newly published Moody's market implied ratings are exhibiting much higher transition frequencies. Therefore, we feel that it is necessary to capture the volatility component and make extensions to existing models to reflect this fact. The theoretical model specification and estimation details are discussed thoroughly in this dissertation. The performance of our models is studied on several simulated data sets and compared to the original model. Finally, the models are applied to both Moody's issuer rating and market implied rating transition data as an application.

CHAPTER 1

INTRODUCTION

Corporate bonds are one of the important ways to sustain the operating activities for most of the businesses. Every year, in the U.S. along, thousands of bonds are issued by firms across all business sectors, from IBM to Cocacola to Gillette. Each of these bonds can be categorized by its issuer, time to maturity, or coupon rate, etc. The buyer of these bonds are usually banks, insurance companies, and pension and mutual funds. They hold these securities as part of their portfolio and hope to earn "fixed income" before the principles are returned. However, there is no free lunch in this situation. Just like the mortgage and personal loan business, the firm that issues the bond could fail to make timely payments and go default. The creditors (who bought the bonds) will then have to line up with others to claim their money which is often in the scale of millions or even billions. Therefore, to estimate the credit risk is an important issue before any transaction occurs. This problem has been widely acknowledged and studied for a long time.

Before purchasing a bond, a creditor will first have to estimate the possible loss in the future. If we temporarily ignore the time effect and only consider a binary event—default or not, the potential loss can be decomposed into three terms: $L = EAD \times LGD \times 1_D$. Here the *exposure at default* (EAD) indicates the amount that subjects to be lost in case of default; the *loss given default* (LGD) is the fraction of the loan's exposure expected to be lost and is often expressed in percentage; 1_D is an indicator variable of the default event D and $P(D) = DP$ is the default probability.

Of the three quantities, the default probability is researched the most as the other two can be estimated with lesser degree of uncertainty. Also, within a portfolio of several hundred bond securities, the correlation of the defaults is usually modeled together with the probability by some statistical models. The most intuitive way to estimate the default

probability is to directly employ historical default data and calculate empirical probabilities. The data availability and quality limits its use as historically there were not many firms went default. Another method is to extract an implied default probability from marketable credit derivatives. Like insurance, the bond holder pay an upfront premium to a counterparty who will then assume the loss in the case of any default. Although this DP can be viewed as market agree-on probability and computed real-time, it may not be an appropriate long term estimator due to its volatility. There have also been several large rating agencies like *Moody's*, *Standard & Poor's* (S&P), and *Fitch* and it's possible to estimate DP from their published ratings.

Some details of the rating process are discussed in the following section. Basically a bond issuer's ratings describe its *creditworthiness*. The rating companies take consideration of firm specific information as well as macroeconomic conditions when making any decision. All three rating agencies use their own categories when publishing their rating reports. However, the categories are quite similar to each other. For example, Table 1.1 and Table 1.2 shows the long term obligations rating systems used by S&P [1] and Moody's [2], respectively. There are also "finer" ratings like Aaa1, Aaa2, etc. to give further information of the credit quality of the issuer.

Generally speaking, each rating of a firm reveals the firm's credit status at the time of evaluation. When the estimated default probability falls in a certain range, a rating is assigned. A single rating does not incorporate the idea when the firm will default. Instead, it only measures the possibility of default in the next unit period of time, which may be a day, a month or a year. This presents challenges to the bond buyers since they would fear the credit status may deteriorate in their holding period. To remedy this situation, rating agencies often (annually) review the latest information of the issuers and issues they cover and make changes as needed.

The change of ratings over time for an issue or an issuer is often called rating transitions. Bond investors pay close attention to this kind of events since they need to constantly measure their potential risk against their counterparty. For larger institutional investors like mutual funds and pension funds, it is even more important to do so since they will be forced to sell bond holdings that fall below certain ratings. There have been various approaches applied to the rating transitions modeling. The most popular ones usually involves Markov chain methods.

Table 1.1: S&P Rating Categories

Ratings	**Descriptions**
AAA	*best credit quality*
	extremely reliable with regard to financial obligations
AA	*very good credit quality*
	very reliable
A	*more susceptible to economic conditions*
	still good credit quality
BBB	*lowest rating in investment grade*
BB	*caution is necessary*
	best sub-investment credit quality
B	*vulnerable to changes in economic conditions*
	currently showing the ability to meet its financial obligations
CCC	*currently vulnerable to nonpayment*
	dependent on favorable economic conditions
CC	*highly vulnerable to a payment default*
C	*close to or already bankrupt*
	payments on the obligation currently continued
D	*payment default on some financial obligation has actually occurred*

1.1 Markov Models

We'll start the discussion by first introducing some basic concepts. Let Ω denote the finite state space of the ratings,

$$\Omega = \{AAA, AA, A, BBB, BB, B, CCC, Default\},$$

and assign every pair (i, j) of states a *transition* or *migration probability*

$$m_{ij} = P[i \to j] \quad (i = 1, \ldots, 8; j = 1, \ldots, 8),$$

where $P[i \to j]$ denotes the probability of a change from rating class i at the beginning of a year to rating class j at the year's end. The Markov property in this context means that the transition probability is independent of past rating transition history. The collection of these probabilities is the one-year transition matrix $\mathbf{M}_1 = (m_{ij})_{i,j=1,\ldots,8}$. Since the default state is the absorbing state, i.e. a firm goes into default will not come back to business,

Table 1.2: Moody's Long-Term Rating Categories

Ratings	**Descriptions**
Aaa	*the highest quality, minimal credit risk*
Aa	*high quality, very low credit risk*
A	*upper-medium grade, low credit risk*
Baa	*medium grade, moderate credit risk may possess certain speculative characteristics*
Ba	*have speculative elements, substantial credit risk*
B	*speculative, high credit risk*
Caa	*poor standing, very high credit risk*
Ca	*highly speculative and are likely in, or very near, default, with some prospect of recovery of principal and interest*
C	*lowest rated class of bonds and are typically in default, with little prospect for recovery of principal or interest*

the last row of $\mathbf{M}_1$ is the vector $(0, \ldots, 0, 1)$. The definition above is assuming that the transition probabilities are time independent. With these assumptions, the multi-year rating transition/migration matrix will be $\mathbf{M}_n = \mathbf{M}_1^n$. We need to bear in mind that this time homogeneity and the Markov property above are usually violated when modeling with real data.

Calculating transition probabilities from $\mathbf{M}_n$ is simple when n is an integer, but it's not the case for time periods that do not contain the whole year. Also, data provided by rating agencies are often for one-year transition only. We can overcome this by considering a continuous time Markov chain with time-dependent transition matrix $\mathbf{M}(t)$ which satisfies the differential equation

$$d\mathbf{M}(t) = \mathbf{Q} \cdot \mathbf{M}(t)dt$$

with the boundary condition $\mathbf{M}(0) = I$. The solution is given by

$$\mathbf{M}(t) = e^{t\mathbf{Q}},$$

where $\mathbf{Q} = (q_{ij})$ is called the *generator* of the Markov chain. The properties of $\mathbf{Q}$ are

$$q_{ii} \leq 0, \quad q_{ij} \geq 0, \quad \sum_{j \neq i} q_{ij} = -q_{ii} \equiv q_i, \quad i, j = 1, \ldots, 8.$$

The following theorem from Bluhm [3] gives a convenient way to constructing the generator matrix. See above reference for the proof.

Theorem 1.1.1. *Let* $\mathbf{Q}$ *be an* 8×8 *generator matrix and define* $\mathbf{\Lambda} \in \mathbb{R}^{8\times 8}$ *as*

$$\mathbf{\Lambda}_{i,j} = \begin{cases} 0 & \text{if } i \neq j \\ \lambda_i > 0 & \text{if } i = j \end{cases}.$$

Then $\mathbf{\Lambda} \cdot \mathbf{Q}$ *is again a generator matrix, i.e., row scaling by constant positive factors is a closed operation in the space of* $\mathbf{Q}$*-matrices.*

1.2 Multi-State Models

Markov property is an important assumption in separating the past and future by giving the present. However, it may prove to be too restrictive when applying to real data. There have been evidences of non-Markovian behavior showed by different authors. One of them, first discussed by Altman [4], is about the rating drift that related to the aging effect. It essentially means that older bonds are prone to make short-term rating change than new issues (or young bonds). Nickell and Varotto [5] in their paper discussed the heterogeneity of rating transition matrices across the industry and domicile of the obligor, and the stage of the business cycle. They showed that the matrices conditioning on the variables above are statistically different from the overall (unconditional) transition matrix. Similar evidence was also given by Kronimus et al. [6]. When taking into account of different business conditions - expansion and contraction, the matrices as well as the loss distribution of credit portfolios are quite different. Lando and Skodeberg [7] applied continuous Markov process to the rating migration data and found out the evidence of transition duration dependence and dependence on previous rating, which is another example of non-Markovian effects.

Several models have been proposed to address different non-Markovian properties. Hansen et al. [8] considered hidden Markov model in dealing with the observed phenomenon of rating change "momentum." When a bond has its rating downgraded to Baa, it is more likely to be further downgraded than another bond which has stayed in the Baa category for a long time. Similar "momentum" applies to the rating upgrade too. By incorporating an

"exciting" state, Hansen [8] concluded that the default probabilities were strongly influenced by the introduction of the hidden excited states.

Another breakthrough in analysis of rating data is the Markov mixture model proposed by Frydman and Schuermann [9]. After studying the discrepancy between estimation results from a Markov model and the actual transition probabilities, Frydman and Schuermann [9] noted that non-Markovian properties may have been caused by the underlying hypothesis of a single homogeneous cohort. By assuming that there exist two groups of issuers with different transition speeds, Frydman and Schuermann [9] estimated the "mixed" stochastic process and concluded that it statically dominates the simple Markov chain model. Their results also show that future of each firm does not only depend on the current rating but on the past rating history as well.

However, the above modifications, like the Markov models before, consider only the aggregated transition probabilities. There is no issuer/issue specific information incorporated. Although various industry information can be used as predictors for the regression-type models, this does not automatically come down to the individual company level. Therefore we need to make further changes to make use of the company specific information. In our research, we will utilize the counting process approach introduced by Aalen [16]in the 70's. Each individual issuer will be treated as a subject. There is one process for each transition type of each subject. When a subject moves from one state to another, the corresponding processes will be activated and all the other processes are deactivated. For example, let us assume that there are totally three ratings A, B, C with A as the highest rating and C the lowest rating. The rating transition state space is then defined by $\mathbb{S} = \{1, 2, \cdots, 6\}$, where $s = 1$ denotes a downgrade from A to B, $s = 2$ denotes a downgrade from A to C, $s = 3$ a downgrade from B to C, $s = 4$ an upgrade from C to A, $s = 5$ an upgrade from B to A, $s = 6$ an upgrade from C to B. When the subject currently has rating A, then it is at risk of transitions from A to B or A to C. This means that the transition processes corresponding to $s = 1, 2$ are possible and all the other transition processes ($s = 3, 4, 5, 6$) are not possible (deactivated).

Given the counting process history $N_s(t)$ for each transition s, we can derive the likelihood using a similar process as that defined in Chapter 2. The likelihood for each process is given

by,

$$\lambda_s(t)^{dN_s(t)} \exp\left(- \int_0^\tau \lambda_s(u)du \right).$$

The likelihood for the whole transition history will be the multiplication of all the individual likelihoods. More details about the likelihood can be found in Chapter 3. Unlike the Markov models, this counting process multi-state model does not restrict itself on the assumption of the history independence. By specifying different forms of the intensity function $\lambda_s(t)$, the model works with all sorts of history related predictors and can be treated as a generalization of the Markov models. If we do want the Markov property, we can always adapt the model by using the Poisson process, which is a special case of the counting process. In this case, the set of the Poisson transition processes works exactly like the Markov chain model. However, the Markov property can be readily relaxed by employing more general counting processes which will easily handle the non-Markovian properties.

This multi-state model has been used widely in the survival analysis in dealing with clinical trials and medical research. However, only few papers (e.g. Lucas et al. [10]) discussed applications of this model on rating transitions. Our contributions include adapting the theoretical framework to the detailed reality problems and also compare the model with the widely adopted Markov chain models. Enlightened by the current credit crunch and criticism of slow action of agency ratings, we will also extend our model to the market implied ratings, as well as the gaps between the two rating types. Note that the study of the rating gaps has been increasingly acknowledged as a better measure than the agency ratings.

The remaining dissertation will be divided into five chapters. Chapter 2 prepares necessary background in survival analysis and the multi-state model. We also give an introduction of the time series model needed to specify the intensity function. Chapter 3 proposed the new model and its properties. Chapter 4 will describe the details of the estimation methods for the model and perform a simulation study. Chapter 5 will apply our proposed model on Moody's market implied rating (MIR) and derive the useful transition matrix. Some issues that we will probe in the future will be listed in chapter 6.

CHAPTER 2

BACKGROUND

2.1 Agency Credit Ratings and Market Implied Ratings

2.1.1 The Agency Rating

Agency ratings have been widely used in the financial industry. The merit of adopting the agency ratings is that the agencies have tremendous resource to do the research and timely review and update the ratings. They are easy to get and fairly stable. Also, in assigning ratings, the agencies not only take consideration of the default probability, but also gauge the possible recovery rate in case the issuer really default. Note the recovery rate equals one minus loss given default which was discussed before in Chapter 1. Therefore two of the most important components in estimating potential loss have been incorporated in the agency ratings. This gives market practitioners an easy way to construct models. Agency ratings are also used to build multi-factor market models as people believe that companies with same ratings have their credit quality highly correlated and their stock and bond prices should move together.

There are three major rating agencies in the United States. They are Standard and Poor's, Moody's Investors Service, and Fitch. All the three agencies have similar processes in assigning ratings. There are two types of rating methodologies used by the agencies. One is analyst driven rating process and the other is model-based rating process. Standard & Poor's is one of the agencies that use the first type of method which can be visualized by the following plot and can be found in the paper [11] published by S&P.

In this rating process, the rating request is initialized by the issuers who want to issue debt to investors. After S&P receives the request, they will conduct an initial evaluation including meeting with issuer management. After more in depth analysis, the analysts will

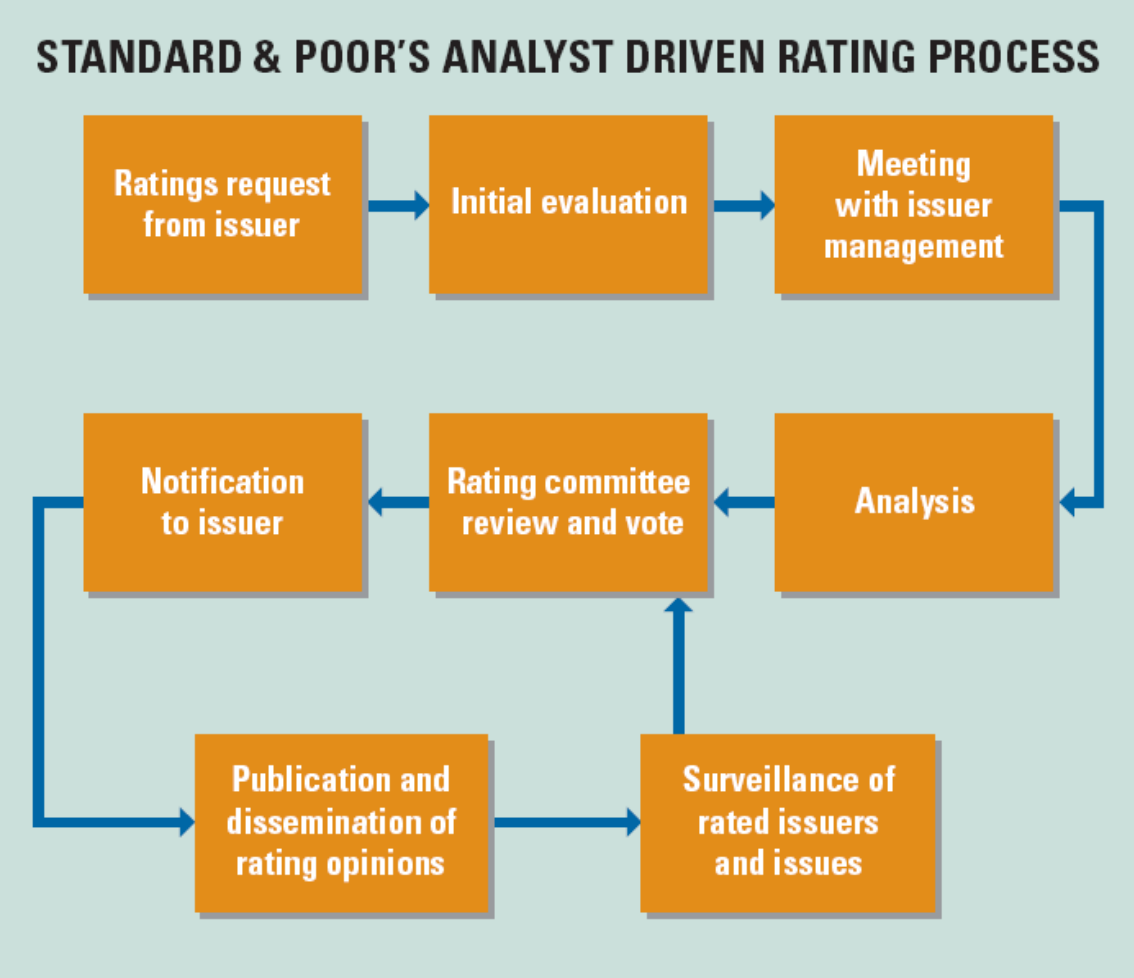

Figure 2.1: S&P Rating Process

submit a report to the internal rating committee in S&P to review and vote. The issuer will be informed about the rating decision later. If the issuer finds out that the rating is based on biased or incorrect information, it can appeal to the agency for a review. The final rating will be published in S&P's monthly report. The analyst driven rating is mainly based on the experience of the analyst as well as the company reports. Therefore, choosing the appropriate analyst candidate is very important. The other type of rating process, the model-driven method, will base the rating mainly on quantitative model and data to avoid human errors and biases. The issuer's asset quality, funding, and profitability are all the input for the mathematical model that employed by the agency. Results of the model will directly yield the rating for the issuer.

There are two types of fee structure among different rating agencies. The agencies are usually paid by either the issuer who wants to be rated or by the subscribers of the published rating reports. If the agency adopts the issuer-pay model, the issuer charges a fee for conducting the analysis and need to provide material non-public information to the agency. The rating decision is usually published to the general public freely. In this model, it is possible that the issuer tries to pressure the agency for a better rating as it has paid for the service. To avoid conflict of interest, the agency usually tries to separate its credit analysis function which is in charge of the rating decisions and the business unit which negotiates with issuers for service fees. The agency also tries to make all the rating decision based on a committee of members to eliminate the possible influence of any single individual. These two measures help to ensure the impartiality of the rating process. The rating agencies that adopt the subscription model will earn fees from the investors and other interested parties that wish to obtain the credit opinions and rating reports. There have been critics claiming that this model limit the information to only the paid investors and not to the general public. The subscribers then may be able to influence the agency for their own benefit. Also the issuer does not pay the agency so it may not provide material non-public information which may be crucial to the rating decisions.

Besides rating the issuers, the rating agencies also rate other type of financial instruments, such as specific debt issues and structured financial instruments. When a issuer, usually corporate or municipal bond, is rated, the rating agency usually starts with the issuer's rating and study more information about the bond, e.g. the terms and conditions, the quality of the collateral, and the seniority level of the debt within the debt structure. The issue's rating is often one or two "notch" above or below the issuer's current rating. The structured financial instruments that can obtain ratings from the agencies are securitized debt through the establishment of certain legal entities called special purpose vehicles (SPE). Typical debt instruments include mortgage-backed securities (MBS), asset-backed securities (ABS), and collateralized debt obligations (CDO). The securitization involves the investment banks buying the asset from debt or loan originator, slice them into standardized securities or repack them into tranches with different return-risk profile, and sell the securities or tranches to investors. In rating this type of instruments, the agencies need to consider not only the original debt or loan that comprise the collateral, but also the legal structure of the SPE and the future cash flows of the tranches. This procedure is typically very hard to be carried

out as most SPEs are designed to be as complicate as possible to avoid being scrutinized.

2.1.2 Market Implied Rating

The concept of market implied rating was first proposed by Goldberg et al. [12] as a report from Barra. The authors used the traded bond prices to construct a distribution of option adjusted yield spreads across different issuers. A basic concept in the bond market is that higher bond price will translate into lower yield for the same bond, given the coupon rate on the bond being the same. Also there are different types of bonds plus various options could be added to them, so it's usually hard to compare the bond price and yield directly as we'll have a lot of underlying factors embedded in the prices or the yields. However, sometimes we only want to look at certain type of risk, and we can compare the two bonds with different risks related to this factor and with the same risks in all other factors. For example, when we want to see the risk associated with certain ratings, we can compute the yield for the corporate bonds with this rating and subtract from it the government bonds with same duration, coupon rates, etc. This way we construct the yield spread which will ideally represent only the risk for this specific rating. Given the inverse relationship between the bond price and the yield, we get a nice representation of the risk using the yield spread. When the spread increase, that means that there is elevated risk for bonds with this rating, and vice versa. In the paper by Goldberg et al. [12], the authors tried to build option adjusted yield spreads (OAS)in which the call or put option contributions are removed and this helps us to separate out only the rating effect. Note that the yield spread was the yield difference between the corporate bond and the government securities (treasuries) with similar characteristics such as maturity and duration. Since the treasury bills and bonds payments were backed by the full faith and credit of the US government, they were traded with no default risk premium embedded. Therefore the yield spread would reflect mainly the market consensus of the default risk of the debt issue or the issuer. Higher yield spread means higher risk premium that people demand to compensate the possible default of the subsequent payments.

Given the OAS distribution, we can plot a histogram for different ratings. The following histogram is taken from the above paper by Goldberg et al. [12].

This plot describes the issuer spreads and spread thresholds for the US dollar and euro domestic market on May 31, 2001. The government bonds used in computing the spread are

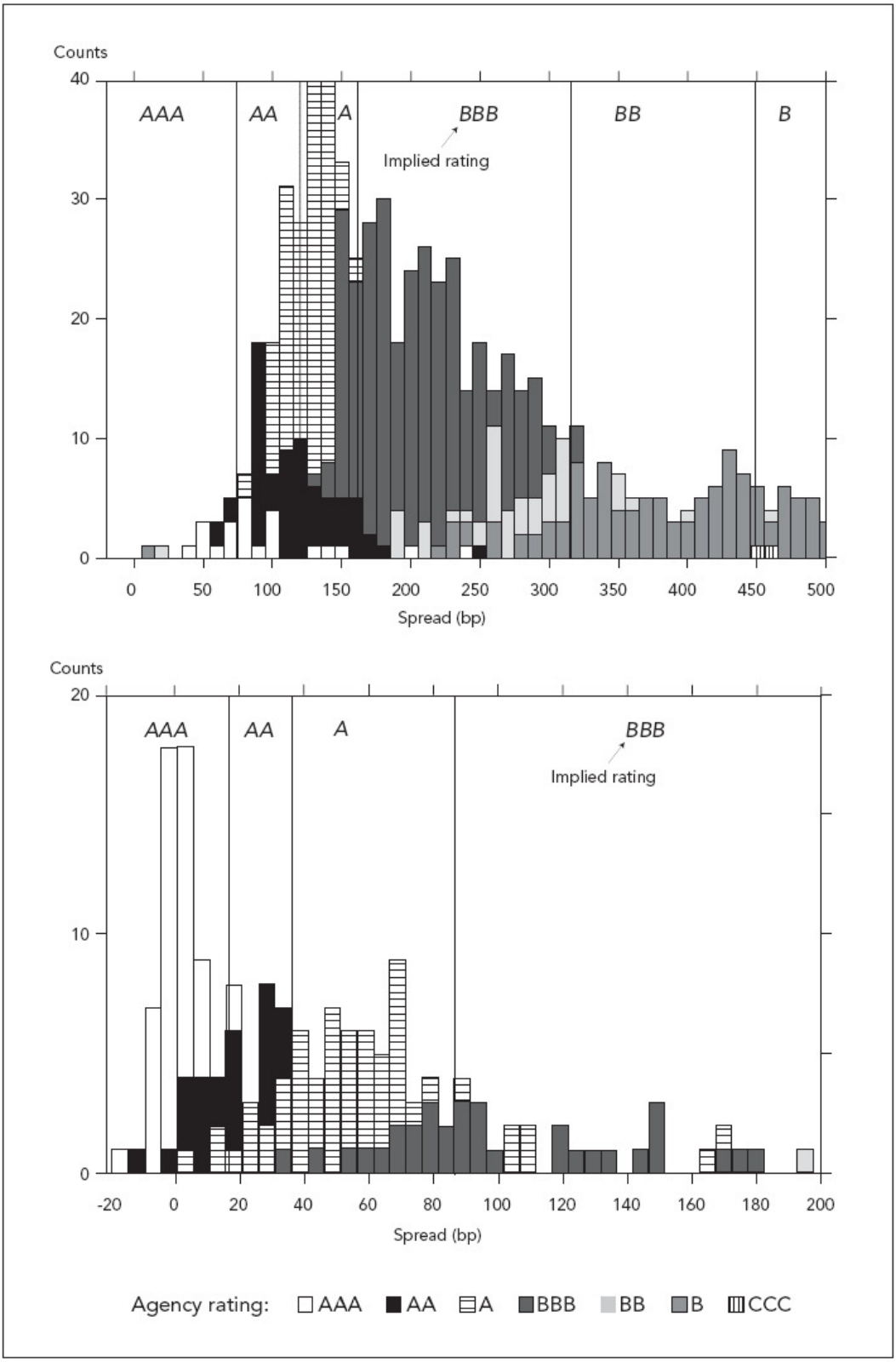

Figure 2.2: Market Implied Rating - Barra

the US Treasury yield curve and the swap curve. We can see that when rating decreases, the credit quality also diminishes so the corresponding spread increases. However, we have many bonds with the same rating categories and there are indeed large overlaps which give us difficulties in distinguish the true underlying implied ratings. Our objective here is to separate the groups of bonds into different implied rating categories such that the power of our identification method is maximized. Here, Goldberg et al. [12] specified the misclassification rate by applying a penalty function, which is defined as a weighted sum of the difference between the spread and the threshold,

$$\text{Penalty} = \sum \text{weight} \times (\text{spread - threshold}),$$

where the weights are chosen so that each rating category's contribution is the same. Finally, the penalty function will be minimized to select the appropriate threshold.

Moody's Investor Services began to build its own market implied rating platform in 2002 and made it public in 2003. The platform shared the same underlying ideas of Barra's paper to construct the market price implied ratings. However, it extended the use of bond price to various traded financial instruments (Mann et al. [13]), like credit default swap (CDS), loan CDS (LCDS), equity price, and Moody's own market default predictor (MDP) which employed the firm's financial ratios. By publishing different implied ratings using multiple instruments, Moody's MIR platform enabled investors to gain more insight on the credit quality of the underlying firms by comparing ratings across different trading markets. For the larger players such as hedge funds, this might also provides them with arbitrage opportunities by identifying inter-market inefficiency. Moody's research (Jiang et al. [14]) mentioned an example by considering the relative higher volatility in the CDS market than in the bond market. When there was adverse news on a specific firm, its CDS tended to be sold off sharply and then rebounded quickly even before any significant reaction of the related bond price. The shrewd investors could exploit this by buying during the sell-off and selling into the strength during the rebound, or even arbitraged using both the CDS and the bonds. Therefore, studying the implied ratings were helpful not only to long term bond investors but also to short term traders as well.

Besides the implied ratings, the gap between MIR and Moody's agency rating is also important to many investors. The gap is defined as the implied rating minus the Moody's rating. A positive or negative gap means that the traded yield spread of the debt issue

is lower or higher than the prevailing yield spread for the issue's rating. In predicting default probability, the rating gaps are generally more favorable than bond spread only since they are calculated to give a relative measure, as claimed by Moody's (Hamilton and Lam [15]). Another Moody's research (Mann et al. [13]) computed the rating transition matrices conditioned on the rating gaps. The analysis showed that in the short term, positive rating gaps usually lead to less default, while negative gaps have seen more defaults. Also, empirical research shows if there is consistent rating gap for a relatively long term, Moody's tend to revise the agency rating toward the implied rating. This shows that the rating gaps could have prediction power of future rating revisions.

2.2 Survival Analysis

2.2.1 Definitions and basic quantities

Survival analysis concerns with an entity, like a person or a company, and the time until some specified event occurs upon this entity. In medical research, the event may be death or appearance of a tumor. In our situation, the bond rating transition is the event. The time to event is also called survival time, which is denoted by a nonnegative random variable T. There are four functions that can be used to characterize the distribution of T. The most intuitive one is the probability density function $f(t)$. Each of the other three functions describes the survival time in a specific way. All the four functions can be derived from each other. We will discuss the other three functions in the following text.

The survival function is the probability that the entity survives to time $T = t$. It is defined as $S(t) = P(T > t)$. We note that the survival function is non-increasing. If T is a continuous random variable, then $S(t) = 1 - F(t) = \int_t^\infty f(s)ds$, where $F(t)$ is the cumulative distribution function of T. Also the probability density function $f(t) = -\frac{dS(t)}{dt}$.

Another function that can be used to describe the survival time is the hazard function or hazard rate. It is defined as,

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t}.$$

From the definition we can see that $h(t)\Delta t$ is approximately the probability of experiencing the event in the next short period $(t, t + \Delta t)$. If T is a continuous random variable, then, $h(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt}$. The cumulative hazard function $H(t)$ is defined by, $H(t) =$

$\int_0^t h(s)ds = -\log S(t)$. Therefore in the continuous case,

$$S(t) = \exp(-H(t)) = \exp\left\{ -\int_0^t h(s)ds \right\}.$$

The last and also useful one is the mean residual life function. Given that the entity survives by time t, this function measures its expected remaining lifetime. The definition is,

$$r(t) = E(T - t|T > t).$$

If T is a continuous random variable, the mean residual function can be written as a function of the survival function, namely,

$$r(t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}.$$

Simple calculation also gives the survival function in $r(t)$,

$$S(t) = \frac{r(0)}{r(t)} \exp\left[-\int_0^t \frac{1}{r(u)}du \right],$$

where $r(0) = E(T) = \int_0^\infty S(u)du$.

When analyzing the time-to-event data, it is usually impossible to observe all the event occurrence due to the time limit or cost considerations. This phenomenon is called censoring and it is the special feature of survival analysis. Sometimes the event of interest occurs before the individual enters the study, which is the censoring time C. This type of left censoring data can be denoted by the pair Y, δ ,where $Y = \max\{T, C\}$, and $\delta = I[Y = T]$. Another type of censoring is the right censoring. If we observe the censoring time C before the event occurs, there is no information on when the event will happen except that $T > C$. In this case, the data can be represented by $Y = \min\{T, C\}$, and $\delta = I[Y = T]$. The third type of censoring is a combination of the above two and we call it the interval censoring. In this case, the only thing we know is that the event happens in a certain interval.

If we look closely on the the relation between the event time T and centering time C, we may notice that sometimes the censoring time is fixed or predetermined. Therefore it is independent of the event time. This situation is called Type I censoring. Also in Type I censoring, we assume that the individuals enter the study at the same time. If it is not the case, then we have the generalized Type I censoring. Now if the study is set to stop when the kth failure occurs, the corresponding censoring is called Type II censoring. In this

case, the event time and censoring time are not independent to each other. However, the censoring at any time depends only on the information so far and does not depend on the information of failure of any single individual. This is an example of the non-informative censoring. In general, when the full data contain both observed and unobserved data, if the censoring only dependents on the observed data, we call it non-informative censoring. Otherwise it is informative censoring. Like generalized Type I censoring, generalized Type II censoring concerns the individuals not entering the study at the same.

In order to perform model specification and estimation, we first need to derive the likelihood function for the survival time. Here we focus on the Type I censoring and consider right-censored observations. This is the scheme used in this thesis. Likelihood of other censoring types can be derived similarly.

Let T be the survival time and C be the censoring time. Define $Y = \min\{T, C\}, \delta = I\{T = Y\}$ as before, the data vector can be written as $(y_i, \delta_i), i = 1, 2, \cdots, n$. Now let's look at the likelihood contribution of (y_i, δ_i), i.e. $P[Y \in [y_i, y_i + dy_i), \delta = \delta_i]$. When $\delta_i = 1$, the event time is known and the likelihood can be written as,

$$\begin{aligned} L_i &\propto P[Y \in [y_i, y_i + dy_i), \delta_i = 1] \\ &\simeq P[T \in [y_i, y_i + dy_i), C > y_i] \\ &= P[T \in [y_i, y_i + dy_i)] \cdot P[C > y_i | T \in [y_i, y_i + dy_i)] \\ &\simeq f_T(y_i) P[C > y_i | T \in [y_i, y_i + dy_i)] \\ &\propto f_T(y_i) \end{aligned} \tag{2.1}$$

If $\delta_i = 0$, we only observe the censoring time and the event time $T > C$,

$$\begin{aligned} L_i &\propto P[Y \in [y_i, y_i + dy_i), \delta_i = 0] \\ &\simeq P[T > y_i, C \in [y_i, y_i + dy_i)] \\ &= P[T > y_i] \cdot P[C \in [y_i, y_i + dy_i) | T > y_i] \\ &\simeq S_T(y_i) P[C \in [y_i, y_i + dy_i) | T > y_i] \\ &\propto S_T(y_i) \end{aligned} \tag{2.2}$$

The last steps above, equation (2.1) and (2.2), are by the assumption of independent censoring. Now,

$$L_i \propto \{f_T(y_i)\}^{\delta_i} \{S_T(y_i)\}^{1-\delta_i} = \{h_T(y_i)\}^{\delta_i} S_T(y_i) \tag{2.3}$$

and the overall likelihood function is the product of the likelihood of each individual i.

The likelihood method is a powerful tool in making inference on the survival data. It is used in some of the models described below. There is another approach that provide insight on the analysis of survival data. Aalen [16] developed his theory by employing the counting process, martingale theory, stochastic integrals, and product-integrals. The counting process is a representation of the stochastic point process. It is defined on $[0, \infty)$ and denoted as $\{N(t), t \in [0, \infty)\}$ with initial state zero, $N(0) = 0$. The sample paths of the counting process is the step function with jump one at the event time. When there are multiple individuals involved in the study, we can define the multivariate counting process as $(N_1(t), \cdots, N_k(t))$ where each component is a univariate counting process defined by $N_i(t) = I[T_i \leq t, \delta_i = 1], t > 0$. Note $N_i(t), i = 1, \cdots, k$ is right continuous and piecewise constant. Let $N(t) = \sum_i N_i(t)$ which is also a counting process as we assume there are no events occurring simultaneously among individuals. Then we denote the history or filtration of the counting process up to time t by $\mathcal{F}_t$. It is a family of σ-algebras on the same measurable space where $N(t)$ is defined. The σ-algebras of the filtration satisfy certain properties as they are increasing, $\mathcal{F}_s \subset \mathcal{F}_t, s < t$, and right-continuous, $\mathcal{F}_s = \bigcap_{t \geqslant s} \mathcal{F}_t$. In the context of survival analysis, the filtration represents the knowledge of all the past failures and individuals at risk at time t, as well as the covariate information so far.

Given $N(t)$, we define the indicator of the event by $dN(t) = N((t + dt)^-) - N(t^-)$. The expectation of this quantity given the history can be expressed by a non-negative process $\lambda(t)$, $E[dN(t)|\mathcal{F}_{t^-}] = \lambda(t)dt$ where $\mathcal{F}_{t^-}$ is the history just prior to t. $\lambda(t)$ is called the intensity process and $\lambda(t)dt$ approximates the probability of an event occurring in the next instant time period. Another important process used in the theory is the cumulative intensity process, defined by $\Lambda(t) = \int_0^t \lambda(s)ds$. It is a nondecreasing predictable process ($\Lambda(t)$ is known given $\mathcal{F}_{t^-}$; left continuous in our case) and has the property of $E[N(t)|\mathcal{F}_{t^-}] = \Lambda(t)$. This property leads to the definition of the counting process martingale $M(t) = N(t) - \Lambda(t)$. A martingale is defined as a stochastic process with expected value at time t equals to the value at some prior time s, given the history at s. In our case, $E[M(t)|\mathcal{F}_s] = M(s), s < t$, or equivalently, $E[dM(t)|\mathcal{F}_{t^-}] = 0$. The definition of martingale hints us that it is just a noise process by removing the predictable trending process $\Lambda(t)$, called the compensator. This subtraction technique is useful in constructing martingales from counting processes.

Although $M(t)$ is a martingale, this is not the case for $M^2(t)$. However, we can make

$M^2(t)$ a martingale by applying the same technique and subtracting its compensator. This compensator is call the predictable variation process of $M(t)$, and denoted as $\langle M \rangle(t)$. To calculate $\langle M \rangle(t)$, we note that $Var[dM(t)|\mathcal{F}_{t^-}] = d\langle M \rangle(t)$ and then $Var[dM(t)|\mathcal{F}_{t^-}] \simeq \lambda(t)$. Now we introduce the concept of stochastic integral of the martingales. Let $K(t)$ be a predictable process. The stochastic integral of $K(t)$ with respect to the martingale $M(t)$ is $\int_0^t K(u)dM(u)$. It is also a martingale and its predictable variation process can be computed from

$$\left\langle \int_0^t K(u)dM(u) \right\rangle = \int_0^t K(u)^2 d\langle M \rangle(u).$$

The above theory can by utilized to derive the non-parametric estimators of the most important quantities of the survival analysis, the hazard rate and survival function. We now focus on a important class of models, the multiplicative intensity model, where the intensity process is given by,

$$\lambda(t) = Y(t)\alpha(t).$$

Here $Y(t)$ is the number of at-risk individuals at time t and it's a predictable process. $\alpha(t)$ is a deterministic function represents the rate of jump occurrence. We want to motivate an estimator of the cumulative rate function $A(t) = \int_0^t \alpha(u)du$. Now $M(t) = N(t) - \int_0^t Y(u)\alpha(u)du$ is a martingale which gives the heuristic estimating equation $0 = dN(t) - Y(t)\alpha(t)dt$. This leads to the famous Nelson-Aalen estimator

$$\hat{A}(t) = \int_0^t Y(u)^{-1} dN(u). \tag{2.4}$$

Since $N(t)$ only jumps when $Y(t)$ is positive, equation (2.4) becomes,

$$\hat{A}(t) = \int_0^t \frac{J(u)}{Y(u)} dN(u),$$

where $J(t)$ is the indicator of whether $Y(t)$ is positive, $J(t) = I[Y(t) > 0]$.

The martingale central limit theorem also enable us to calculate the limiting process as the sample size goes to infinity (or very large). This leads to easy computation of the confidence interval at any fixed time and the confidence band for $A(t)$. The non-parametric Kaplan-Meier estimator for the survival function is derived from the product integral of $1 - d\hat{A}(t)$, i.e.

$$\hat{S}(t) = \prod_{s=0}^{t} [1 - d\hat{A}(s)] = \prod_{s=0}^{t} \left[1 - \frac{dN(s)}{Y(s)}\right].$$

We have seen the likelihood constructed in the traditional way above. The counting process gives another way to write the likelihood. From the intuitive meaning of the intensity process and the jump process $N(t)$, the likelihood for each process is derived as,

$$\lambda(t)^{dN(t)} \exp\left(- \int_0^\tau \lambda(u)du \right).$$

The discussions above have been focusing on the likelihood construction and introduction of the counting process for modeling the survival data. We will now make a short presentation of the popular models and their extensions to accommodate different situations arisen in the real data analysis.

One of the widely used model is the accelerated failure-time (AFT) model proposed by Miller [17]. It assumes that a covariate has multiplicative effect on the event time (or the log of the event time). This is in contrast to the proportional hazard model, which is proposed by Cox [18]. The Cox's model, as we will discuss later, assumes the effect of any covariate is multiplicative to the hazard rate. Another useful comparison between these two models is that the AFT model is a pure parametric model which needs distribution assumption on the parameters while the Cox's model is semiparametric which means the baseline hazard can be left as non-specified and does not affect the result of the (partial) likelihood estimation. The AFT model usually provide robust results and is not very sensitive on the choice of the distributions. Since the effect of a covariate acts directly on the event time (or the life time until the next event), it also has very intuitive explanation on the parameter of interest. For example, it is comprehensible to say that the treatment increase the expected life time of a patient by 10%, comparing to the control group.

The AFT model can be written as the regression form, $\log T = \beta Z + \epsilon$, where Z is the vector of covariates, β is the coefficient parameter, and ϵ is the error term. We can also accommodate the baseline base by adding in the extra term β_0,

$$\log T = \beta_0 + \beta Z + \epsilon.$$

The error term ϵ is assumed to have median 0. In this situation, the survival function has the form of the baseline survival with re-scaled time axis, $S(t|Z) = S_0(te^{-\beta Z})$. And this is where the name accelerated life time comes from.

The proportional odds model is another useful model in the survival analysis. It assumes

the covariate effect acts on the log odds of the survival function, i.e.

$$\frac{S(t|Z)}{1-S(t|Z)} = e^{\beta Z}\frac{S_0(t)}{1-S_0(t)}.$$

Sometimes it is useful and convenient to model the function of interest by an additive model. The additive risk model, proposed by Aalen [19], is a widely used model bearing this feature. Instead of considering the life time till event or the survival function, it focus on the hazard rate directly. The general model expression is,

$$h(t|Z) = \beta Z = \sum_i \beta_i Z_i.$$

Here, as usual, Z represents the covariates and β is the coefficient vector. Often it is useful to center the covariates to compare the hazard rate to a baseline hazard. In this situation, the model breaks into two parts, the baseline hazard and the covariate effects that cause the hazard to deviate from the average person. The model expression is,

$$h(t|Z) = h_0(t) + \beta Z.$$

The additive risk model is very flexible and can accommodate the need for time-varying coefficient and covariates. If the coefficient parameter β is a function of time, it can be assumed to be non-parametric which extends the original model to a greater extent. If we also assume the covariate to be dependent on time, as discussed by Aalen in his originally paper, the counting process approach can be employed and we can use the martingale techniques in the modeling and estimation.

Now let's assume the survival time is subject to right censoring, and let C denote the censoring time which is conditionally independent to T given the covariates. Also let $N(t) = I[Y \leq t, \delta = 1]$ be the counting process associated with the original survival time process, where $Y = \min\{T, C\}$ and δ is the indicator of event, $\delta = I[Y = T]$. Now the intensity process of the counting process $N(t)$ can be modeled in a similar fashion as the hazard rate before,

$$\lambda(t|Z) = \beta(t)Z(t) = \sum_i \beta_i Z_i.$$

In the model estimation, one first considers the integrated functions $A(t) = \int_0^t \beta(s)ds$. Aalen [19] proposed an ordinary least square (OLS) type estimator which is motivated by applying the least square method to the jumps of the counting process $N(t)$. The estimator is

consistent and asymptotically normal. Huffer and McKeague [20] also introduced a weighted least square (WLS) estimator which is an approximate maximum likelihood estimator. It is also consistent and asymptotically normal if the intensity function is bounded away from zero. The WLS estimator has minimal asymptotic variance and, comparing to the OLS estimator, it may provide significant variance reduction.

However, one drawback of the additive risk model is that the positiveness of the hazard rate or the intensity function is not a built-in property. Caution is warranted in the parameter estimation and model diagnostics to avoid non-empirical hazard rates when using this type of model. One way to overcome this deficiency is to use functions that are always positive. The proportional hazard model (or relative risk model), proposed by Cox [18], substitutes the exponential function for the linear function. This means that the covariate effects act on the baseline hazard in a multiplicative way, i.e.

$$h(t|Z) = h_0(t)e^{\beta Z}.$$

One implied assumption under this model is that the hazard ratio for a certain covariate is the same given different values of other covariates. When applying this model in any data analysis, one should make careful assessment on whether this assumption is true. Another important assumption in above model is the common baseline hazard. When this is not the case, one can modify the model to accommodate heterogeneity among different cohorts,

$$h(t|Z) = h_{0j}(t)e^{\beta Z}$$

In the application of the Cox's model, the baseline hazard is usually not assumed to have any specific function form, which leads to the semi-parametric Cox's model. Like the additive risk model, the Cox's model can be extended to account for time-varying coefficient or time-dependent covariates or both, which gives,

$$h(t|Z(t)) = h_0(t)e^{\beta(t)Z(t)}.$$

The inference using the Cox's model is based on the partial likelihood approach introduced by Cox [21]. Recall $h(t|Z) = h_0(t)e^{\beta Z}$, and β is the parameter of interest while $h_0(t)$ is the nuisance parameter. We want to estimate β based on the data we have, i.e. the observed event times. Assume $t_1, \cdots, t_k$ are the event times (not censoring times). Then the partial

likelihood contribution of the jth event is the conditional probability given the risk set,

$$\begin{aligned} PL_i &= P[T_i \in [t, t+dt) | R(t), \text{exactly one failure in } [t, t+dt)] \\ &\simeq \frac{h_0(t)\exp(\beta Z_i(t))dt}{\sum_{j\in R(t)} h_0(t)\exp(\beta Z_j(t))dt} \\ &= \frac{\exp(\beta Z_i(t))}{\sum_{j\in R(t)}\exp(\beta Z_j(t))} \end{aligned}$$

Here the risk set $R(t)$ contains the labels for those individuals that are at risk at time t. The partial likelihood for the observed data is given by,

$$PL \propto \prod_{i=1}^{n} \left\{ \frac{\exp(\beta Z_i(y_i))}{\sum_{j\in R(y_i)}\exp(\beta Z_j(y_j))} \right\}^{\delta_i},$$

where $y_i, i = 1, \cdots, n$ are the observed event and censoring times.

The partial likelihood is the multiplication of conditional probabilities. Unlike the marginal or conditional likelihood, it is not possible to interpret it as a probability statement. The consistency and efficiency criteria need to be constructed for the parameter estimation. Let's define the score function,

$$u_i = u_i(\beta) = \frac{\partial}{\partial\beta}\log(PL_i), i = 1, \cdots, n.$$

It's easy to verify that $E(u_i) = 0$ and,

$$var(u_i) = E\big[-\frac{\partial^2}{\partial\beta^2}\log(PL_i)\big] = I_i.$$

Under certain regularity conditions, the score vector $U = (u_1, \cdots, u_n)$ has the asymptotic property as $n \to \infty$,

$$U \sim N\bigg(0, \sum_{i=1}^{n} E(I_i)\bigg).$$

The consistency and efficiency of the maximum partial likelihood estimate $\hat{\beta}$ can be derived based on above asymptotic normality. We can also have the regular hypothesis testing and confidence intervals using similar technique.

Now we have discussed several models that are among the most frequently used models in survival analysis. We want to make a final note here on modeling the individual or cohort heterogeneity. For example, some people may tend to have more tumors than others even with similar measurable characteristics. Manton et al. [22] introduced the notion

of frailty by adding a random effect term to the hazard rate. Let w be a non-negative random variable with density function $f(w)$ and Laplace transformation $L(s) = E[e^{-sw}]$. Also let $h_0(t), H_0(t), S_0(t)$ denote the baseline hazard rate, cumulative baseline hazard rate and baseline survival function, respectively. Assume the hazard rate given w is,

$$h(t|w) = wh_0(t).$$

The conditional survival function can be derived as,

$$S(t|w) = \exp\left(-\int_0^t h(t|w)dt \right) = \{S_0(t)\}^w.$$

The unconditional survival function,

$$S(t) = E_w[S(t|w)] = E_w[\exp(-w(-\log S_0(t)))] = L(H_0(t)).$$

If the Cox regression model is adopted, we can add the covariate term and get,

$$S(t) = L(e^{\beta Z} H_0(t)).$$

This survival function is not in the form of proportional hazard model except that w is of positive stable distribution. This non-proportionality ensures the identifiability of the distribution of w as long as it has finite mean (Elbers and Ridder [23]). Usually we assume it has mean equals to one for simplicity.

2.2.2 Recurrent Data Analysis

Recurrent data analysis is a special case of the multivariate failure time analysis. There have been two categories of the multivariate failure time analysis. The first category concerns multiple events occurring on a subject, like the relapse of the cancer. The other category focuses on the clustered data which arise when there are natural clustering and data dependence within each subject, like the disease on both eyes of one individual. For the multiple event data analysis, the events could encompass an ordering scheme, which includes the recurrent event analysis as a special case. If there is no ordering presented, the multiple events can be modeled with several concurrent event process.

Recurrent event analysis has been applied mainly in medical research and clinical trials as there is imminent need to deal with relapse of various disease and the subsequent treatment.

The process of recurrent events is often accompanied by a terminal event, which is often the death or the end of study (censoring). There have been two types of models that are widely used. One is the marginal, or partially specified model, which study each recurrence with a separate hazard rate or intensity function. It treats each event as independent or partially dependent on other events. The other type is the fully specified model which considers the joint dependence structure of the events and take into account all the possible information. We will give a brief review on both type of models in the following section. Interested readers are referred to Keles et al.[24] for more details. It is interesting to note that no model fits all the data. This is especially true in the recurrent event analysis. Some models are more useful in certain circumstances while other may be better for a different audience. The results from different models are often not comparable, some are even conflicting to each other. Therefore it is the demand that drives the analysis and builds more advanced models instead of people's imagination.

The counting process notations will be used to describe the models. For the k^{th} event of a randomly selected individual, let T_k be the event time and $N_k(t) = I(T_k \leqslant t)$ be the counting process for this specific event. The history or filtration by time t is denoted by $\mathcal{F}_t^k$ which includes all the information of the process $N_k(t)$ and covariate $Z_k(t)$ up to time t. Let $N(t) = \sum_k N_k(t)$ and the corresponding history by time t to be $\mathcal{F}_t$. We first discuss the partially specified models. Lin et al. [25] proposed a marginal model for the more abroad multiple event analysis and it can be used specifically on the recurrent event analysis. We will here abbreviate it as WLW model.

$$E[dN_k(t)|\mathcal{F}_{t-}^k] = \lambda_k(t|\mathcal{F}_{t-}^k)dt = \lambda_{0k}\exp(\beta_k Z_k(t))dt$$

Here we should emphasis that the intensities are distinct not only for different events, but for different individuals as well, i.e. $\lambda_{ki}(t|\mathcal{F}_{t-}^{ki}) = \lambda_{0k}\exp(\beta_k Z_{ki}(t))$. As the intensities does not dependent on the history of other events, it is possible that an individual will be at risk of having the k^{th} event without having the $(k-1)^{th}$ event. This make the result hard to interpret. Pepe and Cai [26] tried to solve this problem by limiting the risk set on the subjects that have experienced the $(k-1)^{th}$ event,

$$\lambda_k(t|N_{(k-1)}(t-) = 1, \mathcal{F}_{t-}^k) = \lambda_{0k}\exp(\beta_k Z_k(t)).$$

Another partially specified model is proposed by Williams et al. [27] and will be abbreviated as PWP model. This model considers the intensity function given the overall

history of the counting process and the covariate process, $\mathcal{F}(t) = (N(t), Z(t))$. A stratum variable is used to specify the intensity function for each event. Two time scales are employed. One is to consider the total time for the k^{th} event from the start of the study and the other is to only use the time from the immediately preceding event. The models are given below.

$$\begin{aligned} &\text{(Total time model)} && \lambda(t|\mathcal{F}_t) = \lambda_{0k}(t)\exp(\beta_k Z(t)), \quad t \in (t_{k-1}, t_k] \\ &\text{(Gap time model)} && \lambda(t|\mathcal{F}_t) = \lambda_{0k}(t - t_{k-1})\exp(\beta_k Z(t)), \quad t \in (t_{k-1}, t_k] \end{aligned}$$

Here for each k, λ_{0k} is an arbitrary baseline intensity function and β_k is the stratum-specific regression coefficients. The shape of the intensity function depends on the number of events occurring before and the history of the covariates process. The subject moves to the k^{th} stratum right after the $(k-1)^{th}$ event with intensity given above and stay there until the occurrence of the k^{th} event. The problem with the PWP model can be seen by the procedure of the risk set construction for the gap time model. For the i^{th} individual, let $d^i_{k-1,k}$ denote the duration $T_k - T_{k-1}$. The risk set is then $R(t) = \{l : d^l_{k-1,k} \geq d^i_{k-1,k}\}$. This is equivalent to finding the risk set for the regular Cox model by aligning all the $(k-1)^{th}$ event times as if they occur simultaneously. The result is hard to explain if the same events from different individuals happen at vast different times.

Partially specified models are hard to simulate out of their conditions. Therefore they are not real life models. However, they are still widely used as good approximations to the real data and for their simplicity. Both models have been implemented in the commercial software like SAS. To capture the joint structure of the events, we need to model the counting process $N(t)$ as a whole. The following models share this feature.

The multiplicative/proportional intensity model, proposed by Anderson and Gill [28], is to apply a Cox type regression model on the intensity conditional on the history of the process,

$$E[dN(t)|\mathcal{F}_{t-}] = Y(t)\lambda(t)dt = Y(t)\lambda_0(t)\exp(\beta Z(t))dt,$$

where $Y(t)$ is the at-risk indicator. This model assumes that the dependence on the past history $\mathcal{F}_{t-}$ only through the covariate process $Z(t)$. If $Z(t) = Z$, i.e. the covariate is time-invariant, the past events will have no influence on the risk of future recurrence. Sometimes we may want to avoid this type of precise definition of the dependence structure. Lawless and Lawless and Nadeau [29], Nadeau et al. [30] and Yang et al. [31] use the proportional

means/rates model which conditions on the covariate process instead of the overall event history,

$$E[dN(t)|Z(t)] = Y(t)d\mu(t) = Y(t)d\mu_0(t)\exp(\beta Z(t)).$$

Since the model assumes no dependence on the event history, the Markov property holds and the recurrence counting process degenerates to the non-homogeneous Poisson process with the intensity function only depends on t and $Z(t)$. This gives the proportional rates model as above. If the covariate is time-invariant, i.e. $Z(t) = Z$, the integration of the regression equation gives the proportional means model,

$$E[N(t)|Z(t)] = Y(t)\mu(t) = Y(t)\mu_0(t)\exp(\beta Z).$$

When there is a large number of recurrence, the intensity and means/rates models are desired since they can incorporate among different event and extract as much information as possible. However, the marginal models have to consider the events one by one so they are cumbersome in this case. On the contrary, if there are only a few events, the marginal models are better choice than the fully specified models above.

2.3 Time Series Analysis

2.3.1 The Stationarity

Time series $\{X_t\}$ is called strictly stationary if for any combination of k and t, the joint distributions of $\{X_{t_1}, \cdots, X_{t_k}\}$ and $\{X_{t_1+t}, \cdots, X_{t_k+t}\}$ are the same. The time series $\{X_t\}$ is called weakly stationary if it satisfies the following three conditions,

1. $E(X_t) \equiv \mu$;
2. $Var(X_t) \equiv \sigma^2$;
3. $Cov(X_t, X_{t+k}) = \gamma(k)$, for all integers t and k.

These three conditions are equivalent to say that the time series will have constant mean, variance, and the covariance does not depend on the time shift. Generally the strict stationarity is hard to be satisfied empirically. Therefore people tend to focus more on the weak stationarity when deriving the theories over time series.

2.3.2 The ARMA Model

Time series models are one of the most important statistical models used in economic and financial research. The economic data, such as GDP, unemployment rate, and interest rate, and financial data, like debt and equity trading prices are all history-dependent and serially correlated in time, which requires the use of specially designed models. The time series models usually deal with discrete time periods like daily, weekly, monthly, or quarterly data. In high frequency financial modeling, it is also possible to have much shorter periods like minutes, seconds or even milliseconds. Now let X_t be the quantity we are interested in at time t. A basic model is to specify the dependence of X_t on its own history,

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i}, \tag{2.5}$$

where ϕ's are coefficient parameters. The equation (2.5) is called the autoregressive (AR) model. Since the series of X_t is defined recursively, we need to be careful in choosing the multiplicative coefficient so that the whole series will not blow up as t becomes larger. In time series modeling, we have the following stationarity conditions (Tsay [32]). $\{X_t\}$ is said to be *strictly stationary* if the joint distribution of $(X_{t_1}, \cdots, X_{t_k})$ is identical to that of $(X_{t_1+t}, \cdots, X_{t_k+t})$ for all t, where k is an arbitrary positive integer and $(t_1, \cdots, t_k)$ is a collection of k positive integers. Strict stationarity is equivalent to having invariant joint distribution of collection of time series variables. Also, $\{X_t\}$ is *weakly stationary* is both the mean of X_t and the covariance between X_t and X_{t_l} are time-invariant, where l is an arbitrary integer. The stationarity condition for AR(p) model is such that all the solutions of the characteristic equation,

$$1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p = 0$$

are greater than one in modulus. More specifically, the AR(1) model,

$$X_t = \phi_0 + \phi_1 X_{t-1}$$

is stationary if and only if $|\phi_1| < 1$.

Another type of basic time series models is the moving average (MA) model. The quantity of interest X_t is comprised of the sum of shocks, and can be viewed as a simple extension of the white noise series,

$$X_t = c_0 + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \tag{2.6}$$

where $\{a_t\}$ are shocks and c_0 and θ are coefficient parameters. The MA models are always weakly stationary since they are finite linear combinations of a white noise sequence.

The autoregressive and moving average (ARMA) model is a combination of the AR and MA models (equation (2.5) and (2.6)) and depends on both its own history and the history of some shocks, or exogenous impacts,

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i}, \tag{2.7}$$

where the a_t's are the shocks, and ϕ, θ are parameters. The stationarity condition for ARMA model is the same as the AR model by looking at the AR part of the model. This model is useful in many areas besides financial research in providing insight on the time correlations of the data as well as making predictions of the future events.

Many of the economic data exhibit cyclic patterns when monthly or quarterly data are considered. In this case we need to seasonally adjust the time series data so that the underlying trend can be more easily identified. There are also times that the time series are not stationary, but it can be easily transformed into a stationary one by differencing the adjacent terms a couple times. A more general model, ARIMA(p,q,d) model will consider all above cases and is considered the most classical model when fitting time series data. The three parameters p, q, and d represent the order for the AR part, MA part, and the differencing part. For example, ARIMA(p,0,0) would represents an AR(p) model while ARIMA(q) refers to a MA(q) model. The general ARIMA(p,q,d) model can be written as,

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right)(1-L)^d X_t = \left(1 + \sum_{j=1}^{q} \theta_j L^j\right)\epsilon_t, \tag{2.8}$$

where L is the lag operator such that $L^k X_t = X_{t-k}$, and ϕ_i's and θ_j's are defined as the same in ARMA model above equation (2.7).

The ARIMA model (equation (2.8)) can also be combined with linear regression analysis if the residuals have serial correlations,

$$X_{1t} = \alpha + \beta X_{2t} + e_t$$

where X_{1t} is the dependent variable, X_{2t} is the independent variable, α, β are coefficients and e_t are the regression residuals. In the case of observing serial correlations in e_t, we can apply appropriate ARMA models by checking the auto-correlation function (ACF) of the residuals

and determine the order of the ARMA model. Tsay's book [32] provides an excellent review of above subjects.

2.3.3 The GARCH Model

The past two years have seen tremendous volatility in the financial markets, like equity, debt, commodity and even currency markets. Volatility is generally one of the most important part in the financial modeling, together with the return prediction. In the option market, the famous Black-Scholes model need a correctly estimated volatility for the option pricing. In risk management, the value at risk (VaR) calculation is based solely on the volatility estimation. And in portfolio management, the mean-variance structure requires a precise input of the volatility. Finally, various volatility index (VIX, VXO, etc.) have been established to give investors more insight on the market.

However, modeling volatility is not an easy job. It is not directly observed from the trading price or returns, nor can we plug in the last year's volatility to make predictions, as history never repeats itself. A consensus is to use the standard deviation or variance to represent the volatility. There are currently two types of volatility models, one is to use an exact function to define the evolution of the standard deviation, and the other is to use a stochastic equation to estimate the standard deviation. We will briefly discuss both in the following paragraphs. More details can be found in Tsay [32].

The first is the autoregressive conditional heteroscedastic (ARCH) model proposed by Engle [33] in 1982. This Nobel prize winning paper defined the evolution of the volatility in the following manner,

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2,$$

where $\{a_t\}$ is the shock series, $\{\epsilon_t\}$ is white noise series (i.i.d. random variables with mean zero and variance one), α_i are positive coefficients for $i \geq 0$. One of the advantages of the ARCH model is that it gives a_t a heavier tail (positive excess kurtosis) than the normal distribution. This is consistent with empirical observations of asset returns in which there are more outliers than modeling by normal distributions. However, there also weaknesses of the ARCH model. One is that in the ARCH model, the positive and negative shocks are treated equally, while in reality they tend to have asymmetric effects on the asset returns. Another is that the structure of the ARCH model will have restriction on the value of its

parameters, this limits the use of this model on various situations. Moreover, many view the ARCH model as a pure mathematical improvement without real economic or financial meaning on the parameters. This makes the estimation result hard to be explained.

A closer look at the structure of the ARCH model shows that it resembles the characteristics of the MA model. This enable us to guess what will happen if we also add in a similar AR model. Bollerslev [34] proposed a model in 1986 answered this question. This generalized ARCH (GARCH) model extended the ARCH model by adding the autoregressive part,

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2,$$

where $\{a_t\}$, $\{\epsilon_t\}$, and α_i are defined the same as in the ARCH model, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(m,s)}(\alpha_i + \beta_i) < 1$. It is easier to see the ARMA form by defining $\eta_t = a_t^2 - \sigma_t^2$ and rearrange the terms,

$$a_t^2 = \alpha_0 + \sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) a_{t-i}^2 + \eta_t - \sum_{j=1}^{s} \beta_j \eta_{t-j}.$$

Here $\{\eta_t\}$ is a martingale difference series, which means the mean of η_t is zero and the covariance between any two random variables η_t and η_{t-j} for any positive j is zero. Like the ARCH model, the GARCH model also has positive excess kurtosis and "fatter" tails. However, empirical research showed that if $\{\epsilon_t\}$ is normally distributed, the tail for $\{a_t\}$ will not be fat enough to simulate the real data. Therefore, people have suggested to use the student t-distribution for $\{\epsilon_t\}$ for better estimation.

CHAPTER 3

PROPOSED MODEL

In this chapter, we propose a model that considers the credit rating transitions with the presence of observed covariates and unobserved random effect. The model is motivated by the pioneering research of Koopman et al. [10]. Here we extended the use of time series as the random effects by introducing the GARCH model. It is well known that financial data series usually exhibit non-constant volatility or variance. By incorporating the GARCH effect, we hope to provide better estimates of the transition process parameters. In the following sections, we will first specify the model and the theoretical background. Then the estimation procedure and relevant formula are described in details.

3.1 Model Specification

We denote the credit rating transition space by S which is the set of possible rating transitions. Each element $s \in S$ represents a transition from one rating to another. The subjects or issuers are denoted by $k = 1, \cdots, K$. The next step is to define the right-continuous and piecewise constant counting processes $N(t)$, $N_k(t)$ and $N_{sk}(t)$ where $t \in (0, T]$, the observation time interval. The process $N(t)$ makes a jump of size 1 at each time there is a rating transition event for one of the K firms. Similarly, $N_k(t)$ makes a jump of size 1 at each time there is a transition for firm k, and $N_{sk}(t)$ makes a size 1 jump at each time there is a type s transition for firm k. It follows that,

$$N_k(t) = \sum_{s \in S} N_{sk}(t), \qquad N(t) = \sum_{k=1}^{K} N_k(t) = \sum_{k=1}^{K} \sum_{s \in S} N_{sk}(t).$$

Therefore, $N_{sk}(t)$ counts the total transition events of type s for firm k by time t, and similar for $N_k(t)$ and $N(t)$. For each of the counting processes $N_{sk}(t)$, we assume there is a stochastic

intensity process $\lambda_{sk}(t)$ defined as,

$$\lambda_{sk}(t) = \lim_{\Delta t \to 0} \frac{P[N_{sk}((t+\Delta t)^-) - N_{sk}(t^-) = 1 | \mathcal{F}_{t^-}]}{\Delta t},$$

see Andersen and Gill [35]. This intensity is similar to the hazard rate function used in survival analysis. $\lambda_{sk}(t)\Delta t$ is approximately the probability of having a type s transition for firm k during $(t-\Delta t, t]$. However, we need to point out the fact that we are allowing multiple events (transitions) to occur over our observation time span, and it is possible to have multi-type events during this period. The multi-state model lends us tremendous flexibility in modeling this kind of processes.

The transition counting process $N(t)$ is the sum of event-specific counting processes for different subjects. Given a subject with a specific rating, e.g. Aaa, at time t, the subject is only at risk of transitions originated from Aaa instead of from another rating. We account for this at-risk effect by introducing the indicator process $R_{sk}(t)$,

$$R_{sk}(t) = \begin{cases} 1 & \text{if subject } k \text{ is at risk for type } s \text{ transition at time } t^- \\ 0 & \text{otherwise.} \end{cases}$$

We also need to identify and record the transitions when they occur. This is facilitated by another indicator process $Y_{sk}(t)$,

$$Y_{sk}(t) = \begin{cases} 1 & \text{if subject } k \text{ experiences a type } s \text{ transition at time } t \\ 0 & \text{otherwise} \end{cases}$$

The likelihood for the observed path of $N(t)$ is,

$$\prod_{i=1}^{N(T)} \prod_{k=1}^{K} \prod_{s \in S} \exp\left(Y_{sk}(t_i) \ln\{\lambda_{sk}(t_i)\} - R_{sk}(t_i) \int_{t_{i-1}}^{t_i} \lambda_{sk}(t) dt \right), \tag{3.1}$$

where $\{t_i\}_{i=1}^{N(T)}$ are transition times. This is a very general model which can accommodate a wide range of situations. Koopman et al. [10] presented a fully parameterized intensity function with the form of proportional hazard rate,

$$\lambda_{sk}(t) = \lambda_{sk,0}(t) \exp\left(\eta_s + \gamma_s' \omega_k(t) + \alpha_s \psi(t) \right), \qquad t \in (0, T], \tag{3.2}$$

where $\lambda_{sk,0}(t)$ is the baseline intensity function that in the paper (Koopman et al. [10]) has a deterministic form and depends on the past transition durations, $\eta_s, \gamma_s', \alpha_s$ are coefficient parameters, $\omega_k(t)$ are time-dependent covariates and $\psi(t)$ is a latent factor, or random effect.

Since all the data are daily observations, we use a discretized version of the transition process. Recall $(0,T]$ is the time interval that we have observations. Here we assume $T \in \mathbf{N}$ and write $(0,T] = \sum_{i=1}^{T}(i-1,i]$. We further assume the covariates and rating changes are only observed at the end of each day i. Now the condition likelihood of the counting process $N(t)$ given the random effect becomes,

$$\begin{aligned} L(\theta|\Psi_T,\mathcal{F}_T) &= \prod_{i=1}^{N(T)}\prod_{k=1}^{K}\prod_{s\in S}\exp\left(Y_{sk}(t_i)\log\{\lambda_{sk}(t_i)\} - R_{sk}(t_i)\int_{t_{i-1}}^{t_i}\lambda_{sk}(t)dt\right) \\ &= \prod_{i=1}^{T}\prod_{k=1}^{K}\prod_{s\in S}\exp\left(Y_{sk}(i)\log\{\lambda_{sk}(i)\} - R_{sk}(i)\lambda_{sk}(i)\right) \\ &= \prod_{i=1}^{T}\prod_{k=1}^{K}\prod_{s\in S}\exp\Big(Y_{sk}(i)\Big(\log(\lambda_{sk,0}(i)) + \eta_s + \gamma_s'\omega_k(i) + \alpha_s\psi(i)\Big) \\ &\qquad -R_{sk}(i)\lambda_{sk,0}(i)\exp\Big(\eta_s + \gamma_s'\omega_k(i) + \alpha_s\psi(i)\Big)\Big), \end{aligned} \tag{3.3}$$

where $\theta = \{\eta_1,\cdots,\eta_S,\gamma_1,\cdots,\gamma_S,\alpha_1,\cdots,\alpha_S,\rho,\alpha_0,\alpha_1,\beta_1\}$ is the parameter vector, $\Psi_T = (\psi_1,\psi_2,\cdots,\psi_T)$ denote the random effect up to time T, and sigma-field $\mathcal{F}_T$ denote the history of observations up to time T. The second equality is derived from the fact that the intensity function is piecewise constant during each day, so the integration is replaced by the summation. We also need to note that the intensity rates are left-continuous so they are predictable functions.

Koopman et al. [10] proposed a time series model for the random effect $\psi(t)$ that is similar to first order autoregressive (AR(1)) model,

$$\psi_{t_i} = \rho_s^{t_i - t_{i-1}}\psi_{t_{i-1}} + \sigma_{t_i}\varepsilon_{t_i}, \qquad \varepsilon_{t_i} \overset{\text{iid}}{\sim} \mathbf{N}(0,1),$$

where the shocks $\sigma_{t_i}\varepsilon_{t_i}$ are assumed to happen only with the transitions. We believe this time series model is an over-simplified one which does not take into account of the volatility change of the transition processes. Although the credit ratings and their transitions belong to the fixed income side of the financial market and tend not to behave like equity prices, it's become more and more widely accepted that they start to exhibit volatile path with the invention and prosperity of the credit derivatives. An issuer's debt is traded in the bond market daily, just like it's equity. When there is consistent downward pressure on the bond prices, the rating agencies often closely monitor the issuer's financial situation and may make

rating change accordingly. Moreover, starting several years ago, Moody's Investor Services publishes daily the Market Implied Ratings (MIR) that are derived from the traded bond prices. This MIR ratings data enable us to study the transition volatilities which is not feasible when using the very low frequency traditional transition rates. To our knowledge, financial institutions have begun the research on MIR as well as gaps between MIR and agency ratings to predict future rating changes. Based on this analysis, we extend the above time series model by adding the generalized autoregressive conditional heteroscedastic (GARCH) model which has been proved to be able to capture the volatility change,

$$\psi_i = \rho_s \psi_{i-1} + a_i, \qquad a_i = \sigma_i \varepsilon_i, \qquad \sigma_i^2 = \alpha_0 + \sum_{k=1}^{m} \alpha_k a_{i-k}^2 + \sum_{l=1}^{n} \beta_l \sigma_{i-l}^2 \tag{3.4}$$

In our extended model, we omit the baseline intensity function when modeling the transition intensity to avoid identifiability problems that may arise in the presence of the random effects. Also we will use the GARCH(1,1) model which has been proved in literatures to be efficient in modeling volatilities,

$$\psi_i = \rho_s \psi_{i-1} + a_i, \qquad a_i = \sigma_i \varepsilon_i, \qquad \sigma_i^2 = \alpha_0 + \alpha_1 a_{i-1}^2 + \beta_1 \sigma_{i-1}^2. \tag{3.5}$$

Now the joint likelihood of our model becomes,

$$L(\theta, \Psi_T | \mathcal{F}_T) = L(\theta | \mathcal{F}_T, \Psi_T) f(\Psi_T), \tag{3.6}$$

where $L(\theta|\mathcal{F}_T, \Psi_T)$ is given by equation (3.3), and $f(\Psi_T)$ is the likelihood for the AR-GARCH random effect. Our goal is to obtain the MLE using the marginal likelihood which is given by the integration of the joint likelihood over the random effect ψ,

$$L(\theta|\mathcal{F}_T) = \int L(\theta|\mathcal{F}_T, \Psi_T) f(\Psi_T) d\Psi_T. \tag{3.7}$$

There is great difficulty to evaluate above integral as generally the GARCH model can only be derived recursively. In this paper, we will estimate this likelihood by approximating the integral with the Monte Carlo method and employing the concept of the importance sampling.

The Monte Carlo estimation involves first simulating multiple paths of the random effect by using the formula of the GARCH model and then taking the average of the evaluated joint likelihood at at each simulated path. The simulation step requires us to have a

reliable estimation of the the GARCH distribution which can not be obtained analytically. To overcome this difficulty, we employ the importance sampling technique and use the multivariate normal distribution to approximate the GARCH distribution. The following sections will provide details on the importance sampling and Monte Carlo estimation.

3.2 Model Estimation

In this section, we first discuss the procedures necessary to maximize the joint likelihood over the random effect. This provide us the maximizer ψ_{max} and the hessian matrix needed for the importance sampling. Then we gives more detail on how the importance sampling can facilitate the Monte Carlo simulation by using the multivariate normal distribution to substitute the complicated GARCH distribution. The last part is to specify the details of the Monte Carlo estimation and list the important formula.

3.2.1 The Importance Sampling Approximation

The marginal likelihood is give by equation (3.7). The integral is over the distribution of the GARCH(1,1) model which is hard to simulate when we use the Monte Carlo approximation. One way to solve this problem is to use a widely used distribution that is close to our multivariate GARCH distribution. Here we employ the multivariate normal distribution.

$$\begin{aligned} L(\theta|\mathcal{F}_T) &= \int L(\theta|\mathcal{F}_T, \Psi_T) f(\Psi_T) d\Psi_T \\ &= \int L(\theta|\mathcal{F}_T, \Psi_T) \frac{f(\Psi_T)}{G(\Psi_T)} G(\Psi_T) d\Psi_T \\ &= \int \frac{L(\theta, \Psi_T|\mathcal{F}_T)}{G(\Psi_T)} G(\Psi_T) d\Psi_T, \end{aligned}$$

where $G(\Psi_T)$ is the probability density function of the multivariate normal distribution evaluated at Ψ_T. Now the integral becomes the expectation of the ratio $L(\theta, \Psi_T|\mathcal{F}_T)/G(\Psi_T)$ with respect to the multivariate normal density which is much easier to generate. However, we must be careful in choosing the mean and covariance matrix for the normal distribution to make sure the ratio will not become extreme numbers and make the Monte Carlo approximation inefficient. Here we do a little algebraic transformation to the ratio,

$$\begin{aligned} \frac{L(\theta, \Psi_T|\mathcal{F}_T)}{G(\Psi_T)} &= \exp\Big(\log(L(\theta, \Psi_T|\mathcal{F}_T)) - \log(G(\Psi_T)) \Big) \\ &= \exp\Big(l(\theta, \Psi_T|\mathcal{F}_T) - \log(G(\Psi_T)) \Big), \end{aligned}$$

and then maximize the joint log-likelihood $l(\theta, \Psi_T|\mathcal{F}_T)$ over Ψ_T. Let ψ_{max} denote the maximizer,

$$\psi_{max} = \arg\max_{\Psi_T} l(\theta, \Psi_T|\mathcal{F}_T).$$

Now we want the first and second order derivatives of $l(\theta, \Psi_T|\mathcal{F}_T)$ and $\log(G(\Psi_T))$ to be the same, which will help to reduce the chance of extreme values. Therefore, the mean μ and covariance matrix Σ of the multivariate normal distribution will be defined by,

$$\begin{aligned} \mu &= \psi_{max}, \\ \Sigma &= -H(\theta, \psi_{max}|\mathcal{F}_T)^{-1}, \end{aligned} \tag{3.8}$$

where $H(\theta, \psi_{max}|\mathcal{F}_T)$ is the hessian matrix of the joint log-likelihood evaluated at ψ_{max}. Equation (3.8) can be seen by taking the second partial derivative on $\log(G(\Psi_T))$. The following paragraphs will provide more details on the maximization over ψ.

The likelihood of the GARCH(1,1) model can be evaluated with the following formula,

$$\begin{aligned} f(\Psi_T) &= f(\psi_1, \psi_2, \cdots, \psi_T) \\ &= f(\psi_T|\psi_1, \cdots, \psi_{T-1}) f(\psi_{T-1}|\psi_1, \cdots, \psi_{T-2}) \cdots f(\psi_1) \\ &= f(\psi_T|\psi_{T-1}) f(\psi_{T-1}|\psi_{T-2}) \cdots f(\psi_1) \\ &= \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{a_i^2}{2\sigma_i^2} \right), \end{aligned}$$

where $a_i = \psi_i - \rho\psi_{i-1}$.

With the help of the Bayes formula, we can write the joint log-likelihood as,

$$\begin{aligned} & l(\theta, \Psi_T|\mathcal{F}_T) = \log L(\theta, \Psi_T|\mathcal{F}_T) = \log\left(L(\theta|\mathcal{F}_T, \Psi_T) f(\Psi_T) \right) \\ = & \sum_{i=1}^{T} \sum_{k=1}^{K} \sum_{s \in S} \left(Y_{sk}(i) \log\{\lambda_{sk}(i)\} - R_{sk}(i)\lambda_{sk}(i) \right) - T\log(\sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^{T} \left(\log\sigma_i^2 + \frac{a_i^2}{\sigma_i^2} \right) \end{aligned} \tag{3.9}$$

In the optimization process, it is usually more efficient to use analytical gradient and hessian matrix than to use the finite difference method. In our case, this is especially true as there

are dozens of parameters in the model. The gradient can be computed by,

$$\begin{aligned}
&\frac{\partial}{\partial\psi_j} l(\theta, \Psi_T|\mathcal{F}_T) \\
= &\sum_{i=1}^{T}\sum_{k=1}^{K}\sum_{s\in S}\left(\frac{Y_{sk}(i)}{\lambda_{sk}(i)}\frac{\partial\lambda_{sk}(i)}{\partial\psi_j} - R_{sk}(i)\frac{\partial\lambda_{sk}(i)}{\partial\psi_j}\right) - \frac{1}{2}\sum_{i=1}^{T}\left(\frac{1}{\sigma_i^2}\frac{\partial\sigma_i^2}{\partial\psi_j} + \frac{2a_i}{\sigma_i^2}\frac{\partial a_i}{\partial\psi_j} - \frac{a_i^2}{\sigma_i^4}\frac{\partial\sigma_i^2}{\partial\psi_j}\right) \\
= &\sum_{k=1}^{K}\sum_{s\in S}\alpha_s\Big(Y_{sk}(j) - R_{sk}(j)\lambda_{sk}(j)\Big) - \frac{1}{2}\sum_{i=1}^{T}\left(\frac{(\sigma_i^2 - a_i^2)}{\sigma_i^4}\frac{\partial\sigma_i^2}{\partial\psi_j} + \frac{2a_i}{\sigma_i^2}\frac{\partial a_i}{\partial\psi_j}\right). \qquad (3.10)
\end{aligned}$$

From the GARCH(1,1) model,

$$a_i = \psi_i - \rho\psi_{i-1}, \qquad a_i = \sigma_i\varepsilon_i, \qquad \sigma_i^2 = \alpha_0 + \alpha_1 a_{i-1}^2 + \beta_1\sigma_{i-1}^2,$$

we have,

$$\frac{\partial\sigma_i^2}{\partial\psi_j} = 2\alpha_1 a_{i-1}\frac{\partial a_{i-1}}{\partial\psi_j} + \beta_1\frac{\partial\sigma_{i-1}^2}{\partial\psi_j}, \qquad \frac{\partial a_i}{\partial\psi_j} = \delta_{i,j} - \rho\delta_{i-1,j}, \qquad (3.11)$$

where $\delta_{i,j} = 1$ if $i = j$, and is 0 if otherwise. Thus the gradient of the joint log-likelihood can be computed recursively using (3.10) and (3.11).

The hessian matrix of the joint log-likelihood relative to ψ can be calculated using the following formula. When $j \neq m$, the non-diagonal elements $H(j,m)$ of the hessian matrix are,

$$\begin{aligned}
&\frac{\partial^2}{\partial\psi_j\partial\psi_m} l(\theta, \Psi_T|\mathcal{F}_T) \\
= &-\frac{1}{2}\sum_i\left\{\frac{2a_i^2 - \sigma_i^2}{\sigma_i^6}\left(\frac{\partial\sigma_i^2}{\partial\psi_j}\frac{\partial\sigma_i^2}{\partial\psi_m}\right) - \frac{2a_i}{\sigma_i^4}\left(\frac{\partial\sigma_i^2}{\partial\psi_j}\frac{\partial a_i}{\partial\psi_m} + \frac{\partial\sigma_i^2}{\partial\psi_m}\frac{\partial a_i}{\partial\psi_j}\right) + \frac{2}{\sigma_i^2}\left(\frac{\partial a_i}{\partial\psi_j}\frac{\partial a_i}{\partial\psi_m}\right)\right. \\
&\left. + \frac{\sigma_i^2 - a_i^2}{\sigma_i^4}\left(\frac{\partial^2\sigma_i^2}{\partial\psi_j\partial\psi_m}\right) + \frac{2a_i}{\sigma_i^2}\left(\frac{\partial^2 a_i}{\partial\psi_j\partial\psi_m}\right)\right\}. \qquad (3.12)
\end{aligned}$$

When $j = m$, the diagonal elements $H(j,j)$ of the hessian matrix are,

$$\begin{aligned}
&\frac{\partial^2}{\partial\psi_j^2} l(\theta, \Psi_T|\mathcal{F}_T) \\
= &-\frac{1}{2}\sum_i\left\{\frac{2a_i^2 - \sigma_i^2}{\sigma_i^6}\left(\frac{\partial\sigma_i^2}{\partial\psi_j}\right)^2 - \frac{4a_i}{\sigma_i^4}\left(\frac{\partial\sigma_i^2}{\partial\psi_j}\frac{\partial a_i}{\partial\psi_j}\right) + \frac{2}{\sigma_i^2}\left(\frac{\partial a_i}{\partial\psi_j}\right)^2 + \frac{\sigma_i^2 - a_i^2}{\sigma_i^4}\left(\frac{\partial^2\sigma_i^2}{\partial\psi_j^2}\right)\right. \\
&\left. + \frac{2a_i}{\sigma_i^2}\left(\frac{\partial^2 a_i}{\partial\psi_j^2}\right)\right\} - \sum_{k=1}^{K}\sum_{s\in S}\Big(\alpha_s^2 R_{sk}(j)\lambda_{sk}(j)\Big). \qquad (3.13)
\end{aligned}$$

From the GARCH model, we have,

$$\begin{aligned}
\frac{\partial^2 \sigma_i^2}{\partial \psi_j \partial \psi_m} &= 2\alpha_1 \frac{\partial a_{i-1}}{\partial \psi_j}\frac{\partial a_{i-1}}{\partial \psi_m} + 2\alpha_1 a_{i-1}\frac{\partial^2 a_{i-1}}{\partial \psi_j \partial \psi_m} + \beta_1 \frac{\partial^2 \sigma_{i-1}^2}{\partial \psi_j \partial \psi_m} \\
&= 2\alpha_1 \frac{\partial a_{i-1}}{\partial \psi_j}\frac{\partial a_{i-1}}{\partial \psi_m} + \beta_1 \frac{\partial^2 \sigma_{i-1}^2}{\partial \psi_j \partial \psi_m} \qquad (3.14) \\
\frac{\partial^2 a_i}{\partial \psi_j \partial \psi_m} &= 0. \qquad (3.15)
\end{aligned}$$

Now, the hessian matrix can also be computed in an iterative manner by using equations (3.12), (3.13), (3.14), and (3.15).

After the maximization, we obtain the mean and covariance matrix of the multivariate normal distribution and can use them to generate M random samples, $(\psi^1, \psi^2, \cdots, \psi^M)$. The approximated marginal likelihood will then becomes,

$$\begin{aligned}
\hat{L}(\theta|\mathcal{F}_T) &= \frac{1}{M}\sum_{m=1}^{M} \frac{L(\theta, \psi^m|\mathcal{F}_T)}{G(\psi^m)} \\
&= \frac{1}{M}\sum_{m=1}^{M} \exp(\log(L(\theta, \psi^m|\mathcal{F}_T)) - \log(G(\psi^m))) \\
&= \frac{1}{M}\sum_{m=1}^{M} \exp(l(\theta, \psi^m|\mathcal{F}_T) - \log(G(\psi^m))) \\
&= \frac{1}{M}\sum_{m=1}^{M} w_m(\theta) \qquad (3.16)
\end{aligned}$$

where $w_m(\theta) = \exp(l(\theta, \psi^m|\mathcal{F}_T) - \log(G(\psi^m)))$.

3.2.2 Maximize the Monte Carlo Marginal Likelihood over parameter θ

Our objective is to maximize the approximated marginal likelihood (3.16). In order to have the importance sampling approximation work well, we need to restrict the maximizer θ_{max} in a small neighborhood of the initial value θ_0 during the optimization process. It is possible to use constraint optimization which restrict the magnitude of the steps that the optimization routine can walk away from the initial value θ_0. In our model estimation, we take a different approach by introducing a penalized function,

$$p(\theta, \theta_0) = \hat{\delta}(\theta, \theta_0) - c \cdot s(\theta, \theta_0), \qquad (3.17)$$

where $\hat{\delta}(\theta, \theta_0)$ is the difference of the approximated marginal likelihood evaluated at θ relative to the likelihood at θ_0,

$$\hat{\delta}(\theta, \theta_0) \quad = \quad \hat{L}(\theta|\mathcal{F}_T) - \hat{L}(\theta_0|\mathcal{F}_T) = \frac{1}{M}\sum_{m=1}^{M}\Big(w_m(\theta) - w_m(\theta_0)\Big),$$

c is predetermined constant, and $s(\theta, \theta_0)$ is the standard error of $\hat{\delta}(\theta, \theta_0)$,

$$s(\theta, \theta_0) = \sqrt{\frac{1}{M(M-1)}\sum_{m=1}^{M}\Big(w_m(\theta) - w_m(\theta_0)\Big)^2 - \frac{1}{M^2(M-1)}\Big\{\sum_{m=1}^{M}\Big(w_m(\theta) - w_m(\theta_0)\Big)\Big\}^2}.$$

This form of penalized function comes from the concept of the usual confidence interval between the pair test of two means. When M is large, the mean functions will close to normally distributed and the penalized function acts like the lower bound of the confidence interval. If the optimization result gives a positive penalized function, we will know that there is improvement of the approximated marginal likelihood and the above steps will be repeated until the maximum is achieved. Also, the penalty make sure that at any time the optimization routine picks up a parameter θ that is far away from θ_0, we will have larger standard errors and smaller penalized function values, which will force the program to go back and look for an alternate θ. Therefore, we conclude that this penalized function has better statistical meaning than just doing restricted optimization.

3.2.3 Hotelling's T^2, Monte Carlo Sample Size M, and the Penalty Constant c

Recall that in (3.17) we define the penalized likelihood as

$$p(\theta, \theta_0) \quad = \quad \delta(\theta, \theta_0) - c\, s(\theta, \theta_0),$$

where δ is the sample mean of $u_m \equiv w_m(\theta) - w_m(\theta_0)$, $m = 1, \ldots, M$, and $s = \sqrt{V/M}$ where V is the sample variance of $u_1, \ldots, u_M$. For $\theta \approx \theta_0$, we have

$$u_m \approx (\theta - \theta_0)' \frac{\partial w_m(\theta_0)}{\partial \theta},$$

which leads to

$$\delta(\theta, \theta_0) \approx \Delta' \bar{x} \quad \text{and} \quad s^2(\theta, \theta_0) \approx \frac{1}{M}\Delta' \Sigma \Delta,$$

where $\Delta \equiv \theta - \theta_0$, and $\bar{x}$ and Σ are the sample mean vector and sample covariance matrix of the M vectors $\dfrac{\partial w_m(\theta_0)}{\partial \theta}$, $m = 1, \ldots, M$. Thus we have the approximation

$$p(\theta, \theta_0) \approx \Delta' \bar{x} - c\sqrt{\frac{1}{M}\Delta' \Sigma \Delta}$$

when $\Delta = \theta - \theta_0$ is small.

Note that this approximation may be rewritten as

$$p(\theta, \theta_0) \approx \sqrt{\frac{1}{M}\Delta' \Sigma \Delta}\left(\frac{\Delta' \bar{x}}{\sqrt{\frac{1}{M}\Delta' \Sigma \Delta}} - c\right)$$

There is a standard calculation in multivariate analysis (see Mardia, Kent and Bibby (1979), Corollary A.9.2.2, page 480) that shows that

$$\sup_{\Delta} \frac{(\Delta' \bar{x})^2}{\frac{1}{M}\Delta' \Sigma \Delta} \;=\; M\, \bar{x}' \Sigma^{-1} \bar{x}\,.$$

For reasons mentioned above we will define

$$T^2 = M\, \bar{x}' \Sigma^{-1} \bar{x},$$

and take T to be the positive square root of T^2. Then we have shown that for θ near θ_0 we approximately have

$$p(\theta, \theta_0) \leq (T - c)\sqrt{\frac{1}{M}(\theta - \theta_0)' \Sigma (\theta - \theta_0)}\,.$$

If $T < c$, this implies that the penalized likelihood is actually maximized (at least locally) when $\theta = \theta_0$. Thus, if we start the optimization process by taking a small step away from θ_0, we will end up returning to θ_0 as the optimization progresses. This means we must choose M large enough so that $T > c$ (or equivalently $T^2 > c^2$). That is, with a particular choice of the penalty constant c, we must accumulate enough vectors $\psi_1, \psi_2, \psi_3, \ldots$ to guarantee that $T^2 > c^2$ before beginning the optimization.

The quantity $M\, \bar{x}' \Sigma^{-1} \bar{x}$ has the form of the Hotelling's T^2 statistic that would be used for testing $H_0 : \mu = 0$ where μ denotes the population mean vector of the i.i.d. random vectors $\dfrac{\partial w_m(\theta_0)}{\partial \theta}$. This population mean is the limiting value of

$$(1/M) \sum_{m=1}^{M} \frac{\partial w_m(\theta_0)}{\partial \theta}$$

which is the true gradient of the marginal likelihood function at θ_0. Thus, if c is chosen to be a reasonable critical value for the Hotelling's T^2 test, then choosing M large enough so that $T^2 > c^2$ essentially amounts to accumulating evidence until we are reasonably sure that the true gradient of the marginal likelihood is nonzero. In the classical situation with i.i.d. samples of vectors drawn from a multivariate normal population, the distribution of T^2 under H_0 is asymptotically distributed as χ^2_p where p is the dimension of the space. Thus, since the mean and variance of χ^2_p are p and $2p$, respectively, a critical value roughly two standard deviations above the mean of T^2 is given by $c^2 = p + 2\sqrt{2p}$. With p set equal to the total number of parameters in the model (which is the dimension of $\frac{\partial w_m(\theta_0)}{\partial \theta}$), we will used the square root c defined above as the penalty constant in the optimization.

3.2.4 The Penalty on Weights

Recall that in (3.16) the approximation to the marginal likelihood has the form

$$L(\theta) = \frac{1}{M} \sum_{m=1}^{M} w_m(\theta),$$

where $w_m(\theta) = \exp(\ell(\theta, \psi^m) - \log(G(\psi^m)))$. Intuitively, this approximation L seems likely to be inaccurate when the weights w_m are highly variable. Define the relative weights

$$q_m = \frac{w_m}{L},$$

which have a mean of 1. During the optimization it is sometimes observed that $\max q_m$ becomes very large so that the summation in L is dominated by one term (or perhaps by a small number of terms). By introducing an appropriate penalty, we hope to stop this from happening, either by causing the optimization to terminate when this becomes a problem or by causing the optimization to avoid those regions in the parameter space where this happens.

If φ is any strictly convex function, then

$$Q = \sum_{m=1}^{M} \varphi(q_m)$$

is a potential penalty for the weights since Q is minimized when the relative weights q_m are all equal, and increases as they become more unequal. (The function Q is a Schur-convex

function of $q_1, q_2, \ldots, q_M$.) Two simple choices of strictly convex functions are $\varphi(u) = u^2$ and $\varphi(u) = -\log(u)$. Combining this weight penalty with the earlier penalty yields a penalized likelihood of the form

$$p(\theta, \theta_0) = \delta(\theta, \theta_0) - c_1\, s(\theta, \theta_0) - c_2\, Q(\theta)$$

where c_1 and c_2 are positive constants which will be chosen empirically. More specifically, the value of c_1 is defined in the previous section and depends on the number of parameters.

In this thesis, the most analytically convenient choice is $\varphi(u) = -\log(u)$ leading to the penalty

$$\begin{aligned} Q(\theta) &= -\frac{1}{M}\sum_{m=1}^{M} \log q_m \\ &= -\frac{1}{M}\sum_{m=1}^{M} \log w_m + \log L\,. \end{aligned}$$

This is the penalty currently used in our program. This penalty has gradient

$$\frac{\partial Q}{\partial \theta} = -\frac{1}{M}\sum_{m=1}^{M} g_m + \frac{1}{L}\left(\frac{1}{M}\sum_{m=1}^{M} w_m g_m\right), \quad \text{where } g_m(\theta) = \frac{\partial \ell(\theta, \psi^m)}{\partial \theta}$$

consisting of quantities which were already being computed in the program.

3.2.5 The Derivatives and Hessian Matrix

To perform the optimization, we will need to compute the derivative of the penalized function $p(\theta, \theta_0)$,

$$\begin{aligned} \frac{\partial}{\partial \theta} p(\theta, \theta_0) &= \frac{\partial}{\partial \theta}\left(\hat{\delta}(\theta, \theta_0) - c \cdot s(\theta, \theta_0)\right) \\ &= \frac{1}{M}\sum_{m=1}^{M} \frac{\partial}{\partial \theta} w_m(\theta) - \frac{c}{s(\theta, \theta_0)}\left\{\frac{1}{M(M-1)}\sum_{m=1}^{M}\left(w_m(\theta) - w_m(\theta_0)\right)\frac{\partial}{\partial \theta} w_m(\theta)\right. \\ &\qquad \left. -\frac{1}{M^2(M-1)}\left(\sum_{m=1}^{M} w_m(\theta) - w_m(\theta_0)\right)\left(\sum_{m=1}^{M}\frac{\partial}{\partial \theta} w_m(\theta)\right)\right\}, \end{aligned}$$

where $\frac{\partial}{\partial \theta} w_m(\theta) = w_m(\theta)\frac{\partial}{\partial \theta} l(\theta, \psi_m | \mathcal{F}_T)$. The gradient of the joint log-likelihood can be calculated by taking the partial derivatives relative to the parameters specified in the

proportional intensity functions, i.e. η_s, γ_s, and α_s.

$$\begin{aligned}
\frac{\partial}{\partial \eta_s} l(\theta, \Psi_T | \mathcal{F}_T) &= \sum_{i=1}^{T} \sum_{k=1}^{K} \left(\frac{Y_{sk}(i)}{\lambda_{sk}(i)} \frac{\partial \lambda_{sk}(i)}{\partial \eta_s} - R_{sk}(i) \frac{\partial \lambda_{sk}(i)}{\partial \eta_s} \right) \\
&= \sum_{i=1}^{T} \sum_{k=1}^{K} \Big(Y_{sk}(i) - R_{sk}(i) \lambda_{sk}(i) \Big).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \gamma_s} l(\theta, \Psi_T | \mathcal{F}_T) &= \sum_{i=1}^{T} \sum_{k=1}^{K} \left(\frac{Y_{sk}(i)}{\lambda_{sk}(i)} \frac{\partial \lambda_{sk}(i)}{\partial \gamma_s} - R_{sk}(i) \frac{\partial \lambda_{sk}(i)}{\partial \gamma_s} \right) \\
&= \sum_{i=1}^{T} \sum_{k=1}^{K} \omega_k(i) \Big(Y_{sk}(i) - R_{sk}(i) \lambda_{sk}(i) \Big).
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \alpha_s} l(\theta, \Psi_T | \mathcal{F}_T) &= \sum_{i=1}^{T} \sum_{k=1}^{K} \left(\frac{Y_{sk}(i)}{\lambda_{sk}(i)} \frac{\partial \lambda_{sk}(i)}{\partial \alpha_s} - R_{sk}(i) \frac{\partial \lambda_{sk}(i)}{\partial \alpha_s} \right) \\
&= \sum_{i=1}^{T} \sum_{k=1}^{K} \psi(i) \Big(Y_{sk}(i) - R_{sk}(i) \lambda_{sk}(i) \Big).
\end{aligned}$$

Then let's take a look at the partial derivative over parameters specified in the GARCH(1,1) model, i.e. ρ, α_0, α_1, and β_1. Similarly, we have,

$$\begin{aligned}
\frac{\partial}{\partial \rho} l(\theta, \Psi_T | \mathcal{F}_T) &= -\frac{1}{2} \sum_{i=1}^{T} \left(\frac{\left(\sigma_i^2 - a_i^2\right)}{\sigma_i^4} \frac{\partial \sigma_i^2}{\partial \rho} + \frac{2a_i}{\sigma_i^2} \frac{\partial a_i}{\partial \rho} \right) \\
\frac{\partial}{\partial \alpha_0} l(\theta, \Psi_T | \mathcal{F}_T) &= -\frac{1}{2} \sum_{i=1}^{T} \left(\frac{\left(\sigma_i^2 - a_i^2\right)}{\sigma_i^4} \frac{\partial \sigma_i^2}{\partial \alpha_0} + \frac{2a_i}{\sigma_i^2} \frac{\partial a_i}{\partial \alpha_0} \right). \\
\frac{\partial}{\partial \alpha_1} l(\theta, \Psi_T | \mathcal{F}_T) &= -\frac{1}{2} \sum_{i=1}^{T} \left(\frac{\left(\sigma_i^2 - a_i^2\right)}{\sigma_i^4} \frac{\partial \sigma_i^2}{\partial \alpha_1} + \frac{2a_i}{\sigma_i^2} \frac{\partial a_i}{\partial \alpha_1} \right). \\
\frac{\partial}{\partial \beta_1} l(\theta, \Psi_T | \mathcal{F}_T) &= -\frac{1}{2} \sum_{i=1}^{T} \left(\frac{\left(\sigma_i^2 - a_i^2\right)}{\sigma_i^4} \frac{\partial \sigma_i^2}{\partial \beta_1} + \frac{2a_i}{\sigma_i^2} \frac{\partial a_i}{\partial \beta_1} \right).
\end{aligned}$$

From the GARCH(1,1) model,

$$a_i = \psi_i - \rho \psi_{i-1}, \qquad a_i = \sigma_i \varepsilon_i, \qquad \sigma_i^2 = \alpha_0 + \alpha_1 a_{i-1}^2 + \beta_1 \sigma_{i-1}^2,$$

we have,

$$\begin{aligned}
\frac{\partial \sigma_i^2}{\partial \rho} &= 2\alpha_1 a_{i-1}\frac{\partial a_{i-1}}{\partial \rho} + \beta_1 \frac{\partial \sigma_{i-1}^2}{\partial \rho}, \qquad \frac{\partial a_i}{\partial \rho} = -\psi_{i-1}; \\
\frac{\partial \sigma_i^2}{\partial \alpha_0} &= 1 + \beta_1 \frac{\partial \sigma_{i-1}^2}{\partial \alpha_0}, \qquad \frac{\partial a_i}{\partial \alpha_0} = 0; \\
\frac{\partial \sigma_i^2}{\partial \alpha_1} &= a_{i-1}^2 + \beta_1 \frac{\partial \sigma_{i-1}^2}{\partial \alpha_1}, \qquad \frac{\partial a_i}{\partial \alpha_1} = 0; \\
\frac{\partial \sigma_i^2}{\partial \beta_1} &= \sigma_{i-1}^2 + \beta_1 \frac{\partial \sigma_{i-1}^2}{\partial \beta_1}, \qquad \frac{\partial a_i}{\partial \beta_1} = 0.
\end{aligned}$$

Therefore the partial derivatives of the penalized function can be computed recursively using above formulas.

CHAPTER 4

SIMULATION STUDY

In this section, we first introduce the notations that are necessary to explain the simulation procedure and results. In our model, the joint log-likelihood function is specified by equation (3.9). We have also given the details about the discretized intensity used in the joint log-likelihood,

$$\lambda_{sk}(i) = \exp\Big(\eta_s + \gamma_s'\omega_k(i) + \alpha_s\psi(i)\Big). \tag{4.1}$$

In the intensity function (4.1), there are three sets of parameters η_s, γ_s, α_s. They share the common subscript s which represent different rating transitions. We will give more details about the ratings and rating transitions below. The data we use include both the agency credit rating data and bond implied rating data from Moody's Investors Service. The Moody's rating system for long term debts of companies has nine major rating categories (Table 1.2), {Aaa, Aa, A, Baa, Ba, B, Caa, Ca, C}. There are also finer rating categories like Aa1, Aa2, Aa3, etc. which are nested in the major ratings. However, studying the finer ratings will lead us to too many rating transitions and much more complicated transition structure, and they are not considered in the current study. For the major ratings, we will label them with numbers such that the highest rating corresponds to 1 and the lowest rating to 9.

Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
1	2	3	4	5	6	7	8	9

Excluding the same-rating movements, i.e. no transitions, the 9 major rating categories lead to 72 different types of transitions. They can be organized in the matrix form (table 4.1).

Table 4.1: All possible transitions

	Aaa (1)	Aa (2)	A (3)	Baa (4)	Ba (5)	B (6)	Caa (7)	Ca (8)	C (9)
Aaa (1)	*	T_{12}	T_{13}	T_{14}	T_{15}	T_{16}	T_{17}	T_{18}	T_{19}
Aa (2)	T_{21}	*	T_{23}	T_{24}	T_{25}	T_{26}	T_{27}	T_{28}	T_{29}
A (3)	T_{31}	T_{32}	*	T_{34}	T_{35}	T_{36}	T_{37}	T_{38}	T_{39}
Baa (4)	T_{41}	T_{42}	T_{43}	*	T_{45}	T_{46}	T_{47}	T_{48}	T_{49}
Ba (5)	T_{51}	T_{52}	T_{53}	T_{54}	*	T_{56}	T_{57}	T_{58}	T_{59}
B (6)	T_{61}	T_{62}	T_{63}	T_{64}	T_{65}	*	T_{67}	T_{68}	T_{69}
Caa (7)	T_{71}	T_{72}	T_{73}	T_{74}	T_{75}	T_{76}	*	T_{78}	T_{79}
Ca (8)	T_{81}	T_{82}	T_{83}	T_{84}	T_{85}	T_{86}	T_{87}	*	T_{89}
C (9)	T_{91}	T_{92}	T_{93}	T_{94}	T_{95}	T_{96}	T_{97}	T_{98}	*

Each element $T_{ij}, i, j = 1, \cdots, 9, i \neq j$ represents a transition from rating i at the beginning of the period to rating j at the end of the period. Here the period could mean one year, one quarter, one month, or even one day, depending on the context of the data. In my dissertation, daily transitions are studied. Looking at the above matrix, the diagonal elements are represented by asterisks to highlight the fact that there are no parameters for same-rating transitions in the model as these parameters will not be needed. This matrix shows that there are transitions between all possible pairs of different ratings. However, in reality, there are rarely "large jumps" in rating transitions. Figure 4.1 and 4.2 show us the empirical transitions frequencies for the agency ratings and market implied ratings, respectively.

From figure 4.1, which shows the agency rating transition frequencies, the majority of the transitions will not move more than three categories away. For market implied ratings, figure 4.2 shows that the transitions tend to move no more than four categories away. Since our simulation study will generate data that should bear the characteristics of the real data, we believe it is necessary to restrict the number of the transition parameters such that they will follow the same pattern of the real transition frequencies. This has another advantage of cutting down the number of parameters that need to be estimated and reducing the computation load. Our simulation study will force all three intensity parameters η_s, γ_s, and

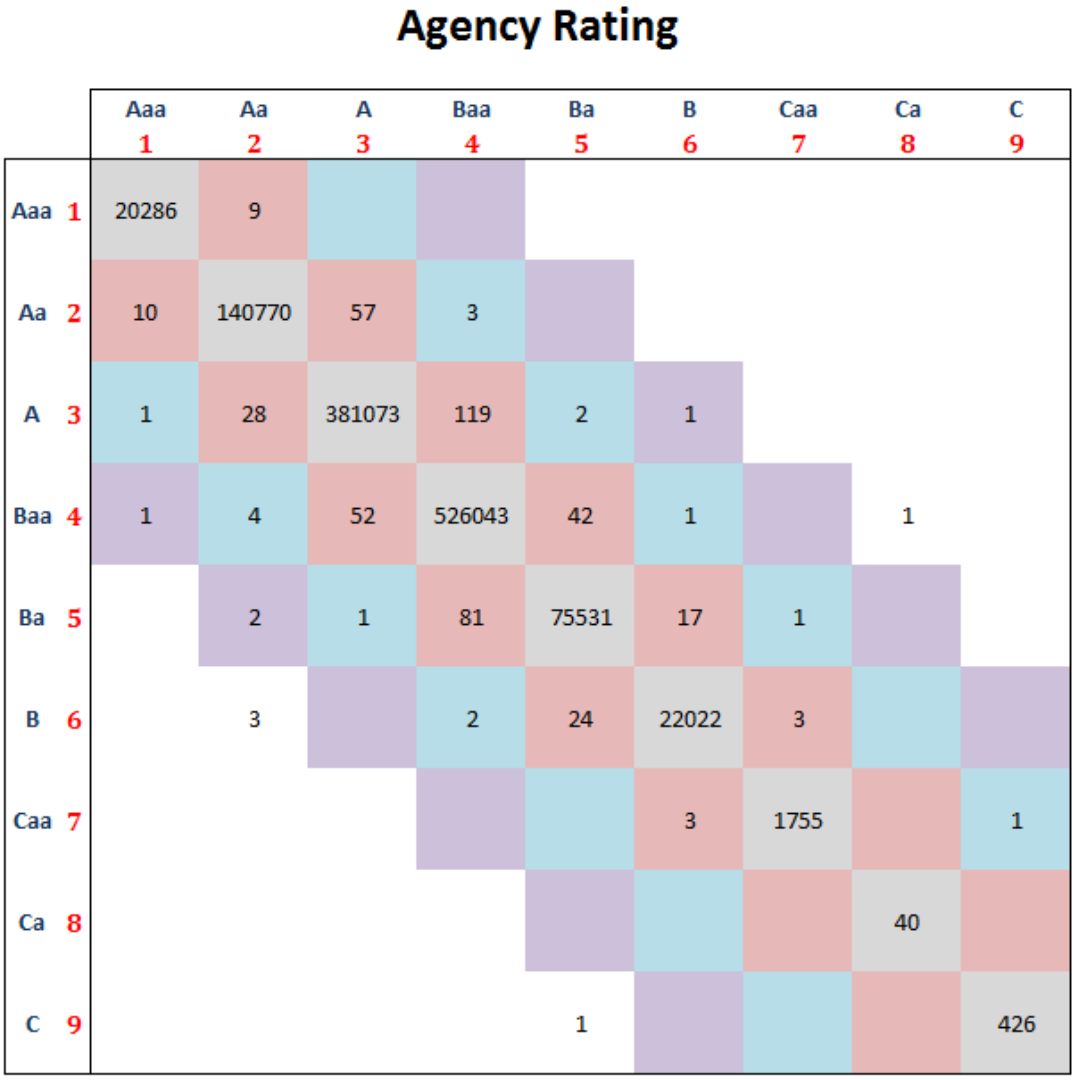

Figure 4.1: Agency Credit Rating: the no-transition movements are also included.

α_s to follow a common pattern matrix $\mathbf{X}$,

$$\mathbf{X} = \begin{pmatrix} * & d_{12} & d_{13} & d_{14} & * & * & * & * & * \\ d_{21} & * & d_{23} & d_{24} & d_{25} & * & * & * & * \\ d_{31} & d_{32} & * & d_{34} & d_{35} & d_{36} & * & * & * \\ d_{41} & d_{42} & d_{43} & * & d_{45} & d_{46} & d_{47} & * & * \\ * & d_{52} & d_{53} & d_{54} & * & d_{56} & d_{57} & d_{58} & * \\ * & * & d_{63} & d_{64} & d_{65} & * & d_{67} & d_{68} & d_{69} \\ * & * & * & d_{74} & d_{75} & d_{76} & * & d_{78} & d_{79} \\ * & * & * & * & d_{85} & d_{86} & d_{87} & * & d_{89} \\ * & * & * & * & * & d_{96} & d_{97} & d_{98} & * \end{pmatrix}, \tag{4.2}$$

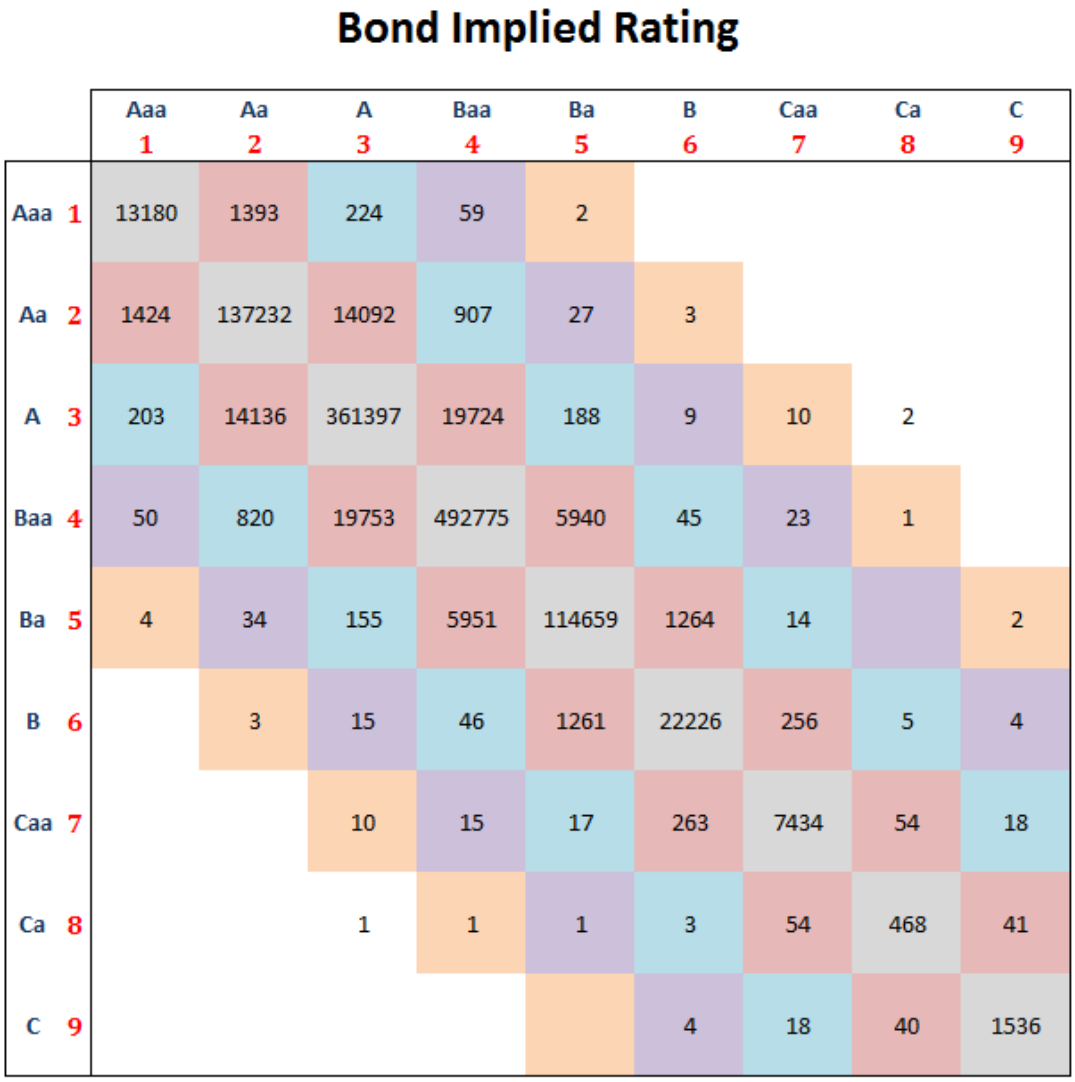

Figure 4.2: Bond Implied Rating: the no-transition movements are also included.

where $d_{ij}, i, j = 1, \cdots, 9$ represent allowed transitions and asterisks represents disallowed transition types. This pattern matrix will limit the distance of the potential transition movements.

Another practical issue could arise when we apply the estimation procedure on the simulated or the application datasets. Since all the transitions are guided by the probabilities, some types of the transitions may never be observed in the data even if the corresponding probabilities is strictly positive. If a particular transition s' is never observed, the likelihood is

maximized when $\eta_{s'}$ is negative infinity which makes the intensity $\lambda_{s'k}(i)$ equal to zero for all k and i. That is, if a transition is never observed, the likelihood is maximized by making that transition impossible. Unfortunately, the value $\eta_{s'} = -\infty$ will be approached (if at all) only in the limit as the number of iterations of the estimation procedure goes to infinity. To avoid this difficulty, since we know in advance that the maximum likelihood estimate will make the transition s' impossible, we simply eliminate s' from the set of allowed transitions. This means that, when computing the maximum likelihood estimates for a particular data set, we replace the set S in equation (3.3) for the likelihood by the smaller set S^* of transitions which are actually observed in the data. In terms of the pattern matrix (4.2), we are replacing the parameters corresponding to the unobserved transitions by asterisks.

Besides the restricted number of transition types, we make another simplification to the transition intensity function by omitting the fixed covariate effect. Our model bases on Koopman's multistate latent-factor intensity model [10] by extending the time series structure of the random effect in the intensity function. The focus of the current simulation study is to capture the time-varying volatility effect within the transition processes and try to improve the precision of the model estimation, instead of determining the contribution of the covariate effect. Therefore, this modification does not detract from the main purpose of our simulation.

Now the intensity function becomes,

$$\lambda_s(i) = \exp\Big(\eta_s + \alpha_s \psi(i)\Big), \tag{4.3}$$

where η_s and α_s have the matrix form of (4.2). Comparing to equation (4.1), the above intensity (equation (4.3)) assumes that all the firms share the same rating transition intensity process.

After reviewing the modifications to the possible transition types and intensity functions, we can specify the details of the simulation process. The first step is to generate the simulated data for estimation. In this procedure, we need to randomly assign rating moves based on the current rating of the company and the probability transition matrix specified by the intensity function given the true parameters η_s and α_s. The necessary steps are detailed below.

The first step is to generate AR(1)-GARCH(1,1) time series random effects $\psi(i), i =$

$1, \cdots, T$ by recursively applying the formula,

$$\begin{aligned} \psi(i) &= \rho_s \psi(i-1) + a(i), \\ a(i) &= \sigma(i)\epsilon(i), \\ \sigma(i)^2 &= \alpha_0 + \alpha_1 a(i-1)^2 + \beta_1 \sigma(i-1)^2, \end{aligned}$$

where $\epsilon(i)$ are iid $N(0,1)$ and the initial values of the time series are set as, $\sigma(1) = \sqrt{\alpha_0}$ and $\psi(1) = a(1)$.

After obtaining the random effects, we will focus on generating the intensity functions. Before going into the details of this process, we want to point out the way we select the true and initial values for the transition intensity parameter matrix of η_s and α_s. Recall that they both follow the same pattern matrix as in (4.2). There are 42 parameter values in the pattern matrix. If η_s and α_s are different matrices, which is generally the case, there will be totally 84 parameters. We also need the parameters to follow certain patterns such that it is more likely to move to more adjacent ratings than far away ratings. It is possible to assign the 84 parameters randomly, but we decide to take a shortcut. Our simulation will choose the η_s and α_s parameters with the following pattern matrix,

$$\mathbf{X} = \begin{pmatrix} * & d_1 & d_2 & d_3 & * & * & * & * & * \\ d_1 & * & d_1 & d_2 & d_3 & * & * & * & * \\ d_2 & d_1 & * & d_1 & d_2 & d_3 & * & * & * \\ d_3 & d_2 & d_1 & * & d_1 & d_2 & d_3 & * & * \\ * & d_3 & d_2 & d_1 & * & d_1 & d_2 & d_3 & * \\ * & * & d_3 & d_2 & d_1 & * & d_1 & d_2 & d_3 \\ * & * & * & d_3 & d_2 & d_1 & * & d_1 & d_2 \\ * & * & * & * & d_3 & d_2 & d_1 & * & d_1 \\ * & * & * & * & * & d_3 & d_2 & d_1 & * \end{pmatrix}. \tag{4.4}$$

Now we only need to pick six parameter values by hand, i.e. one set of $\{d_1, d_2, d_3\}$ for η_s and another set for α_s. Remember that we still have 84 intensity function parameters in our simulation. We just set some of the true and initial values to be the same to reduce the hardship of having to choose many parameter values by hand. This assumption will not change the nature of our model.

Given the pattern matrix (equation (4.4)) and the true parameters described below, we can define η_s and α_s. The intensity function values for each company at every time point t will then be specified by using the formula,

$$\lambda_s(i) = \exp\Big(\eta_s + \alpha_s \psi(i)\Big). \tag{4.5}$$

Given the intensity functions above, our next task is to specify when rating transitions occur and their types. The following two-stage process is involved. For each company k at time t, all intensity functions of possible transition types for that company are aggregated to form a single intensity function. For example, at time t, let's assume company k bears the rating 2 (or AA before the conversion). Then at time $t+1$, the only possible transitions it could make are $2 \to 1$, $2 \to 3$, $2 \to 4$, or $2 \to 5$. We add up the intensity functions of these four transitions at time t. Our intention is to generate a Poisson process using this aggregated intensity function to assign the next rating movement. In the simulation, we take an approximately equivalent approach by looking at the interarrival times which should follow the exponential distribution. Given the intensity at the end of the previous day, we first determine the time to the next event. If it is larger than one day, we decide there is no transition in the current day and the end-of-day rating will be the same as that in the previous day. On the other hand, if the time to the next transition is less than one day, a transition should occur and we will go to the next stage to determine the transition type. In the second stage, a multinomial random number is generated to randomly assign the transition type according to the proportion of the intensity function to the sum of all at-risk intensity functions. This concludes the second stage of the rating assignment. The whole process will repeat for all companies and for each time point during the observation time period.

However, we should be careful with the above data generation procedure. By only considering the interarrival time, we are actually counting not only the probability of the first arrival, but the sum of probabilities of all possible arrivals that could happen in one day. This two quantities only approximately equal to each other if the values of the intensities are small. More specifically, given the simulated data, the likelihood will include a factor of $\frac{\lambda_s}{U}(1-e^{-U})$ for transitions, and a factor of e^{-U} if no transition occurs. Here the quantity U is defined to be $U = \sum_t \lambda_t$ where t represents all transitions at risk. Now if the sum U is relatively small, $\frac{\lambda_s}{U}(1-e^{-U})$ will be approximately equal to λ_s and this likelihood closely match the likelihood we proposed in chapter 3. However, this approximation will fail if U is not close to zero and the two likelihood will not match. Therefore, we need to carefully choose the true and initial parameters for the simulation such that they generate the intensities that are close to zero. We will revisit this when discussing the parameter setup.

4.1 The Practical Issues in the Simulation Study

4.1.1 The Computation Speed

We have encountered a speed problem at the early stage of the simulation study. The program run very slow even with relatively small dataset. In this section, we will analyze and propose a method to solve this problem. Our computer program have been implemented in both R statistical package and Matlab under the PC environment, while neither of them performs significantly better than the other in terms of the computation time.

We have carefully studied the speed problem and found several factors that may contribute to this situation. Recall that the importance sampling technique is used to approximate the integration calculation of the marginal likelihood,

$$L(\theta|\mathcal{F}_T) = \int L(\theta|\mathcal{F}_T, \Psi_T) f(\Psi_T) d\Psi_T,$$

with the sum of a Monte Carlo sample,

$$\hat{L}(\theta|\mathcal{F}_T) = \frac{1}{M} \sum_{m=1}^{M} \frac{L(\theta, \psi^m|\mathcal{F}_T)}{G(\psi^m)}.$$

Here the Monte Carlo sample includes a large number of simulated paths of the random effects over time which in our model is characterized by the AR-GARCH time series. Since the time series will change each time when we generate the simulated paths, the likelihood of the model has to be computed repeatedly using the updated time series random effect. The Monte Carlo sample size M is usually several thousands or tens of thousands. This means each time we want to evaluate the approximated marginal likelihood, it will be necessary to generate a Monte Carlo sample with thousands of simulated paths and compute the likelihood millions of times. Even when the data set size is relatively small as in the pilot study, the above process usually takes many hours to finish, if we do not make further restrictions.

Besides the computation of the approximated marginal likelihood, we also need to count in the time to compute the gradient which is necessary in performing the optimization. Usually the optimization function in R or Matlab could either take the user defined gradient or calculate an approximated gradient by applying the finite difference method to the objective function. In our case, the latter method is overwhelmingly time consuming because it requires thousands of marginal likelihood calculation which will translate into millions of

Monte Carlo sampling and approximations. Therefore we choose to provide the specific gradient function to the optimization routine. Since the gradient is defined based on the simulated Monte Carlo marginal likelihood, we are required to compute the gradient in a similar fashion to the approximated marginal likelihood, which again is time consuming.

Based on the above discussion, we will need to speed up the program as fast as possible while thinking of other reasonable ways to reduce the computation time. After carefully analyzing the programming code, we have switched from using explicit loops to fully vectorizing most of the program code. This helps us to use matrix manipulations which is much faster than the loop operations in both R and Matlab.

Besides vectorizing the codes, we have also collapsed the data into a condensed data structure that makes the data scanning faster while sacrificing some functionality. We will discuss this in the next paragraphs.

In the simulation, we merge the data from different companies into an aggregated dataset. The intensity functions will have the form as follows,

$$\lambda_s(i) = \exp\Big(\eta_s + \alpha_s \psi(i)\Big).$$

From this equation, it is easy to see that none of the components of the intensity function will depend on the subscript k. In other words, the intensity function will be the same across all firms, given that we observe at the same time. Therefore we can aggregate the data to help expedite the computation of the likelihood. This means in equation (3.3), the triple product is replaced by a double product in i and s in which $Y_{sk}(i)$ is replaced by $Y_{s\cdot}(i)$ and $R_{sk}(i)$ is replaced by $R_{s\cdot}(i)$ where the dot represents the summation over k,

$$\begin{aligned} L(\theta|\Psi_T, \mathcal{F}_T) \;=\; \prod_{i=1}^{T}\prod_{s\in S} \exp\Big(Y_{s\cdot}(i)\Big(\log(\eta_s + \gamma_s'\omega(i) + \alpha_s\psi(i)\Big) \\ - R_{s\cdot}(i)\exp\Big(\eta_s + \gamma_s'\omega(i) + \alpha_s\psi(i)\Big)\Big). \end{aligned} \tag{4.6}$$

Since the size of the aggregated data structure no longer depends the number of companies, we are granted the ability to handle much larger data sets than the ones in the pilot study without increasing the computation time by a noticeable amount. This means we will be able to employ many more companies while keeping the number of time points at a reasonable level. In the simulation study which is detailed in the following section, we manage to have hundreds of time points while increasing the number of companies ten folds

from fifty to five hundred. The time used for each of these datasets is almost independent of the number of companies.quite similar. Therefore it lends us great flexibility when testing our simulation programs and determining the precision of the simulation results.

Besides the data structure and computation speed issue, we have designed our own line search optimization program. At the beginning of the model implementation, the model estimation procedure uses standard optimization routine to obtain the maximum likelihood estimator. In R, we use the "optim" function. This function use the optimization method named L-BFGS-B, which is a classical quasi-Newton maximization method that allows the box type of constraints. The "BFGS" algorithm is named after four people, Broyden, Fletcher, Goldfarb, and Shanno, for their contribution of providing an update method for the approximated Hessian matrix. The prefix "L" represents the limit memory feature of the algorithm which suits particularly well for optimization problems with higher dimensions. The suffix "B" means the algorithm can be applied on box constraints in contrast to the unconstrained L-BFGS method. The box constraints means that we can set the lower bound (LB) and the upper bound (UB) for all or part of the variables.

The L-BFGS-B optimization program works in an iterative manner. It starts with an user provided initial "guess" of the unknown parameters and march in a carefully calculated direction that the program thinks would best increase the objective function value. Calculating this direction requires the gradient function evaluated at the current parameter values. This is similar to other quasi-Newton methods. The optimization routine would also use a carefully calculated step length to determine how far away we will go relative to the current position. The current step parameter values, the search direction, and the step length together will produce a new position in the parameter space with updated parameter values. This ends the current iteration and also provide the starting point for the next search. However, as we mentioned before, determining the search direction and step length will require the value of the likelihood and corresponding gradient given the current parameter value. Each of these calculation would take a lot of time to finish. This poses a big problem for us since the optimization may require many iterations before converging to the final maximizer. The overall time for the optimization could last many days.

In Matlab, the function we used for the constraint optimization is "fmincon". Depends on the problem, Matlab will automatically choose one of three algorithms for the optimization, the trust-region-reflective algorithm, the active-set algorithm, and the interior-point

algorithm.

The trust-region-reflective algorithm is a subspace trust-region method and is based on the interior-reflective Newton method. Each iteration involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients (PCG). The details can be found in Coleman and Li [36] and [37]. The active-set algorithm uses a sequential quadratic programming (SQP) method. In this method, the function solves a quadratic programming (QP) subproblem at each iteration. It updates an estimate of the Hessian of the Lagrangian at each iteration using the BFGS formula. More details can be found in Powell [38] and [39]. The interior-point algorithm is described in Nocedal et al. [40].

The matlab function can dynamically choose the appropriate algorithm but still can not change the iterative manner. Therefore we do not get a big boost in the speed when switching to Matlab. We have tested the program using both R and Matlab and the results confirm this conclusion.

The "optim" function in R and "fmincon" function in Matlab are both optimized for various situations. However, in our case they don't always produce good results. Given the box constraints, both functions like to try the corner of the constrains which usually leads to undesired situations like infinite loop or overflow. Therefore we decide to step back and write a simple gradient ascent program with the box constraints. In each step, a new set of ψ random effects is generated based on the previous step estimated parameters. These ψ vectors are then used by the Monte Carlo estimation of the marginal likelihood to produce a new search direction. Amazingly, this procedure produces much better results comparing to the more sophisticated R and Matlab optimization functions.

4.1.2 The Initial Values for η

Another improvement we have implemented is to obtain a better starting point for the parameter optimization than the random guess. Choosing good initial parameters could help to both reduce the computation time and increase the chance of finding the global maxima. The method we use is very simple. It takes estimates of the η parameters based on the data and use those as the initial values in the optimization. These are the MLE's for a simple model which includes only the η parameters.

Recall the likelihood for our proposed model is,

$$L(\theta|\Psi_{N(T)},\mathcal{F}_T) = \prod_{i=1}^{T}\prod_{k=1}^{K}\prod_{s\in S}\exp\Big(Y_{sk}(i)\Big(\eta_s+\alpha_s\psi(i)\Big)-R_{sk}(i)\exp\Big(\eta_s+\alpha_s\psi(i)\Big)\Big) \quad (4.7)$$

By omitting the AR-GARCH random effect, we get the simple model,

$$L(\eta|\mathcal{F}_T) = \prod_{i=1}^{T}\prod_{k=1}^{K}\prod_{s\in S}\exp\Big(Y_{sk}(i)\eta_s-R_{sk}(i)\exp\big(\eta_s\big)\Big).$$

Then we take the derivative and set it to zero. This results in the MLE for η_s for any given $s\in S$,

$$\eta_s=\log\left(\frac{\sum_i\sum_k Y_{sk}(i)}{\sum_i\sum_k R_{sk}(i)}\right). \quad (4.8)$$

Empirical testing shows that this derived initial η value performs much better than the random guesses. The initial values are usually close to the true values and the optimization routine tends to terminate after small amount of iterations.

4.2 The Simulation Study

Our simulation study is implemented in both R and Matlab programming language. All the testing is under the PC environment. Since we have aggregated the transitions across different companies in the condensed data structure, we are able to implement the optimization codes in a vectorized form. All the data, ordinary and penalized likelihood, gradient, and hessian matrix are then stored as matrix structures and calculated using matrix manipulations. Comparing to our previous programming codes which use the nested loops when scanning the data, the current vectorization serves to boost the performance by several orders of magnitude in terms of the computation time. Therefore we are able to validate and present simulation results on relatively large datasets.

However, when enjoying the better performance, we need to be aware of the limitations of our program. Currently we are able to expand the data set to include thousands of companies without affecting the overall computation time. This nice performance is achieved by sacrificing the ability to incorporating the firm specific covariates. Although the macroeconomic variables can explain a big part of the rating transition performance, there might be situations that people want to see more details at the company level. It would

also be useful and sometimes important to look at the difference among companies in terms of their transition performance. With the company specific information incorporated, the more detailed analysis would lend us a useful tool to study any particular company's health comparing to other Markov chain transition models. The only problem of including the firm specific covariates is that we will need to go back to the original data structure. Then the vectorization of the optimization code may no longer work and we will need to go back to the nested loops. This would mean much more computation time if we want to keep the current level of dataset size. Since we don't have more powerful computer hardware and software to perform this type of analysis, currently we will temporarily omit these covariates and make sure our program generate desired results. We believe for those who have access to more powerful computational resource, they can take a closer look at further computational speed improvement and estimate the full model with all types of covariates included.

4.2.1 The Parameter Setup

Now let's look at the parameters we use for the simulation. The true parameters used to generate data are as following,

- Intensity parameters:
  - For η_s: $d_1 = -5$, $d_2 = -6$, $d_3 = -7$;
  - For α_s: $d_1 = 0.1$, $d_2 = 0.2$, $d_3 = 0.3$.
- AR-GARCH parameters: $\rho = 0.5$, $\alpha_0 = 0.1$, $\alpha_1 = 0.05$, $\beta_1 = 0.8$.

We have tested different sets of values for the intensity parameters. Our goal is twofold. First we need to make sure the intensity function values, which are approximately the probability of having a transition in the next day, are close to zero so that the continuous time process can effectively approximate the discrete daily rating transitions. To achieve this, negative η_s are desired. At the same time, we will set α_s to be positive because we could have either positive or negative random effect ψ_t and the sign of α_s is irrelevant.

Besides the signs of the parameters, we also need to determine their magnitude. First the value of η_s should dominate the product $\alpha_s \cdot \psi$. Otherwise there is no guarantee of small intensity values after taking the exponential function. Our second goal is to generate simulated data that have similar transition patterns/frequencies as in the real data. This is

achieved by carefully balancing and testing the AR-GARCH parameters. Also, as mentioned in the data generation procedure, the intensity functions should be close to zero so that the proposed likelihood closely match the real likelihood induced from the data. The intensity and AR-GARCH parameters are chosen so that the generated intensity function is in the order of 10^{-3}. We will show the plot of the intensity functions later in this section.

One characteristic of the intensity parameters is that the three values d_1, d_2, and d_3 should exhibit an order to represent the decreasing trend of probabilities when moving to a farther rating. For η_s, we choose to decrease the values by one when move to the next farther rating category, i.e. $d_1 - d_2 = 1$ and $d_2 - d_3 = 1$. This implies the intensity will decrease by a factor of $\exp(1) = 2.72$ given other things being the same. This pattern approximately matches what we see from the real data.

For the AR-GARCH parameters, we first want to put a restriction on them so that the generated time series process is stationary. Recall that the stationarity conditions for the AR-GARCH process are,

$$
\begin{aligned}
&|\rho| < 1, \\
&\alpha_0 > 0, \quad \alpha_1 \geq 0, \quad \beta_1 \geq 0, \\
&\alpha_1 + \beta_1 < 1.
\end{aligned}
$$

Based on our discussion of the rating momentum [8] in chapter 1, the company could get the initial rating upgrade or downgrade and then the further rating changes will be along the same direction. To capture this momentum effect, it is reasonable to set the AR parameter ρ as a positive number. Under this setting, one large random effect generally leads to another large random effect at the next moment. This will make intensity values larger over consecutive periods which implies the momentum effect.

The other three parameters α_0, α_1, and β_1 of the GARCH part are chosen so that they are consistent with our observation. α_0 generally associates with the long run process variance or volatility. We set α_0 to be 0.1 which approximately match our observations from the data. We also set α_1 to be a small number 0.05 and β_1 to be relatively larger 0.8. The objective here is to make the generated data show enough volatility cluster effect. This will help us to determine if the program work well under this condition.

Now we have specified all the true parameters values that are used to generate the data for the simulation program. Let's look at some visual examples of the intensity functions. The

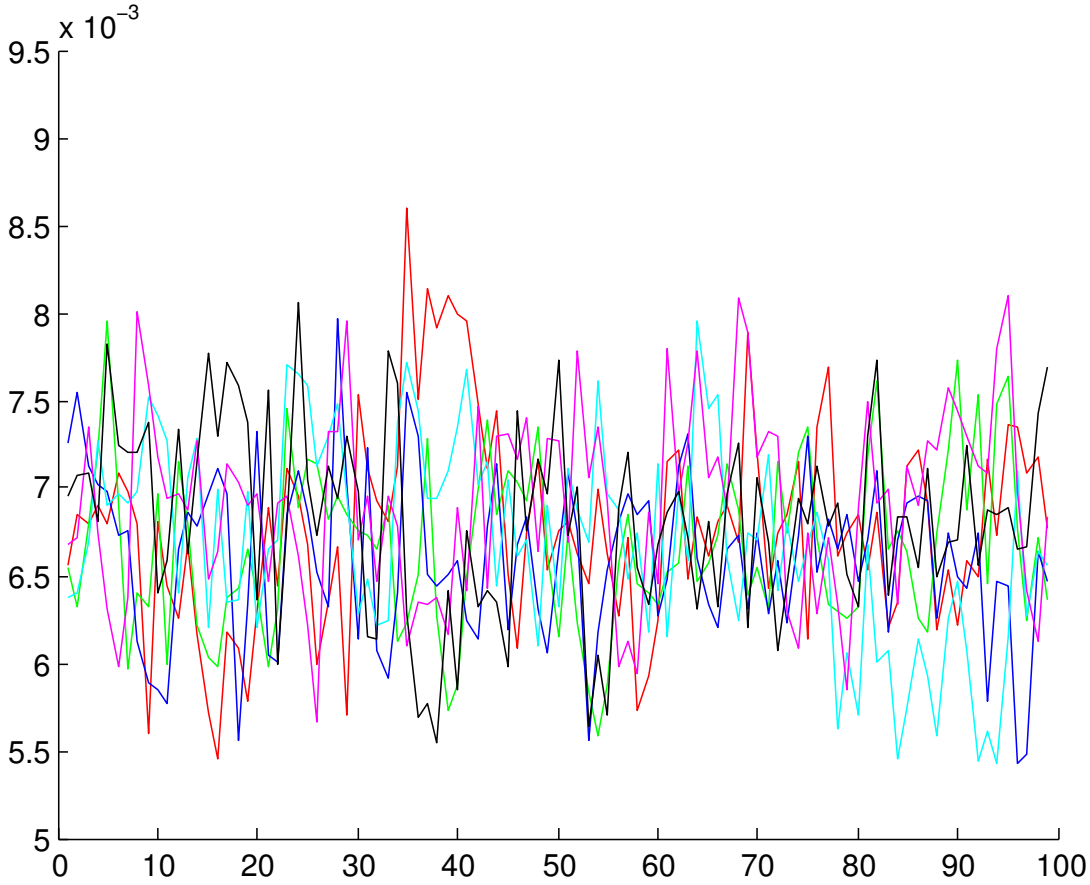

Figure 4.3: Simulated intensity function over 100 time points.

following graph 4.3 shows the intensity values over the time period $[0, 100]$ for six different simulations. The x-axis is the time points and the y-axis shows the intensity values. Note we have only drawn the intensities with $\eta_s = -5$ and $\alpha_s = 0.1$. The other two types of intensities, $\eta = -6$ $\alpha_s = 0.2$ and $\eta_s = -7$ $\alpha_s = 0.3$, will have essentially the same pattern but shift downward toward zero by a factor of approximately 2.72 and 7.34, respectively.

In this plot, we simulated the described intensity function for 6 times over 100 time points (days). The values of the intensity functions are in the range of 10^{-3} which is deemed small enough to guarantee the likelihood associated to the data generation process closely match the model likelihood.

The initial parameters used in the estimation algorithm are described below.

- The intensity parameters:
  - For η_s, the initial parameters will be derived from the data using formula 4.8. Section 4.1.2 has the detailed explanation.
  - For α_s: $d_1 = 0.1$, $d_2 = 0.1$, $d_3 = 0.1$.
- The AR-GARCH parameters: $\rho = 0.1$, $\alpha_0 = 0.1$, $\alpha_1 = 0.1$, $\beta_1 = 0.1$.

Here we want to note a small change made to the intensity parameters matrix $\{\alpha_s\}$. We need to avoid the non-identifiability issue for α_s when multiplying it to the random effect. This issue can be seen easily by increasing α_s by k times and decreasing the random effects to $\frac{1}{k}$ of the original value. This operation does not affect the intensity function values. Therefore we need to put a restriction either on the coefficient parameter α_s or on the AR-GARCH parameters. Our method is to set one of the AR-GARCH parameters α_0 to a predetermined constant.

The initial parameters are set such that they are reasonable guesses according to our knowledge of the data. We use 0.1 for all α_s parameters and AR-GARCH parameters. Their order of magnitude can be seen from the plot of the data but we keep them as the same value so that our knowledge of the true value doesn't lead to any bias.

For the Monte Carlo simulation, we generate M samples of multivariate Gaussian paths to compute the approximated integral for the importance sampling and the marginal likelihood. As stated in section 3.2.3, M should be large enough so that the Hotelling's T^2 statistic should be larger than the square of the penalty constant c^2. In the current simulation study, we choose the penalty constant c also based on section 3.2.3

$$c = \sqrt{p + 2\sqrt{2p}},$$

where p is the number of parameters to be estimated.

During the simulation, the optimization will pick up M automatically based on the above formula. In practice, we have observed M ranges from 5,000 to 50,000. Although M is usually large, we have managed to obtain simulation results in a reasonable time due to the vectorization of the programming code.

We have tested our program using different combinations of the number of firms and time points. Table 4.2 gives a summary of this simulation study under the above initial parameter

setup,

Table 4.2: The Simulation Study Results

Time		Model Parameters		GARCH Parameters					
Points	Firms	$RMSE_\eta$	$RMSE_\alpha$	$mean_\rho$	$RMSE_\rho$	$mean_{\alpha_1}$	$RMSE_{\alpha_1}$	$mean_{\beta_1}$	$RMSE_{\beta_1}$
50	200	0.8186	0.8120	0.2549	0.4373	0.0368	0.0433	0.4866	0.4558
	300	0.6454	0.7911	-0.0015	0.5940	0.0269	0.0482	0.7414	0.2741
	400	0.7055	0.8577	0.2162	0.4848	0.0102	0.0509	0.8292	0.1554
	500	0.7080	0.7226	0.1405	0.5661	0.0315	0.0516	0.6147	0.3725
100	200	0.5031	0.6667	0.1924	0.3437	0.0622	0.0328	0.3769	0.5053
	300	0.4779	0.6961	0.4753	0.2456	0.0756	0.0600	0.4130	0.5299
	400	0.4537	0.6298	0.3007	0.3207	0.0668	0.0630	0.4435	0.4907
	500	0.4188	0.6598	0.3843	0.2111	0.0782	0.0299	0.4833	0.3996
150	200	0.4362	0.6048	0.2472	0.3353	0.0562	0.0434	0.3736	0.5094
	300	0.3816	0.5693	0.1785	0.3230	0.0322	0.0449	0.5305	0.4442
	400	0.3764	0.6170	0.2409	0.3177	0.0636	0.0713	0.5150	0.3978
	500	0.3849	0.6070	0.3766	0.3293	0.0486	0.0344	0.6886	0.2558
200	200	0.3859	0.5307	0.1778	0.3377	0.0498	0.0299	0.3213	0.5629
	300	0.3299	0.6636	0.2186	0.3035	0.0374	0.0352	0.4271	0.5085
	400	0.3551	0.5194	0.2492	0.2823	0.0309	0.0463	0.6798	0.1748
	500	0.3259	0.6051	0.2902	0.2401	0.0634	0.0188	0.5813	0.2998
True Value				0.5000		0.0500		0.8000	

Ideally, there are totally eighty-seven parameters for the intensity functions based on the pattern matrix (4.4) and the AR-GARCH random effects. Forty-two of them are η parameters and forty-two of them are α parameters. Then we have three freely varying AR-GARCH parameters ρ, α_1, and β_1 after α_0 is fixed to avoid the non-identifiability issues. In our simulation, we have run 10 replications for each of the T and K combination above. All individual parameters are estimated and compared to the true values. However, reporting the results for each parameters and all 10 replications would distract the readers from having a overall feeling of how the simulation program works in general. Therefore we only include the root mean squared error (RMSE) of the intensity parameters over the 10 simulations for each combination of time points and firms. Interested readers are referred to appendix A for more detailed results. Here we want to note that for some of the replications, some of the η or α parameters may not be missing due to the unobserved transitions. Therefore we have to take this into consideration by using appropriate n in the formula below.

RMSE is defined as the square root of the average of the squared difference between the true and estimated parameters. For example, if we let $\beta_j, j = 1, \cdots, n$ denotes the true parameters and $\hat{\beta}_j$ denotes the estimated parameters, then the RMSE of β parameters will be given as the following formula,

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{\beta}_j - \beta_j)^2}.$$

The RMSE is given for both the η and α intensity parameters in table 4.2. The last six columns are the sample means and RMSE for AR-GARCH parameters over the 10 simulations for each T and K combination. The true values of the AR-GARCH parameters are also listed in the last row of table 4.2 for comparison purpose.

We can see that the RMSE of the estimated η parameters are on average getting smaller when the number of time points and the number of firms increase. This means the estimated η parameters are closer to the true values as T increase. This is desirable because we expect the estimation error to decrease with larger datasets. However, the RMSE of the α parameters does not show a clear decreasing pattern when the number of time points or firms increase.

The different performance behavior for η and α parameters can partially be explained by the using of the simplified model estimator for η as the initial values. As we mentioned before, this step helps to restrict the initial η values in a region that is close to the true value. Due to the non-linear structure of the AR-GARCH model, there could be multiple local maxima that the optimization will converge to. This method helps to increase the odds of finding the global maxima and reduce the computation speed. On the other hand, also due to the complication of the AR-GARCH structure, it is hard find good initial values for α parameters. This makes the estimation of α much harder than that of η. The optimization tends to find a local maxima that may or may not close to the true values. This helps to explain the variability of the RMSE shown in table 4.2.

For the AR-GARCH parameters, the RMSE of the rho parameter generally decrease when T increases. For the α_1 and β_1 parameters, we don't observe similar patterns.

In summary, we believe our simulation procedure shows some potential to be developed into a useful method when the volatility change is concerned. There are some difficulties that still need to be addressed before any actual use of this algorithm. This will be part of the future work.

CHAPTER 5

APPLICATION STUDY

In the last chapter, we discussed the details on implementing our proposed credit rating transition models which extended the idea from the paper by Koopman et al. [10]. The simulation studies were carried out thereafter with datasets of various sizes to test the performance of the program. The results of these studies were promising and the conclusion was that we would start to apply this model to the market implied rating (MIR) transition study.

Before we discuss the model application, let's first take a look at the market implied rating data that we will deal with. In chapter 4, we have summarized the rating transition frequencies into the table 4.2. In the table, the company's ratings can move to as far as five rating categories away. However, we also notice that there are only few transitions that move more than three categories. Therefore in the simulation studies, we have restricted the model to only consider transitions that move one, two, or three categories. When we apply the model to the real data, there are similar restrictions. This means in the model likelihood function (4.7), we only include the contribution from transition that move within three categories. At the same time, the at-risk and transition indicators Y and R will still include all possible transitions, no matter how far they move. Although this treatment creates some inconsistencies among the components of the likelihood function, we believe it is tolerable and wouldn't make the parameter estimation inaccurate since the number of transition over four and more categories is extremely small comparing to the total number of transitions. On the other hand, if all existing transition types are included, the parameter estimates for those of fewer transitions will not be reliable.

The data we use contain 743 U.S. companies that cover a wide range of industries. There are both public and private companies but all of them have their issued debt trading in the

bond market. All the agency and market implied ratings are published by Moody's in a period from Oct 1, 2002 to Nov 20, 2008. There are totally 1549 trading days during this period. Since our program cannot handle this large time series, we have selected the most recent 100 periods as our sample window. The rating transitions are calculated from this raw data and will be used in our model application.

Now we will apply the model described in chapter 3 and 4 to the market implied rating transitions. One problem of the estimation process is that we need to have the appropriate initial parameter values to start the program. Since the model use the importance sampling technique to approximate the integral for the marginal likelihood, the optimization routine is not allowed to search for a solution that is far away from the initial parameters. Otherwise the importance sampling method will not lead to the desired approximation result. Therefore we designed the penalized likelihood method and restricted the optimization to only search a neighborhood of the initial parameters. Now if the initial values are far away from the true values, we will need to perform a series of the optimization routines such that the estimated parameters from the previous optimization search being the initial parameters of the next. The time spent by this whole procedure will becomes unbearably long if we choose inappropriate initial values, i.e. if they are too far away from the true values and we have to run too many iterations of the optimizations. This is not a problem for the simulation study since we know what the true values are. However, when we want to apply the model to the real data, we need to be careful when assigning the initial values.

The choice of the initial values depends on our observations of the actual aggregated transition process. In the simulation study in chapter 4, we discussed a way to aggregate the data into a condensed data structure to speed up the program. The intensity functions then become,

$$\lambda_s(i) = \exp\Big(\eta_s + \gamma_s'\omega(i) + \alpha_s\psi(i)\Big).$$

This implies that we need to select initial parameter values that only base on different transition types and do not worry about different companies. Now the definition of the intensity process is given by equation (3.1). Roughly speaking, $\lambda_s(t)$ means given the past information, the probability of having a transition of type s in the next moment. This reminds us to look at the actual transition probabilities. If the empirical transition probabilities can be plotted for each transition type, we will be able to have a good idea of

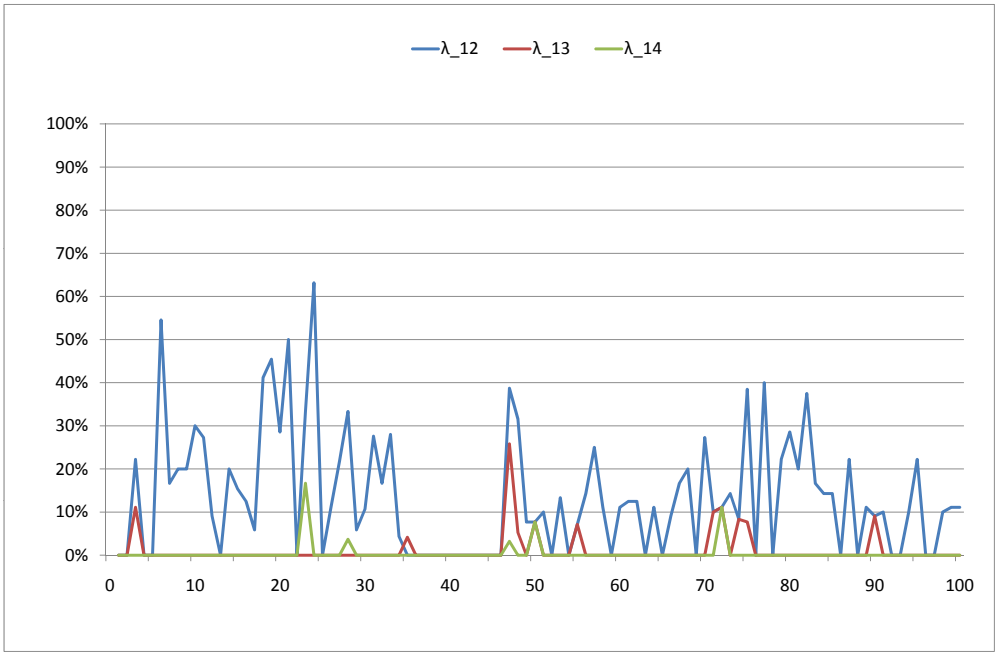

Figure 5.1: MIR: empirical transition probability time trend for rating 1

how to choose the initial values.

The following nine figures 5.1-5.9 show the trend of the transition probabilities. We are looking at transitions for 100 time points and 743 companies. At each time point, all of the transitions are aggregated to calculate the probabilities. For example, at time 5, there are 237 out of the 734 companies that bear rating 3. In the next day, 11 of them move to rating 2 and 12 of them move to rating 4. The empirical probabilities are then 4.64% for the transition type $3 \to 2$ and 5.06% for the transition type $3 \to 4$.

In each of the plots, the curves are the illustration of the probability time trends for each specific rating transition type. For example, figure 5.1 shows three curves that correspond to transition types $1 \to 2$, $1 \to 3$, and $1 \to 4$. We label them with the name of their intensity functions, λ_{12}, λ_{13}, and λ_{14}. The curves ranges from time point 1 to 100. Each point on the curve is the empirical transition probability for that time point.

Here we want to note that for some transition types, there are few transitions observed in the data. This leads to large variability in the calculation of the empirical transition probabilities. For example, in the last two plots 5.8 and 5.9, empirical probability can reach

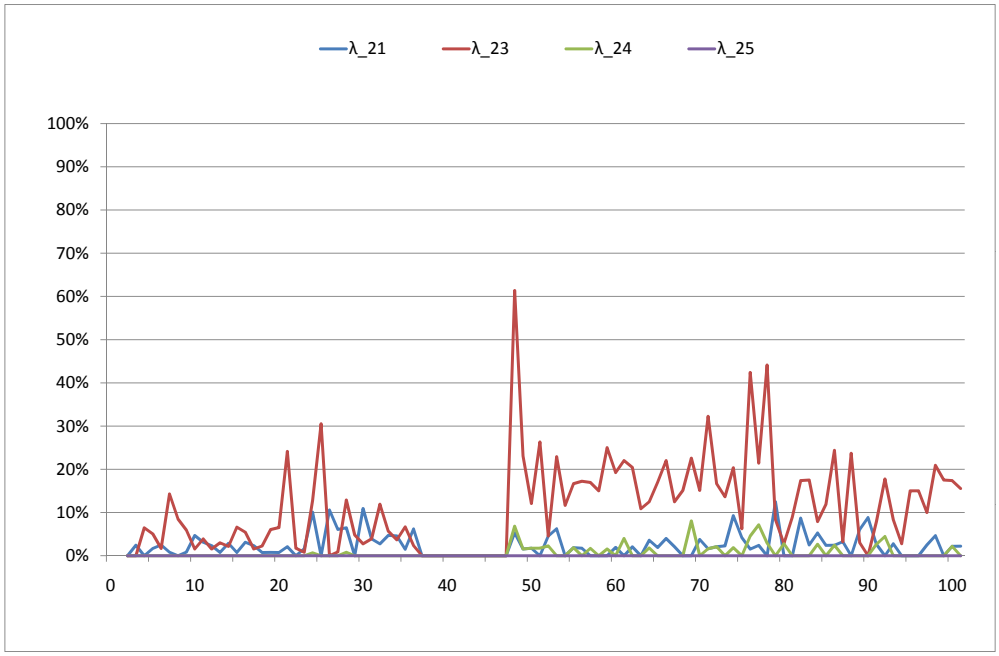

Figure 5.2: MIR: empirical transition probability time trend for rating 2

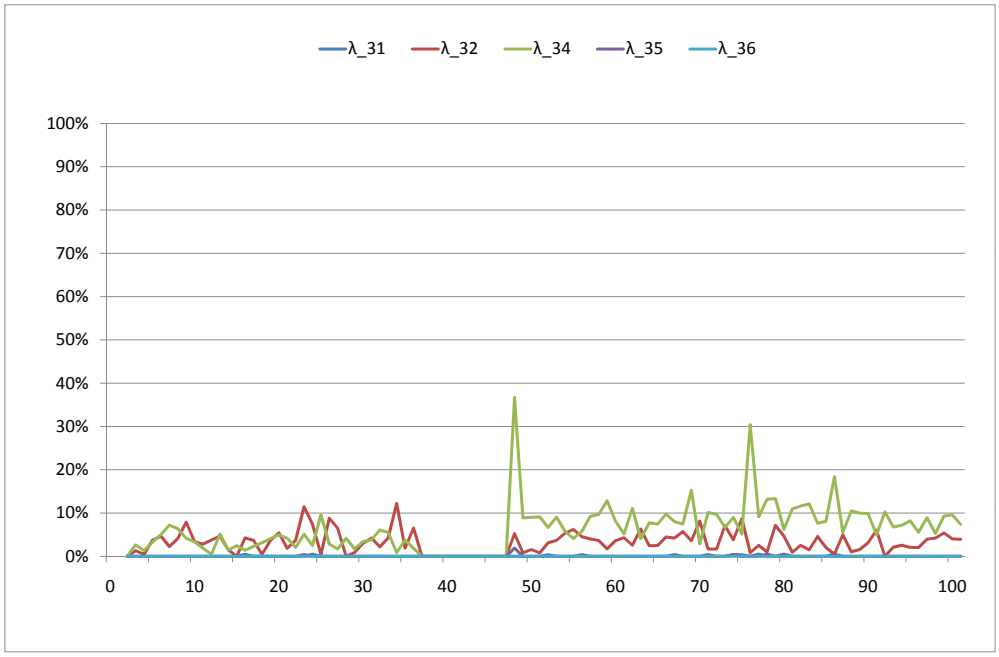

Figure 5.3: MIR: empirical transition probability time trend for rating 3

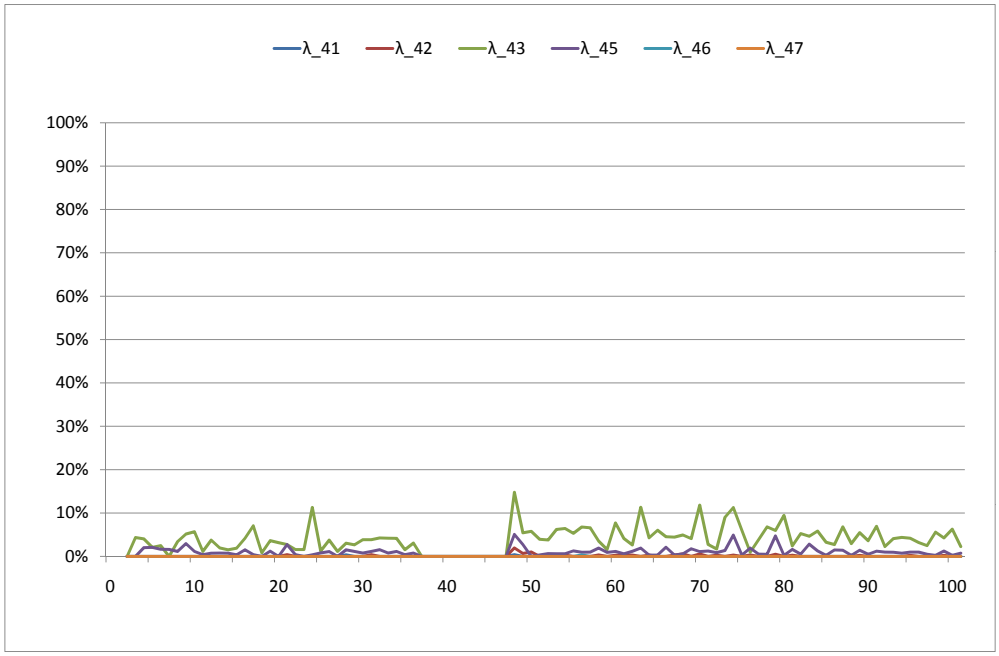

Figure 5.4: MIR: empirical transition probability time trend for rating 4

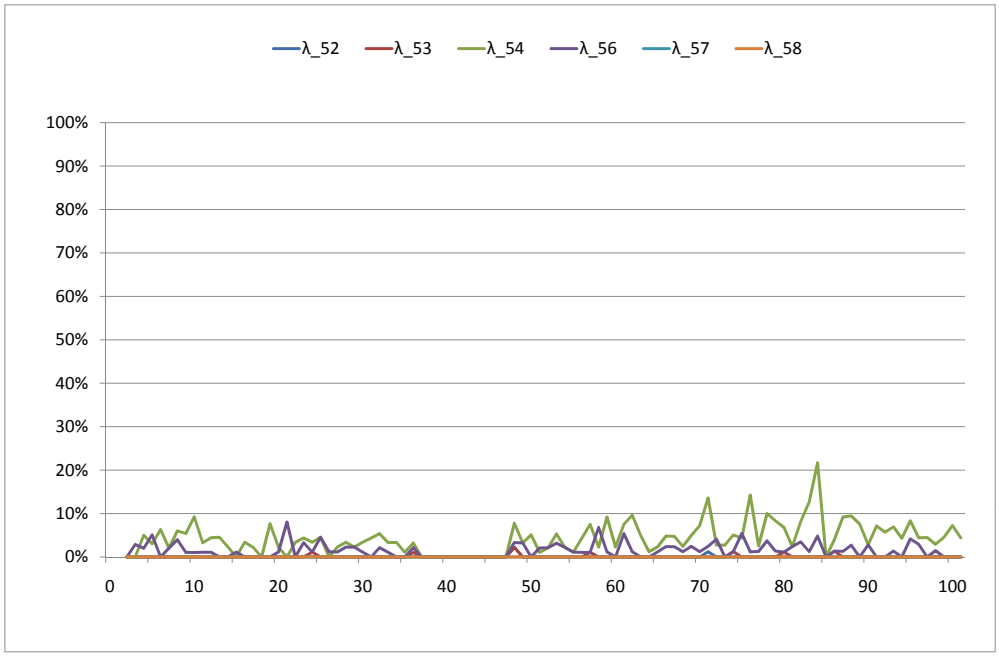

Figure 5.5: MIR: empirical transition probability time trend for rating 5

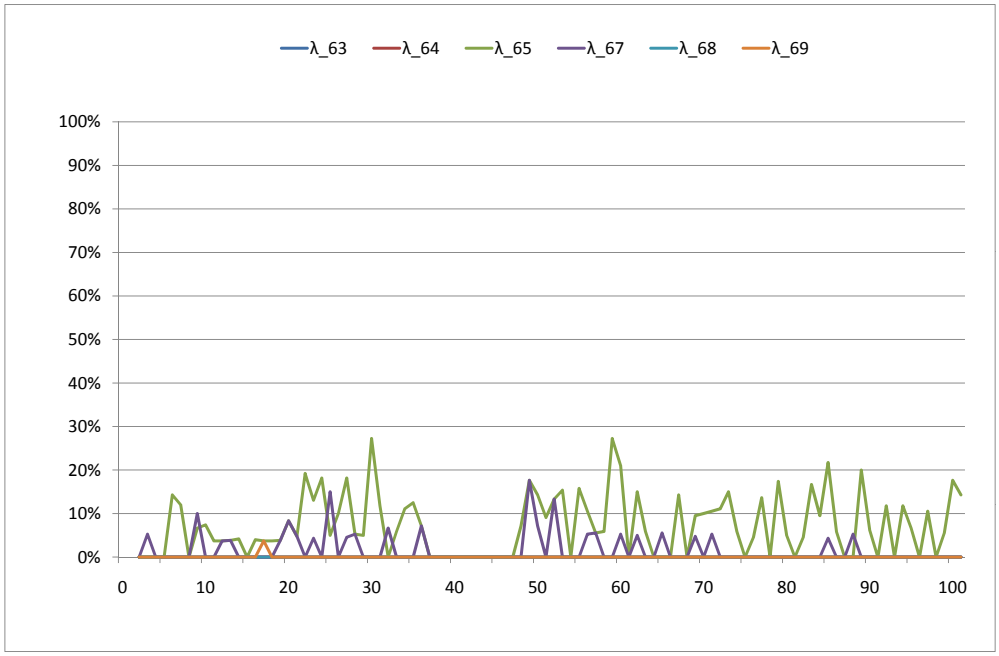

Figure 5.6: MIR: empirical transition probability time trend for rating 6

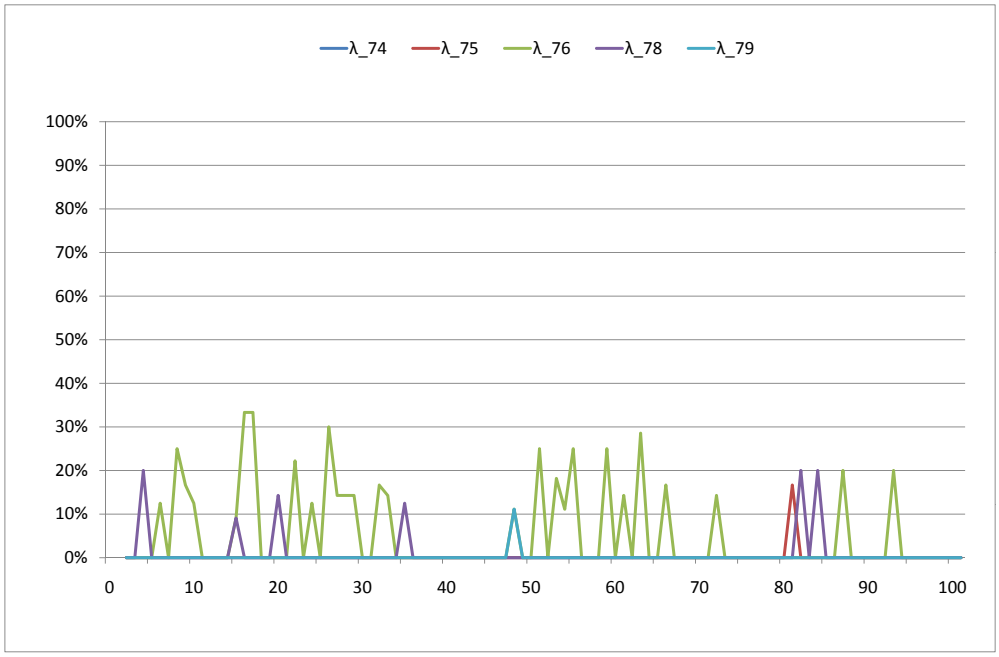

Figure 5.7: MIR: empirical transition probability time trend for rating 7

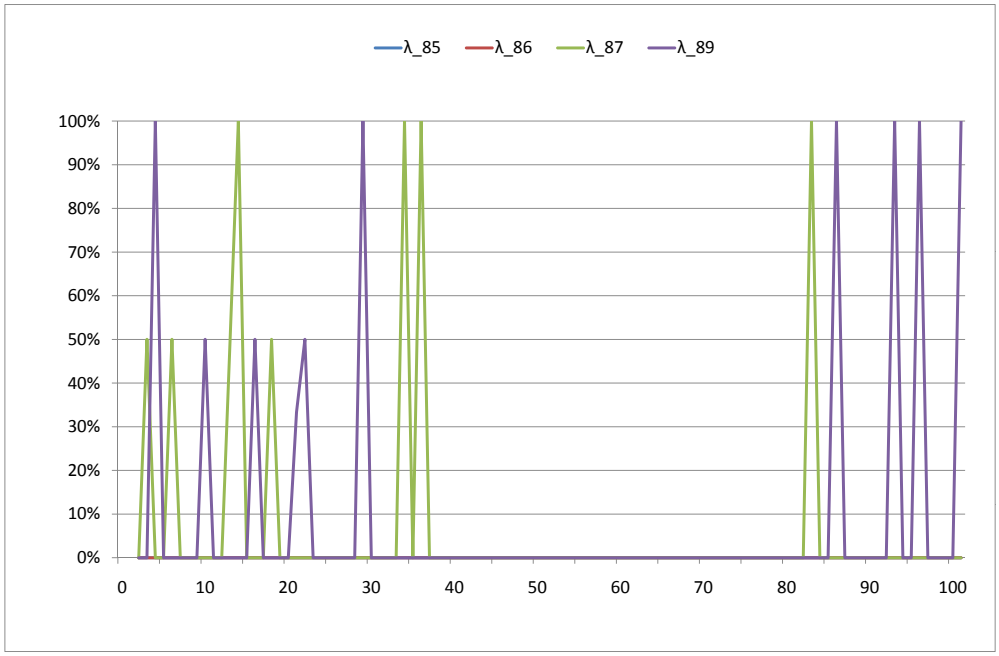

Figure 5.8: MIR: empirical transition probability time trend for rating 8

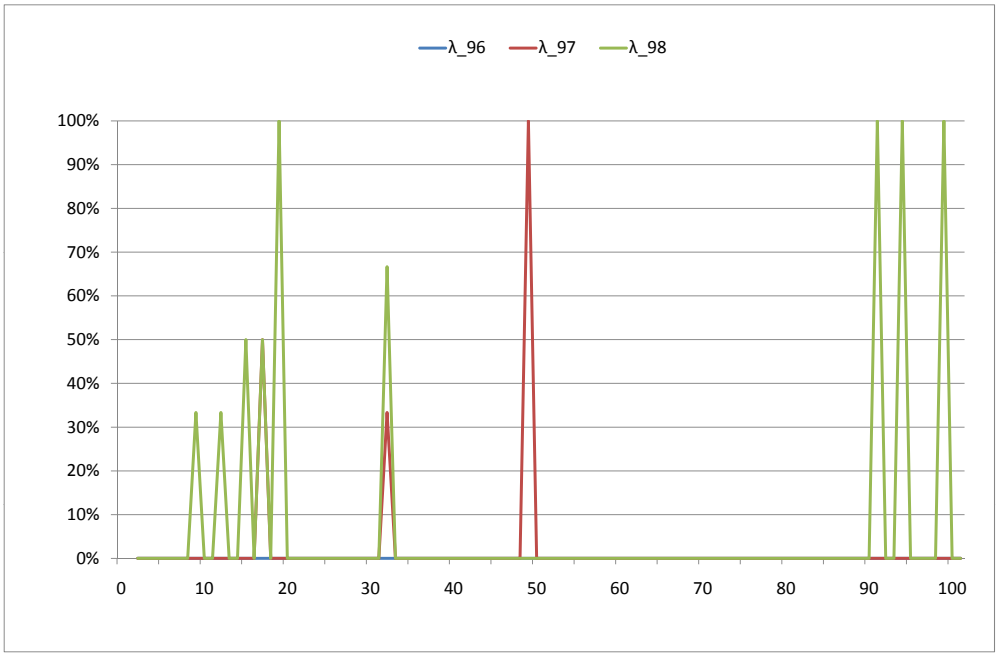

Figure 5.9: MIR: empirical transition probability time trend for rating 9

100but there are only 2 to 3 firms that are at risk and all of them make transitions.

Choosing the initial parameters by simply looking at the charts is a little arbitrary. However, we do not expect to pick a very good entry point but only want to make the initial values as close to the true values as possible. Our method of determining the parameters are as following.

For the intensity parameters η_s and α_s, we decide to estimate the values for only the transition types that have enough meaningful rating movements. Therefore only the following 18 transition types will have their likelihood contribution counted in the final model, $1 \to 2$, $1 \to 3$, $2 \to 1$, $2 \to 3$, $2 \to 4$, $3 \to 2$, $3 \to 4$, $4 \to 3$, $4 \to 5$, $5 \to 4$, $5 \to 6$, $6 \to 5$, $6 \to 7$, $7 \to 6$, $7 \to 8$, $8 \to 7$, $8 \to 9$, and $9 \to 8$.

For the intensity parameters η, we will use the same method as in section 4.1.2 to get their initial values. This includes taking the simplified model by omitting the AR-GARCH random effect,

$$L(\eta|\mathcal{F}_T) \quad = \quad \prod_{i=1}^{T}\prod_{k=1}^{K}\prod_{s\in S}\exp\Big(Y_{sk}(i)\eta_s - R_{sk}(i)\exp\left(\eta_s\right)\Big),$$

and then use the MLE as the initial parameters for η_s, given $s \in S$,

$$\eta_s = \log\left(\frac{\sum_i \sum_k Y_{sk}(i)}{\sum_i \sum_k R_{sk}(i)}\right). \tag{5.1}$$

This has proved to be a very effective way to get a good sense of the true parameters. Finally, the initial values for η are,

Table 5.1: Initial Intensity Parameter Values for η

*	-2.06	-4.33	*	*	*	*	*	*
-3.79	*	-2.38	-5.26	*	*	*	*	*
*	-3.42	*	-2.75	*	*	*	*	*
*	*	-3.18	*	-4.60	*	*	*	*
*	*	*	-3.16	*	-4.17	*	*	*
*	*	*	*	-2.58	*	-4.05	*	*
*	*	*	*	*	-2.83	*	-4.57	*
*	*	*	*	*	*	-1.73	*	-1.63
*	*	*	*	*	*	*	-1.97	*

For the α and AR-GARCH parameters, we do not have this luxury. Therefore we choose the following initial parameters by simulating the data using these values and find that they superficially resemble the real data.

The chosen initial values for α are listed as below,

Table 5.2: Initial Intensity Parameter Values for α_s

*	0.1	0.1	*	*	*	*	*	*
0.1	*	0.1	0.1	*	*	*	*	*
*	0.1	*	0.05	*	*	*	*	*
*	*	0.05	*	0.1	*	*	*	*
*	*	*	0.1	*	0.1	*	*	*
*	*	*	*	0.1	*	0.1	*	*
*	*	*	*	*	0.1	*	0.1	*
*	*	*	*	*	*	0.1	*	0.1
*	*	*	*	*	*	*	0.1	*

For the AR-GARCH parameters, we have (excluding α_0),

- AR-GARCH initial parameters: $\rho = 0.5$, $\alpha_1 = 0.0833$, $\beta_1 = 0.3333$.

Now given the initial parameter setup, we can perform the optimization routine for the parameter estimates. Before this, we want to make a final note on the covariates of the intensity processes. In our program, there will be no covariates involved in the estimation, just like in the simulation studies in chapter 4. The reason of this decision is as following. The data we are working with are the market implied rating transitions. These daily data require us to put in covariates that also change daily, or at least every couple days. One possible choice of the covariates is to use the companies' stock prices, but some of the companies are private and there is no way we can find the daily change of their value from the published data. On the other hand, we can also use the bond prices since all of them should have traded bonds otherwise there will no market implied ratings. However, the bond prices is not publicly available like the stock prices and we can not afford buying these prices for our research. Therefore, we temporarily ignore the covariates. This has no impact on the quality of the model as it can be readily applied on any appropriate covariates when available.

Now the final estimated parameter results are as following.

- AR-GARCH estimated parameters: $\rho = 0.6014$, $\alpha_1 = 0.3359$, $\beta_1 = 0.3630$.

Comparing these parameters with the initial values,

- AR-GARCH initial parameters: $\rho = 0.5$, $\alpha_1 = 0.0833$, $\beta_1 = 0.3333$,

We can see that the two sets of values are not quite close, but they are generally with the same order of magnitude.

The estimated intensity parameters are listed in the following two tables.

Table 5.3: Estimated Intensity Parameter Values for η_s

*	-2.4367	-4.2255	*	*	*	*	*	*
-3.6762	*	-2.0619	-5.2588	*	*	*	*	*
*	-3.7002	*	-2.8761	*	*	*	*	*
*	*	-2.9747	*	-4.6993	*	*	*	*
*	*	*	-3.4558	*	-3.3822	*	*	*
*	*	*	*	-2.8891	*	-3.7502	*	*
*	*	*	*	*	-2.7907	*	-4.3313	*
*	*	*	*	*	*	-2.6543	*	-1.5389
*	*	*	*	*	*	*	-2.4738	*

Table 5.4: Estimated Intensity Parameter Values for α_s

*	0.1000	0.1466	*	*	*	*	*	*
0.0523	*	0.1592	0.1678	*	*	*	*	*
*	0.0401	*	0.1401	*	*	*	*	*
*	*	0.0805	*	0.1026	*	*	*	*
*	*	*	0.0879	*	0.0887	*	*	*
*	*	*	*	0.0703	*	0.0856	*	*
*	*	*	*	*	0.0773	*	0.0909	*
*	*	*	*	*	*	0.0897	*	0.1001
*	*	*	*	*	*	*	0.0951	*

Comparing these two parameter matrices (table 5.3 and 5.4) with their initial values (table 5.1 and 5.2), we can see that the estimated parameters are generally close to the initial parameter, although there are still some gaps between them.

The final goal of our model building is to construct a reasonable set of transition probability estimates from the market implied ratings. Now that we have obtained the estimated intensity parameters, we can start to work on this transition matrix.

Recall that the estimated intensity process $\widehat{\lambda}_s(t)$ is defined by,

$$\widehat{\lambda}_s(t) = \exp\left(\widehat{\eta}_s + \widehat{\alpha}_s \psi(t)\right).$$

Since the intensity at time t is roughly the probability of recording a transition at the next short moment $(t, t+\Delta t]$, this coincides with the generator matrix concept in the continuous Markov chain. Therefore, we are able to form the generator matrix $Q(t)$ by defining,

$$Q(t) = \begin{pmatrix} -q_1(t) & \widehat{\lambda}_{12}(t) & \widehat{\lambda}_{13}(t) & \cdots & \widehat{\lambda}_{19}(t) \\ \widehat{\lambda}_{21}(t) & -q_2(t) & \widehat{\lambda}_{23}(t) & \cdots & \widehat{\lambda}_{29}(t) \\ \widehat{\lambda}_{31}(t) & \widehat{\lambda}_{32}(t) & -q_3(t) & \cdots & \widehat{\lambda}_{39}(t) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \widehat{\lambda}_{91}(t) & \widehat{\lambda}_{92}(t) & \widehat{\lambda}_{93}(t) & \cdots & -q_9(t), \end{pmatrix} \tag{5.2}$$

where for the diagonal elements, $q_i(t)$ is the sum of all the off-diagonal elements for the i^{th} row,

$$q_i(t) = \sum_{\substack{j=1 \\ j \neq i}}^{9} \widehat{\lambda}_{ij}(t), \qquad i = 1, \cdots, 9,$$

such that all the rows sum to 0. Also here $\hat{\lambda}_{ij}(t)$ denotes the estimated intensity function for the transition from rating i to rating j.

With the generator matrix $Q(t)$ at hand, we can derive the transition matrix over any time interval $[T_1, T_2]$, where $0 \leq T_1 < T_2 \leq T$. This transition matrix is given by the product integral on the generator matrix, i.e.

$$M(T_1, T_2) = \prod_{T_1}^{T_2} (I + Q(t)dt)$$

where I is the 9 identity matrix.

One can think of the product integral just like the regular Riemann integral, but with the sum in the definition replaced by a product. The formal definition of the product integral is,

$$\prod_0^t (1 + dA(s)) = \lim_{\max |t_i - t_{i-1}| \to 0} \prod (1 + (A(t_i) - A(t_{i-1}))),$$

where the limit is taken over a series of partitions $0 = t_0 < t_1 < \cdots < t_k = t$ such that each partition is finer than the predecessor. The matrix version of the product integral is similarly defined as,

$$\prod_0^t (1 + dX(s)) = \lim_{\max|t_i - t_{i-1}| \to 0} \prod (1 + (X(t_i) - X(t_{i-1}))),$$

where $X(t)$ is a matrix-valued function.

In our model, since we use the discretized time, the generator matrix becomes an approximation of the product integral by setting $dt = 1$,

$$M(T_1, T_2) = \prod_{T_1}^{T_2} (I + Q(t)).$$

The generator matrix is dependent on the random effect which is unobserved. Therefore we compute the transition matrix by taking the unconditional expectation of $M(T_1, T_2)$,

$$\bar{M}(T_1, T_2) = E\Big(M(T_1, T_2)\Big) = E\left(\prod_{T_1}^{T_2} (I + Q(t))\right).$$

Now we can estimate the above unconditional expectation by using the Monte Carlo approximation. In this process we set $M = 20,000$. These $20,000$ i.i.d. AR-GARCH paths are simulated using equation (3.5) with the final parameter estimates obtained from the real transition data. The transition matrices are computed for each generated path. The approximated transition matrix will be the average of these entities, i.e.

$$\widehat{\bar{M}}(T_1, T_2) = \frac{1}{M} \sum_{m=1}^{M} \prod_{T_1}^{T_2} (I + Q^k(t)),$$

where $Q^k(t)$ is the generator matrix constructed using the random effects and intensity functions from the k^{th} path.

The following table 5.5 is an example of the estimated transition matrix over $[60, 90]$ using the simulated AR-GARCH paths. We can see that for the higher ratings like Aaa, Aa and A, the probabilities of moving to another rating state is low. However, for lower ratings, especially Ca and C, the companies tend to move frequently from their current low ratings to an adjacent low ratings.

Table 5.5: Estimated Transition Matrix

Ratings	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	0.9744	0.0248	0.0008	0	0	0	0	0	0
Aa	0.0042	0.9837	0.0019	0.0002	0	0	0	0	0
A	0	0.0009	0.9936	0.0055	0	0	0	0	0
Baa	0	0	0.0065	0.9916	0.0019	0	0	0	0
Ba	0	0	0	0.0103	0.9878	0.0019	0	0	0
B	0	0	0	0	0.0217	0.9765	0.0018	0	0
Caa	0	0	0	0	0	0.0086	0.9869	0.0045	0
Ca	0	0	0	0	0	0	0.0102	0.9861	0.0037
C	0	0	0	0	0	0	0	0.0245	0.9755

Finally, we have derived a multi-state intensity model with AR-GARCH random effect that builds on the existing models with only AR time series random effects. It has the enhanced ability of handling reasonablely higher frequencies of the rating transitions. We have tested the model with multiple simulation studies and then applied the model on the more recent market implied rating transition data. The estimated transition matrix can be computed from the estimated model parameters. We believe the model can enrich the toolbox of the researchers and provide an alternative tool when the classical Markov chain model does not apply while studying the daily transitions.

CHAPTER 6

Future Work

Although we have build the model, specified all the details and performed the simulation and application study, there are some works that can be done to improve and extend the model in several aspects. We will summarize them in the following sections.

6.1 The Fixed Covariates

In chapter 3, we proposed the multi-state intensity models and specified the intensity function to have the form of proportional hazard rate without the baseline function,

$$\lambda_{sk}(t) = \exp\left(\eta_s + \gamma_s'\omega_k(t) + \alpha_s\psi(t)\right), \tag{6.1}$$

where s represents different transition types and k represents different companies. Currently our simulation and application studies do not involve the fixed covariates $\omega_k(t)$ which may be macro economic variables like the interest rates, or firm specific variables like asset quality, financial leverage, and profitability, etc. Therefore, we have only considered the reduced form of the intensity function,

$$\lambda_{sk}(t) = \exp\left(\eta_s + \alpha_s\psi(t)\right). \tag{6.2}$$

In the application study, we also did not consider the covariates. The reason is that the bond prices that we intend to include at the beginning of the study are hard to obtain. Furthermore, due to the computation limit, we have to speed up the program by using a condensed form of data structure which does not allow firm specific variables. Although omitting the covariates do not alter the nature of the model, we still want to test their effects and compare the results to our current ones. Therefore, we plan to include them in the future application studies of the proposed model. However, we should point out that

including any firm specific covariates will invalidate the current usage of the condensed data structure and it will be necessary to find a better way to speed up the program. This could be time consuming. On the other hand, including more covariates will make the model more complete and will definitely increase the accuracy of the model forecasting. Therefore, we should carefully weigh the benefit tradeoff before approach to the next step.

6.2 The Model Application to Agency Rating Transition Data and Comparison with the Markov Chain Model

There is a major difference between the Markov chain models and our counting process models. The former is a cohort level model which estimate each individual transition probability using the rating history of a group of companies of the same ratings. In this case, it is not feasible to consider the firm level characteristics. On the contrary, our models have the advantage of considering the firm level variables and build the transition matrices for each separate company. Our second plan is to build the Markov Chain model on the same market implied rating dataset and compare the result with our current transition matrices. Since the MIR data contains daily transitions, our model should have the advantage of capturing the volatility structure of the time series effects. Even if we do not get the improvement as expected, we still get a model that is theoretically more appropriate than the Markov model in cases the Markov property fails to hold.

Another plan is to also apply our model on the classical agency rating transitions. We need to be careful with this type of data since they tend to contain very few transitions over a short time period. Since our model is designed to fit the high frequency daily transitions, the result may not be as good as those of the Markov chain model.

APPENDIX A

The Simulation Results on Individual η and α parameters

In this appendix we present the simulation results for individual η and α estimators. Recall that in chapter 4 there are totally sixteen different combinations of time points T and firms K. For each combination, we performs 10 replications using the same true model parameters. For each parameter, we take the average of the 10 estimators and also compute the RMSE from those 10 estimators. These steps result in 16 matrices that are shown below. Each of them corresponds to a combination of T and K. The elements in each matrix are the average of the estimators and the numbers in the parenthesis are the corresponding RMSE calculated using the sample means of η and α as well as the number of parameters that are present (non-missing) from the 10 replications. Note that if any η or α are missing, the calculation will omit it's contribution to the mena and RMSE.

Table A.1: The simulation results for η for $T = 50$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.359	-6.107	-6.902	*	*	*	*	*
		(0.491,10)	(0.608,10)	(0.394,4)					
Aa	-5.539	*	-5.049	-6.174	-7.252	*	*	*	*
	(0.762,10)		(0.354,10)	(0.593,10)	(0.263,2)				
A	-6.100	-5.347	*	-5.133	-6.254	-6.828	*	*	*
	(0.666,10)	(1.048,10)		(0.308,10)	(0.696,9)	(0.523,5)			
Baa	-7.038	-6.088	-5.225	*	-5.189	-6.539	-6.429	*	*
	(0.715,8)	(0.728,10)	(0.680,10)		(0.643,10)	(0.965,8)	(0.571,1)		
Ba	*	-7.059	-6.111	-5.093	*	-5.253	-6.383	-6.775	*
		(0.568,7)	(0.535,10)	(0.235,10)		(0.465,10)	(0.622,9)	(0.552,6)	
B	*	*	-6.544	-5.969	-5.125	*	-5.184	-6.500	-6.425
			(0.642,8)	(0.544,9)	(0.554,10)		(0.494,10)	(0.776,10)	(0.789,6)
Caa	*	*	*	-7.053	-5.738	-5.114	*	-4.896	-6.737
				(0.925,9)	(0.537,9)	(0.270,10)		(0.382,10)	(1.432,10)
Ca	*	*	*	*	-6.771	-6.347	-5.289	*	-4.958
					(0.453,8)	(0.946,10)	(0.523,10)		(0.416,10)
C	*	*	*	*	*	-6.971	-6.024	-5.159	*
						(0.198,7)	(0.465,10)	(0.570,10)	

Table A.2: The simulation results for α for $T = 50$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.335	0.204	0.341	*	*	*	*	*
		(0.712,10)	(0.673,10)	(0.339,4)					
Aa	0.150	*	-0.134	-0.019	-0.425	*	*	*	*
	(0.649,10)		(0.743,10)	(0.923,10)	(0.855,2)				
A	-0.160	0.384	*	0.400	0.670	0.284	*	*	*
	(0.833,10)	(1.179,10)		(0.499,10)	(1.094,9)	(0.468,5)			
Baa	0.356	0.273	0.230	*	0.270	0.190	0.748	*	*
	(1.260,8)	(0.547,10)	(0.573,10)		(0.874,10)	(0.813,8)	(0.448,1)		
Ba	*	0.597	0.406	0.079	*	0.326	0.059	0.190	*
		(1.075,7)	(1.024,10)	(0.389,10)		(0.715,10)	(0.742,9)	(0.958,6)	
B	*	*	0.190	0.304	0.297	*	-0.067	0.080	0.003
			(0.669,8)	(0.387,9)	(0.583,10)		(0.633,10)	(0.814,10)	(0.323,6)
Caa	*	*	*	-0.013	0.056	-0.017	*	-0.074	0.590
				(1.609,9)	(0.834,9)	(0.438,10)		(0.420,10)	(1.605,10)
Ca	*	*	*	*	-0.098	0.520	0.233	*	0.308
					(0.506,8)	(0.557,10)	(0.550,10)		(0.531,10)
C	*	*	*	*	*	0.008	-0.005	0.411	*
						(0.802,7)	(0.706,10)	(0.942,10)	

Table A.3: The simulation results for η for $T = 50$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.129	-6.051	-6.671	*	*	*	*	*
		(0.341,10)	(0.323,9)	(0.423,5)					
Aa	-5.170	*	-5.233	-6.345	-7.080	*	*	*	*
	(0.414,10)		(0.493,10)	(0.868,10)	(0.375,6)				
A	-6.469	-5.324	*	-5.300	-6.701	-7.400	*	*	*
	(0.845,10)	(0.450,10)		(0.344,10)	(0.983,10)	(0.768,6)			
Baa	-7.458	-6.548	-5.015	*	-5.056	-6.095	-7.153	*	*
	(0.583,6)	(0.700,10)	(0.264,10)		(0.271,10)	(0.613,9)	(0.710,5)		
Ba	*	-7.441	-6.385	-5.249	*	-5.299	-6.440	-7.406	*
		(0.757,9)	(0.667,10)	(0.675,10)		(0.460,10)	(0.910,9)	(1.325,8)	
B	*	*	-7.218	-6.371	-5.142	*	-5.023	-6.315	-7.105
			(0.488,9)	(1.015,10)	(0.607,10)		(0.378,10)	(0.565,10)	(0.854,6)
Caa	*	*	*	-7.099	-6.037	-5.232	*	-5.080	-6.197
				(0.729,9)	(0.340,10)	(0.413,10)		(0.427,10)	(0.603,10)
Ca	*	*	*	*	-6.957	-6.346	-5.225	*	-4.966
					(0.888,10)	(0.759,10)	(0.592,10)		(0.352,10)
C	*	*	*	*	*	-7.223	-6.404	-5.206	*
						(0.538,8)	(0.910,10)	(0.401,10)	

Table A.4: The simulation results for α for $T = 50$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.409	0.351	-0.348	*	*	*	*	*
		(0.465,10)	(0.525,9)	(0.899,5)					
Aa	0.210	*	0.347	0.249	-0.379	*	*	*	*
	(0.656,10)		(0.405,10)	(1.009,10)	(0.986,6)				
A	0.111	0.401	*	-0.105	0.234	-1.097	*	*	*
	(0.820,10)	(0.554,10)		(0.628,10)	(0.457,10)	(1.526,6)			
Baa	1.047	0.320	0.007	*	0.298	0.063	0.469	*	*
	(0.995,6)	(0.953,10)	(0.432,10)		(0.483,10)	(0.576,9)	(1.135,5)		
Ba	*	-0.106	0.636	0.558	*	0.105	0.871	-0.069	*
		(0.977,9)	(0.856,10)	(0.720,10)		(0.810,10)	(1.027,9)	(1.468,8)	
B	*	*	0.230	0.542	0.526	*	-0.037	0.417	0.257
			(0.758,9)	(0.967,10)	(0.907,10)		(0.341,10)	(0.575,10)	(0.917,6)
Caa	*	*	*	-0.273	0.223	0.279	*	0.081	0.604
				(0.676,9)	(0.466,10)	(0.402,10)		(0.355,10)	(0.875,10)
Ca	*	*	*	*	-0.223	-0.156	-0.007	*	0.125
					(1.246,10)	(0.676,10)	(0.810,10)		(0.591,10)
C	*	*	*	*	*	-0.022	0.306	0.021	*
						(0.868,8)	(0.995,10)	(0.406,10)	

Table A.5: The simulation results for η for $T = 50$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-4.974	-6.799	-7.257	*	*	*	*	*
		(0.147,10)	(1.156,10)	(0.652,8)					
Aa	-5.122	*	-4.909	-6.560	-6.952	*	*	*	*
	(0.255,10)		(0.283,10)	(0.734,10)	(0.565,7)				
A	-5.971	-5.016	*	-5.041	-6.058	-7.143	*	*	*
	(0.401,10)	(0.245,10)		(0.148,10)	(0.467,10)	(0.698,7)			
Baa	-7.712	-6.381	-5.277	*	-5.094	-6.645	-6.902	*	*
	(1.130,10)	(0.589,10)	(0.400,10)		(0.268,10)	(1.288,10)	(0.440,10)		
Ba	*	-8.175	-6.170	-5.007	*	-5.176	-6.332	-7.578	*
		(1.367,10)	(0.420,10)	(0.190,10)		(0.388,10)	(0.803,10)	(1.047,10)	
B	*	*	-6.941	-6.087	-5.078	*	-5.188	-6.055	-6.975
			(0.793,10)	(0.412,10)	(0.762,10)		(0.286,10)	(0.373,10)	(1.006,10)
Caa	*	*	*	-7.284	-6.328	-5.052	*	-4.964	-6.129
				(0.835,10)	(0.724,10)	(0.157,10)		(0.153,10)	(0.616,10)
Ca	*	*	*	*	-7.881	-6.301	-5.223	*	-5.227
					(1.409,10)	(0.557,10)	(0.339,10)		(0.345,10)
C	*	*	*	*	*	-7.956	-6.170	-5.254	*
						(1.434,10)	(0.551,10)	(0.520,10)	

Table A.6: The simulation results for α for $T = 50$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.037	0.645	-0.210	*	*	*	*	*
		(0.346,10)	(0.810,10)	(1.092,8)					
Aa	0.204	*	0.175	0.599	0.214	*	*	*	*
	(0.395,10)		(0.327,10)	(0.734,10)	(0.477,7)				
A	-0.244	0.361	*	0.009	0.019	0.473	*	*	*
	(0.612,10)	(0.530,10)		(0.445,10)	(0.880,10)	(1.072,7)			
Baa	0.845	0.169	0.178	*	0.156	-0.066	0.594	*	*
	(1.089,10)	(0.489,10)	(0.356,10)		(0.423,10)	(1.281,10)	(0.757,10)		
Ba	*	-0.034	0.718	0.341	*	0.246	-0.174	0.069	*
		(1.641,10)	(0.671,10)	(0.417,10)		(0.471,10)	(0.909,10)	(1.236,10)	
B	*	*	0.397	0.171	-0.117	*	0.242	-0.077	-0.598
			(0.897,10)	(0.833,10)	(0.737,10)		(0.325,10)	(0.559,10)	(1.427,10)
Caa	*	*	*	0.258	-0.077	-0.014	*	-0.140	-0.082
				(0.965,10)	(1.097,10)	(0.330,10)		(0.369,10)	(0.792,10)
Ca	*	*	*	*	-0.462	0.172	0.632	*	-0.024
					(1.780,10)	(0.798,10)	(0.642,10)		(0.713,10)
C	*	*	*	*	*	1.008	0.193	-0.264	*
						(1.416,10)	(0.723,10)	(0.770,10)	

Table A.7: The simulation results for η for $T = 50$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-4.978	-6.204	-8.423	*	*	*	*	*
		(0.248,10)	(0.615,10)	(1.953,10)					
Aa	-5.232	*	-5.086	-5.964	-7.194	*	*	*	*
	(0.271,10)		(0.129,10)	(0.391,10)	(0.959,10)				
A	-6.074	-5.054	*	-5.140	-6.233	-7.506	*	*	*
	(0.112,10)	(0.241,10)		(0.290,10)	(0.791,10)	(0.635,8)			
Baa	-7.799	-6.316	-5.009	*	-5.011	-6.377	-7.403	*	*
	(1.142,10)	(0.551,10)	(0.192,10)		(0.231,10)	(0.582,10)	(1.311,10)		
Ba	*	-6.481	-6.116	-5.003	*	-4.869	-6.884	-7.374	*
		(0.677,9)	(0.247,10)	(0.246,10)		(0.228,10)	(1.131,10)	(1.023,10)	
B	*	*	-6.884	-6.075	-5.082	*	-5.237	-6.120	-6.884
			(0.614,10)	(0.570,10)	(0.202,10)		(0.323,10)	(0.375,10)	(0.490,10)
Caa	*	*	*	-7.852	-6.068	-5.234	*	-5.112	-6.390
				(1.594,10)	(0.705,10)	(0.374,10)		(0.232,10)	(0.708,10)
Ca	*	*	*	*	-7.529	-5.881	-5.203	*	-4.996
					(1.130,10)	(0.400,10)	(0.508,10)		(0.260,10)
C	*	*	*	*	*	-7.322	-6.154	-4.991	*
						(0.701,10)	(0.675,10)	(0.215,10)	

Table A.8: The simulation results for α for $T = 50$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.221	0.230	1.006	*	*	*	*	*
		(0.187,10)	(0.620,10)	(1.923,10)					
Aa	0.371	*	0.117	0.182	0.300	*	*	*	*
	(0.436,10)		(0.188,10)	(0.437,10)	(0.969,10)				
A	0.062	-0.068	*	-0.001	0.099	-0.106	*	*	*
	(0.411,10)	(0.326,10)		(0.220,10)	(0.658,10)	(0.790,8)			
Baa	0.234	0.044	-0.059	*	0.196	0.155	-0.199	*	*
	(0.987,10)	(0.872,10)	(0.478,10)		(0.221,10)	(0.511,10)	(1.422,10)		
Ba	*	-0.117	-0.041	0.245	*	0.260	0.296	0.218	*
		(0.721,9)	(0.654,10)	(0.256,10)		(0.245,10)	(0.571,10)	(0.900,10)	
B	*	*	0.417	0.219	-0.082	*	0.223	-0.017	0.152
			(0.693,10)	(0.791,10)	(0.382,10)		(0.456,10)	(0.584,10)	(0.214,10)
Caa	*	*	*	0.224	-0.150	0.059	*	0.047	-0.166
				(1.490,10)	(0.536,10)	(0.473,10)		(0.366,10)	(0.799,10)
Ca	*	*	*	*	0.955	0.212	0.081	*	0.196
					(1.134,10)	(0.713,10)	(0.487,10)		(0.199,10)
C	*	*	*	*	*	0.223	0.210	0.036	*
						(0.568,10)	(0.865,10)	(0.314,10)	

Table A.9: The simulation results for η for $T = 100$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.012	-6.394	-7.307	*	*	*	*	*
		(0.269,10)	(0.630,10)	(0.640,5)					
Aa	-5.072	*	-5.191	-5.938	-7.373	*	*	*	*
	(0.223,10)		(0.363,10)	(0.330,10)	(0.681,7)				
A	-6.567	-5.114	*	-5.045	-6.544	-7.483	*	*	*
	(1.063,10)	(0.331,10)		(0.206,10)	(0.755,10)	(0.736,8)			
Baa	-6.889	-6.226	-4.952	*	-5.235	-6.413	-6.477	*	*
	(0.508,10)	(0.521,10)	(0.295,10)		(0.318,10)	(0.646,10)	(0.751,7)		
Ba	*	-6.814	-6.082	-5.189	*	-5.155	-6.301	-7.052	*
		(0.385,9)	(0.402,10)	(0.423,10)		(0.320,10)	(0.597,10)	(0.380,10)	
B	*	*	-7.033	-6.154	-4.894	*	-5.250	-6.258	-6.454
			(0.454,10)	(0.614,10)	(0.383,10)		(0.422,10)	(0.594,10)	(0.756,8)
Caa	*	*	*	-7.263	-5.967	-5.162	*	-5.214	-6.082
				(0.546,10)	(0.457,10)	(0.405,10)		(0.330,10)	(0.226,10)
Ca	*	*	*	*	-7.124	-6.149	-5.139	*	-5.128
					(0.475,9)	(0.384,10)	(0.287,10)		(0.358,10)
C	*	*	*	*	*	-7.134	-6.265	-4.951	*
						(0.701,10)	(0.606,10)	(0.217,10)	

Table A.10: The simulation results for α for $T = 100$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.415	0.409	0.590	*	*	*	*	*
		(0.679,10)	(0.617,10)	(0.637,5)					
Aa	0.148	*	-0.185	-0.026	0.065	*	*	*	*
	(0.414,10)		(0.719,10)	(0.560,10)	(0.443,7)				
A	0.908	0.585	*	-0.218	-0.128	-0.148	*	*	*
	(1.334,10)	(0.840,10)		(0.608,10)	(0.826,10)	(0.590,8)			
Baa	-0.418	0.562	0.114	*	0.460	0.152	0.209	*	*
	(1.103,10)	(0.844,10)	(0.286,10)		(0.670,10)	(0.218,10)	(0.534,7)		
Ba	*	0.086	-0.153	0.467	*	0.425	0.304	0.256	*
		(0.337,9)	(0.671,10)	(0.676,10)		(0.528,10)	(0.415,10)	(0.635,10)	
B	*	*	0.004	0.214	0.051	*	0.319	0.241	0.150
			(0.924,10)	(0.556,10)	(0.205,10)		(0.491,10)	(0.261,10)	(0.576,8)
Caa	*	*	*	-0.159	0.209	0.420	*	0.027	-0.071
				(0.686,10)	(0.879,10)	(0.634,10)		(0.325,10)	(0.483,10)
Ca	*	*	*	*	-0.417	-0.125	-0.027	*	0.548
					(1.148,9)	(0.767,10)	(0.453,10)		(1.016,10)
C	*	*	*	*	*	-0.106	0.258	0.097	*
						(0.631,10)	(0.256,10)	(0.317,10)	

Table A.11: The simulation results for η for $T = 100$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.140	-5.966	-7.301	*	*	*	*	*
		(0.264,10)	(0.312,10)	(0.796,10)					
Aa	-4.982	*	-4.885	-6.068	-7.310	*	*	*	*
	(0.074,10)		(0.234,10)	(0.486,10)	(0.377,10)				
A	-6.272	-5.035	*	-4.853	-6.202	-7.374	*	*	*
	(0.506,10)	(0.195,10)		(0.302,10)	(0.442,10)	(0.625,10)			
Baa	-6.948	-6.156	-4.927	*	-5.070	-6.012	-7.656	*	*
	(0.255,10)	(0.203,10)	(0.216,10)		(0.296,10)	(0.360,10)	(0.881,10)		
Ba	*	-7.131	-6.299	-4.992	*	-5.158	-6.501	-7.031	*
		(0.538,10)	(0.612,10)	(0.152,10)		(0.265,10)	(0.553,10)	(0.462,10)	
B	*	*	-7.082	-5.763	-5.081	*	-4.958	-6.439	-6.855
			(0.727,10)	(0.426,10)	(0.310,10)		(0.282,10)	(0.714,10)	(1.007,8)
Caa	*	*	*	-7.075	-6.086	-5.047	*	-4.845	-5.975
				(0.491,10)	(0.359,10)	(0.134,10)		(0.167,10)	(0.301,10)
Ca	*	*	*	*	-7.367	-6.338	-5.315	*	-5.091
					(0.714,10)	(0.580,10)	(0.319,10)		(0.217,10)
C	*	*	*	*	*	-7.467	-5.900	-5.170	*
						(1.066,10)	(0.115,10)	(0.230,10)	

Table A.12: The simulation results for α for $T = 100$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.027	0.155	-0.148	*	*	*	*	*
		(0.584,10)	(0.437,10)	(0.823,10)					
Aa	-0.304	*	0.099	0.241	0.282	*	*	*	*
	(0.663,10)		(0.315,10)	(0.231,10)	(0.178,10)				
A	0.264	-0.062	*	0.136	-0.567	0.113	*	*	*
	(0.484,10)	(0.255,10)		(0.118,10)	(1.234,10)	(0.926,10)			
Baa	0.040	0.286	0.130	*	0.059	-0.086	0.113	*	*
	(0.567,10)	(0.705,10)	(0.295,10)		(0.424,10)	(0.532,10)	(0.651,10)		
Ba	*	-0.340	0.138	0.221	*	0.547	0.067	0.133	*
		(1.019,10)	(0.575,10)	(0.263,10)		(0.658,10)	(0.791,10)	(0.416,10)	
B	*	*	-0.893	0.245	0.019	*	-0.008	-0.146	0.663
			(1.745,10)	(0.423,10)	(0.226,10)		(0.245,10)	(1.110,10)	(0.930,8)
Caa	*	*	*	0.394	-0.239	0.181	*	-0.181	-0.025
				(0.526,10)	(0.919,10)	(0.216,10)		(0.429,10)	(0.663,10)
Ca	*	*	*	*	0.287	0.753	0.181	*	0.410
					(0.740,10)	(0.850,10)	(0.457,10)		(0.373,10)
C	*	*	*	*	*	0.900	0.436	0.245	*
						(1.540,10)	(0.401,10)	(0.252,10)	

Table A.13: The simulation results for η for $T = 100$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.048	-6.274	-7.396	*	*	*	*	*
		(0.194,10)	(0.537,10)	(0.660,10)					
Aa	-5.193	*	-5.036	-6.130	-7.179	*	*	*	*
	(0.326,10)		(0.244,10)	(0.377,10)	(0.667,10)				
A	-6.084	-5.193	*	-5.019	-6.142	-7.316	*	*	*
	(0.261,10)	(0.306,10)		(0.148,10)	(0.307,10)	(0.665,10)			
Baa	-7.348	-5.972	-4.966	*	-5.015	-6.035	-7.269	*	*
	(0.679,10)	(0.249,10)	(0.176,10)		(0.136,10)	(0.337,10)	(0.629,10)		
Ba	*	-7.449	-6.016	-4.979	*	-5.266	-6.306	-7.206	*
		(0.920,9)	(0.213,10)	(0.157,10)		(0.306,10)	(0.515,10)	(0.726,10)	
B	*	*	-7.294	-6.093	-5.116	*	-5.106	-6.063	-7.108
			(0.796,10)	(0.425,10)	(0.220,10)		(0.193,10)	(0.292,10)	(0.292,10)
Caa	*	*	*	-7.165	-6.390	-5.096	*	-5.108	-6.132
				(0.600,10)	(0.627,10)	(0.240,10)		(0.246,10)	(0.236,10)
Ca	*	*	*	*	-7.374	-6.151	-5.001	*	-5.073
					(0.773,10)	(0.325,10)	(0.189,10)		(0.215,10)
C	*	*	*	*	*	-7.385	-6.242	-5.057	*
						(0.722,10)	(0.490,10)	(0.217,10)	

Table A.14: The simulation results for α for $T = 100$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.108	0.509	-0.063	*	*	*	*	*
		(0.155,10)	(0.818,10)	(1.168,10)					
Aa	0.470	*	0.284	0.039	0.326	*	*	*	*
	(0.625,10)		(0.329,10)	(0.195,10)	(0.652,10)				
A	0.431	0.256	*	0.152	0.221	-0.113	*	*	*
	(0.613,10)	(0.401,10)		(0.253,10)	(0.498,10)	(0.793,10)			
Baa	0.263	-0.146	0.031	*	-0.111	0.153	-0.096	*	*
	(0.730,10)	(0.835,10)	(0.384,10)		(0.401,10)	(0.549,10)	(0.869,10)		
Ba	*	0.202	0.267	0.017	*	0.437	0.192	0.226	*
		(1.093,9)	(0.524,10)	(0.412,10)		(0.760,10)	(0.737,10)	(0.921,10)	
B	*	*	0.550	0.026	0.125	*	0.017	0.159	0.084
			(1.036,10)	(0.676,10)	(0.460,10)		(0.272,10)	(0.585,10)	(0.705,10)
Caa	*	*	*	-0.069	0.037	0.150	*	0.040	0.017
				(0.555,10)	(0.643,10)	(0.427,10)		(0.194,10)	(0.695,10)
Ca	*	*	*	*	0.351	0.274	0.128	*	-0.116
					(0.397,10)	(0.524,10)	(0.299,10)		(0.449,10)
C	*	*	*	*	*	0.081	0.056	0.005	*
						(0.793,10)	(0.637,10)	(0.449,10)	

Table A.15: The simulation results for η for $T = 100$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-4.964	-5.956	-6.987	*	*	*	*	*
		(0.089,10)	(0.235,10)	(0.394,10)					
Aa	-4.978	*	-4.955	-6.052	-7.248	*	*	*	*
	(0.183,10)		(0.056,10)	(0.237,10)	(0.990,10)				
A	-6.008	-5.068	*	-5.017	-6.073	-7.014	*	*	*
	(0.161,10)	(0.130,10)		(0.035,10)	(0.362,10)	(0.432,10)			
Baa	-6.953	-5.989	-5.048	*	-5.034	-6.373	-7.328	*	*
	(0.094,10)	(0.255,10)	(0.154,10)		(0.200,10)	(0.775,10)	(0.387,10)		
Ba	*	-7.469	-6.061	-5.076	*	-5.174	-5.979	-6.994	*
		(1.076,10)	(0.100,10)	(0.116,10)		(0.236,10)	(0.276,10)	(0.309,10)	
B	*	*	-7.735	-6.016	-5.186	*	-4.935	-6.150	-7.337
			(0.846,10)	(0.146,10)	(0.275,10)		(0.082,10)	(0.399,10)	(0.437,10)
Caa	*	*	*	-6.773	-6.077	-5.096	*	-4.961	-6.276
				(0.346,10)	(0.223,10)	(0.142,10)		(0.168,10)	(0.638,10)
Ca	*	*	*	*	-7.867	-6.102	-5.044	*	-5.050
					(1.036,10)	(0.358,10)	(0.113,10)		(0.099,10)
C	*	*	*	*	*	-6.927	-6.430	-5.216	*
						(0.339,10)	(0.460,10)	(0.270,10)	

Table A.16: The simulation results for α for $T = 100$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-0.014	0.045	-0.219	*	*	*	*	*
		(0.205,10)	(0.382,10)	(0.729,10)					
Aa	-0.221	*	-0.006	0.277	-0.460	*	*	*	*
	(0.348,10)		(0.193,10)	(0.580,10)	(0.966,10)				
A	0.392	0.265	*	0.182	0.584	-0.029	*	*	*
	(0.392,10)	(0.353,10)		(0.362,10)	(0.795,10)	(0.526,10)			
Baa	-0.025	0.096	0.165	*	0.267	-0.496	0.643	*	*
	(0.455,10)	(0.472,10)	(0.153,10)		(0.308,10)	(1.007,10)	(0.553,10)		
Ba	*	-0.343	0.079	0.053	*	0.447	0.362	0.011	*
		(1.230,10)	(0.313,10)	(0.092,10)		(0.527,10)	(0.422,10)	(0.293,10)	
B	*	*	0.055	-0.001	-0.252	*	0.229	0.415	-1.271
			(1.225,10)	(0.807,10)	(0.543,10)		(0.240,10)	(0.527,10)	(1.870,10)
Caa	*	*	*	0.058	0.032	0.334	*	0.246	-0.937
				(0.588,10)	(0.716,10)	(0.280,10)		(0.415,10)	(1.625,10)
Ca	*	*	*	*	0.737	0.571	-0.031	*	0.118
					(0.764,10)	(0.512,10)	(0.223,10)		(0.308,10)
C	*	*	*	*	*	0.048	-0.058	0.136	*
						(0.687,10)	(1.218,10)	(0.312,10)	

Table A.17: The simulation results for η for $T = 150$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.065	-5.759	-7.200	*	*	*	*	*
		(0.144,10)	(0.299,10)	(0.556,9)					
Aa	-5.134	*	-5.149	-6.022	-7.213	*	*	*	*
	(0.210,10)		(0.362,10)	(0.345,10)	(0.631,8)				
A	-6.248	-5.075	*	-5.003	-6.164	-7.391	*	*	*
	(0.536,10)	(0.135,10)		(0.242,10)	(0.445,10)	(0.627,8)			
Baa	-7.111	-6.129	-5.059	*	-5.213	-5.939	-6.865	*	*
	(0.390,8)	(0.247,10)	(0.235,10)		(0.239,10)	(0.266,10)	(0.620,9)		
Ba	*	-6.949	-6.194	-5.124	*	-5.063	-6.219	-7.227	*
		(0.639,10)	(0.354,10)	(0.218,10)		(0.383,10)	(0.587,10)	(0.604,10)	
B	*	*	-7.335	-6.056	-5.083	*	-5.100	-6.000	-7.148
			(0.627,10)	(0.217,10)	(0.231,10)		(0.217,10)	(0.517,10)	(0.609,10)
Caa	*	*	*	-7.236	-6.100	-5.066	*	-5.202	-6.495
				(0.660,10)	(0.419,10)	(0.208,10)		(0.325,10)	(0.810,10)
Ca	*	*	*	*	-7.196	-6.019	-5.256	*	-5.017
					(0.674,9)	(0.275,10)	(0.322,10)		(0.222,10)
C	*	*	*	*	*	-7.054	-5.973	-4.906	*
						(0.607,10)	(0.380,10)	(0.224,10)	

Table A.18: The simulation results for α for $T = 150$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.150	0.234	0.305	*	*	*	*	*
		(0.147,10)	(0.439,10)	(0.324,9)					
Aa	0.099	*	-0.015	0.377	0.160	*	*	*	*
	(0.445,10)		(0.434,10)	(0.354,10)	(0.740,8)				
A	0.332	0.005	*	0.362	-0.110	0.054	*	*	*
	(0.566,10)	(0.343,10)		(0.509,10)	(0.451,10)	(0.267,8)			
Baa	-0.049	0.393	-0.011	*	0.420	0.329	0.029	*	*
	(0.612,8)	(0.539,10)	(0.733,10)		(0.580,10)	(0.372,10)	(0.615,9)		
Ba	*	0.378	0.316	0.083	*	0.476	0.226	0.169	*
		(0.560,10)	(0.348,10)	(0.272,10)		(0.783,10)	(0.376,10)	(0.563,10)	
B	*	*	-0.103	0.036	0.413	*	0.355	0.061	-0.093
			(0.629,10)	(0.458,10)	(0.525,10)		(0.402,10)	(0.319,10)	(0.512,10)
Caa	*	*	*	0.145	-0.101	0.092	*	0.070	0.172
				(0.746,10)	(0.637,10)	(0.231,10)		(0.195,10)	(0.489,10)
Ca	*	*	*	*	0.060	0.286	0.282	*	0.007
					(0.635,9)	(0.325,10)	(0.506,10)		(0.447,10)
C	*	*	*	*	*	0.195	0.266	0.081	*
						(0.890,10)	(0.387,10)	(0.414,10)	

Table A.19: The simulation results for η for $T = 150$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.045	-6.135	-7.215	*	*	*	*	*
		(0.228,10)	(0.181,10)	(0.702,10)					
Aa	-5.081	*	-5.000	-6.149	-7.305	*	*	*	*
	(0.164,10)		(0.137,10)	(0.306,10)	(0.848,10)				
A	-6.074	-5.047	*	-5.044	-5.914	-6.330	*	*	*
	(0.263,10)	(0.201,10)		(0.167,10)	(0.211,10)	(0.264,9)			
Baa	-7.195	-6.157	-5.062	*	-5.113	-6.159	-7.396	*	*
	(0.453,10)	(0.245,10)	(0.144,10)		(0.179,10)	(0.255,10)	(0.672,10)		
Ba	*	-6.205	-5.962	-5.066	*	-5.162	-6.153	-6.590	*
		(0.338,9)	(0.272,10)	(0.352,10)		(0.253,10)	(0.388,10)	(0.506,9)	
B	*	*	-7.204	-6.257	-4.987	*	-5.040	-6.204	-6.982
			(0.510,10)	(0.412,10)	(0.199,10)		(0.096,10)	(0.418,10)	(0.711,10)
Caa	*	*	*	-7.374	-5.949	-5.123	*	-5.080	-6.190
				(0.664,10)	(0.182,10)	(0.221,10)		(0.233,10)	(0.408,10)
Ca	*	*	*	*	-7.079	-5.984	-5.082	*	-5.083
					(0.291,10)	(0.335,10)	(0.160,10)		(0.202,10)
C	*	*	*	*	*	-7.252	-6.106	-5.101	*
						(0.738,10)	(0.267,10)	(0.175,10)	

Table A.20: The simulation results for α for $T = 150$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.163	0.419	-0.032	*	*	*	*	*
		(0.269,10)	(0.538,10)	(0.674,10)					
Aa	0.025	*	-0.025	-0.124	0.607	*	*	*	*
	(0.234,10)		(0.263,10)	(0.820,10)	(1.226,10)				
A	0.284	0.239	*	-0.050	0.241	0.013	*	*	*
	(0.438,10)	(0.376,10)		(0.312,10)	(0.273,10)	(0.712,9)			
Baa	0.241	-0.471	0.082	*	0.128	0.072	0.369	*	*
	(0.590,10)	(0.850,10)	(0.440,10)		(0.221,10)	(0.354,10)	(0.968,10)		
Ba	*	0.170	0.057	0.285	*	0.007	0.041	0.104	*
		(0.384,9)	(0.409,10)	(0.462,10)		(0.507,10)	(0.810,10)	(0.631,9)	
B	*	*	0.202	0.427	-0.018	*	0.192	-0.002	-0.182
			(0.729,10)	(0.652,10)	(0.270,10)		(0.348,10)	(0.508,10)	(0.744,10)
Caa	*	*	*	0.042	0.257	0.016	*	0.069	-0.229
				(0.607,10)	(0.409,10)	(0.369,10)		(0.388,10)	(0.765,10)
Ca	*	*	*	*	0.313	0.112	0.066	*	0.090
					(0.564,10)	(0.472,10)	(0.403,10)		(0.307,10)
C	*	*	*	*	*	0.011	0.033	0.012	*
						(0.964,10)	(0.350,10)	(0.240,10)	

Table A.21: The simulation results for η for $T = 150$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.149	-6.072	-7.205	*	*	*	*	*
		(0.234,10)	(0.230,10)	(0.488,10)					
Aa	-4.966	*	-4.988	-6.339	-7.004	*	*	*	*
	(0.135,10)		(0.119,10)	(0.464,10)	(0.466,9)				
A	-6.024	-5.027	*	-5.001	-6.209	-7.179	*	*	*
	(0.382,10)	(0.206,10)		(0.194,10)	(0.416,10)	(0.569,10)			
Baa	-7.104	-6.066	-5.006	*	-5.160	-6.059	-7.152	*	*
	(0.433,10)	(0.322,10)	(0.120,10)		(0.221,10)	(0.293,10)	(0.419,10)		
Ba	*	-7.082	-6.064	-5.039	*	-5.056	-6.232	-7.204	*
		(0.512,10)	(0.247,10)	(0.180,10)		(0.124,10)	(0.489,10)	(0.641,10)	
B	*	*	-7.443	-5.996	-5.083	*	-5.046	-6.127	-7.193
			(0.842,10)	(0.229,10)	(0.168,10)		(0.124,10)	(0.259,10)	(0.473,10)
Caa	*	*	*	-7.337	-6.050	-5.069	*	-5.042	-6.020
				(0.682,10)	(0.346,10)	(0.099,10)		(0.161,10)	(0.431,10)
Ca	*	*	*	*	-7.209	-6.163	-4.976	*	-4.999
					(0.584,10)	(0.388,10)	(0.181,10)		(0.161,10)
C	*	*	*	*	*	-7.163	-6.034	-5.068	*
						(0.504,10)	(0.228,10)	(0.170,10)	

Table A.22: The simulation results for α for $T = 150$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.157	0.162	0.074	*	*	*	*	*
		(0.537,10)	(0.637,10)	(0.836,10)					
Aa	0.071	*	-0.090	0.391	0.364	*	*	*	*
	(0.483,10)		(0.402,10)	(0.925,10)	(0.431,9)				
A	0.244	0.139	*	0.226	0.609	0.563	*	*	*
	(0.362,10)	(0.544,10)		(0.272,10)	(0.835,10)	(1.177,10)			
Baa	0.719	0.015	-0.036	*	0.057	0.480	-0.112	*	*
	(0.869,10)	(0.860,10)	(0.388,10)		(0.595,10)	(0.584,10)	(1.006,10)		
Ba	*	-0.019	0.161	-0.093	*	0.199	0.644	0.567	*
		(0.674,10)	(0.769,10)	(0.466,10)		(0.418,10)	(0.742,10)	(1.347,10)	
B	*	*	-0.092	0.151	0.328	*	0.220	0.357	-0.275
			(1.031,10)	(0.332,10)	(0.487,10)		(0.429,10)	(0.605,10)	(1.027,10)
Caa	*	*	*	-0.141	0.284	0.145	*	-0.057	0.545
				(1.016,10)	(0.815,10)	(0.482,10)		(0.484,10)	(0.799,10)
Ca	*	*	*	*	-0.197	0.370	0.255	*	-0.065
					(1.016,10)	(0.584,10)	(0.396,10)		(0.315,10)
C	*	*	*	*	*	-0.122	0.320	0.181	*
						(1.055,10)	(0.526,10)	(0.394,10)	

Table A.23: The simulation results for η for $T = 150$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.065	-6.174	-7.318	*	*	*	*	*
		(0.152,10)	(0.344,10)	(0.601,10)					
Aa	-4.970	*	-5.007	-6.086	-6.997	*	*	*	*
	(0.132,10)		(0.116,10)	(0.208,10)	(0.376,10)				
A	-6.183	-4.966	*	-5.090	-6.078	-7.240	*	*	*
	(0.577,10)	(0.129,10)		(0.231,10)	(0.402,10)	(0.524,10)			
Baa	-7.357	-6.077	-5.055	*	-5.065	-6.212	-7.660	*	*
	(0.585,10)	(0.217,10)	(0.151,10)		(0.141,10)	(0.525,10)	(0.883,10)		
Ba	*	-7.298	-6.057	-5.062	*	-5.065	-6.072	-7.167	*
		(0.506,10)	(0.356,10)	(0.172,10)		(0.093,10)	(0.249,10)	(0.608,10)	
B	*	*	-7.148	-6.101	-5.022	*	-5.071	-5.997	-6.907
			(0.577,10)	(0.295,10)	(0.144,10)		(0.183,10)	(0.209,10)	(0.569,10)
Caa	*	*	*	-7.421	-6.099	-5.029	*	-5.078	-6.177
				(0.763,10)	(0.332,10)	(0.172,10)		(0.210,10)	(0.239,10)
Ca	*	*	*	*	-7.309	-6.213	-5.027	*	-5.080
					(0.562,10)	(0.318,10)	(0.130,10)		(0.130,10)
C	*	*	*	*	*	-7.209	-6.153	-5.043	*
						(0.345,10)	(0.219,10)	(0.204,10)	

Table A.24: The simulation results for α for $T = 150$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.030	-0.149	0.144	*	*	*	*	*
		(0.336,10)	(0.956,10)	(1.007,10)					
Aa	-0.085	*	0.071	0.331	0.095	*	*	*	*
	(0.316,10)		(0.354,10)	(0.449,10)	(0.692,10)				
A	0.110	-0.040	*	0.085	0.518	-0.405	*	*	*
	(0.946,10)	(0.352,10)		(0.365,10)	(0.741,10)	(1.289,10)			
Baa	0.654	0.148	-0.043	*	0.248	0.385	0.285	*	*
	(0.879,10)	(0.269,10)	(0.468,10)		(0.518,10)	(0.892,10)	(1.338,10)		
Ba	*	0.973	0.406	0.067	*	-0.114	0.345	0.407	*
		(0.858,10)	(0.669,10)	(0.319,10)		(0.407,10)	(0.390,10)	(0.837,10)	
B	*	*	-0.179	0.319	0.179	*	0.245	-0.048	0.323
			(0.740,10)	(0.560,10)	(0.464,10)		(0.582,10)	(0.618,10)	(0.483,10)
Caa	*	*	*	0.133	-0.028	0.088	*	0.083	0.014
				(1.403,10)	(0.533,10)	(0.387,10)		(0.308,10)	(0.849,10)
Ca	*	*	*	*	0.726	-0.051	-0.101	*	0.002
					(0.641,10)	(0.863,10)	(0.313,10)		(0.598,10)
C	*	*	*	*	*	-0.011	-0.349	0.197	*
						(0.781,10)	(0.907,10)	(0.410,10)	

Table A.25: The simulation results for η for $T = 200$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-4.979	-6.046	-7.034	*	*	*	*	*
		(0.204,10)	(0.329,10)	(0.500,8)					
Aa	-5.060	*	-5.037	-6.102	-7.262	*	*	*	*
	(0.131,10)		(0.214,10)	(0.198,10)	(0.598,10)				
A	-6.267	-4.982	*	-5.074	-6.160	-7.162	*	*	*
	(0.461,10)	(0.120,10)		(0.333,10)	(0.393,10)	(0.458,10)			
Baa	-7.218	-6.282	-5.101	*	-4.995	-6.182	-7.056	*	*
	(0.723,10)	(0.460,10)	(0.169,10)		(0.156,10)	(0.429,10)	(0.517,10)		
Ba	*	-7.160	-5.977	-5.039	*	-5.088	-6.184	-6.992	*
		(0.395,10)	(0.344,10)	(0.191,10)		(0.168,10)	(0.648,10)	(0.404,10)	
B	*	*	-6.767	-6.098	-4.972	*	-5.147	-6.399	-6.862
			(0.512,10)	(0.264,10)	(0.127,10)		(0.221,10)	(0.572,10)	(0.449,10)
Caa	*	*	*	-7.211	-6.198	-5.143	*	-4.923	-6.063
				(0.666,10)	(0.448,10)	(0.289,10)		(0.158,10)	(0.287,10)
Ca	*	*	*	*	-7.120	-6.117	-5.006	*	-5.002
					(0.488,10)	(0.523,10)	(0.134,10)		(0.135,10)
C	*	*	*	*	*	-6.956	-6.024	-5.038	*
						(0.319,10)	(0.243,10)	(0.283,10)	

Table A.26: The simulation results for α for $T = 200$, $K = 200$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.293	0.117	0.093	*	*	*	*	*
		(0.309,10)	(0.283,10)	(0.607,8)					
Aa	0.279	*	0.040	-0.004	0.090	*	*	*	*
	(0.304,10)		(0.324,10)	(0.560,10)	(0.325,10)				
A	0.330	0.228	*	0.478	0.043	0.193	*	*	*
	(0.399,10)	(0.421,10)		(0.666,10)	(0.615,10)	(0.420,10)			
Baa	-0.173	-0.159	0.195	*	0.196	0.208	-0.050	*	*
	(0.674,10)	(0.609,10)	(0.142,10)		(0.401,10)	(0.322,10)	(0.668,10)		
Ba	*	0.042	-0.262	0.072	*	0.186	-0.017	0.111	*
		(0.562,10)	(0.766,10)	(0.281,10)		(0.189,10)	(0.459,10)	(0.446,10)	
B	*	*	0.145	0.188	0.110	*	0.258	0.318	0.112
			(0.190,10)	(0.408,10)	(0.176,10)		(0.392,10)	(0.387,10)	(0.353,10)
Caa	*	*	*	0.410	0.233	0.117	*	0.159	0.169
				(0.507,10)	(0.527,10)	(0.162,10)		(0.144,10)	(0.338,10)
Ca	*	*	*	*	-0.079	0.186	0.086	*	0.132
					(0.514,10)	(0.507,10)	(0.204,10)		(0.226,10)
C	*	*	*	*	*	0.073	0.096	0.137	*
						(0.426,10)	(0.142,10)	(0.416,10)	

Table A.27: The simulation results for η for $T = 200$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-4.934	-6.198	-7.129	*	*	*	*	*
		(0.194,10)	(0.297,10)	(0.474,10)					
Aa	-5.081	*	-5.014	-6.043	-7.121	*	*	*	*
	(0.120,10)		(0.072,10)	(0.148,10)	(0.411,10)				
A	-6.063	-4.968	*	-4.948	-6.375	-7.057	*	*	*
	(0.298,10)	(0.111,10)		(0.140,10)	(0.530,10)	(0.549,10)			
Baa	-7.332	-6.069	-5.019	*	-5.109	-6.099	-7.192	*	*
	(0.621,10)	(0.204,10)	(0.134,10)		(0.170,10)	(0.299,10)	(0.452,10)		
Ba	*	-6.902	-6.058	-5.108	*	-4.993	-6.160	-7.310	*
		(0.497,10)	(0.153,10)	(0.228,10)		(0.151,10)	(0.424,10)	(0.596,10)	
B	*	*	-7.201	-6.029	-5.074	*	-5.001	-6.053	-7.075
			(0.576,10)	(0.191,10)	(0.178,10)		(0.169,10)	(0.299,10)	(0.381,10)
Caa	*	*	*	-6.978	-5.985	-4.979	*	-5.024	-5.987
				(0.409,10)	(0.254,10)	(0.163,10)		(0.111,10)	(0.205,10)
Ca	*	*	*	*	-6.910	-5.840	-4.983	*	-5.000
					(0.512,10)	(0.233,10)	(0.143,10)		(0.145,10)
C	*	*	*	*	*	-7.045	-5.892	-4.915	*
						(0.507,10)	(0.193,10)	(0.179,10)	

Table A.28: The simulation results for α for $T = 200$, $K = 300$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-0.005	0.124	0.562	*	*	*	*	*
		(0.260,10)	(0.474,10)	(0.709,10)					
Aa	0.133	*	0.117	-0.081	0.208	*	*	*	*
	(0.365,10)		(0.346,10)	(0.497,10)	(0.673,10)				
A	-0.036	0.070	*	0.156	-0.372	-0.158	*	*	*
	(0.413,10)	(0.316,10)		(0.166,10)	(0.973,10)	(0.836,10)			
Baa	-0.030	0.223	0.148	*	0.082	0.296	0.043	*	*
	(1.191,10)	(0.695,10)	(0.218,10)		(0.394,10)	(0.802,10)	(0.575,10)		
Ba	*	0.248	0.246	0.087	*	0.232	-0.193	0.430	*
		(0.500,10)	(0.244,10)	(0.406,10)		(0.287,10)	(0.713,10)	(0.745,10)	
B	*	*	0.526	-0.067	0.201	*	-0.008	0.238	0.285
			(0.657,10)	(0.527,10)	(0.614,10)		(0.410,10)	(0.665,10)	(0.761,10)
Caa	*	*	*	0.124	-0.004	0.164	*	0.249	0.148
				(0.628,10)	(0.621,10)	(0.384,10)		(0.300,10)	(0.230,10)
Ca	*	*	*	*	-0.151	0.155	0.224	*	0.281
					(0.914,10)	(0.226,10)	(0.307,10)		(0.244,10)
C	*	*	*	*	*	0.281	0.100	0.051	*
						(0.516,10)	(0.477,10)	(0.272,10)	

Table A.29: The simulation results for η for $T = 200$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.023	-6.003	-7.109	*	*	*	*	*
		(0.121,10)	(0.228,10)	(0.459,10)					
Aa	-5.111	*	-5.021	-6.089	-7.415	*	*	*	*
	(0.169,10)		(0.078,10)	(0.213,10)	(0.608,10)				
A	-6.210	-5.092	*	-5.034	-6.099	-7.197	*	*	*
	(0.343,10)	(0.151,10)		(0.105,10)	(0.238,10)	(0.573,10)			
Baa	-7.043	-6.103	-5.036	*	-5.044	-6.069	-7.192	*	*
	(0.350,10)	(0.214,10)	(0.162,10)		(0.183,10)	(0.210,10)	(0.492,10)		
Ba	*	-7.311	-6.127	-5.120	*	-5.101	-6.334	-7.262	*
		(0.604,10)	(0.391,10)	(0.208,10)		(0.194,10)	(0.591,10)	(0.536,10)	
B	*	*	-7.135	-6.076	-5.013	*	-5.096	-6.169	-7.166
			(0.329,10)	(0.305,10)	(0.142,10)		(0.192,10)	(0.332,10)	(0.567,10)
Caa	*	*	*	-7.290	-6.061	-5.131	*	-5.013	-6.121
				(0.379,10)	(0.258,10)	(0.194,10)		(0.113,10)	(0.234,10)
Ca	*	*	*	*	-7.346	-6.115	-5.110	*	-5.045
					(0.623,10)	(0.290,10)	(0.181,10)		(0.100,10)
C	*	*	*	*	*	-7.373	-6.093	-5.129	*
						(0.775,10)	(0.329,10)	(0.233,10)	

Table A.30: The simulation results for α for $T = 200$, $K = 400$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.162	0.176	0.407	*	*	*	*	*
		(0.484,10)	(0.660,10)	(0.902,10)					
Aa	0.092	*	-0.009	0.159	-0.332	*	*	*	*
	(0.411,10)		(0.473,10)	(0.472,10)	(1.347,10)				
A	0.387	0.270	*	0.134	0.303	0.457	*	*	*
	(0.804,10)	(0.359,10)		(0.454,10)	(0.571,10)	(0.963,10)			
Baa	0.299	0.040	0.433	*	-0.044	0.205	0.029	*	*
	(0.985,10)	(0.364,10)	(0.669,10)		(0.425,10)	(0.453,10)	(0.985,10)		
Ba	*	0.143	0.247	-0.008	*	0.283	0.463	0.325	*
		(0.962,10)	(0.837,10)	(0.352,10)		(0.492,10)	(0.720,10)	(1.179,10)	
B	*	*	0.209	-0.040	0.383	*	0.185	0.072	-0.512
			(1.012,10)	(0.802,10)	(0.478,10)		(0.482,10)	(0.547,10)	(1.181,10)
Caa	*	*	*	0.550	0.173	0.091	*	0.381	-0.248
				(0.694,10)	(0.456,10)	(0.400,10)		(0.385,10)	(0.667,10)
Ca	*	*	*	*	0.411	0.142	0.480	*	0.231
					(0.956,10)	(0.924,10)	(0.558,10)		(0.333,10)
C	*	*	*	*	*	0.570	0.278	-0.021	*
						(0.870,10)	(0.549,10)	(0.519,10)	

Table A.31: The simulation results for η for $T = 200$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	-5.009	-6.066	-7.129	*	*	*	*	*
		(0.105,10)	(0.235,10)	(0.326,10)					
Aa	-4.963	*	-5.026	-6.060	-6.989	*	*	*	*
	(0.130,10)		(0.102,10)	(0.143,10)	(0.369,10)				
A	-6.055	-4.973	*	-5.053	-6.132	-7.323	*	*	*
	(0.169,10)	(0.104,10)		(0.100,10)	(0.344,10)	(0.501,10)			
Baa	-7.100	-6.250	-5.071	*	-5.089	-6.135	-7.405	*	*
	(0.555,10)	(0.411,10)	(0.149,10)		(0.193,10)	(0.223,10)	(0.700,10)		
Ba	*	-7.259	-6.168	-5.095	*	-5.053	-6.151	-7.317	*
		(0.525,10)	(0.279,10)	(0.159,10)		(0.147,10)	(0.218,10)	(0.532,10)	
B	*	*	-7.185	-5.958	-5.061	*	-5.057	-6.108	-7.189
			(0.506,10)	(0.171,10)	(0.085,10)		(0.164,10)	(0.278,10)	(0.379,10)
Caa	*	*	*	-7.313	-6.052	-5.058	*	-5.060	-6.039
				(0.464,10)	(0.268,10)	(0.097,10)		(0.132,10)	(0.247,10)
Ca	*	*	*	*	-7.166	-6.137	-5.042	*	-5.165
					(0.475,10)	(0.319,10)	(0.116,10)		(0.221,10)
C	*	*	*	*	*	-7.292	-6.219	-5.067	*
						(0.683,10)	(0.313,10)	(0.087,10)	

Table A.32: The simulation results for α for $T = 200$, $K = 500$

	Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
Aaa	*	0.115	0.077	0.161	*	*	*	*	*
		(0.556,10)	(0.821,10)	(0.827,10)					
Aa	0.294	*	0.100	0.322	0.061	*	*	*	*
	(0.273,10)		(0.359,10)	(0.485,10)	(1.323,10)				
A	0.110	0.106	*	0.054	0.113	0.846	*	*	*
	(0.677,10)	(0.258,10)		(0.337,10)	(0.687,10)	(1.140,10)			
Baa	0.517	-0.086	0.348	*	0.148	-0.147	0.089	*	*
	(0.942,10)	(0.935,10)	(0.453,10)		(0.596,10)	(0.661,10)	(1.218,10)		
Ba	*	0.211	0.497	0.059	*	0.368	0.075	0.153	*
		(0.906,10)	(0.717,10)	(0.317,10)		(0.493,10)	(0.771,10)	(1.173,10)	
B	*	*	0.495	0.104	0.063	*	-0.009	-0.157	0.634
			(1.161,10)	(0.516,10)	(0.475,10)		(0.553,10)	(0.857,10)	(0.976,10)
Caa	*	*	*	0.010	0.276	0.086	*	-0.074	0.398
				(1.260,10)	(0.637,10)	(0.521,10)		(0.541,10)	(0.541,10)
Ca	*	*	*	*	0.159	0.243	0.285	*	0.073
					(0.872,10)	(0.988,10)	(0.348,10)		(0.523,10)
C	*	*	*	*	*	0.548	-0.100	0.078	*
						(1.156,10)	(0.834,10)	(0.480,10)	

REFERENCES

[1] Standard & poor's corporate ratings criteria. Report, Standard & Poor's, 1998. 1

[2] Moody's rating symbols & definitions. Report, Moody's Investor Services, 2003. 1

[3] L. Overbeck C. Bluhm and C. Wagner. *An Introduction to Credit Risk Modeling*. Chapman & Hall, 2003. 1.1

[4] E. Altman. The importance and subtlety of credit rating migration. *Journal of Banking & Finance*, 22:1231–1247, 1998. 1.2

[5] W. Perraudin P. Nickell and S. Varotto. Stability of rating transitions. *Journal of Banking & Finance*, 24:203–277, 2000. 1.2

[6] A. Kronimus C. Schagen A. Bangia, F.X. Diebold and T. Schuermann. Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking & Finance*, 26:445–474, 2002. 1.2

[7] D. Lando and T.M. Skodeberg. Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance*, 26:423–444, 2002. 1.2

[8] E. Hansen J.H.E. Christensen and D. Lando. Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance*, 28:2575–2602, 2004. 1.2, 4.2.1

[9] H. Frydman and T. Schuermann. Credit rating dynamics and markov mixture models. *Working Paper*, 2007. 1.2

[10] A. Lucas S.J. Koopman and A. Monteiro. The multi-state latent factor intensity model for credit rating transitions. *Tinbergen Institute Discussion Paper*, 2005. 1.2, 3, 3.1, 3.1, 3.1, 4, 5

[11] Guide to credit rating essentials. Report, Standard & Poor's. 2.1.1

[12] L. Goldberg L. Breger and O. Cheyette. Market implied ratings. *Barra*, 2003. 2.1.2, 2.1.2

[13] C. Mann M. Woolley D. Dwyer J. Gibbon S. Qu E. Fingerman D.W. Munves, D.T. Hamilton and J. King. Moody's market implied ratings description, methodology, and analytical applications. *Moody's Investors Service*, 2007. 2.1.2

[14] S. Jiang R. Cantor C. Mann D.A. Russell D.W. Munves, E. Mulic and J. Homan. Mir strategies. *Moody's Investors Service*, 2005. 2.1.2

[15] D.T. Hamilton and C. Lam. Predicting defaults with bond spreads and bond-implied ratings. *Moody's Investors Service*, 2008. 2.1.2

[16] O.O. Aalen. Statistical inference for a family of counting processes. *Ph.D. Dissertation, University of California, Berkeley*, 1975. 1.2, 2.2.1

[17] R.G. Miller. Least squares regression with censored data. *Biometrika*, 63:449–464, 1976. 2.2.1

[18] D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972. 2.2.1

[19] O.O. Aalen. A model for nonparametric regression analysis of counting processes. In *Lecture Notes Statistics*, volume 2, pages 1–25. Springer-Verlag, New York, 1980. 2.2.1

[20] F.W. Huffer and I.W. McKeague. Weighted least squares estimation for aalenś additive risk model. *Journal of the American Statistical Association*, 86:114C129, 1991. 2.2.1

[21] D.R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975. 2.2.1

[22] K.G. Manton J.W. Vaupel and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of individual mortality. *Demography*, 16:439–454, 1979. 2.2.1

[23] C. Elbers and G. Ridder. True and spurious duration dependence: the identifiability of the proportional hazards model. *The Review of Economic Studies*, XLIX:403–409, 1982. 2.2.1

[24] S. Keles M. Miloslavsky and M.J. Laan. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society B*, 66:239–257, 2004. 2.2.2

[25] D.Y. Lin. L.J. Wei and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association*, 84:1065–1073, 1989. 2.2.2

[26] M. Pepe and J. Cai. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, 88:811–820, 1993. 2.2.2

[27] B.J. Williams R.L. Prentice and A.V. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68:373–379, 1981. 2.2.2

[28] P.K. Andersen and R.D. Gill. Cox's regression model counting process: A large sample study. *Annals of Statistics*, 10:1100–1120, 1982. 2.2.2

[29] J. Lawless and C. Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37:158–168, 1995. 2.2.2

[30] C. Nadeau J. Lawless and R. Cook. Analysis of mean and rate functions for recurrent events. In *Lecture Notes in Statistics*, volume 123, pages 37–49. 1997. 2.2.2

[31] I. Yang D.Y. Lin, L.J. Wei and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society B*, 62:711–730, 2000. 2.2.2

[32] R. Tsay. *Analysis of Financial Times Series*. John Wiley & Sons, Inc., New Jersey, 2005. 2.3.2, 2.3.2, 2.3.3

[33] R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflations. *Econometrica*, 50:987–1007, 1982. 2.3.3

[34] T. Bollerslev. Generalized autoregressive conditional heteroscedasticity. *Econometrics*, 31:307–327, 1986. 2.3.3

[35] R.D. Gill P.K. Andersen, ∅. Borgan and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993. 3.1

[36] T.F. Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6:418–445, 1996. 4.1.1

[37] T.F. Coleman and Y. Li. On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, 67:189–224, 1994. 4.1.1

[38] M.J.D. Powell. *A Fast Algorithm for Nonlinearly Constrained Optimization Calculations*. Springer Verlag, 1978. 4.1.1

[39] M.J.D. Powell. *The Convergence of Variable Metric Methods For Nonlinearly Constrained Optimization Calculations*. Academic Press, 1978. 4.1.1

[40] J. Nocedal R.A. Waltz, J.L. Morales and D. Orban. An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical Programming*, 107:391–408, 2006. 4.1.1

BIOGRAPHICAL SKETCH

Zhi Li

Zhi Li was born in the north of China. He received his Bachelors degree from Tsinghua University in China. He came to US in 2001 and obtained his Master's degree in Applied Mathematics from Stanford University in June 2004. In 2005, he enrolled into the Phd program in the Department of Statistics of the Florida State University. He began working on his dissertation under the guidance of Professor Fred Huffer and Professor Xufeng Niu in 2006.

Zhi's research interests include time series analysis, stochastic processes, optimization, and financial data modeling.