Dependence modelling in insurance and finance using copulas

Xie, Wenjun

2017

Xie, W. (2017). Dependence modelling in insurance and finance using copulas. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/69657>

<https://doi.org/10.32657/10356/69657>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**DEPENDENCE MODELLING IN INSURANCE AND
FINANCE USING COPULAS**

XIE WENJUN

NANYANG BUSINESS SCHOOL

2016

**DEPENDENCE MODELLING IN INSURANCE AND
FINANCE USING COPULAS**

XIE WENJUN

Nanyang Business School

A thesis submitted to Nanyang Technological University in partial
fulfilment of the requirement for the degree of Doctor of Philosophy

2016

ACKNOWLEDGEMENTS

There are a number of people without whom this thesis might not have been written, and to whom I am sincerely grateful.

First of all, I would like to thank my supervisor Professor Wu Yuan, whose expertise, understanding, generous guidance, and support made it possible for me to work on a topic that was of great interest to me. What I learnt from him is not only academic knowledge but also attitude, determination, and pursuit towards life. There is no word that can show my gratefulness for him.

I would also like to thank my thesis committee members, Professor Shinichi Kamiya and Professor Xiang Liming. I received great help from them in writing my thesis. They are all very busy but are always patient to all my enquiries. Professor Xiang was also my supervisor for the undergraduate final year project, who taught me how to do a research project for the first time.

It was also my pleasure to meet Professor Christine Lim and worked on several joint projects together during these years. Professor Lim helped me a lot in the academic writing and inspired me many research ideas. I benefited a lot from her rigorousness and professionalism, which I really appreciate.

I also like to thank all the professors that taught me during my student's life. I would never come to this point without their help. I also like to express many thanks to all officers in the NBS Ph.D. office, without whom my life would be

much harder. In addition, to all my friends, thanks a lot and my life cannot go on without yours accompany.

Last but not least, I like to thank my parents for all the support during my studies. Nothing is better than being their son. Moreover, to my fiancée, Sushuang, I love you and am looking for opening a new chapter of my life with you!

ABSTRACT

Understanding and quantifying dependence is at the core of all modelling efforts in the areas of insurance and finance. Insurance and financial variables are usually non-normal and have excessive skewness and kurtosis. Therefore, the conventional linear regression models may not be suitable to use in these cases. In this thesis, I introduce the use of the copula modelling technique in modelling the dependence in insurance and financial problems. A copula is an effective tool to model the joint distribution of random variables. The advantages of the copula modelling technique are twofold. First, it estimates the marginal distributions of each random variable separately, which removes the assumptions of normal marginal distributions at the beginning, and it can allow specifications of distributions with heavy tails. Second, it allows flexibility in choosing the functional form that best describes the dependency structure between cumulative densities of random variables. This enables us to not only investigate the linear associations but also any kind of non-linear associations.

This thesis contains four studies. The first study discusses the use of copulas in a financial trading strategy, pairs trading. This study tries to generalize the pairs trading strategy using the copula technique to explicitly capture the marginal distributions as well as the dependency structure between the stock returns. The second study proposes a benchmark measurement of the loss ratio within a loss triangle using copulas to take consideration of dependency structure between different insurance business lines and discusses their implications on loss

reserving. The third study proposes a copula-based residual bootstrapping procedure to deal with data paucity problem faced in cyber insurance pricing analysis. In the fourth study, I extend the current analysis into the multi-dimensional scenario. Current copula regression analysis is mainly restricted to two-dimension due to lack of choice of copulas at multi-dimension. In this study, I attempt to build a framework of multi-dimensional copula regression analysis using a non-parametric copula called Bernstein copula. It provides a closed-form solution and can be adapted to any dependency structure. I demonstrate the usefulness of the proposed method using two examples in insurance and finance.

Table of Contents

ACKNOWLEDGEMENTS.....	2
ABSTRACT.....	4
CHAPTER I. INTRODUCTION.....	10
1.1 BACKGROUND	10
1.2 INTRODUCTION TO COPULAS	12
1.2.1 Elliptical Copulas	13
1.2.2 Archimedean Copulas.....	14
1.2.3 Copula Regression Analysis	15
1.3 OVERVIEW OF THESIS	16
CHAPTER II. APPLICATION OF COPULAS TO PAIRS TRADING STRATEGY	19
2.1 BACKGROUND OF PAIRS TRADING STRATEGY	19
2.2 CHARACTERIZING MISPRICING USING THE COPULA TECHNIQUE	22
2.3 COPULA-BASED PAIRS TRADING STRATEGY	24
2.3.1 Formation Period	25
2.3.2 Trading Period	25
2.4 EMPIRICAL ANALYSIS	27
2.4.1 One-Pair-One-Cycle Illustration.....	27
2.4.2 Large Sample Analysis using Utility Industry Data	29
2.5 REMARKS	36
CHAPTER III. APPLICATION OF COPULAS TO LOSS TRIANGLE BENCHMARKS	
CONSTRUCTION	45
3.1 BACKGROUND OF LOSS RESERVING	45
3.2 CONSTRUCTION OF LOSS TRIANGLE BENCHMARKS.....	49
3.2.1 Conditional Distribution of Small Business Line	50
3.2.2 Copula-Based Benchmark	52
3.3 SIMULATION ANALYSIS ON THE PERFORMANCE OF PROPOSED	
BENCHMARKS	54
3.4 EMPIRICAL ANALYSIS ON LEADING U.S. P&C INSURANCE COMPANIES ...	59
3.4.1 Data Description	59
3.4.2 Benchmarks	60
3.4.3 Implications for Loss Reserving.....	63

3.5 REMARKS	65
CHAPTER IV. COPULA REGRESSION WITH BOOTSTRAPPING: A CYBER INSURANCE CONTEXT	78
4.1 BACKGROUND OF CYBER INSURANCE	78
4.2 COPULA REGRESSION WITH BOOTSTRAPPING.....	83
4.2.1 Copula Regression: Cyber Loss Variable versus its Explanatory Covariates	83
4.2.2 Copula-Based Bootstrapping Procedure.....	85
4.2.3 Steps of Pricing Cyber Insurance	88
4.3 SIMULATION ANALYSIS ON THE PERFORMANCE OF COPULA-BASED RESIDUAL BOOTSTRAPPING	89
4.4 PRICING CYBER INSURANCE USING AN INDUSTRY DATASET AND ITS IMPLICATIONS	93
4.4.1 Data Description	93
4.4.2 Policy Types	93
4.4.3 Modelling of the Cyber Loss Variable	94
4.4.4 Calculations of Premiums.....	95
4.4.5 Findings and Implications	97
4.5 REMARKS	98
CHAPTER V. EXTENSION TO MULTI-DIMENSIONAL COPULA REGRESSION ANALYSIS IN INSURANCE AND FINANCE	112
5.1 EXISTING METHODOLOGIES	112
5.2 BERNSTEIN COPULA REGRESSION MODEL	115
5.3 EXAMPLE 1: SIMULATION ANALYSIS ON THE PERFORMANCE OF MULTI-DIMENSIONAL COPULA-BASED LOSS TRIANGLE BENCHMARKS	118
5.3.1 Benchmark Construction	118
5.3.2 Simulation Analysis.....	119
5.4 EXAMPLE 2: MULTI-DIMENSIONAL PAIRS TRADING STRATEGY.....	123
5.4.1 Group Selection	124
5.4.2 Group Trading Strategy	126
5.4.3 Data Description and Settings	128
5.4.4 Main Results	129
5.5 REMARKS	132

CHAPTER VI. CONCLUDING REMARKS	142
REFERENCES.....	145
APPENDIX.....	156
Table A1. Claim Triangles for the Personal Auto Line (1989-2007)	156
Table A2. Claim Triangles for the Commercial Auto Line (1989-2007)	157

List of Figures

Figure 2.1 Distance Strategy	38
Figure 2.2 Copula Fitting Result (Formation Period)	39
Figure 2.3 Copula Fitting Result (Trading Period)	39
Figure 2.4 Cumulative Returns for the Top Five Pairs Portfolios (Without One-Day Waiting)	39
Figure 3.1 Multiple Time Series Plots of Loss Ratios	68
Figure 3.2 Cumulative Probabilities of the Personal versus Commercial Auto Lines	69
Figure 3.3 Reserve Estimation Errors at Development Lag One (Accident Year 2006).....	70
Figure 4.1 Comparison of the Q-Q plots between the samples generated with or without copula-based residual bootstrapping to the original sample.....	101
Figure 4.2 Linear Regression Fit.....	103
Figure 4.3 Scatter Plot of Cumulative Densities	104
Figure 4.4 Copula Regression Fit.....	105
Figure 4.5 Sampling Distribution of Premium for Bootstrapped Resamples	106
Figure 5.1 Accuracies of the Benchmark with different sample size.....	134
Figure 5.2 Accuracies of the Bernstein Benchmark, Gaussian Benchmark and the Reported Loss Ratio (ABS: Absolute Error).....	136
Figure 5.3 Loss Reserving Accuracies (ABS: Absolute Error)	137

List of Tables

Table 2.1 Information Criteria of Copula Fitting Result.....	41
Table 2.2 Results of Pairs Trading Strategies on Stock Pair BKD-ESC.....	41
Table 2.3 Excess Returns of Pairs Trading Strategies	42
Table 2.4 Trading Statistics and Composition of Pairs Portfolio.....	43
Table 2.5 Systematic Risk of Pairs Trading Strategies	44
Table 3.1 Simulation Analysis on Assuming Different Copula Structures....	71
Table 3.2 Simulation Analysis on Assuming Different Degrees of Associations	72
Table 3.3 Summary Statistics of Five Leading U.S. Insurance Companies...	73
Table 3.4 Copula Fitting Result for a Leading U.S. Insurance Company.....	74
Table 3.5 Benchmarks at Development Lag Zero (Accident Year 2006).....	74
Table 3.6 Reserve Estimations at Development Lag One (Accident Year 2006).....	75
Table 3.7 Reserve Estimation Using Copula-Based Approach.....	76
Table 4.1 Kolmogorov–Smirnov tests between Population, without bootstrapping and bootstrapping samples	107
Table 4.2 2003 ICSA Surveyed Computer Virus Data	108
Table 4.3 Policy Types and Premiums (Herath and Herath, 2011).....	109
Table 4.4 Copula Fitting Result for the Original Sample	110
Table 4.5 Premium for Different Types of Policies	111
Table 5.1 Trade Conditions and Positions for Multi-Dimensional Copula..	138
Table 5.2 Summary of Returns for Different Strategies and Different Markets	139
Table 5.3 Overall Comparison of Trading Statistics of Different Strategies	141

CHAPTER I. INTRODUCTION

1.1 BACKGROUND

Nowadays, the world becomes a tremendous well-organized system with every component of it coordinated and associated with each other. This is even more obvious in the context of risk modelling in the insurance and financial markets. For example, the 2008 global financial crisis starts from the U.S. housing market, but it is eventually spread to all kinds of industry sectors and countries all over the world (Itzhak et al., 2012, Sandoval and Franca, 2012). Similarly, the recent outbreak of Ebola virus disease is firstly reported in Guinea by the World Health Organization (WHO) but then spread rapidly to other Western African countries and the rest of the world with frequent geographic movement (Walsh et al., 2005). These facts suggest that it is of equal importance to consider the interdependence of risk events rather than examining them alone. In other words, it is of great interest to study a group of risk factors together and get a decent understanding of the associations between them, instead of treating all factors separately.

In the literature, classical linear regression analysis is the most adopted method to study associations between different variables, subject to its simple implementation and clear interpretation of the model. Besides, there are numerous extensions to this baseline model, such as generalized linear regression, logistic model (Cox, 1958) and Tobit model (Tobin, 1958). These great

extensions are helpful to deal with non-normal data, or count data which the classical linear regression model cannot work on.

However, there are still certain limitations on the classical linear regression analysis and its extensions. First, there is a limitation on the marginal distribution assumption for the dependent variable. Classical linear regression analysis requires the dependent variable to follow a normal distribution, and the generalized linear regression also requires it to follow a distribution from exponential families. However, the current consensus is that insurance and financial data are hardly normally distributed and usually have heavy tails (Cont, 2001). This fact suggests that classical linear regression analysis and its extensions may not work well in these cases. Second, the linear association is only one type of associations that is of our interest. In fact, non-linear associations, especially tail dependence, is of great importance to insurance and financial studies. For example, the tail dependence of default probabilities of large financial institutions may cause systematic risk in the financial system, and this is what we observed in the 2008 financial crisis. Thus, understanding the non-linear associations is as important as linear associations in insurance and financial studies, and the linear regression analysis may not be able to accomplish this goal.

To deal with the aforementioned limitations of linear regression analysis in insurance and financial studies, I used the copula modelling technique to study associations between different variables in this thesis. A copula is a function that joins multivariate distribution function to its one-dimensional marginal

distributions. In other words, a copula is a multivariate distribution function in which one-dimensional margins are uniform on the interval $[0, 1]$. The advantages of the copula modelling technique are twofold. First, it estimates the marginal distributions of each random variable separately, which removes the assumptions of normal marginal distributions at the beginning, and it can allow specifications of distributions with heavy tails. Second, it allows flexibility in choosing the functional form that best describes the dependency structure between cumulative densities of random variables. This enables us to not only investigate the linear associations but also any kind of non-linear associations. To summarize, the copula modelling technique can be seen as a generalization of linear regression models. For example, if the underlying marginal distributions of variables are normal and their dependency structure is normal, then the copula modelling technique is equivalent to the classical linear regression model.

1.2 INTRODUCTION TO COPULAS

The word “copula” comes from the Latin language, and it is a noun meaning “link” or “tie”. In statistical theory, a copula is a function linking the joint distribution of random variables to their separate marginal distributions. The foundation of the copula modelling technique relies on the Sklar’s theorem (Sklar, 1959). If H is the joint cumulative distribution function for the random variables $Y, X_1, \dots, X_k$ with continuous marginal distributions $G, F_1, \dots, F_k$, then there exists a copula function C , such that

$$H(y, x_1, \dots, x_k) = C(G(y), F_1(x_1), \dots, F_k(x_k)). \quad (1.1)$$

Interchangeably, the joint distribution H can also be derived by combining the marginal distributions $G, F_1, F_2, \dots, F_k$ and the copula function C that links them.

From this point of view, the copula modelling technique allows different specification of the dependency structure by giving flexibility in choosing the functional form of the copula function C . There are some commonly used families of copulas in the literature, representing different types of dependency structures.

1.2.1 Elliptical Copulas

Elliptical copulas mainly describe symmetric dependency structures. There are two commonly used elliptical copulas: Gaussian copula and Student-T copula. Gaussian copula, which is also known as normal copula, describes the basic symmetric dependency structure. When the marginal distributions of the variables follow normal distributions and the copula joining them is normal, then it is equivalent to assume the multivariate normal distribution of variables.

The bivariate Gaussian copula C_G is defined as

$$C_G(u, v; \rho) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho), \quad (1.2)$$

Where u and v represent the cumulative densities for the two random variables. Φ^{-1} is the inverse cumulative distribution function of a standard normal

distribution, and Φ_2 is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and correlation $\rho \in [-1, 1]$.

Compared to Gaussian copula, Student-T copula places more concentrations on the tail dependence, but it still describes symmetric dependency structure. The bivariate Student-T copula C_T is defined as

$$C_T(u, v; \theta, df) = T_{\theta, df}(T_{df}^{-1}(u), T_{df}^{-1}(v); \theta, df), \quad (1.3)$$

where T_{df}^{-1} is the inverse of the scalar standard Student-T cumulative distribution with degree of freedom df , and $T_{\theta, df}$ is the bivariate Student-T cumulative distribution with dependence parameter $\theta \in [-1, 1]$ and degree of freedom df .

1.2.2 Archimedean Copulas

Besides elliptical copulas, another important family of copulas is the Archimedean copulas. It has a general representation form as

$$C_A(u, v; \theta) = \varphi^{-1}(\varphi(u; \theta), \varphi(v; \theta); \theta), \quad (1.4)$$

where φ is called the generating function which is continuous, strictly decreasing and convex. The most commonly used Archimedean copulas are Clayton, Frank, and Gumbel copulas.

Clayton copula specializes in describing strong left tail dependence and soft right tail dependence, and it is defined as

$$C_C(u, v; \theta) = [\max(u^{-\theta} + v^{-\theta} - 1, 0)]^{-1/\theta}, \theta \in [-1, \infty) \setminus \{0\}. \quad (1.5)$$

Frank copula emphasizes on strong central dependence and soft tail dependence, and it is defined as

$$C_F(u, v; \theta) = -\frac{1}{\theta} \log \left[\frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{(\exp(-\theta) - 1)} \right], \theta \in \mathbb{R} \setminus \{0\}. \quad (1.6)$$

Gumbel copula best describes dependency structure with strong right tail dependence but soft left tail dependence. It is defined as

$$C_G(u, v; \theta) = \exp \left\{ - \left[(-\log u)^\theta + (-\log v)^\theta \right]^{\frac{1}{\theta}} \right\}, \theta \in [1, \infty). \quad (1.7)$$

1.2.3 Copula Regression Analysis

After obtaining the joint distribution of variables from the copula modelling technique, the next task of interest is to derive the conditional distribution of the dependent variable given its explanatory covariates, and this can be done under the copula framework by taking the first derivative of the copula function. According to Crane and Van Der Hoek (2008), if I denote $u = G(y)$ and $v_i = F_i(x_i)$, then

$$\text{Prob}(Y \leq y | X_1 = x_1, \dots, X_k = x_k) = \frac{\partial C(u, v_1, v_2, \dots, v_p)}{\partial v_1 \partial v_2 \dots \partial v_p} \bigg/ \frac{\partial C(1, v_1, v_2, \dots, v_p)}{\partial v_1 \partial v_2 \dots \partial v_p}. \quad (1.8)$$

Besides the conditional mean, median, or variance, there is more information that we can obtain from this conditional distribution of the dependent variable, such as Value at Risk (VaR) or Expected Shortfall (ES) that are commonly used in insurance and financial studies.

1.3 OVERVIEW OF THESIS

There is a plethora of literature about copula modelling technique and its applications. However, there are several milestones that cannot be ignored in the insurance and financial studies. Though the concept of copula was raised in Sklar (1959), it only attracts tremendous attention from researchers in the recent decades. Frees and Valdez (1998) provided an inspiring description of the use of the copula technique to model the joint dependence of random variables in the actuarial context. After that, Embrechts et al. (2002) and Embrechts and Puccetti (2006) introduced the copula modelling technique in risk management. Besides, copula regression is a relatively new methodology proposed recently (Crane and Van Der Hoek, 2008; Parsa and Klugman, 2011).

The contribution of this thesis is twofold. First, I further promote the applications of copula modelling technique into insurance and financial businesses with innovations in the methodology. Second, I try to provide an alternative method to deal with the multi-dimensional copula regression analysis which is under-developed in the literature.

The rest of the thesis is organized as follows. Chapter II discusses the use of copulas in a financial hedging strategy, pairs trading. Pairs trading is a well-acknowledged speculative investment strategy, with the distance method the most commonly adopted method. However, it is noted that the profitability of this approach decreases in the recent years. As current consensus recognizes that

stock returns are rarely jointly normal, this study tries to generalize the pairs trading strategy using copula technique to explicitly capture the marginal distributions as well as the dependency structure between the stock returns. With a better understanding of the joint distribution of the two stocks, practitioners could gain preferential entry positions and have more trading opportunities. The overall empirical results verify the proposed strategy's ability to generate higher profits compared with the conventional distance method.

Chapter III proposes a benchmark measurement of the loss ratio within a loss triangle, which uses copulas to take consideration of the dependency structure between different insurance business lines. Its implications on the loss reserving accuracy are also discussed. Loss reserving plays a central role in property and casualty insurance. However, this is not an easy task for a small business line in which the large sample theory may not be applied. The difference between the reported loss ratio and its expected value could make the subsequent reserving estimation inaccurate. In this study, I propose to mitigate this issue by examining the historical dependence between the target small business line and its related large business line using the copula technique. I then construct a benchmark measurement to approximate the expected loss ratio of the small line using the information from its related large line loss ratio, which is assumed to converge to its expected value predicted by the large sample theory. Both simulation and empirical analysis suggest that this benchmark could be a better approximation for the expected loss ratio of the small line than its reported loss ratio. By using data from personal and commercial auto lines of leading U.S.

insurance firms, I have found significant improvements in the reserve accuracies by using the derived benchmark instead of the reported loss ratio in the loss triangle for subsequent reserving analysis. Besides, this finding should be robust regardless of the reserving method adopted.

Chapter IV proposes a copula-based residual bootstrapping procedure to deal with data paucity problem faced in cyber insurance pricing analysis. Cyber insurance has received overwhelming attention in the recent decade. However, the pricing of it faces difficulties due to the complex inter-dependencies under the digital world and the data paucity problem. I propose a copula framework with a bootstrapping approach to deal with the aforementioned difficulties. Copula has the advantage of estimating the joint distribution of random variables and bootstrapping can essentially relief the data paucity concern. Simulation results suggest the advantages of the proposed model in capturing the population characteristics when the sample size is small. Empirical analysis using an industry dataset implies that insurers may gain competitiveness using the proposed model.

Chapter V extends the current copula regression analysis into multi-dimension. The main restriction of the current copula regression analysis is lack of choice of copulas at multi-dimension. In this study, I attempt to build a framework of multi-dimensional copula regression analysis using a non-parametric copula called Bernstein copula. It has closed-form solution and can

be adapted to any dependency structure. I demonstrate the usefulness of the proposed method using practical examples in insurance and finance.

Finally, Chapter VI gives the concluding remarks.

CHAPTER II. APPLICATION OF COPULAS TO PAIRS

TRADING STRATEGY

2.1 BACKGROUND OF PAIRS TRADING STRATEGY

Pairs trading refers to a long-short, delta neutral investment strategy, with the distance method the most commonly used algorithm. The idea is to identify a pair of stocks whose prices have moved together historically, and subsequently, construct long-short positions when the two prices diverge. Gatev et al. (2006) were pioneers in investigating this simple trading strategy, documenting significant excess returns of this strategy in the large-sample analysis. Later, its effectiveness was documented in different time periods and markets as well (Andrada et al., 2005; Perlin, 2009; Pizzutilo, 2013). However, Do and Faff (2010) adopted a longer sample period, and identified a significant drop in the profitability of this simple strategy in the recent period (0.33% mean excess return per month for 2003-09 versus 1.24% mean excess return per month for 1962-88), leading to the consideration of how we can improve the profitability of this simple strategy.

The high profitability of this pairs trading strategy mainly rests on two factors: the correct identification of high-quality pairs and optimal modeling of the associations between the two stocks within a pair. The former ensures that this simple strategy is market-risk free, and the latter helps to decide the optimal trading positions. Do and Faff (2010) addressed the first factor, and concluded that it is the arbitrage risk (De Long et al., 1999), which refers to deterrence from participating in arbitrage activities owing to the risk of further divergence, that explains most of the decrease in the profitability observed in their sample period. Thus, they suggested employing pairs with better convergence properties, such as the pairs from the same industry classifications. In this chapter, I try to focus on the second factor, which was previously ignored. As pairs trading is to trade the temporary mispricing within a pair of stocks, if the associations between the two stocks are not correctly understood, then the judgment on the mispricing can be misleading and thus produce undesirable trades.

The conventional pairs trading strategy analyzed in all the aforementioned papers is called the distance method. It uses the distance between normalized prices¹ as the criterion to judge the degree of mispricing between stocks. Although not made explicitly, this distance measurement is equivalent to

¹ I refer to Gatev et al. (2006) for our definition of normalized prices, which can essentially be seen as measurements of cumulative returns. The normalized prices for the first day of the formation period and trading period are set to one.

linear correlation analysis, which is able to fully reveal the information about the associations between the two stocks under the assumption of multivariate normal stock returns. However, the current consensus is that stock returns are rarely joint normal (Cont, 2001) and thus non-linear associations, such as tail dependence, also play an important role in the modeling of stock returns (Ane and Kharoubi, 2003). Given this concern, the distance method no longer appears optimal, as it may cause loss of dependency information between the two stocks within a pair.

In this chapter, I propose to use the copula technique to generalize the pairs trading strategy under the consensus of non-normal stock returns. By recognizing that the stock returns are rarely jointly normal, the copula technique is an effective tool in modeling their joint distribution. The copula technique has two main advantages. First, it allows estimations of the marginal distributions followed by the copula, which removes the assumption of normal marginal distributions from the beginning of the process. Second, it can provide explicit functions to describe the dependence, which gives more precise understanding of the dependency structure between variables. By combining the estimated marginal distributions and the dependency structure, copula technique allows estimation of the joint distribution of individual variables. Therefore, in this chapter, to best understand the associations between the two stocks within a pair, I use the copula technique to capture the joint distribution of them.

It is noticeable that there are some studies that conduct preliminary trials of the application of copulas to pairs trading (Ferreira, 2008; Liew and Wu, 2013;

Stander et al., 2013). Ferreira (2008) and Liew and Wu (2013) directly used static copula to model stock prices or cumulative returns. However, the static copula cannot be applied to model nonstationary time-series (Nelson, 2006), and stock prices/cumulative returns are hardly stationary, making the usage of their proposed strategies limited. Besides, Stander et al. (2013) considered the daily mispricing between the two stocks within a pair but missed the time structure or the cumulative effect of the mispricing, which makes their strategy fundamentally different from the principle of pairs trading. Moreover, these studies did not provide large-sample results, therefore their given examples suffer from data snooping criticisms and are not convincing for portfolio management. Capitalizing on the inspiring ideas in these studies, this chapter introduces a generalized copula-based approach that can be used for any selected pair in the market, with a demonstration of its effectiveness using a large-sample analysis.

2.2 CHARACTERIZING MISPRICING USING THE COPULA TECHNIQUE

In resolving the difficulties that the distance method encounters, the main priority is to accurately capture the joint distribution of stock returns, and then decide the entry positions accordingly. Thus, I utilize a copula as a tool to model the joint distribution of random variables in developing my new measurements of mispricing.

Assume that stocks X and Y are candidates for pairs trading. Their daily closing prices are P_t^X and P_t^Y , and their daily simple returns are recorded as R_t^X and R_t^Y . I denote F_X and F_Y as the marginal distribution functions of R_t^X and R_t^Y , and H as their joint distribution function. Then, according to Sklar's theorem, there must exist a copula function C such that

$$H(r_t^X, r_t^Y) = C\left(F_X(r_t^X), F_Y(r_t^Y)\right). \quad (2.1)$$

After obtaining the joint distribution of daily returns for stock X and stock Y, I construct measurements to denote the degree of mispricing utilizing the idea of conditional probabilities. If R_t^X and R_t^Y represent the random variables of the daily returns of stocks X and Y on day t, and the realizations of those returns on day t are r_t^X and r_t^Y , I define my mispricing indexes $MI_t^{X|Y}$ and $MI_t^{Y|X}$ as

$$\begin{aligned} MI_t^{X|Y}(r_t^X, r_t^Y) &= P(R_t^X < r_t^X | R_t^Y = r_t^Y). \\ MI_t^{Y|X}(r_t^X, r_t^Y) &= P(R_t^Y < r_t^Y | R_t^X = r_t^X). \end{aligned} \quad (2.2)$$

The defined mispricing indexes range from 0 to 1 as they represent the cumulative conditional probabilities. Generally speaking, a value of $MI_t^{X|Y}$ or $MI_t^{Y|X}$ equal to 0.5 represent that one stock is neither overvalued or undervalued compared to the other stock, considering their historical joint distribution. Nevertheless, a value higher than 0.5 would indicate that the underlying stock is overvalued conditional on the other stock, and vice versa. A value closer to 1

indicates a higher degree of overvaluation while a value closer to 0 indicates a higher degree of undervaluation.

To calculate the mispricing indexes from Equation (2.2), I take the first derivative of the copula function.² If I define $u = F_X(r_t^X)$ and $v = F_Y(r_t^Y)$, then

$$\begin{aligned} MI_t^{X|Y} &= \frac{\partial C(u,v)}{\partial v}. \\ MI_t^{Y|X} &= \frac{\partial C(u,v)}{\partial u}. \end{aligned} \tag{2.3}$$

Hence, instead of performing a linear correlation analysis in the conventional distance method, the mispricing indexes proposed in this research contains more accurate information about the associations between the two stocks derived from the estimated joint distribution of them. Thus, they should characterize the mispricing of the two stocks more accurately and provide better trading opportunities. One thing to note here is that these mispricing indexes only represent the degree of mispricing for each single day, to conduct the proposed pairs trading strategy, I need to sum them up to get an overall degree of mispricing. In the next section, I introduce the trading strategy based on these mispricing indexes.

2.3 COPULA-BASED PAIRS TRADING STRATEGY

² The formal proof can be found in Nelson (2006).

The trading strategy comprises the following two stages: formation and trading.

2.3.1 Formation Period

Daily return series r_t^X and r_t^Y are calculated during the formation period. Then, the best fitted marginal distributions of r_t^X and r_t^Y are estimated, after which the optimal dependency structure is estimated using different categories of copulas (Gumbel, Frank, Clayton, Normal and Student-T).³ That with the highest likelihood value is chosen.

2.3.2 Trading Period

The daily returns of stock X and stock Y are calculated during the trading period. $MI_t^{X|Y}$ and $MI_t^{Y|X}$ are also calculated using the estimated copula from the formation period. Trading indicators are defined as FlagX and FlagY, and they are set to 0 before the commencement of the trading period. During the trading period, $(MI_t^{X|Y} - 0.5)$ and $(MI_t^{Y|X} - 0.5)$ are added to FlagX and FlagY, respectively, on a daily basis. In addition, D is defined as the trigger point and S as the stop-

³ I acknowledge that only a limited number of copulas are considered in this study. Owing to the computational complexity involved, only the five most commonly used copulas were chosen for demonstration purposes. I believe that these five cover the majority of the dependency structures commonly encountered in practice. To improve accuracy, practitioners can incorporate more copulas in the fitting process if they wish.

loss position. The following are the four possible cases for an open position (assuming that no trades are open).

1. When FlagX reaches D, we short-sell stock X and buy stock Y in equal amounts.
2. When FlagX reaches negative D, we short-sell stock Y and buy stock X in equal amounts.
3. When FlagY reaches D, we short-sell stock Y and buy stock X in equal amounts.
4. When FlagY reaches negative D, we short-sell stock X and buy stock Y in equal amounts.

If trades are opened based on FlagX, then they are closed if FlagX returns to zero or reaches stop-loss position S or negative S. If they are opened based on FlagY, then they are closed if FlagY returns to zero or reaches stop-loss position S or -S. After trades are closed, both FlagX and FlagY are reset to zero, and all opening trades are closed at the end of the trading period regardless of the values of FlagX and FlagY.

Note that D and S are pre-specified values, of which there are endless combinations. Practitioners can perform back-testing to choose the optimal trigger points. For illustration purposes, in this chapter, I choose to conduct analysis using $D = 0.6$ and $S = 2$, which is perceived as a good combination during my back-testing process. To avoid data snooping criticisms, I also perform additional analysis that is further explained.

2.4 EMPIRICAL ANALYSIS

2.4.1 One-Pair-One-Cycle Illustration

I first use a single pair to illustrate the idea. The stock pair considered is Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC). This highly correlated stock pair is listed in the healthcare sector of www.pairslog.com and is also cited as an example in Liew and Wu (2013). Both companies are listed on the New York Stock Exchange (NYSE) and are very similar in terms of their business operations. Both are in the long-term care facilities industry of the health care sector. The time periods considered in this example are January 2 to December 30, 2008 (formation period) and December 31, 2008 to July 1, 2009 (trading period). The stock pair is verified to be highly correlated during the formation period, as seen in Figure 2.1(a). The correlation coefficient between them is 0.942.

Because of the close relation between the normalized prices of BKD and ESC, the spread during the formation period (shown in Figure 2.1(b)) fluctuates around zero most of the time and demonstrates a tendency to revert to zero. However, the relation changes dramatically during the trading period because the normalized prices drift farther apart over the period and the spread moves away from its original mean value of zero. This shift suggests a potential change in the structure of the stock pair or the conventional distance strategy's inability to fully capture the information of associations between the two stocks.

I carry out a comparative analysis using the copula approach. Figure 2.2 illustrates the dataset from the formation period, with the copulas fitted to it. The Student-t copula best captures the general structure, particularly the upper and lower tails of the formation period data, which is further verified by the test values reported in Panel A of Table 2.1. At the same time, Figure 2.3 and Panel B of Table 2.1 demonstrate that Student-t remains the best fitting copula for the dataset for the trading period. This result suggests that the dependency structure of the stock returns does not change over time, and the distance method cannot capture this time-invariant dependency structure due to rigid pre-assumptions about the stock returns. In contrast, the proposed copula approach can provide a better knowledge of the associations between the two stocks and may provide better trading opportunities.

Overall, the results in Table 2.2 show that the copula approach undisputedly performs much better than the distance strategy, regardless of the one-day wait.⁴ Assuming \$10,000 in initial capital, the distance method results in a -\$592 loss, whereas the proposed copula method results in an \$847 gain in the case without a one-day wait. In the one-day wait case, the distance method generates a loss of up to -\$1526, whereas the copula method produces a gain of \$1060. In this example, the copula approach is clearly more profitable and

⁴ Gatev et al. (2006) recommended investigating the one-day-wait case to deal with the bid-ask spread bounce.

generates greater excess returns. Its consistent superiority over the distance approach is further verified in an industry-wide dataset in the next section.

2.4.2 Large Sample Analysis using Utility Industry Data

I used stocks from utility industry to conduct my large-sample analysis. I chose the utility industry because Do and Faff (2010) raised the point that choosing pairs within a more confined industry increases the profitability of pairs trading, and stocks in the utility industry have better co-movements.

2.4.2.1 Data

My daily stock price data are extracted from the Center for Research in Security Prices (CRSP) database. The sample period is January 1, 2003 to December 31, 2012. All of the stocks selected are publicly traded on the NYSE, AMEX or NASDAQ. To form the initial dataset, I filtered the stocks to identify those with a Standard Industrial Classification (SIC) code beginning with 49 (Utility). I subsequently dropped penny stocks and stocks with missing values, to give a final sample comprising 89 stocks.

2.4.2.2 Excess Return Computation

As previously noted, Gatev et al. (2006) identify two return measurements: the return on committed capital and the fully invested return. The former considers the return on actual employed capital, which means that all of the money prepared for potential trades is considered in the principal amount even if the position for the particular stock pair is not open. In contrast, the fully invested return

considers only the money currently being traded as the principal amount to calculate the return. The first measurement is clearly much more conservative and practical because the opportunity cost of the money set aside for potential trades is taken into account. As the use of such a conservative measurement increases the credibility of my results, I adopt the return on committed capital in this chapter. The formula for calculating the excess return is exactly the same as that defined in Gatev et al. (2006):

$$r_{P,t} = \sum_{i \in P} w_{i,t} r_{i,t} / \sum_{i \in P} w_{i,t}. \quad (2.4)$$

$$w_{i,t} = w_{i,t-1} (1 + r_{i,t-1}) = (1 + r_{i,1}) \dots (1 + r_{i,t-1}). \quad (2.5)$$

where r defines returns and w defines weights, and daily returns are compounded to obtain monthly returns.

2.4.2.3 Main Results

I follow the analysis conducted by Gatev et al. (2006) to provide a comparable evaluation. After cleaning up the data, I ran them across 17 time periods, including both formation and trading periods. A formation period is defined as a 12-month period, taken to be 252 days.⁵ Historical data from the formation period

⁵ Although the actual number of trading days for each month varies, I assume 21 trading days for each. Hence, “12-month formation period” and “six-month trading period” may not be strictly accurate.

are used for pairs formation and estimation of the distributions and parameters. A trading period is defined as a six-month period, taken to be 126 days.

During each formation period, every possible combination of the 89 stocks is considered, and the sum of squared deviations between the two normalized price series is calculated. The top 5 stock pairs chosen are those with the lowest sum of squared deviations between the two series. The same also applies to the top 20 and top 101-120 pairs.

This pair-formation approach is in accordance with that adopted in Gatev et al. (2006). Applying the minimum-distance criterion ensures that the stock pairs chosen have relatively smaller values for their sum of squared deviations than the other possible combinations. In other words, the stock pairs selected share a close relationship during the formation period, with contemporaneously similar upward and downward movements. This approach is found to best approximate the way in which actual traders choose pairs (Gatev et al., 2006).

Table 2.3 presents the excess returns of the pair portfolios generated by the distance and copula strategies. Panel A shows the results of the two trading strategies when the positions are opened at the end of the day on which prices diverge and closed at the end of the day on which they converge. The value of average excess returns, t-statistics, and percentage of observations with negative excess returns clearly show the copula strategy to outperform the conventional distance strategy. The excess returns produced by this strategy are consistently higher than those generated by the distance strategy for the top 5, top 20 and 101-

120 stock pairs. Moreover, the values of the t-statistics indicate that the excess returns are positive and significant for the proposed copula method but not for the distance method (for all three sets of portfolios tested). In addition, the percentage of observations with negative excess returns is smaller for the copula than distance strategy across-the-board, that is, for the top 5, top 20 and top 101-120 stock pairs.

Despite the positive results shown in Panel A, I acknowledge the concern regarding the bid-ask spread (Jegadeesh, 1990; Jegadeesh and Titman, 1995; Conrad and Kaul, 1989), which is mentioned in Gatev et al., 2006 as a practical issue faced in trading. To alleviate this concern, I also implemented one-day waiting, as recommended by Gatev et al. (2006). Panel B of Table 2.3 reports the trading strategy results when the positions of each stock pair are opened on the day following price divergence and closed on the day following convergence. Although the average excess returns are affected by the one-day wait for both strategies, those generated by the copula strategy are still generally better than those produced by the distance strategy, thus further strengthening our confidence in the proposed strategy.

Table 2.4 summarizes the trading statistics and composition of the pair portfolios for both the distance and copula strategies. From the average number of pairs traded per six-month period, we observe that the copula strategy identifies trades for all of the top pairs selected. In addition, the average number of round-trip trades per pair is higher across the board for this strategy. Although

the proposed strategy is clearly able to identify more trading opportunities than the traditional distance strategy, it requires the pairs to be open for a greater number of months. However, the proposed copula strategy also brings about a smaller standard deviation in the time open per pair in months. Accordingly, the strategy affords greater certainty about the amount of time that traders will be in an open position.

The foregoing cross-sectional analysis of U.S. utility industry stocks using both strategies provides clear evidence of the proposed copula strategy's superior performance relative to the conventional distance strategy. The copula strategy exhibits very satisfactory excess returns even after the application of one-day waiting. In addition, I have also verified that this strategy is able to identify more trading opportunities, an extremely important finding for the market, as every potential market signal constitutes an opportunity to profit.

2.4.2.4 Systematic Risks

In order to examine that the superior returns generated by the copula method are not due to any risk factors, I ran the return series against Fama-French three factors (Fama and French, 1993) and Carhart momentum factor (Carhart, 1997) for both distance method and copula method, and the results are shown in Table 2.5. Similarly, I ran for all top 5, top 20 and 101-120 pairs portfolio, considering both the cases of no waiting and waiting for one day.

The results generally show that the excess returns generated by both distance method and copula method are rarely explained by risk factors. Considering various portfolios, most of the coefficients for the four factors are insignificant. The intercepts, which represent the risk-adjusted excess returns, are generally consistent with the raw returns shown in Table 2.3, with copula method generating higher excess returns than distance method. If we consider the case of one-day waiting, then only the excess return for the top 5 pairs portfolio using copula method is significant while the excess returns for different portfolios using distance method are all insignificant.

2.4.2.5 Transactions Costs and Data Snooping

There are two main sources of transaction costs: bid-ask spread costs and short-selling costs. Getczy et al. (2002) noted that the short-selling costs of the rebate rate on short sales are low for large traders, just 4-15 basis points (bp) per year. Therefore, the main focus for transaction costs is bid-ask spread costs. Keim and Madhavan (1997) found the average effective spread for stocks in the CRSP database in 1991 to be 37 bp, and more recent NYSE research showed the average effective bid-ask spread in 2001 ranges from 14 to 18 bp for NYSE and NASDAQ common stocks. This finding is reasonable, as the liquidity of the stock market has increased over the years, and thus the current average bid-ask spread may be even lower. The six-month excess return, with one-day waiting for the top 5 pairs, is 264 bp (44 bp per month). Considering the average of 6.8 transactions per six-month period stated in Table 2.4, the transactions costs would

be around 110 bp. Hence, the excess returns obtained using the proposed copula strategy, and assuming one-day waiting, remain positive and significant. However, for other portfolios, and particularly for the results produced by the distance strategy, the excess returns generated may be insufficient to cover the transaction costs.

In order to relief data snooping criticisms, I also conduct two more tests. First, I altered the trigger points for the distance method in search of the optimal profits. The results of this optimal analysis differed little from the results reported in Table 2.3, and I thus concluded that the results of the proposed copula method with trigger points (0.6/2) were better than the optimal result from the distance method. This means that although the trigger points used in this chapter may not be optimal, it is still better than the optimal result that can be obtained by the distance method, which undoubtedly fulfill the illustration purpose.

Second, to avoid concern that this set of trigger points was good only for a certain period or a certain pair, I conducted subsample analysis and observed consistent results over time. Figure 2.4 shows the cumulative returns along the 17 trading periods for the top five pairs portfolios using the copula strategy. We observe a steady increase in the cumulative returns, with especially higher returns during the financial crisis period. This is reasonable, as Li (2014) documented that overall stock market downturns will lead to stronger co-movements in stock prices than market upturns. Overall, I believe that the results of these two

additional analyses should alleviate data snooping concerns about the proposed method.

2.5 REMARKS

Pairs trading is a well-acknowledged technique in the financial industry, and its effectiveness has been consistently documented over time and across markets. However, a marked decline in the profitability of this simple strategy has been noted in recent years. In this chapter, I try to generalize the strategy under the current consensus of non-normal stock returns, which use the copula technique to optimally capture the dependency structure between two stocks, thus getting rid of the pre-assumption of multivariate normal stock returns inferred by the conventional distance method. I employed a large-sample analysis using utility industry data to demonstrate the proposed strategy's effectiveness and superior profitability.

I also used a one-pair-one-cycle example to illustrate the importance of identifying the correct dependency structure between two stocks. When distance method cannot observe the time-invariant structure between two stocks due to its limitations, the proposed copula approach can better capture this time-invariant property. Although linear correlation may change from time to time, the big picture of the dependency structure between two stocks is more robust and not easy to shift over time. Thus, the proposed copula approach, which captures this point, gives us more confidence in finding better trading opportunities.

Despite the superiority of the results obtained from copulas, there are certainly areas that could be further improved and developed. For example, the stock pairs selected for both methods in this chapter are identified according to minimum distance, an approach that focuses only on the linear association between two stocks. Accordingly, a number of good candidates with strong non-linear associations may have been neglected. Because copulas are able to capture both linear and non-linear relations between random variables, it is possible that the proposed method might have performed even better if we had selected the stock pairs by incorporating non-linear association measurements such as Kendall's tau and Spearman's rho. Moreover, copulas can capture the dependency structure of more than two stocks. Hence, it should be possible to construct a multi-dimensional pairs trading framework. By incorporating more information from additional stocks, such a framework would likely be more accurate in determining the relatively undervalued or overvalued positions of stocks, thereby yielding higher profits.

Finally, it should be noted that the proposed trading strategy could also be applied to high-frequency trading. Although this chapter considers daily trading for illustration purposes, the overall procedure is not dependent on the time frame used, and the strategy is also applicable to a smaller time interval, such as an hour. This and the other aforementioned unresolved issues constitute interesting research topics for further investigation. I believe that investigation of these issues will help to improve the proposed strategy and provide a broader perspective on the use of copulas in pairs trading.

Figure 2.1 Distance Strategy

Note:

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Formation Period: January 2 to December 30, 2008

Trading Period: December 31, 2008 to July 1, 2009

Figure 2.2 Copula Fitting Result (Formation Period)

Note:

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Formation Period: January 2 to December 30, 2008

Figure 2.3 Copula Fitting Result (Trading Period)

Note:

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Trading Period: December 31, 2008 to July 1, 2009

Figure 2.4 Cumulative Returns for the Top Five Pairs Portfolios (Without One-Day Waiting)

Note:

Sample Period: 2003-2012

Number of Trading Periods: 17 (102 months)

Table 2.1 Information Criteria of Copula Fitting Result

	<i>SIC (-)</i>	<i>AIC (-)</i>	<i>HQIC (-)</i>
<i>Panel A: Formation Period</i>			
Student-T	103.730	110.732	107.943
Clayton	98.906	102.415	101.012
Normal	93.926	97.435	96.032
Frank	87.236	90.745	89.342
Gumbel	85.985	89.495	88.092
<i>Panel B: Trading Period</i>			
Student-T	62.821	68.396	66.189
Gumbel	64.041	66.845	65.725
Normal	57.436	60.240	59.120
Frank	53.658	56.462	55.342
Clayton	50.397	53.201	52.081

Note:

Pair: Brookdale Senior Living Inc. and Emeritus Corporation (BKD-ESC)

Formation Period: January 2 to December 30, 2008

Trading Period: December 31, 2008 to July 1, 2009

SIC = Schwarz information criterion; AIC = Akaike information criterion; HQIC = Hannan-Quinn information criterion.

Table 2.2 Results of Pairs Trading Strategies on Stock Pair BKD-ESC

	Distance Strategy	Copula Strategy
<i>Panel A: Trading strategies implemented as intended</i>		
Total Profit (Capital: \$10,000)	\$-592.99	\$847.27
Total Number of Transactions	3	8
Average Monthly Excess Returns	-0.009	0.014
<i>Panel B: Trading strategies implemented with one-day waiting</i>		
Total Profit (Capital: \$10,000)	\$-1526.23	\$1060.01
Total Number of Transactions	3	7
Average Monthly Excess Returns	-0.011	0.017

Note:

Formation Period: January 2 to December 30, 2008

Trading Period: December 31, 2008 to July 1, 2009

Table 2.3 Excess Returns of Pairs Trading Strategies

Pair portfolio	Distance Strategy			Copula Strategy		
	Top 5	Top 20	101 - 120	Top 5	Top 20	101 - 120
<i>Panel A: Excess return distribution (no waiting)</i>						
Average excess return (committed)	0.0011	0.0008	0.0006	0.0078***	0.0022**	0.0022*
t-Statistic	0.775	1.100	0.846	4.184	2.497	1.870
Excess return distribution						
Median	-0.0000	0.0002	0.0007	0.0062	0.0024	0.0015
Standard deviation	0.0138	0.0073	0.0077	0.0187	0.0089	0.0116
Skewness	1.1047	0.0309	0.1671	1.082	-0.0991	0.3133
Kurtosis	8.3501	3.3476	4.0179	5.7403	3.8341	4.6376
Minimum	-0.0372	-0.0187	-0.0233	-0.0415	-0.0248	-0.0355
Maximum	0.0700	0.0214	0.0255	0.0769	0.0293	0.0419
Observations with excess returns < 0 (%)	50%	49%	46%	35%	39%	44%
<i>Panel B: Excess return distribution (one-day waiting)</i>						
Average excess return (committed)	0.0008	0.0004	0.0006	0.0044***	0.0006	0.0001
t-Statistic	0.5877	0.5822	0.7758	3.0500	0.6640	0.1082
Excess return distribution						
Median	0.0000	-0.0006	0.0006	0.0045	0.0004	0.0005
Standard deviation	0.0144	0.0063	0.0076	0.0147	0.0088	0.0123
Skewness	2.0471	0.2664	0.0009	-0.3860	-0.2221	0.4018
Kurtosis	12.850	3.152	4.392	6.501	5.293	4.996
Minimum	-0.0342	-0.0156	-0.0223	-0.0599	-0.0264	-0.0342
Maximum	0.0814	0.0187	0.0260	0.0542	0.0339	0.0449
Observations with excess returns < 0 (%)	48%	52%	45%	34%	47%	49%

*, **, *** represent significances at the 5%, 1%, and 0.1% levels respectively.

Table 2.4 Trading Statistics and Composition of Pairs Portfolio

Pair portfolio	Distance Strategy			Copula Strategy		
	Top 5	Top 20	101 - 120	Top 5	Top 20	101 - 120
<i>Panel A: Trading statistics (no waiting)</i>						
Average price deviation trigger for opening pairs	0.0416	0.0495	0.0740			
Average number of pairs traded per six-month period	4.82	18.52	17.70	5.00	20.00	20.00
Average number of round-trip trades per pair	1.74	1.54	1.26	6.83	6.60	6.67
Standard deviation of number of rounds trips per pair	1.13	1.03	0.82	2.27	2.11	2.14
Average time pairs are open in months	2.14	1.95	2.00	4.40	4.45	4.51
Standard deviation of time open, per pair, in months	1.49	1.38	1.41	0.52	0.49	0.50
<i>Panel B: Trading statistics (one-day waiting)</i>						
Average price deviation trigger for opening pairs	0.0416	0.0495	0.0740			
Average number of pairs traded per six-month period	4.82	18.52	17.70	5.00	20.00	20.00
Average number of round-trip trades per pair	1.74	1.54	1.26	6.80	6.55	6.62
Standard deviation of number of rounds trips per pair	1.13	1.03	0.82	2.24	2.08	2.13
Average time pairs are open in months	2.12	1.93	1.99	4.36	4.41	4.47
Standard deviation of time open, per pair, in months	1.48	1.37	1.40	0.51	0.49	0.50

Table 2.5 Systematic Risk of Pairs Trading Strategies

Pair portfolio	Distance Strategy			Copula Strategy		
	Top 5	Top 20	101 - 120	Top 5	Top 20	101 - 120
<i>Panel A: Systematic Risks (no waiting)</i>						
Intercept	0.0011	0.0007	0.0006	0.0080***	0.0023*	0.0022*
t-statistics	(0.8360)	(1.0140)	(0.8090)	(4.4690)	(2.6020)	(1.9720)
Market Risk	0.0513	0.0273	0.0343	-0.1521***	0.0187	0.0235
t-statistics	(1.4000)	(1.3740)	(1.6750)	(-3.1450)	(0.7850)	(0.7540)
SMB	-0.0764	-0.0131	-0.0303	0.1935*	-0.0675	-0.0688
t-statistics	(-1.1430)	(-0.3610)	(-0.8110)	(2.1910)	(-1.5570)	(-1.2080)
HML	-0.1206*	-0.0248	-0.0330	-0.0668	-0.0271	-0.0784
t-statistics	(-1.9930)	(-0.7560)	(-0.9790)	(-0.8360)	(-0.6920)	(-1.5220)
Momentum	-0.0296	0.0103	-0.0241	-0.0552	-0.0253	0.0122
t-statistics	(-1.0330)	(0.6600)	(-1.5010)	(-1.4570)	(1.3610)	(0.5000)
<i>Panel B: Excess return distribution (one-day waiting)</i>						
Intercept	0.0008	0.0003	0.0005	0.0044***	0.0005	0.0003
t-statistics	(0.6100)	(0.5050)	(0.7250)	(3.0450)	(0.6390)	(0.6090)
Market Risk	0.0481	0.0185	0.0484**	-0.0723*	0.0589**	0.0324
t-statistics	(1.2560)	(1.0850)	(2.4830)	(-1.8320)	(2.5940)	(0.9710)
SMB	-0.0373	-0.0019	-0.0375	0.1049	-0.0795*	-0.0313
t-statistics	(-0.5340)	(-0.0640)	(-1.0530)	(1.4570)	(-1.9180)	(-0.5140)
HML	-0.1410*	-0.0231	-0.0413	0.0666	-0.0186	-0.0662
t-statistics	(-2.2260)	(-0.8190)	(-1.2830)	(1.0230)	(-0.4980)	(-1.2010)
Momentum	-0.0285	0.0093	-0.0332*	-0.0291	-0.0228	0.0085
t-statistics	(-0.9500)	(0.6970)	(-2.1750)	(-0.9410)	(-1.2810)	(0.3230)

*, **, *** represent significances at the 5%, 1%, and 0.1% levels respectively.

CHAPTER III. APPLICATION OF COPULAS TO LOSS

TRIANGLE BENCHMARKS CONSTRUCTION

3.1 BACKGROUND OF LOSS RESERVING

Establishing reserves for incurred yet unpaid claims is essential in property and casualty insurance. Claims often take time to be reported and processed, and insurers must reserve sufficient provisions to pay for those that are made late. Under- and over-reserving are both undesirable, as the former increases the risk of insolvency and the latter decreases the efficiency of allocating capital.

Reserving for a small business line is not an easy task, mainly due to its small number of policies. Large sample theory states that when the number of policies goes to infinity, the average loss per policy converges to its theoretical mean value. However, the large sample theory may not be applicable for a small business line, and thus there is no certainty that the reported loss ratio, which is calculated as loss over premium, can converge to its expected value. For example, Figure 3.1 shows the ten years of loss ratios for both personal and commercial auto lines of a U.S. insurance company examined in Shi and Frees (2011). Each curve in the graph corresponds to loss ratios at different development lags for a specific accident year. It is observed that the curves of the loss ratios for the personal line are nicely ordered, but that is not the case for the commercial auto line in which the curves exhibit many crosses with each other. The explanation for this observation comes from the large sample theory. The premium size of

the personal auto line, which is a proxy for the number of policies, is around 30 times larger than the premium size of the commercial auto line. For the personal auto line, its loss ratios are supposed to converge to its expected values. A higher loss ratio at an early development lag may indicate that the total expected loss at that specific accident year is larger, leading to higher loss ratios at subsequent development lags. Thus, we observe that the curves are all ordered. On the contrary, a higher loss ratio at an early development lag for the commercial auto line may represent two possibilities, either a larger expected total loss for that accident year or just a random variation from the expected value that happened to be positive. If it is the latter case, then reserving analysis following the same procedure of the large business line may be problematic, as it will over-estimate the reserve for the commercial auto line.

There is a plethora of literature about loss reserving and its methodologies, but few studies addressed the issue of reserving for a small business line. Earlier studies conducted estimations for a single line of business (Verrall, 1996; Taylor, 2000; England and Verrall, 2002; De Alba, 2002; Wüthrich and Merz, 2008; Antonio and Beirlant, 2008). In later studies, researchers paid attention to the correlation of claims between different business lines and established loss reserving models that incorporate interdependence between different lines. One group of researchers modelled this interdependence using distribution-free approaches, which mainly include the multivariate chain-ladder method (Braun, 2004; Schmidt, 2008; Merz and Wüthrich, 2008; Zhang, 2010) and the multivariate additive loss-reserving method (Hess et al., 2006; Merz and

Wüthrich, 2009b). Merz and Wüthrich (2009a) also considered models built on combinations of the two methods. Another group of researchers used parametric methods based on distributional families. The Bayesian model is an effective approach in this category (De Alba, 2006; Shi and Valdez, 2014). Copula-based models have also received increasing attention recently (Brehm, 2002; De Jong, 2012; Shi and Frees, 2011; Zhang and Dukic, 2012). Nevertheless, these methodologies proposed in the literature are built upon the reported loss triangle. For the small business line, the reported loss ratio contains two pieces of information, the expected loss ratio and a random variation that cannot be neglected due to a small sample. If we directly apply existing methodologies to the small business line, then our reserve will probably be over-estimated (under-estimated) if the random term is positive (negative).

In this chapter, I propose to solve the reserving problem for a small business line by observing the historical dependence between the small business line and its related large business line, and use the information from the large line loss ratio to provide a benchmark for the small line loss ratio, as an approximation to the expected loss ratio of the small business line. Due to its large number of policies, the loss ratio of a large business line can be treated as being fairly close to its expected value, and this available information could be very useful to determine the expected loss ratio of the small line, and help to differentiate it from the random variation. For example, if the loss ratio of the related large business line is high, then the high loss ratio of the small line will probably be

caused by excessive claims made in that particular accident year, instead of a positive random variation.

In this chapter, I use the copula modelling technique to model the dependence between the small line and its related large line. The use of copulas in risk management and actuarial modelling has increased rapidly in recent decades. Frees and Valdez (1998) provided an inspiring description of the use of the copula technique to model the joint dependence of random variables in the actuarial context. There are two advantages to use the copula technique. First, it allows individual marginal distributions to be estimated separately and thus permits researchers to use specific distributions to account for the heavy tails commonly seen in insurance data. Second, it allows the flexibility in choosing the functional form to describe the dependence. This is crucial to actuarial data as they usually exhibit strong tail dependence. Using the commercial and personal auto lines as an example, one can expect a strong right-tail dependence between their loss ratios given circumstances such as storms or downpours that significantly affect both lines of insurance.

The contribution of this chapter is twofold. First, I propose a copula regression model to provide a benchmark for the small line loss ratio in order to differentiate its expected loss ratio from the random variation term. Second, I demonstrate how this derived benchmark can be used to improve the reserve accuracy of the small business line. I first conduct simulation analysis to demonstrate the effectiveness of the proposed benchmark. Subsequently, I

perform empirical analysis using data from U.S. leading insurance companies. Significant improvements in the reserve estimations are obtained when I modify the loss ratio triangle with the proposed benchmark at the bottom left corner, using either the chain-ladder method or an advanced copula approach proposed in Shi and Frees (2011) to calculate the reserves.

3.2 CONSTRUCTION OF LOSS TRIANGLE BENCHMARKS

The aim of this study is to provide a benchmark to approximate for the expected loss ratio of the small business line using information from its related large line. It is noted that the “large vs. small” concept may refer to various combinations of business lines, such as “personal vs. commercial auto lines,” “large vs. small insurance companies” and “large vs. small country businesses.” Before introducing the procedure to construct the benchmark, below is a list of the terminology and notations that are used throughout this chapter.

- **Reported Loss Triangle:** This assembles all reported incremental loss ratios $\theta_{i,j}^{(n)}$, where i refers to the accident year with the maximum value I and j refers to the development lag up to J . The superscript (n) , $n \in \{S, L\}$, indicates the small line and the large line.
- **Total Loss Ratio:** $\theta_{i,T}^{(n)}$ represents the total loss ratio at accident year i for the n^{th} business line. It is calculated as $\theta_{i,T}^{(n)} = \sum_{t=0}^J \theta_{i,t}^{(n)}$.
- **Total Undeveloped Loss Ratio:** $\theta_{i,j,T}^{(n)}$ represents the total undeveloped loss ratio at accident year i and development lag j for the n^{th} business line. It is

calculated as $\theta_{i,j,T}^{(n)} = \sum_{t=j}^I \theta_{i,t}^{(n)}$

- **Benchmark:** $M_{i,j}^{(S)}$ represents the proposed benchmark for $\theta_{i,j}^{(s)}$.
- **Modified Loss Triangle:** It is equivalent to the reported loss triangle except that $\theta_{I,0}^{(s)}$ is replaced by $M_{I,0}^{(S)}$.

Though the proposed benchmark can be used to justify any position in the loss triangle, throughout the discussion of this chapter, I aim at the left bottom corner of the loss triangle to illustrate the idea of constructing the benchmark (e.g. In the Appendix A1, the left bottom corner refers to the accident year 2007, development lag 0), as the latest accident year has minimum information available but largest uncertainties needed to be estimated. Thus, for the rest of the chapter, the goal is to provide a benchmark $M_{I,0}^{(S)}$ for $\theta_{I,0}^{(s)}$ and explore the implications of it on reserve estimation.

The random vector of the multivariate incremental loss ratio θ_{ij} can be expressed as

$$\theta_{ij} = (\theta_{ij}^{(S)}, \theta_{ij}^{(L)}), i \in \{0, \dots, I\} \text{ and } j \in \{0, \dots, J\}, \quad (3.1)$$

It is conventionally assumed that $I > J$ in most loss reserving models. However, our discussion is not restricted to this limitation, and can be further applied to imbalanced cases.

3.2.1 Conditional Distribution of the Small Line Loss Ratio

To provide a reliable benchmark for the loss ratio of a small business line using information from its related large line, I estimate the conditional distribution of the loss ratio of a small business line on the loss ratio of its related large line.

The conditional distribution function $D_{\theta_{i,j}^{(S)}|\theta_{i,j}^{(L)}}$ of the loss ratio $\theta_{i,j}^{(S)}$ given $\theta_{i,j}^{(L)}$ is defined as

$$D_{\theta_{i,j}^{(S)}|\theta_{i,j}^{(L)}}(\theta_{i,j}^{(S)}) = P(\theta_{i,j}^{(S)} < \theta_{i,j}^{(S)} | \theta_{i,j}^{(L)} = \theta_{i,j}^{(L)}). \quad (3.2)$$

According to the large sample theory, when the number of policies tends to infinity, $\theta_{i,j}^{(L)}$ will converge to $E[\theta_{i,j}^{(L)}]$, i.e., $\theta_{i,j}^{(L)} = E[\theta_{i,j}^{(L)}]$, then the conditional distribution function becomes

$$D_{\theta_{i,j}^{(S)}|\theta_{i,j}^{(L)}}(\theta_{i,j}^{(S)}) = P(\theta_{i,j}^{(S)} < \theta_{i,j}^{(S)} | \theta_{i,j}^{(L)} = E[\theta_{i,j}^{(L)}]). \quad (3.3)$$

Either the mean or median of this conditional variable can serve as a good candidate of the benchmark for the small line loss ratio. I adopt the median measurement in this study, as it can be solved explicitly using most of the copula functions, while the mean measurement would require numerical methods to deal with the integration problem. Moreover, the distribution of loss ratio usually has heavy tails, so the median measurement is a better choice in this case. I define the median of the conditional random variable $\theta_{i,j}^{(S)} | \theta_{i,j}^{(L)} = E[\theta_{i,j}^{(L)}]$ as $M_{i,j}^{(S)}$.

One prerequisite for making this benchmark valid is that the small and large business lines perform strongly in a dependent fashion. Otherwise, the

conditional distribution of the loss ratio of the small line based on the large line would have little inferential power. In an extreme case where the small line is independent of the large line, the conditional distribution of the loss ratio of the small line would be the same as its unconditional distribution. In contrast, with a strong dependence such as that exhibited by a highly positive correlation, a higher loss ratio of a large business line would indicate unusual higher excessive claims in that particular accident year. The defined benchmark $M_{i,j}^{(S)}$, which is the median of the conditional distribution of $\theta_{i,j}^{(S)} | \theta_{i,j}^{(L)} = E[\theta_{i,j}^{(L)}]$, would be higher in this case.

3.2.2 Copula-Based Benchmark

In order to estimate the conditional distribution $D_{\theta_{i,j}^{(S)} | \theta_{i,j}^{(L)}}(\theta_{i,j}^{(S)})$, I adopt the copula regression model in this study. In the loss-reserving context, there are numerous studies that implement copula techniques (Brehm, 2002; Frees and Wang, 2005; De Jong, 2012; Shi and Frees, 2011; Zhang and Dukic, 2012). However, among the various applications of the copula technique, copula regression is a relatively new methodology (Crane and Van Der Hoek, 2008; Parsa and Klugman, 2011). Compared with conventional regression methods such as simple and generalized linear regression models, the copula regression model eliminates rigid assumptions about the residuals and allow flexibility in the functional form describing the dependence between different variables. Thus, it can be considered as a generalization of the conventional regression models. Crane and Van Der Hoek (2008) provided formulas for the copula regression

model in two-dimensional cases. Parsa and Klugman (2011) extended the model to multi-dimensional cases using Gaussian and Student-T copulas.

The aim of the copula regression is to first model the joint distribution of the dependent variable and its covariates, followed by the derivation of the conditional distribution of the dependent variable on the covariates. Denote the marginal probability distribution functions of $\boldsymbol{\theta}_{i,j}^{(S)}$ and $\boldsymbol{\theta}_{i,j}^{(L)}$ as $F_{i,j}^{(S)}$ and $F_{i,j}^{(L)}$, i.e.,

$$F_{i,j}^{(S)}(\boldsymbol{\theta}_{i,j}^{(S)}) = \text{Prob}(\boldsymbol{\theta}_{i,j}^{(S)} < \boldsymbol{\theta}_{i,j}^{(S)}) \quad (3.4)$$

$$F_{i,j}^{(L)}(\boldsymbol{\theta}_{i,j}^{(L)}) = \text{Prob}(\boldsymbol{\theta}_{i,j}^{(L)} < \boldsymbol{\theta}_{i,j}^{(L)}). \quad (3.5)$$

Denote $\mathbf{u}_{i,j}^{(S)} = F_{i,j}^{(S)}(\boldsymbol{\theta}_{i,j}^{(S)})$ and $\mathbf{u}_{i,j}^{(L)} = F_{i,j}^{(L)}(\boldsymbol{\theta}_{i,j}^{(L)})$. Assume that $H_{i,j}$ is the joint distribution function of $\boldsymbol{\theta}_{i,j}^{(S)}$ and $\boldsymbol{\theta}_{i,j}^{(L)}$. According to Sklar's Theorem (Sklar, 1959), there must exist a copula C_j that satisfies

$$H_{i,j}(\boldsymbol{\theta}_{i,j}^{(S)}, \boldsymbol{\theta}_{i,j}^{(L)}) = C_j(\mathbf{u}_{i,j}^{(S)}, \mathbf{u}_{i,j}^{(L)}) = C_j[F_{i,j}^{(S)}(\boldsymbol{\theta}_{i,j}^{(S)}), F_{i,j}^{(L)}(\boldsymbol{\theta}_{i,j}^{(L)})]. \quad (3.6)$$

In this chapter, I assume that the copula is invariant of accident year i for simplification purpose of my model. Although there exist some exogenous factors, it is reasonable to assume that the dependency structure between the large and small lines are not easily shifted over time. Meanwhile, I am not excluding exogenous factors such as accident year and development year, as the associations of these effects are exactly part of the dependence that we want to describe between the two lines. Nevertheless, precise modelling of the time-

variant dependence or the exogenous factors may increase the quality of the benchmark, but it is not the main interest in this study.

The conditional distribution function thus can be derived by taking the first derivative of the copula function with respect to $\mathbf{u}_{i,j}^{(L)}$.

$$\begin{aligned}
D_{\boldsymbol{\theta}_{i,j}^{(S)}|\boldsymbol{\theta}_{i,j}^{(L)}}(\boldsymbol{\theta}_{i,j}^{(S)}) &= P(\boldsymbol{\theta}_{i,j}^{(S)} < \boldsymbol{\theta}_{i,j}^{(S)} | \boldsymbol{\theta}_{i,j}^{(L)} = \boldsymbol{\theta}_{i,j}^{(L)}) \\
&= P\left(F_{i,j}^{(S)}(\boldsymbol{\theta}_{i,j}^{(S)}) < F_{i,j}^{(S)}(\boldsymbol{\theta}_{i,j}^{(S)}) \middle| F_{i,j}^{(L)}(\boldsymbol{\theta}_{i,j}^{(L)}) = F_{i,j}^{(L)}(\boldsymbol{\theta}_{i,j}^{(L)})\right) \\
&= P(\mathbf{u}_{i,j}^{(S)} < \mathbf{u}_{i,j}^{(S)} | \mathbf{u}_{i,j}^{(L)} = \mathbf{u}_{i,j}^{(L)}) \\
&= \partial C_j(u_{i,j}^{(S)}, u_{i,j}^{(L)}) / \partial u_{i,j}^{(L)}. \tag{3.7}
\end{aligned}$$

Then the benchmark $M_{i,j}^{(S)}$ can be derived by solving the following equation that

$$D_{\boldsymbol{\theta}_{i,j}^{(S)}|\boldsymbol{\theta}_{i,j}^{(L)}}(M_{i,j}^{(S)}) = 0.5. \tag{3.8}$$

3.3 SIMULATION ANALYSIS ON THE PERFORMANCE OF PROPOSED BENCHMARKS

In this section, simulation analysis is conducted in order to examine the ability of the proposed benchmark in approximating the expected loss ratio of the small business line as well as its implications on the subsequent loss reserving. To accomplish this goal, we first have to simulate reported loss triangles representing two different business lines, with one of them having characteristics

of a small business line and the other similar to a large line. Moreover, the two lines should be simulated in a way that is correlated with each other.

Assume that the two lines are denoted as L and S standing for the large line and the small line, respectively. I simulate loss ratios for the two lines up to accident year I and development lag J. The detailed simulation procedure is as follows.

- Given a selected copula structure C with parameter value θ , I simulate I pairs $(u_{iT}^{(L)}, u_{iT}^{(S)})$ simultaneously. $u_{iT}^{(L)}$ and $u_{iT}^{(S)}$ represent the cumulative densities of the total loss ratios at accident year i for the large line and the small line respectively.
- Assume that $F_T^{(L)}$ and $F_T^{(S)}$ represent the marginal cumulative distribution function for the total loss ratios of the two lines. Then the total loss ratios $\theta_{i,T}^{(L)}$ and $\theta_{i,T}^{(S)}$ for accident year i are derived from

$$\theta_{i,T}^{(L)} = F_T^{(L)-1}(u_{iT}^{(L)}) \quad (3.9)$$

$$\theta_{i,T}^{(S)} = F_T^{(S)-1}(u_{iT}^{(S)}) \quad (3.10)$$

$\theta_{i,T}^{(L)}$ and $\theta_{i,T}^{(S)}$ are assumed to follow log-normal distributions with mean $\mu^{(L)}$ and $\mu^{(S)}$ and standard deviation $\sigma^{(L)}$ and $\sigma^{(S)}$.

- Loss ratios for each development lag j for the two lines are simulated as

$$\theta_{ij}^{(L)} = \theta_{i,j,T}^{(L)}(a_{ij}^{(L)} + \delta_{ij}^{(L)}) \quad (3.11)$$

$$\theta_{ij}^{(S)} = \theta_{i,j,T}^{(S)}(a_{ij}^{(S)} + \delta_{ij}^{(S)}) \quad (3.12)$$

where $a_{ij}^{(L)}$ and $a_{ij}^{(S)}$ are the expected development ratios at accident year i and development lag j , and $\delta_{ij}^{(L)}$ and $\delta_{ij}^{(S)}$ are assumed to follow uniform random distributions from $(-\varepsilon_{ij}^{(L)}, \varepsilon_{ij}^{(L)})$ and $(-\varepsilon_{ij}^{(S)}, \varepsilon_{ij}^{(S)})$.

- The proposed benchmark $M_{I,0}^{(S)}$ is then calculated using the method described in Section 3.2. The expected loss ratio $\overline{\theta_{I,0}^{(S)}}$ is calculated as $\theta_{I,T}^{(S)} a_{I,0}^{(S)}$.
- Using the simulated reported loss triangle, I adopt the chain-ladder method and calculate the reserve for accident year I , denoted as R_I . Then I use the derived benchmark to replace $\theta_{I,0}^{(S)}$ in the loss triangle, and calculate the reserve using the chain-ladder method again, denoted as R'_I . The simulated total undeveloped loss ratio $\theta_{I,1,T}^{(S)}$ for the remaining development lags for accident year I is calculated as $(\theta_{I,T}^{(S)} - \theta_{I,0}^{(S)})$.

Through the simulation analysis, some values of the parameters are fixed as: $I = 100$, $J = 9$, $\mu^{(L)} = 0.8$, $\mu^{(S)} = 0.7$, $\sigma^{(L)} = 0.4$, $\sigma^{(S)} = 0.07$, $\varepsilon_{ij}^{(L)} = 0.01$ and $\varepsilon_{ij}^{(S)} = 0.1$. $\sigma^{(S)}$ and $\varepsilon_{ij}^{(S)}$ are set to be larger than $\sigma^{(L)}$ and $\varepsilon_{ij}^{(L)}$ in order to reflect that small business line has a larger variation in the reported loss ratio from its expected value. Despite the fixed parameters, I allow changes in the copula types and parameters to test the effectiveness of the proposed benchmark.

Each group of simulation is run for 1000 times. After that, the mean squared error and the absolute error of the benchmark and the simulated reported loss ratio from the expected loss ratio are recorded, as well as for the mean squared error (MSE) and the absolute error (ABS) of the two reserves from the actual simulated total undeveloped loss. It is noticed that the performance of the benchmark may be affected, but should not be subject to the reserving method used. The benchmark construction is considered as a general step before any reserving methods are applied. In the empirical analysis, I also incorporate other models to demonstrate this point.

The first group of simulations shows the performance of the proposed benchmark under different copula structures. Commonly used structures including Gumbel, Clayton, Frank and normal are included. The parameter value is set at 10 for Gumbel, Clayton and Frank copula, and 0.8 for normal Copula. Panel A of Table 3.1 shows the performance of the proposed benchmark in approximating the expected loss ratio of the small business line compared with the reported loss ratio in these scenarios. Panel B of Table 3.1 shows the loss reserving accuracies obtained by either adopting the proposed benchmark or directly using the reported loss ratios in the simple chain-ladder method. It is observed that regardless of the copula structure used, the benchmark is always closer to the pre-determined expected loss ratio with smaller ABS and MSE compared with the reported loss ratios. Moreover, this advantage is transferred to the loss reserving accuracy estimation, where we observe significant gain in

the loss reserving accuracies by adopting the proposed benchmark in the loss reserving calculation than directly using the reported loss ratios.

The second group of simulations shows the performance of the proposed benchmark under different degrees of associations. Theoretically, the benchmark should only be effective when there is a strong association between the two lines, otherwise, the conditional distribution of the small line loss ratio on the large line will be identical to the unconditional distribution. In this simulation, I fix the copula structure to be Gumbel copula and vary the copula parameter from 2, 3, 5, 10, 20 to 50. A larger value of the parameter for the Gumbel copula represents higher degree of associations. The results are shown in Table 3.2. In Panel A, it is observed that there is a decreasing trend in the ABS and MSE of the benchmark errors when the copula parameter goes from 2 to 50. A Gumbel copula with parameter 2 indicates weak associations, thus the performance of the proposed benchmark is not any different from the reported loss ratios in approximating the expected loss ratio, though we still can see marginal improvement in the loss reserving accuracies shown in Panel B. When the associations are strong, both errors of the benchmark and the subsequent loss reserving accuracies are largely improved, which is in line with my theory.

In other words, the simulation analysis has demonstrated the effectiveness of the proposed benchmark in approximating the expected loss ratio of the small business line, especially when there is a strong association between the small line and its related large line. By replacing the loss triangle with the derived

benchmark, the loss reserving accuracies can also be increased significantly. In the next section, I will continue my analysis using real data from U.S. leading property and casualty insurance companies.

3.4 EMPIRICAL ANALYSIS ON LEADING U.S. P&C

INSURANCE COMPANIES

My empirical analysis is mainly based on five leading U.S. property and casualty insurance companies, investigating both the personal and commercial auto lines. I choose the five leading insurance companies to ensure that the premium size for the personal auto line, a proxy for the number of policies, is large enough to conduct my analysis. For some insurance companies, even the premium size for the personal line is much larger than the commercial auto line, it is still relatively small for their loss ratios to converge to the expected values. In this case, the personal auto line is the large business line, and the commercial auto line is the small business line. The two lines are strongly associated with each other. It is important to note that the proposed model is not restricted to the area of personal and commercial auto lines. These lines are chosen because the related data are publicly available. The proposed model can be applied to any correlated small line and large line.

3.4.1 Data Description

The data used in this study were taken from Schedule P of the National Association of Insurance Commissioners (NAIC) database, which provides

annual firm-level paid claims for major business lines of property and casualty insurers. I extract the paid claims for the personal and commercial auto lines of five leading U.S. insurance companies from Schedule P - Parts 3B and 3C, covering a sample period of 1989-2007. I choose leading insurance companies in term of size to make sure that their personal auto line is large enough to conduct my analysis. I also choose five companies instead of only one to avoid any data-snooping criticisms. The companies are chosen according to the premium size of the personal auto line in 1989, with the requirement that no entries of the paid claims triangles could be missing for either line until 2007. The paid claims are then normalized by the earned premiums for the corresponding accident years. For this particular example, the development lag is up to nine years.

Some of the summary statistics for the two lines of the five insurance companies are shown in Table 3.3. On average, the premium size of the personal auto line is 20 times the size of the commercial auto line, with a range of 3 to 40. The correlation between the loss ratios of the personal and commercial lines is relatively high at development lags zero or one, with six out of ten above 0.7. This is consistent with the requirement that the two lines are highly related.

3.4.2 Benchmarks

I use data from the first company to illustrate the benchmark calculation. The paid claims triangles for its personal and commercial auto lines are shown in Appendix A1 and A2. The annual average premium size for the personal line is \$13,860,997, which is around 40 times the premium size of the commercial line

(\$314,933). The correlation between the loss ratios is 0.77 at development lag zero and 0.75 at development lag one. I intend to provide the benchmark for the commercial line loss ratio at development lag zero for the accident year 2006. As my data is up to 2007, I still can obtain real reported loss data of the development lag one for the accident year 2006, and this will be used to make comparisons with my estimated reserves later.

The loss ratios at development lag zero for the different accident years are denoted as

$$\boldsymbol{\theta}_{i,0} = \left(\boldsymbol{\theta}_{i,0}^{(C)}, \boldsymbol{\theta}_{i,0}^{(P)} \right), \quad i \in \{k+0, k+1, \dots, k+19\}, \quad (3.13)$$

where k is equal to the initial year 1989. The aim is then to provide a benchmark for $\boldsymbol{\theta}_{2006,0}^{(C)}$. First, $\boldsymbol{\theta}_{i,0}^{(C)}$ and $\boldsymbol{\theta}_{i,0}^{(P)}$, $i \in \{k+0, k+1, \dots, k+19\}$ are fitted separately for the optimal marginal distributions. In this example, they are best fitted via generalized extreme value distribution among a handful of selections according to likelihood values, denoted as $F^{(C)}$ and $F^{(P)}$. The data are then transformed into their cumulative probability measurements $u_{i,0}^{(C)}$ and $u_{i,0}^{(P)}$, where

$$u_{i,0}^{(C)} = F^{(C)}\left(\boldsymbol{\theta}_{i,0}^{(C)}\right) \quad (3.14)$$

$$u_{i,0}^{(P)} = F^{(P)}\left(\boldsymbol{\theta}_{i,0}^{(P)}\right). \quad (3.15)$$

The scatter plot of $u_{i,0}^{(C)}$ versus $u_{i,0}^{(P)}$ can be found in Figure 3.2. They are highly positively correlated, with a strong right-tail dependence. The positive

correlation is expected, as they both refer to auto insurance. The strong right-tail dependence is also reasonable, as they are more commonly affected by extreme shocks such as bad weather, making their dependence stronger at the right tail.

Five commonly used copulas (Gumbel, Frank, Clayton, Normal and Student-T) are then fitted to the cumulative probability measurements. The general fitting result is shown in Table 3.4. With a parameter θ equal to 2.223, the Gumbel copula has the highest likelihood value and is chosen as the best fitted copula. This is not surprising, as it best describes the strong right-tail dependence among the five candidates.

Once the copula is estimated, the resulting conditional distribution function can be derived. In this example, the conditional distribution function of $\theta_{i,0}^{(C)} | \theta_{i,0}^{(P)}$ can be calculated as

$$D_{\theta_{i,0}^{(C)} | \theta_{i,0}^{(P)}}(\theta_{i,0}^{(C)}) = C_{\theta}(u_{i,0}^{(C)}, u_{i,0}^{(P)}) * \left[\left(-\ln u_{i,0}^{(C)} \right)^{\theta} + \left(-\ln u_{i,0}^{(P)} \right)^{\theta} \right]^{\left(\frac{1-\theta}{\theta} \right)} \\ * \left(-\ln u_{i,0}^{(P)} \right)^{\theta-1} / u_{i,0}^{(P)}, \quad (3.16)$$

where

$$C_{\theta} = \exp\{ - [\left(-\ln u_{i,0}^{(P)} \right)^{\theta} + \left(-\ln u_{i,0}^{(C)} \right)^{\theta}]^{1/\theta} \}. \quad (3.17)$$

Deriving the benchmark $M_{2006,0}^{(C)}$ for the commercial line is equivalent to solve the following equation:

$$D_{\theta_{2006,0}^{(C)}|\theta_{2006,0}^{(P)}}\left(M_{2006,0}^{(C)}\right)=0.5, \quad (3.18)$$

which calculates $F^{(C)}\left(M_{2006,0}^{(C)}\right)$ at 0.1798 and $M_{2006,0}^{(C)}$ at 0.1918.

For the remaining four companies, the reported loss ratios and benchmarks of the commercial line at development lag zero for the accident year 2006 are shown in Table 3.5. Two of the companies reported loss ratios are higher than the benchmarks, and the remaining three reported ratios are all lower than their benchmarks respectively.

3.4.3 Implications for Loss Reserving

I adopt two methods in my reserve calculation. The first is the basic chain-ladder method and the second is the copula-based model proposed by Shi and Frees (2011). I argue that modifications to the loss triangle should not depend on the loss-reserving method used, and should be a critical step prior to any kind of reserving model. As a sound data source is required for precise estimations, the modification is a process of providing a better data source, and all kinds of reserving methods can be considered as channels for obtaining results from such a data source. Thus, the reserving model used should not affect this adjustment to the loss triangle. As it is almost impossible to examine all loss reserving models simultaneously, I choose two representatives: the simple chain-ladder method and the copula-based model to illustrate my idea. The former deals with a single line at one time and the latter accounts for the interdependence between different lines simultaneously. Furthermore, the chain-ladder method is a

distribution-free approach, and the copula-based model relies on parametric representation.

For each company, I modify the loss triangle by replacing $\theta_{2006,0}^{(C)}$ with $M_{2006,0}^{(C)}$, which are otherwise equal, and I aim to calculate the reserve of development lag one for the accident year 2006. The results of the reserves derived from both methods with and without modifications to the loss triangles using derived benchmarks for the five companies are shown in Table 3.6, and more direct comparisons of the error terms are shown in Figure 3.3. It is obvious that on average, the estimation errors resulting from the modifications using benchmarks are much smaller than the estimation errors of the reserves derived without modification to the loss triangles, regardless of whether the chain-ladder method or copula-based model is used. Specifically, Table 3.7 shows the fitting results using the copula-based model following the same procedure in Shi and Frees (2011). We can see that the modification of the loss triangle using the derived benchmark leads to a significant change in the regression coefficient derived for the accident year 2006, while almost no changes to any other coefficients. This suggests the importance of using the benchmark to approximate the expected loss ratio of the small business line. Otherwise, the coefficient will be overstated in this case.

The overall results show an average improvement of 6% using the chain-ladder method and 5% using the copula-based method in the loss reserving accuracies. Among the five insurance companies, only the reserving accuracy of

the second company performs a bit worse after modifying loss triangles with the benchmarks derived. There are two possible explanations for this. First, because the proposed benchmark is very close to the reported loss ratio, there is not much difference between using the derived benchmark or not. Second, because the estimated copula, in this case, is not consistent with our intuition of right-tail dependence, I do not rule out the possibility of a potential structure change during my 20-year sample period. Apart from this company, a modified loss triangle can lead to an improvement in reserve estimation accuracy of around 8% on average using the chain-ladder method and 7% using the copula-based method. I then conclude that modifying loss triangles using my proposed benchmark could increase the loss-reserving accuracy of the small insurance line if we are able to identify a highly correlated big line.

3.5 REMARKS

The importance of loss reserving has been widely recognized, and numerous models have been proposed to increase the efficiency and accuracy of reserve estimation. This chapter, which recognizes the abilities of different models in estimating reserves, provides a new view of the problem by looking at the small lines and its interdependence with associated large lines.

Large sample theory predicts that a sample mean will converge to a population mean when the sample size tends to infinity. However, for a small business line, a small insurance company, or even a small country business, the

number of policies is not large enough to ensure the convergence of the sample mean. In this case, there is no way of knowing whether a loss ratio is high (low) due to high (low) expected loss ratio or just random variations. If an expected loss ratio is indeed high or low, then the related reserves present no problems. However, if the loss ratio deviates from its expected value, then the related reserves could be inaccurate. Considering that neither over-reserving nor under-reserving is desirable for insurance companies, it is necessary to approximate an expected loss ratio.

This chapter solves the preceding problem by implementing a copula regression model to provide a benchmark for the loss ratio of a small business line using information from its related large business line. I assume that the reported loss ratio of the large business line is close to its expected value. Thus, the conditional mean or median of the small business line based on the information of the large line can be treated as a reliable benchmark for the expected loss ratio.

I use both simulations and empirical analysis to demonstrate the effectiveness of the proposed benchmark in approximating the expected loss ratios of the small line and its subsequent loss reserving. My empirical analysis implies that regardless of which model is used to estimate reserves, replacing the bottom left corner of the original loss ratio using the proposed benchmark could essentially increase the loss reserving accuracies, though it is noted that the

benchmark is only effective when the two lines are strongly associated with each other.

Moreover, the benchmark can be applied to many other aspects. First, it is not restricted to two-dimensional cases, and can be made more accurate by incorporating more information from various related lines. For example, to estimate the reserves for a business from a small European country, the adjustment can be applied according to similar lines of more than one large country, such as Germany and France. Second, it can be applied beyond small business lines. Although the sample mean of large business lines may be much closer to the population mean, other related large lines can nevertheless be used to further improve the estimation. By involving more information from other lines, the estimation accuracy can be increased. All of these topics are interesting for future research.

Figure 3.1 Multiple Time Series Plots of Loss Ratios

Note: This is a replication of Figure 1 in the study of Shi and Frees (2011). Each line corresponds to an accident year.

Figure 3.2 Cumulative Probabilities of the Personal versus Commercial Auto Lines

Figure 3.3 Reserve Estimation Errors at Development Lag One (Accident Year 2006)

Table 3.1 Simulation Analysis on Assuming Different Copula Structures

Panel A: Performance of Proposed Benchmark							
Copula Parameter		Error (Benchmark)		Error (Reported Loss Ratio)		Difference	p-value
		ABS	MSE	ABS	MSE		
Clayton	10	0.0309	0.0420	0.0353	0.0413	-0.0044	0.0000
Gumbel	10	0.0271	0.0337	0.0368	0.0431	-0.0098	0.0000
Frank	10	0.0337	0.0418	0.0357	0.0418	-0.0020	0.0035
Normal	0.8	0.0311	0.0409	0.0354	0.0414	-0.0043	0.0000
Student-T	0.8, df=2	0.0305	0.0399	0.0358	0.0416	-0.0053	0.0000

Panel B: Loss Reserving Accuracies							
Copula Parameter		Error (Benchmark)		Error (Reported Loss Ratio)		Difference	p-value
		ABS	MSE	ABS	MSE		
Clayton	10	0.0447	0.0579	0.0704	0.0826	-0.0257	0.0000
Gumbel	10	0.0436	0.0537	0.0739	0.0864	-0.0302	0.0000
Frank	10	0.0497	0.0620	0.0723	0.0848	-0.0226	0.0000
Normal	0.8	0.0472	0.0596	0.0712	0.0835	-0.0240	0.0000
Student-T	0.8, df=2	0.0470	0.0592	0.0714	0.0840	-0.0244	0.0000

Table 3.2 Simulation Analysis on Assuming Different Degrees of Associations

Panel A: Performance of Proposed Benchmark						
Parameter	Error (Benchmark)		Error (Reported Loss Ratio)		Difference	p-value
	ABS	MSE	ABS	MSE		
2	0.03590	0.04310	0.03530	0.04160	0.0006	0.6027
3	0.03230	0.04060	0.03500	0.04130	-0.0027	0.0311
5	0.02779	0.03431	0.03556	0.04169	-0.0078	0.0000
10	0.02709	0.03366	0.03684	0.04307	-0.0098	0.0000
20	0.02551	0.03138	0.03596	0.04200	-0.0105	0.0000
50	0.02502	0.03023	0.03568	0.04201	-0.0107	0.0000

Panel B: Loss Reserving Accuracies						
Parameter	Error (Benchmark)		Error (Reported Loss Ratio)		Difference	p-value
	ABS	MSE	ABS	MSE		
2	0.05300	0.06700	0.07000	0.08320	-0.0170	0.0000
3	0.04650	0.05811	0.07010	0.08271	-0.0236	0.0000
5	0.04580	0.05613	0.07181	0.08418	-0.0260	0.0000
10	0.04364	0.05371	0.07387	0.08638	-0.0302	0.0000
20	0.04261	0.05249	0.07188	0.08416	-0.0293	0.0000
50	0.04192	0.05122	0.07200	0.08480	-0.0301	0.0000

Table 3.3 Summary Statistics of Five Leading U.S. Insurance Companies

Serial No.	Average Premium Size (Personal Line)	Average Premium Size (Commercial Line)	Correlation of Loss Ratio at Development Lag Zero	Correlation of Loss ratio at Development Lag One
1	\$13,860,997	\$314,933	0.77	0.75
2	\$3,809,353	\$148,883	0.81	0.39
3	\$1,090,900	\$446,517	0.43	0.33
4	\$3,093,401	\$439,070	0.32	0.76
5	\$14,203,659	\$320,025	0.76	0.78
Average	\$7,211,662	\$333,886	0.62	0.60

Note: The data used for this table were taken from Schedule P of the National Association of Insurance Commissioners (NAIC) database, which provides annual firm-level paid claims for major business lines of property and casualty insurers. I extracted the paid claims for the personal and commercial auto lines from Schedule P - Parts 3B and 3C, covering a sample period of 1989-2007.

Table 3.4 Copula Fitting Result for a Leading U.S. Insurance Company

	Fitted Parameter	Standard Deviation	Likelihood
Gumbel	$\theta = 2.223$	0.814	7.638
Frank	$\theta = 6.129$	2.881	6.288
Clayton	$\theta = 1.697$	1.074	5.344
Normal	$\rho = 0.792$	0.135	7.557
Student-T	$\rho = 0.790$	0.197	7.523
	d.f.=35.459	N.A	

Table 3.5 Benchmarks at Development Lag Zero (Accident Year 2006)

Serial No.	Reported Loss Ratio	Proposed Benchmark
1	0.2142	0.1918
2	0.1226	0.1258
3	0.0997	0.1091
4	0.1314	0.1390
5	0.2121	0.1935

Table 3.6 Reserve Estimations at Development Lag One (Accident Year 2006)

Serial No.	Actual Loss Ratio	Chain Ladder	Chain Ladder (Modification)	Improvement	Copula	Copula (Modification)	Improvement
1	0.145	0.184	0.163	14.44%	0.172	0.154	12.34%
2	0.121	0.124	0.127	-2.66%	0.151	0.155	-3.56%
3	0.130	0.107	0.113	4.14%	0.105	0.110	3.93%
4	0.171	0.149	0.158	5.27%	0.140	0.148	4.51%
5	0.143	0.181	0.166	10.37%	0.171	0.157	9.78%

Note: I adopt both the simple chain-ladder method and the copula-based model proposed by Shi and Frees (2011) to conduct the reserve estimations. Moreover, I apply both the methods to the reported loss triangle, in addition to a modified loss triangle with the benchmark. The improvement is calculated as

$$\text{Improvement} = (|\text{Estimated Reserve}(\text{without modification}) - \text{Actual Loss Ratio}| - |\text{Estimated Reserve}(\text{with modification}) - \text{Actual Loss Ratio}|) / \text{Actual Loss Ratio}.$$

Table 3.7 Reserve Estimation Using Copula-Based Approach

	Frank Copula				Product Copula			
	No Modification		Modification		No Modification		Modification	
	Personal	Commercial	Personal	Commercial	Personal	Commercial	Personal	Commercial
Intercept	-1.164	-1.766	-1.164	-1.766	-1.164	-1.766	-1.164	-1.766
AY=1998	0.005	0.279	0.005	0.279	0.005	0.279	0.005	0.279
AY=1999	0.098	0.255	0.098	0.255	0.098	0.255	0.098	0.255
AY=2000	0.155	0.485	0.156	0.485	0.156	0.485	0.156	0.485
AY=2001	0.172	0.268	0.172	0.268	0.172	0.268	0.172	0.268
AY=2002	0.114	0.343	0.114	0.343	0.115	0.343	0.115	0.343
AY=2003	-0.031	0.265	-0.031	0.265	-0.031	0.266	-0.031	0.266
AY=2004	-0.089	0.131	-0.089	0.130	-0.089	0.131	-0.089	0.131
AY=2005	-0.073	0.203	-0.073	0.203	-0.073	0.203	-0.073	0.203
AY=2006	-0.059	0.225	-0.059	0.114	-0.059	0.225	-0.059	0.115
Dev=1	-0.457	-0.231	-0.457	-0.232	-0.457	-0.230	-0.457	-0.230
Dev=2	-1.302	-0.677	-1.302	-0.678	-1.302	-0.676	-1.302	-0.676
Dev=3	-1.879	-1.182	-1.879	-1.182	-1.879	-1.181	-1.879	-1.181
Dev=4	-2.520	-1.750	-2.520	-1.750	-2.52	-1.750	-2.520	-1.750
Dev=5	-3.227	-2.524	-3.227	-2.524	-3.227	-2.524	-3.227	-2.524
Dev=6	-3.935	-3.203	-3.935	-3.203	-3.935	-3.203	-3.935	-3.203
Dev=7	-4.604	-3.437	-4.604	-3.437	-4.604	-3.437	-4.604	-3.437
Dev=8	-5.283	-4.100	-5.283	-4.100	-5.283	-4.099	-5.283	-4.099
Dev=9	-5.250	-4.744	-5.250	-4.744	-5.250	-4.743	-5.250	-4.743
Scale	0.032	0.143	0.032	0.143	0.032	0.143	0.032	0.143
Dependence	-0.107		-0.107		-		-	
Loglik	464.56		464.66		464.59		464.62	

Note: This table shows the regression fitting result of using the copula-based model proposed by Shi and Frees (2011) for a leading U.S. insurance company. The Frank copula fitting exhibits the highest likelihood values among the five commonly used copulas (Clayton, Gumbel, Frank, Normal, and Student-T). Using the product copula is equivalent to assuming the independence between the personal and commercial lines. “No Modification” refers to use the reported loss triangle and “Modification” refers to use the modified loss triangle.

CHAPTER IV. COPULA REGRESSION WITH BOOTSTRAPPING: A CYBER INSURANCE CONTEXT

4.1 BACKGROUND OF CYBER INSURANCE

Cyber security has received an increasing amount of attention in the last decade. According to Verizon's 2014 Data Breach Investigations Report, in 2013 there were as many as 1,367 confirmed data breaches and 63,437 security incidents in the 50 participating firms. A report by the Centre for Strategic and International Studies shows that 3,000 US companies were hacked in 2013, costing the global economy more than \$400 billion and revealing that both the frequency and severity of cyber-attacks exhibit rapidly increasing trends. These reports strongly demonstrate the need for an insurance mechanism that transfers the risks of cyber-attacks.

Cyber insurance emerged in the 1990s, but developed rapidly in the last decade corresponding to the increasing demand. However, the popularity of this product remains subdued relative to the increasing frequency and severity of cyber-attacks. The insurance brokerage arm of Marsh & McLennan Companies estimates that the US cyber insurance market was only worth \$1 billion in 2013 and \$2 billion in 2014 in gross written premiums, which is far less than the \$400 billion loss induced by cyber-attacks in 2013. This relatively subdued market may be due to the lack of convincing pricing models, which makes it hard for

both insurance companies and firms to evaluate the benefits and costs of selling or purchasing this product.

There are several major problems in pricing cyber insurance products, including the complex cross-sectional dependency structure of risk measures and the paucity of data. First, the relationships between different cyber risks and their associated explanatory covariates can be complicated. As mentioned in Herath and Herath (2011), although problems like moral hazard and adverse selection are commonly faced by all insurance products, IT security has certain unique characteristics. The risks associated with the Internet in terms of location, degree, and visibility are similar to those associated with cyber-attacks, and thus firms usually adopt similar measures for these risks, which make the risk exposures of different firms interdependent of each other, creating difficulties in pricing cyber insurance products. Second, multiple researchers have recognized that pricing insurance products is complicated by the paucity of data (Gordon et al., 2003; Baer and Parkinson, 2007; Herath and Herath, 2011). The prices of insurance products are traditionally based on actuarial tables constructed from historical records. However, due to the short history of cyber insurance products, it is hard to obtain sufficient data to perform pricing analyses. Moreover, due to reputation concerns, companies may not be willing to reveal that they have been attacked or the true loss data due, which makes it even more difficult to collect sufficient data to perform pricing analyses.

Few studies have attempted to solve these difficulties in cyber insurance pricing. Gorden et al. (2003) provided a sound qualitative analysis of a framework for applying cyber insurance and described how an insurance mechanism can be used as a tool to manage cyber risks. However, quantitative discussions of cyber insurance pricing are still under-developed. One research stream focuses on the theoretical development of cyber insurance pricing. For example, Mukhopadhyay et al. (2005) proposed a utility model to price cyber insurance products. It started with the utility functions of individuals and calculated the premium that gives them incentives to buy cyber insurance. Barrachini and Addressi (2014) also proposed a multistate model to deal with cyber losses. It focused on the development of different stages of cyber risks and studied the transition probabilities between them. Ogut et al. (2011) discussed the optimal level of cyber insurance taken by a firm under correlated cyber risks, but it had no implications for cyber insurance pricing. To summarize, these theoretical models are useful for understanding the instincts of cyber insurance, but they are difficult to implement due to data difficulties. Another research stream focuses on statistical development in cyber insurance products using historical data. Böhme and Kataria (2006) studied models and measurements of correlations between global cyber risks. They proposed a new classification for the correlation properties of cyber-risks based on a twin-tier approach, but they did not provide a way to deal with the data paucity problem. Herath and Herath (2011) used a copula model to deal with cyber insurance pricing. However, their model was restricted to the specific correlation between the loss and the number

of computers affected, and it also could not solve the data paucity problem. In sum, although pricing has been partially addressed by some researchers, there is no convincing and executable pricing model for cyber insurance that can deal with both complex inter-dependencies and data paucity.

In this chapter, I propose a copula framework with a bootstrapping approach to deal with the cyber insurance pricing. I first use a copula regression model to study the associations between the cyber losses and its explanatory covariates, and then implement a copula-based residual bootstrapping procedure to deal with the data paucity problem.

Copula regression serves as a generalization of the conventional linear regression models to study how the dependent variable, which is the cyber loss variable in this research, is related to its explanatory covariates. Casualty losses are usually skewed to the right, and other than mathematical simplicity, there is no logical argument to support a linear relationship between the dependent variable and its explanatory covariates (Parsa and Klugman, 2011). Compared to the conventional linear regression models, copula regression has the advantage of estimating the joint distribution of the dependent variable and its explanatory covariates, taking into consideration both the linear and non-linear associations between them. It also imposes no restriction and allows separate estimation of the marginal distributions for every interested variable.

I further implement a copula-based residual bootstrapping procedure based on the copula regression model. Bootstrapping is a method for assigning

measures of accuracy to sample estimates (Efron and Tibshirani, 1979). This technique allows estimation of the sampling distribution of almost any statistics using simple methods (Varian, 2005). Such a technique is especially useful when handling limited data since it provides a way to account for the distortions caused by the specific sample that may not be fully representative of the population. To simultaneously bootstrap the dependent variable and its associated explanatory variables, one of the solutions is to use the residual bootstrapping (Pardoe, 2000), which calculates the residuals between the raw data and the expected value, and generates pseudo-data by resampling the residuals. This method is based on the conventional linear regression models, which requires linear assumptions between the dependent variable and its associated explanatory covariates, and imposes restrictions on the marginal distributions of the random variables to be normal. In this research, these assumptions are most likely not valid in the context of cyber insurance. Thus, I propose a copula-based residual bootstrapping procedure instead to accommodate for the necessity of understanding the dependency structure between the loss variable and its explanatory covariates.

Overall, the main contribution of this chapter is to provide a general copula framework with a bootstrapping approach to deal with the difficulties commonly encountered in cyber insurance pricing, including the complex interdependencies and the data paucity problem. To support my proposed model, I first conduct simulation analysis, and the results suggest that the proposed model can better capture the population characteristics when the sample size is small. Specifically, the tail information is easily lost when the sample size is small, but

the proposed copula-based bootstrapping procedure can compensate for this information loss. I further perform additional empirical data analysis using a real industry dataset. The result is consistent with the simulation analysis, which indicates that insurers may gain a better understanding of the tail distribution of the cyber loss variable using the proposed model. This is essential, as if the right tail is underestimated, there is a higher probability of insolvency, whereas if the left tail is underestimated, then the estimated premium could be higher than it should be, leading to the potential loss of competitiveness in the market.

4.2 COPULA REGRESSION WITH BOOTSTRAPPING

My proposed framework mainly consists of two components, i.e., the copula regression model and the copula-based residual bootstrapping procedure. In the first sub-section, I introduce the copula regression model to study the relationships between the loss variable and its associated explanatory variables. After that, I describe the copula-based residual bootstrapping procedure to deal with the data paucity problem. In the last sub-section, I summarize the pricing steps of my proposed model.

4.2.1 Copula Regression: Cyber Loss Variable versus its Explanatory Covariates

Insurance data are usually non-normal and exhibit excessive skewness and kurtosis (Frees and Valdez, 1998). In the cyber insurance context, the dependent variable is the loss variable, which has complex associations with its explanatory covariates such as firm characteristics, security measures, and computer use. As

a result, a linear regression model may not be appropriate, as it only investigates the linear associations between the dependent variables and its associated explanatory covariates and imposes rigid assumptions on their marginal distributions. Thus, I use the copula regression model to investigate the relations between the cyber loss variable and its explanatory covariates.

Suppose Y is the dependent variable (loss variable), with $X_1, X_2, \dots, X_p$ being the main explanatory covariates. Assume G is the cumulative density function of Y , and $F_i, i = 1, 2, \dots, p$, are the cumulative density functions of the explanatory covariates. According to Sklar's theorem (Sklar, 1959), there exists a copula C linking the joint distribution function H of $Y, X_1, X_2, \dots, X_p$ to their separate cumulative densities, i.e.

$$H(y, x_1, x_2, \dots, x_p) = C(G(y), F_1(x_1), F_2(x_2), \dots, F_p(x_p)). \quad (4.1)$$

Denote $u = G(y)$ and $v_j = F_j(x_j), j = 1, 2, \dots, p$.

Crane and Van Der Hoek (2008) demonstrated that the formula of the conditional probability distribution $\text{Prob}(Y \leq y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$ is

$$\begin{aligned} & \text{Prob}(Y \leq y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) \\ &= \text{Prob}(U \leq u | V_1 = v_1, V_2 = v_2, \dots, V_p = v_p) \\ &= \frac{\partial C(u, v_1, v_2, \dots, v_p)}{\partial v_1 \partial v_2 \dots \partial v_p} \bigg/ \frac{\partial C(1, v_1, v_2, \dots, v_p)}{\partial v_1 \partial v_2 \dots \partial v_p}. \end{aligned} \quad (4.2)$$

The intention of the copula regression is to get the probability distribution of Y conditional on the values of its explanatory covariates. After obtaining the conditional distribution, the conditional mean, median, standard deviation, or the estimated confidence interval of the dependent variable Y can be easily derived for statistical inference.

4.2.2 Copula-Based Bootstrapping Procedure

Copula regression can address the complex inter-dependencies among cyber risks; however, the statistical inferences based on the regression may not be reliable due to the data paucity problem. Thus, I further implement a copula-based bootstrapping procedure to improve the accuracy and reliability of my model.

Bootstrapping procedures resample the original sample data and assign measures of accuracy to sample estimates based on the resampled data (Efron and Tibshirani, 1994). By resampling multiple times, one can get a fairly good idea of the sampling distribution of a particular variable (Kesar and Xie, 2005). Moreover, assuming that all of the resampled data consist of a reasonable approximation of the original population, the qualities of the bootstrapped inference statistics can be assumed to be acceptable.

A commonly used method for dealing with a simultaneous simulation of the dependent variable (loss variable) and its associated explanatory covariates is the residual bootstrapping method.

Suppose the original sample of Y is $(y_1, y_2, \dots, y_n)$ with p explanatory variables. The residual bootstrapping requires running a linear regression at the first stage and find all fitted value $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$. Then all residuals $(\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_n)$ from the regression analysis are calculated by

$$\hat{\epsilon}_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n. \quad (4.3)$$

After that, a resample $(y_1^*, y_2^*, \dots, y_n^*)$ is calculated by

$$y_i^* = \hat{y}_i + \hat{\epsilon}_i^*, \quad i = 1, 2, \dots, n. \quad (4.4)$$

where $\hat{\epsilon}_i^*$ is randomly picked from $\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_n$ with replacement.

The residual bootstrapping technique is a good choice when dealing with correlated variables. However, its limitation is that it only accounts for the linear associations between the dependent variable and its associated explanatory covariates, and imposes normal marginal distribution assumption on the dependent variable. As under the context of cyber insurance, there exist complex inter-dependencies, which indicates that simple residual bootstrapping may not be desirable. Thus, I propose a copula-based residual resampling technique. Instead of using linear regression to get the residuals, I use copula regression model to obtain the conditional cumulative density of the dependent variable and make resamples based on that.

To conduct the copula-based residual bootstrapping, I first estimate $u = G(y)$ as the cumulative density function of Y , and $v_j = F_j(x_j), j = 1, 2, \dots, p$, as

the cumulative densities of the explanatory variables. Second, I estimate a copula C linking the joint distribution function H of $Y, X_1, X_2, \dots, X_p$ to their separate cumulative densities. Then the conditional distribution of $\text{Prob}(Y \leq y | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$ can be derived using Equation (4.2).

I denote the conditional cumulative density of y_i as $\hat{p}_i$ where

$$\hat{p}_i = \text{Prob}(U \leq u_i | V_{1,i} = v_{1,i}, V_{2,i} = v_{2,i}, \dots, V_{p,i} = v_{p,i}), \quad i = 1, 2, \dots, n. \quad (4.5)$$

After that, a resample $(u_1^*, u_2^*, \dots, u_n^*)$ is calculated by solving the equation below

$$\hat{p}_i^* = \text{Prob}(U \leq u_i^* | V_{1,i} = v_{1,i}, V_{2,i} = v_{2,i}, \dots, V_{p,i} = v_{p,i}), \quad i = 1, 2, \dots, n. \quad (4.6)$$

where $\hat{p}_i^*$ is randomly picked from $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_n$ with replacement. Therefore, the corresponding resample $(y_1^*, y_2^*, \dots, y_n^*)$ are calculated by

$$y_i^* = G^{-1}(u_i^*), \quad i = 1, 2, \dots, n. \quad (4.7)$$

Repeating the above process many times allow us to get a series of resamples of the dependent variable Y and its associated explanatory covariates $X_1, X_2, \dots, X_p$. After that, we can conduct the calculation of the premiums of different policies based on each resample, and approximates a sampling distribution of the premium. Based on this approximated sampling distribution, we may obtain more accurate and reliable premiums of the cyber insurance policies.

4.2.3 Steps of Pricing Cyber Insurance

With the copula regression model and the copula-based residual bootstrapping procedure, below is a summary of the pricing process for the cyber insurance products.

Step 1: Identify the loss variable Y and its associated explanatory covariates $X_1, X_2, \dots, X_p$.

Step 2: Perform copula regression analysis and use Equation (4.5) to obtain the series of the conditional cumulative densities $\hat{p}_i$, $i = 1, 2, \dots, n$.

Step 3: Produce S resamples denoted as R_i , $i = 1, 2, \dots, S$ ($S=1,000$ in the simulation and empirical analysis) using Equation (4.6) and (4.7).

Step 4: Given a resample R_i , assume $T: R_i \rightarrow P_i$ is the mapping from any resample to its derived premium.

Step 5: Given P_i , $i = 1, 2, \dots, S$ derived from S resamples, we can approximate the sampling distribution of the premium P . The mean and standard deviation of the premium P are calculated as follows

$$\hat{P} = \frac{1}{S} \sum_{i=1}^S P_i, \quad (4.8)$$

$$\sigma(\hat{P}) = \sqrt{\frac{1}{S} \sum_{i=1}^S (P_i - \hat{P})^2}. \quad (4.9)$$

4.3 SIMULATION ANALYSIS ON THE PERFORMANCE OF COPULA-BASED RESIDUAL BOOTSTRAPPING

The purpose of this simulation study is to test the performance of the proposed model in capturing the population characteristics of the dependent variable Y (loss variable) with respect to its explanatory covariates $X_1, X_2, \dots, X_p$, especially in the case that the sample size is small.

To accomplish the purpose, I first pre-assign the marginal distributions of the dependent variable Y and its associated explanatory covariates $X_1, X_2, \dots, X_p$ as well as the copula linking them. After that, I generate the following samples using different methods.

Population: I randomly generate 10,000 combinations of $Y, X_1, X_2, \dots, X_p$ simultaneously using the pre-specified marginal distributions and copula. Since it is generated with accurate population characteristics pre-assigned, and the sample size of it is 10,000, which is considered as large, so it is treated as a good approximation of the population of Y . For simplicity, I just call this sample Population in the remaining part of this section.

Sample 1 (sample generated without bootstrapping): From the 10,000 combinations of $Y, X_1, X_2, \dots, X_p$ generated for the Population, I randomly select a sub-sample of it of size n . Then I re-fit the marginal distributions of the selective sub-sample and the copula. From the fitted marginal distributions and the copula,

I randomly generate another 10,000 combinations of $Y, X_1, X_2, \dots, X_p$ again, and consist of Sample 1 of the loss variable Y .

Sample 2 (sample generated with copula-based residual bootstrapping):

Given the sub-sample used in the calculation of Sample 1, I perform the copula-based residual bootstrapping procedure on it described in Section 4.2, and repeat for 1000 times. Then a random selection of 10,000 observations from all of the resamples derived consist of Sample 2 of the loss variable Y .

I adopt both Q-Q plots and Kolmogorov–Smirnov (K-S) tests to compare the Samples 1 and 2 to the Population. If the proposed model is beneficial in understanding the Population characteristics of Y , so we should observe that Sample 2, which utilizes the copula-based residual bootstrapping approach, should be a closer fit to the Population compared to Sample 1, which does not employ any bootstrapping procedure.

To align with the empirical analysis presented later, I adopt the following simulation settings. I assume there is only one explanatory covariate, i.e., $p = 1$. I assume that X and Y follow normal distributions. Y has a mean of 100,000 and a standard deviation of 20,000, while X has a mean of 200 and a standard deviation of 30. The size n of the selective sub-sample is varying from 15, 50 to 100, representing different degrees of data paucity problem. I also test for different copulas which represent different data dependency structures among loss variable and the explanatory covariate. Namely, I adopt the most commonly used copulas including normal, Clayton, Gumbel and Frank copulas. For each set

of simulation, I run for 100 times and compute the average value of the K-S statistics. Table 4.1 summarizes the K-S test statistics for the simulations under different pre-assumptions.

A first look at the average values of K-S statistics indicate that the null hypothesis, i.e., both the Sample 1 generated without bootstrapping procedure and the Sample 2 generated with the proposed copula-based residual bootstrapping procedure have the same distribution of the Population, are rejected when n is 15 regardless of the type of the copula, and cannot be rejected when the sample size is 50 or 100. However, there are still two general trends that can be observed in Table 4.1. First, with an increasing sample size n , the K-S statistics for both Sample 1 and Sample 2 are decreasing. This is intuitive as with a larger initial sample size, the Sample 1 and Sample 2 will both be closer to the Population in terms of the distribution. The second observation is that although the significances of the test statistics vary, the K-S statistics testing between Sample 2 and the Population are smaller than the test statistics testing between Sample 1 and the Population in most cases. Moreover, this difference is statistically significant when the sample size is small, regardless of the copula pre-assumed. A lower K-S statistics indicate the Sample 2 generated with the proposed copula-based residual bootstrapping procedure is a better representative of the Population than the Sample 1, and this advantage is more significant when the sample size is small. That is to say, the proposed bootstrapping procedure can increase the accuracy of the estimation of the loss distributions, and thus gains an advantage in the cyber insurance pricing process.

I further show the representative Q-Q plots between Sample 1 and Sample 2 to the Population in Figure 4.1. Figure 4.1(a) shows the plots when the initial sample size is as small as 15. Since cyber insurance develops rapidly only after the new century, so this number roughly corresponds to the size of the annual data that insurers will have. Figure 4.1(b) shows the plots when the initial sample size n is as big as 100, where it can be considered as there is no data paucity problem in this case. We notice that the results using the proposed bootstrapping model have big improvements from the results without bootstrapping procedure, in the case that the sample size n is 15, regardless the copula structure pre-assumed. More specifically, since the sample size is small, the sample generated without bootstrapping procedure has a high tendency to overlook the tail distribution, including both the right tail and the left tail. To the pricing context, the right tail can be very important, as it may represent the catastrophe losses that will be imposed by the policies. Overlooking so may result in high insolvency risk for the insurance companies. However, we observe that the proposed bootstrapping procedure could relief this concern of the small sample without bootstrapping. As bootstrapping allows variation of the sample parameters and gives a more comprehensive study on the possible scenarios, and thus can compensate for some of the tail distribution that is neglected by the study without bootstrapping.

In the other case that the sample size n is large enough, then it is already sufficient to capture the distribution information of the population, and thus further bootstrapping procedure will not have many improvements, and produce

similar results as the case without bootstrapping. Thus, Figure 4.1(b) shows that both the samples generated with bootstrapping or not are very close to the original sample. However, the bootstrapping procedure could still help in understanding the sampling distribution of the statistical inferences in this case.

In the next section, I will further illustrate the results using a real industry dataset.

4.4 PRICING CYBER INSURANCE USING AN INDUSTRY DATASET AND ITS IMPLICATIONS

4.4.1 Data Description

Cyber insurance data set are hardly publicly available. Thus, I adopt a dataset used in Herath and Herath (2011) for empirical illustration purpose, i.e., the 2003 International Cyber Security Academy (ICSA) survey data. It includes two variables, number of computers affected and the direct dollar amount of loss. Table 4.2 shows the original data. The sample size is only 15 in this case, which is considered as insufficient for straightforward statistical inference.

4.4.2 Policy Types

General insurance usually has different policy types associated with different deductibles or policy limits, which may exhibit different properties. For comparison purpose, I adopt the settings specified in Herath and Herath (2011) which is summarized in Table 4.3. They assume that the firm likely loss L can be

pro-rated based on the number of computers affected, and the final premium for each computer also takes time value of money into consideration.⁶

4.4.3 Modelling of the Cyber Loss Variable

The complex nature of the interdependencies between different risk exposures and individual characteristics make linear regression models insufficient. Thus, I adopt the copula regression model in describing the associations between the loss variable and the explanatory covariate. In this particular example, I use copula regression model to investigate the direct dollar amount of loss variable Y with the number of computers affected X .

To better illustrate the advantage of copula regression model, I also conduct a linear regression analysis, considering both the loss amount and the logarithm of the loss amount as the dependent variable. The fitted results are shown in Figure 4.2. Figure 4.2(a) shows the result using the logarithm of the loss amount as the dependent variable. It is observed that it fits poorly as most of the actual points are far away from the fitted line. Figure 4.2(b) directly use loss amount as the dependent variable, and we see it is a better fit than the Figure

⁶ The detail explanation of this calculation formula can refer to Herath and Herath (2011). The purpose of this empirical section is to demonstrate the usefulness of our proposed model. The calculation formulas are not of our major concerns, so I just follow the setting in Herath and Herath (2011).

4.2(a). However, we can see directly from the graph, the relation between the loss amount and the number of computers affected is not linear, and the actual points most likely consist of a convex curve.

I next conduct copula regression model. It turns out that the marginal distributions are fitted by Weibull distributions. Figure 4.3 shows the scatter plot between the cumulative densities of the direct dollar amount of loss versus the number of computers affected. It is observed that they exhibit a strong positive correlation.

I fit five commonly used copulas to the data, and the best-fitted copula is the normal copula according to the likelihood values. Estimated parameters and likelihood values are shown in Table 4.4.

Then the fitted result of the copula regression model is shown in Figure 4.4. The first impression from the graph shows the better fit of copula regression model when the number of computers affected is generally large. Next, by zooming into the area where the number of computers affected is generally small, we still observe obvious advantage of the copula regression model against the linear regression model, revealing that the copula regression model better captures the dependency structure between the interested variables, and thus provides better estimates.

4.4.4 Calculations of Premiums

The next step is to illustrate the necessities of using copula-based residual bootstrapping analysis in pricing cyber insurance product. First, I derive the premiums for different policies aforementioned without bootstrapping procedure. The parameters in the pricing formulas are assigned values as below, where $l=10$, $m=500$, $a_1=400$, $a_2=125$, $a_3=300$, $\delta=2\%$. T is assumed to be a Poisson process with parameter 2. Premiums are calculated from $n = 10,000$ times simulated data from the selected copula using the formulas below. $E(P)$ is the expected claim amount, $\sigma(P)$ is the standard deviation of the expected claim, and P^k is the k^{th} simulated present value of loss, C^k is the k^{th} simulated claim, and T^k is the k^{th} simulated time elapse between now and claim happens. The premiums for different types of policies are shown in Table 4.5.

$$E(P) = \frac{1}{n} \sum_{k=1}^n P^k = \frac{1}{n} \sum_{k=1}^n C^k e^{-\delta T^k} \quad (4.10)$$

$$\sigma[P] = \sqrt{\frac{1}{n} \sum_{k=1}^n (P^k)^2 - [E(P)]^2}. \quad (4.11)$$

I then conduct the copula-based residual bootstrapping procedure introduced in section 2. I simulated $S = 1000$ resamples accordingly by randomly assigning conditional cumulative densities derived using equation (4.2). Then for every resample, I conduct the same pricing analysis as for the original sample without bootstrapping analysis, and finally, obtain 1000 sets of premiums for all types of policies. The mean and standard deviation of the premium for different types of policies are shown in Table 4.5 side by side with the result generated without bootstrapping procedure for comparison purpose.

4.4.5 Findings and Implications

Take a closer look at Table 4.5, it is worth noting that the mean of the premium derived from 1000 resamples for Type I and Type II policies without policy limit are larger than the premium obtained from the original sample without bootstrapping, while it is the opposite for Type III policies with policy limits. For example, the premium for Policy I derived from the original sample is \$286.39 while it is \$289.11 with the bootstrapping method. On the other hand, the premium for Policy III with a policy limit \$10000, co-insurance ratio 10% and deductible \$1000 derived with the original sample is \$78.89, higher than the \$77.48 derived from the bootstrapping method. To investigate this point, I plot the sampling distribution of the premium for Type I and Type III, shown in Figure 4.5. The results are very consistent with the simulation results. It can be observed that the histogram of Type I premium has much stronger right tail than the Type III premium, while Type III premium has stronger left tail than Type I. This indicates that if we do not consider the sampling distribution of the premium in the bootstrapped samples, we may underestimate the possibility that a Type I premium could be a right tail extreme case, resulting in a final premium estimate that is lower than it should be. Furthermore, we may underestimate the possibility that a Type III premium could be a left tail extreme case, resulting in a final premium estimate that is higher than it should be. The underlying explanation for this observation is that the original sample may underestimate both the right and left tails of the loss distribution. Without a policy limit, the right tail will probably

be underestimated, whereas with a policy limit, the right tail is cut, and only the left tail is observed.

The implication of these results is obvious. Practitioners may lower the current premium of Type III policies to increase competitiveness, but they should increase the premium of Type I policies to avoid insolvency risk. As the bootstrapping result is a sampling distribution rather than a point estimator, it should provide practitioners with the confidence to set more reasonable premiums. At the same time, it should also provide insurers with credibility to introduce the products to clients, and thus boost the cyber insurance industry.

Moreover, if we look at the standard deviation of premiums, it is no wonder that the sampling distribution of premium with bootstrapping could provide a narrower band. The average standard deviations as percentages of the premiums calculated from the original sample without bootstrapping are around 60% to 100%, while they are all below 10% for the premiums derived from the bootstrapping method. This is even more crucial for practitioners as they may have a clearer picture of the risk profile and can better determine the best premium they want to employ considering their risk positions.

4.5 REMARKS

In this chapter, I propose a general copula-based framework with a bootstrapping approach to develop cyber insurance pricing. The complex associations between the variables of interest are addressed by a copula regression model and the data

paucity problem is overcome by a copula-based residual bootstrapping procedure. On the one hand, copula regression is a generalization of the conventional linear regression models, and can improve the fitting result when there are non-linear associations between the dependent variable and the explanatory covariates. On the other hand, the copula-based residual bootstrapping procedure provides a way to account for the distortions caused by a specific sample that may not be fully representative of the population.

I first use simulation studies to demonstrate the advantages of the proposed model; specifically, its ability to compensate for the information loss on the tail distribution when the sample size is small. I then use an empirical analysis of a real industry dataset to investigate the relationship between the cyber loss variable and the number of computers affected, and demonstrate that adopting a copula regression model may lead to a better understanding of the dependency structure between the variables of interest, and thus produce more accurate and reliable estimations. As the residual bootstrapping procedure is based on the residuals obtained from the first step of the regression analysis, the copula regression model is more accurate and indirectly improves the reliability of the bootstrapping analysis. The proposed copula-based residual bootstrapping analysis is essential for solving the data paucity problem faced by insurers when they price cyber insurance products. In this example, I show that insurers can use copula-based residual bootstrapping procedure to obtain a reliable and robust estimation of the premium. When this derived premium is higher than the premium obtained from the original sample, the insurance company could

increase the existing premium to decrease insolvency risk; similarly, if this derived premium is lower, the insurance company may lower the existing premium to increase competitiveness. Moreover, the bootstrapping analysis gives a much narrower confidence interval for the premiums, and this can help insurers to evaluate their risk.

However, the proposed model has a potential limitation that should be studied further. Although the model is multi-dimensional, there are a limited number of multi-dimensional copula. As mentioned in Parsa and Klugman (2011), among the commonly used copulas only the Gaussian and Student-T copula allow asymmetric degrees of associations between different dimensions, but both may overlook important tail dependence. An interesting future research problem would be to apply some non-parametric copulas such as a Bernstein copula (Sancetta and Satchell, 2004) to conduct multi-dimensional copula regression analysis.

Figure 4.1 Comparison of the Q-Q plots between the samples generated with or without copula-based residual bootstrapping to the original sample.

(a) $n = 15$

(b) $n = 100$

Figure 4.2 Linear Regression Fit

(a) The logarithm of Loss v.s. No. of Computers Affected

(b) Loss v.s. No. of Computers Affected

Figure 4.3 Scatter Plot of Cumulative Densities

Figure 4.4 Copula Regression Fit

Figure 4.5 Sampling Distribution of Premium for Bootstrapped Resamples

Table 4.1 Kolmogorov–Smirnov tests between Population, without bootstrapping and bootstrapping samples

		Normal	Clayton	Gumbel	Frank
n=15	Population vs. without bootstrapping	0.109	0.1823	0.1752	0.211
	Population vs. with bootstrapping	0.0964	0.1524	0.1433	0.1831
	Difference	-0.0126**	-0.0299***	-0.0319***	-0.0279***
n=50	Population vs. without bootstrapping	0.0496	0.0903	0.0824	0.1107
	Population vs. with bootstrapping	0.0581	0.0864	0.0731	0.0934
	Difference	0.0085	-0.0039	-0.0093*	-0.0173**
n=100	Population vs. without bootstrapping	0.0212	0.0378	0.0351	0.0813
	Population vs. with bootstrapping	0.0273	0.0316	0.033	0.0692
	Difference	0.0061	-0.0062	-0.0021	-0.0121*

*, **, *** represent significances at the 5%, 1%, and 0.1% levels respectively.

Critical values for K-S test: 10% - 0.0546; 5% - 0.0608; 1% - 0.0729; 0.1% - 0.0872

Table 4.2 2003 ICSA Surveyed Computer Virus Data

	Virus	X	Y
1	W32/Blaster	1291	355,648.42
2	W32/Slammer	849	339,832.66
3	W32/Sobig	238	115,729.51
4	W32/Klez	140	65,090.38
5	W32/Yaha	118	45,402.25
6	W32/Swen	108	66,053.73
7	W32/Dumaru	87	39,182.88
8	W32/Mimail	70	19,556.82
9	W32/Nachi	63	20,087.13
10	W32/Fizzer	58	20,465.35
11	W32/BugBear	50	10,180.13
12	W32/Lirva	47	11,769.29
13	W32/Sober	21	6,944.48
14	W32/SirCam	21	5,339.08
15	W32/Ganda	19	7,547.77
X represents number of computers affected			
Y represents corresponding financial loss (Unit: \$)			

Table 4.3 Policy Types and Premiums (Herath and Herath, 2011)

Stage 1: Firm Level Likely Loss L l is the lower limit of the number of computers affected and m is the higher limit. a_1, a_2, a_3 are the pre-specified fixed losses $L = g(X, Y) = \begin{cases} a_1, & X < l \\ a_2 + \left(\frac{X-l}{X}\right)Y, & l \leq X < m \\ a_3 + \left(\frac{X-m}{X}\right)Y, & m \leq X \end{cases}$	
Stage 2: Policy Types and Claim P	
Policy Type	Loss Calculation Formulas
I. Basic first party damage policy with no deductible	$P = L = g(X, Y)/X$
II. First party damage policy with a deductible d	$P = \begin{cases} 0, & \text{if } L < d \\ (L - d)/X, & \text{if } L \geq d \end{cases}$
III. First party damage policy with deductible d , co-insurance a , and limit k	$P = \begin{cases} 0, & \text{if } L < d \\ (1 - a)(L - d)/X, & \text{if } d \leq L < d + p/(1 - a) \\ k/X, & \text{if } L \geq d + k/(1 - a) \end{cases}$
Stage 3: Premium C T is the time elapsed between now and the time the claim happens; δ is the discount rate. $C = Pe^{-\delta T}$	

Table 4.4 Copula Fitting Result for the Original Sample

Copula	Estimated Parameter	Likelihood values
Frank	$\hat{\theta}=29.14$	22.34
Clayton	$\hat{\theta}=8.666$	18.93
Gumbel	$\hat{\theta}=6.229$	20.67
Normal	$\hat{\rho}=0.9793$	23.03
Student-T	$\hat{\rho}=0.9789, \widehat{d.f}=46.22$	22.94

Table 4.5 Premium for Different Types of Policies

Policy I: Basic first party damage policy with no deductible

Original	Average Premium	\$286.39
	St.dev (%)	40.32%
Bootstrapping	Average Premium	\$289.11
	St.dev (%)	7.54%

Policy II: First party damage policy with a deductible d

	Deductible(d)	0	500	1000	1500	2000
Original	Average Premium	\$286.39	\$260.25	\$253.94	\$248.03	\$242.42
	St.dev (%)	49.07%	61.34%	63.11%	64.78%	66.39%
Bootstrapping	Average Premium	\$289.11	\$264.14	\$257.02	\$251.41	\$246.29
	St.dev (%)	7.54%	8.29%	9.12%	8.48%	8.81%

Policy III: First party damage policy with deductible d , co-insurance a , and limit k

			Deductible(d)	0	1000	2000
k=25000	Original	a=10%	Average Premium	\$162.57	\$136.16	\$128.58
			St.dev(%)	66.64%	80.92%	83.61%
		a=20%	Average Premium	\$153.13	\$129.65	\$122.92
			St.dev(%)	65.49%	80.08%	83.10%
	Bootstrapping	a=10%	Average Premium	\$158.57	\$132.46	\$124.38
			St.dev(%)	5.35%	6.48%	6.85%
		a=20%	Average Premium	\$149.16	\$125.66	\$118.19
			St.dev(%)	5.53%	6.65%	6.98%
k=10000	Original	a=10%	Average Premium	\$102.70	\$78.89	\$73.92
			St.dev(%)	84.55%	97.33%	100.14%
		a=20%	Average Premium	\$97.95	\$76.79	\$72.37
			St.dev(%)	82.61%	96.85%	100.23%
	Bootstrapping	a=10%	Average Premium	\$101.13	\$77.48	\$71.70
			St.dev(%)	3.84%	5.30%	5.83%
		a=20%	Average Premium	\$96.41	\$75.03	\$69.56
			St.dev(%)	4.04%	5.34%	5.95%

CHAPTER V. EXTENSION TO MULTI-DIMENSIONAL COPULA REGRESSION ANALYSIS IN INSURANCE AND FINANCE

5.1 EXISTING METHODOLOGIES

Understanding and quantifying dependence is at the core of all modelling efforts in the areas of insurance and finance (Frees and Valdez, 1998). However, conventional linear regression analysis may not be suitable for insurance and financial studies as the relationships are often nonlinear and the probability distribution of the dependent variable could be non-normal (Parsa and Klugman, 2011). Therefore, copula modelling technique has received tremendous attention in the recent literature. Nevertheless, due to lack of choice of copulas at the multi-dimensional level, most of the studies are restricted to two-dimensional analysis. Thus, the aim of this chapter is to provide an alternative method to extend the current copula regression analysis to a multi-dimensional level. This will allow for a significantly wider application of this statistical tool.

Frees and Valdez (1998) provided a good guidance in applying the copula modeling technique to risk modelling in insurance. After that, copula technique has been applied to various areas in insurance and finance, such as loss reserving (De Jong, 2012; Shi and Frees, 2011; Zhang and Dukic, 2012), mortality modelling (Shemyakin and Youn, 2006; Spreeuw, 2006; Wang et al., 2015), and

return and volatility analysis for stock and forex markets (Yang and Hamori, 2013; Luo et al., 2015; Wu et al., 2014). However, it is noticed that most of these studies are restricted to two-dimension, or only Gaussian and Student-T copulas are investigated for multi-dimensional cases.

The reason is that two-dimensional copula models are well developed in the literature with numerous copulas at hand to choose from. Among all of them, Archimedean copulas are the most widely used family. For example, Clayton copula is a good representative of left tail dependence, and Gumbel copula can better capture the right tail dependence while Frank copula captures central dependence. Other than that, elliptical copulas including Gaussian and Student-T copulas are both commonly used copulas for modelling two-dimensional cases when dealing with symmetric dependence.

The extension of the copula regression analysis into multi-dimension faces difficulties in the current literature, as few copulas can allow different degrees of associations between any two dimensions (Parsa and Klugman, 2011). For example, a three-dimensional Archimedean copula will require the associations between any two of the three variables to be the same. This is rare when dealing with regression analysis in the real world. Though elliptical copulas are available for multi-dimensional copula regression analysis, they can only describe symmetric dependence and have small concentrations on the tail dependence, which limit their usage (Parsa and Klugman, 2011).

Some solutions are proposed to deal with the difficulties faced in the multi-dimensional case. For example, one of the methods is to use a parametric model called a hierarchy tree. It decomposes the multi-dimensional dependence modelling into several bivariate dependence modelling (Bürgi et al., 2008; Savu and Trede, 2010; Okhrin et al., 2013; Zhu et al. 2015). However, by the name of this method, a well-justified hierarchy network is needed, and thus may not have adaption to the general case. Another solution is to use the vine copula. The idea of a regular vine is to decompose the multi-dimensional distribution into many layers of two-dimensional conditional distribution. Therefore, it may become very complicated as the dimension increases. The factor copula model (Chen et al., 2015; Oh and Patton, 2015) is an effective tool in dimension reduction. Hence, it can become a very appealing tool when the dimension is high. Other than that, it does not provide a closed-form solution to the probability distribution and there may be a loss of information in the process of dimension reduction.

To deal with the aforementioned difficulties, the primary contribution of this chapter is to propose the use of a non-parametric copula called Bernstein copula to perform multi-dimensional copula regression analysis. There are two main advantages that make this copula an ideal candidate for multi-dimensional copula regression analysis. First, it uses Bernstein polynomials to smooth the copula density, and thus can deal with any dependency structure, either symmetric or asymmetric. Second is its ease of implementation due to the ability to obtain a closed-form solution for the conditional distribution of the dependent

variable on its covariates. This is essential to ensure that the proposed model is computational friendly in the multi-dimensional case.

5.2 BERNSTEIN COPULA REGRESSION MODEL

Denote Y as the dependent variable, and $X_1, X_2, \dots, X_k$ as the associated covariates. To obtain the conditional expectation $E(Y|X_1 = x_1, \dots, X_k = x_k)$, the copula regression model estimates the conditional distribution of Y on its covariates. If H is the joint cumulative distribution function for the random variables $Y, X_1, X_2, \dots, X_k$ with continuous marginal distributions $G, F_1, F_2, \dots, F_k$, then there exists a copula function C , such that

$$H(y, x_1, \dots, x_k) = C(G(y), F_1(x_1), \dots, F_k(x_k)). \quad (5.1)$$

Similarly, if I denote $u = G(y)$ and $v_j = F_j(x_j)$ for $j = 1, \dots, k$.

According to Sklar's theorem, a copula C can also be identified such that

$$C(u, v_1, \dots, v_k) = H(y, x_1, \dots, x_k). \quad (5.2)$$

Crane and Van Der Hoek (2008) proves that the formula of the conditional probability distribution of the dependent variable is

$$\text{Prob}(Y \leq y | X_1 = x_1, \dots, X_k = x_k) = \frac{\partial C(u, v_1, v_2, \dots, v_k)}{\partial v_1 \partial v_2 \dots \partial v_k} / \frac{\partial C(1, v_1, v_2, \dots, v_k)}{\partial v_1 \partial v_2 \dots \partial v_k}. \quad (5.3)$$

Bernstein copula is a technique to approximate the true copula density. Sancetta and Satchell (2004) indicated that Bernstein copula can be used to approximate an arbitrary copula. The estimation accuracy can be improved by

altering the order of the polynomials, which controls the rate of convergence. I follow Sancetta and Satchell (2004) to construct the Bernstein copula. Let $\alpha\left(\frac{r_u}{m_u}, \frac{r_1}{m_1}, \dots, \frac{r_k}{m_k}\right)$ be a real-valued constant indexed by $(r_u, r_1, r_2, \dots, r_k)$, $r_u \in N_+$ and $r_j \in N_+$, such that $0 \leq r_u \leq m_u$ and $0 \leq r_j \leq m_j$, $j = 1, 2, \dots, k$.

The Bernstein copula is defined as $C_B: [0,1]^{k+1} \rightarrow [0,1]$, where

$$C_B(u, v_1, \dots, v_k) = \sum_{r_u=0}^{m_u} \sum_{v_1=0}^{m_1} \dots \sum_{v_k=0}^{m_k} \alpha\left(\frac{r_u}{m_u}, \frac{r_1}{m_1}, \dots, \frac{r_k}{m_k}\right) P_{r_u, m_u}(u) P_{r_1, m_1}(v_1) \dots P_{r_k, m_k}(v_k), \quad (5.4)$$

$$\text{and } P_{r_j, m_j}(x) = \binom{m_j}{r_j} x^{r_j} (1-x)^{m_j-r_j}. \quad (5.5)$$

From the definition, $m_u, m_1, m_2, \dots, m_k$ do not necessarily be the same. However, for practical use, we usually assume $m_1 = m_2 = \dots = m_k = m$ for simplicity and m is a pre-specified value that controls the rate of convergence.

Assume that the sample size is n , and the s^{th} sample is recorded as $(u_s, v_{1,s}, \dots, v_{k,s})$. Let $C_n\left(\frac{r_u}{m}, \frac{r_1}{m}, \dots, \frac{r_k}{m}\right)$ be the empirical copula at $\left(\frac{r_u}{m}, \frac{r_1}{m}, \dots, \frac{r_k}{m}\right)$, which is defined as $\frac{1}{n} \sum_{s=1}^n I\left\{[\cap_{j=1}^k (v_{j,s} \leq \frac{r_j}{m})] \cap (u_s \leq \frac{r_u}{m})\right\}$. The empirical Bernstein copula density is defined in a similar manner as equation (5.4) with

$$\alpha\left(\frac{r_u}{m}, \frac{r_1}{m}, \dots, \frac{r_k}{m}\right) = C_n\left(\frac{r_u}{m}, \frac{r_1}{m}, \dots, \frac{r_k}{m}\right) \quad (5.6)$$

An important property of Bernstein copula is that it has a closed-form representation of the conditional distribution function derived in Equation (5.3).

The Bernstein copula is constructed in a way that is a product of many P functions defined in Equation (5.5). Each P function only contains one random variable, and thus, it allows us to differentiate every random variable one by one, resulting in a closed-form solution to Equation (5.3). This important feature enables an efficient computational process in forecasting the conditional distribution of the dependent variable given multiple covariates.

The conditional probability distribution function in Equation (5.3) can be written in the following form under the Bernstein copula framework:

$$C'_B = \frac{\sum_{r_u=0}^m \dots \sum_{r_1=0}^m \dots \sum_{r_k=0}^m \alpha\left(\frac{r_u}{m}, \frac{r_1}{m}, \dots, \frac{r_k}{m}\right) P_{r_u,m}(u_1) D_{r_1,m}(v_1) \dots D_{r_k,m}(v_k)}{\sum_{r_u=0}^m \dots \sum_{r_1=0}^m \dots \sum_{r_k=0}^m \alpha\left(\frac{r_u}{m}, \frac{r_1}{m}, \dots, \frac{r_k}{m}\right) P_{r_u,m}(1) D_{r_1,m}(v_1) \dots D_{r_k,m}(v_k)} \quad (5.7)$$

where the function D is defined as:

$$D_{r_j,m}(x) = \binom{m}{r_j} [r_j x^{r_j-1} (1-x)^{m-r_j} - (m-r_j) x^{r_j} (1-x)^{m-r_j-1}]. \quad (5.8)$$

It is noted that m is used to adjust the rate of convergence for the Bernstein copula to the true density. If the sample size is small, m cannot be set to a large number considering the overfitting problem. On the contrast, if the sample size is large, we can increase the value of m to obtain a higher estimation accuracy, though there is always a trade-off between estimation accuracy and computational complexities. In the following sections, I illustrate the usage of this proposed method in practical insurance and financial problems using both simulation studies and empirical analysis.

5.3 EXAMPLE 1: SIMULATION ANALYSIS ON THE MULTI-DIMENSIONAL COPULA-BASED LOSS TRIANGLE BENCHMARKS

In Chapter III, it is mentioned that the central limit theorem ensures the convergence of the reported loss ratio to its theoretical mean value when the number of policies is large. However, this may not be applicable to the small business line, in which the number of policies is small. This creates difficulties in loss reserving for the small business line, as the reported loss ratio may be significantly different from its expected value.

In Chapter III, the empirical analysis is restricted to two business lines. The small business line is the commercial auto line, and the related large business line is the personal auto line. It is also mentioned that this benchmark can be used in many contexts, such as big insurance companies versus small insurance companies, or similar business lines in big countries versus small countries. With respect to this idea, an extension of the current model from two-dimension to multi-dimension is a must. For example, reinsurance companies usually need to deal with business from multiple countries within a region. For a business line in a small country, it will be ideal to use the information from several big countries around it simultaneously. This will require a consideration of the multi-dimensional model.

5.3.1 Benchmark Construction

The construction of the multi-dimensional benchmark is not much different from the two-dimensional case. Suppose the target small business line is S, and the related large lines are $L_1, L_2, \dots, L_n$. I use $\theta_{ij}^{(n)}$ to indicate the incremental loss ratios of the n^{th} business line at development lag i of accident year j . The random vector of multivariate incremental loss ratio θ_{ij} can be expressed by

$$\theta_{ij} = \left(\theta_{ij}^{(S)}, \theta_{ij}^{(L_1)}, \theta_{ij}^{(L_2)}, \dots, \theta_{ij}^{(L_n)} \right), i \in \{0, \dots, I\} \text{ and } j \in \{0, \dots, J\}, \quad (5.9)$$

where I denotes the latest accident year and J denotes the longest development lag.

Similar to Chapter III, the target is to provide a benchmark for the loss ratio of the small business line at the left bottom corner. The conditional probability distribution $D_{\theta_{ij}^{(S)} | \theta_{ij}^{(L_1)}, \theta_{ij}^{(L_2)}, \dots, \theta_{ij}^{(L_n)}}$ of the loss ratio $\theta_{ij}^{(S)}$ given $\theta_{ij}^{(L_1)}, \theta_{ij}^{(L_2)}, \dots, \theta_{ij}^{(L_n)}$ is defined as

$$D_{\theta_{ij}^{(S)} | \theta_{ij}^{(L_1)}, \theta_{ij}^{(L_2)}, \dots, \theta_{ij}^{(L_n)}}(\theta_{ij}^{(S)}) = P(\theta_{ij}^{(S)} < \theta_{ij}^{(S)} | \theta_{ij}^{(L_1)} = \theta_{ij}^{(L_1)}, \dots, \theta_{ij}^{(L_n)} = \theta_{ij}^{(L_n)}). \quad (5.10)$$

Either the conditional mean or median can serve as a good benchmark to the loss ratio of the small business lines. In this example, I use the median measurement following Chapter III. To get the conditional probability distribution, we need to perform multi-dimensional copula regression analysis.

5.3.2 Simulation Analysis

In order to see the performance of the proposed method under different sample size, I conduct the simulation analysis in this section. The data are simulated following Chapter III with the below characteristics. I assume that there are three business lines. The first two business lines L_1 and L_2 are assumed to be large. This means that the reported loss ratios are closer to their expected values and have smaller random variations. The third business line S is assumed to be a small business line and its reported loss ratios can have large variations from their expected values. The yearly total loss ratios of the three business lines are assumed to be correlated with a defined copula structure, either Gaussian or non-Gaussian (I use Clayton, Frank, and Gumbel copula structures for non-Gaussian structures) in the later analysis. Multiple years' total loss ratios of the three business lines are simulated in the way following Chapter III.

The aim is to provide a benchmark for the left bottom corner of the loss triangle of the small line and explore the implication of this benchmark on loss reserving. To fulfil this goal, I derive the benchmark for the small business line in each simulation and compare it to the actual expected value determined by the pre-assumptions. Moreover, I look at the reserve derived by using this benchmark and compared it to the actual simulated loss.

The sample size is set to vary from 20 to 1000, and m is fixed at 10 for the Bernstein copula. When the sample size is 1000, I investigate an additional case where m is equal to 20. I adopt both Gaussian copula and Bernstein copula in the benchmark construction in order to give a comparison. Figure 5.1 shows

the results of the accuracies of the provided benchmark by assuming different copula structures and sample size. The error is calculated as the average of the absolute errors between the benchmark/simulated reported loss ratio to the simulated expected loss ratio of 1000 repeated times of simulations. It is clear that both the benchmarks calculated using the Gaussian copula or the Bernstein copula are more accurate than the simulated reported loss ratio, as seen from the smaller errors compared to the simulated theoretical mean of the loss ratio. However, it is observed that when the simulated structure is from Clayton, Frank or Gumbel copula, the performance of the Bernstein benchmark exceeds the performance of Gaussian copula when the sample size is 200 or larger. This differs from the case when the simulated structure is normal, where the performance of the Bernstein Benchmark only converges to the performance of the Gaussian Benchmark when the sample size increases. This observation suggests that when the underlying structure is normal, the Gaussian copula is sufficient, and the Bernstein copula can only work well when the sample size is large. However, when the underlying structure is not normal (such as strong tail dependence) and the sample size is not small, the Bernstein copula is better to capture the intrinsic dependence structure than the Gaussian copula. It is also worth noting that when the sample size is large, the performance of Bernstein copula is better with a larger m .

Figure 5.2 shows the distribution of the errors of the benchmark from the simulated theoretical mean of the loss ratio when the sample size is 1000 and the structure is from the Clayton copula. Figure 5.2(a) refers to the errors between

the Bernstein benchmark proposed and the actual simulated loss ratio. Figure 5.2(b) refers to the errors between the Gaussian benchmark proposed and the actual simulated loss ratio. Finally, Figure 5.2(c) refers to the errors between the simulated reported loss ratio and the actual simulated loss ratio. It can be seen that the errors do not exhibit a strong association with the theoretical mean using the Bernstein copula. On the contrary, we observe an obvious pattern that errors are associated with the theoretical mean of loss ratio for the benchmark calculated using Gaussian copula or the simulated reported loss ratio. This is another evidence that Bernstein copula is able to better capture the underlying dependence structure than the Gaussian copula when it is non-normal.

Figure 5.3 reports the loss reserving accuracies using the proposed benchmarks or the simulated reported loss ratio in the loss triangle for loss reserving. The reserving method used is the chain-ladder method. The result is consistent with chapter III, and we observe that the proposed benchmark can significantly increase the loss reserving accuracies. Moreover, adopting the benchmark calculated using the Bernstein copula in the reserving analysis is better than using the Gaussian Benchmark in this case with a large sample size.

In conclusion, through this simulation analysis, we can see that the Bernstein copula can be a good candidate to model the associations between different variables when we have a plentiful number of samples. When the dependency structure is non-normal, Bernstein copula may be better to capture the dependency information than simply using the Gaussian copula. In the next

section, I further perform an empirical analysis using real stock prices to demonstrate the effectiveness of the Bernstein copula in the application of multi-dimensional copula regression analysis.

5.4 EXAMPLE 2: MULTI-DIMENSIONAL PAIRS TRADING STRATEGY

In Chapter II, Do and Faff (2010) mentioned that it is the increased arbitrage risk that makes the distance method less profitable. Arbitrage risk refers to the risk of further divergence between the prices of the stock pairs. In other words, if we view one stock price as a benchmark to another stock within a pair, arbitrage risk implies that the benchmark is not a good choice as the other stock continuously moves away from the benchmark. From this point of view, a search for a good benchmark is essential.

One way of obtaining a good benchmark is to use more information than that provided by one stock to construct the benchmark. For example, if we can discover a series of stocks which move together in the literature, then ideally, using the information of all the other stocks to construct a benchmark for a specific stock should be more accurate than using just any one of the stocks. This may be referred to as a “group trading” concept which is also mentioned in Gatev et al. (2006). They described that pairs trading may not be limited to two stocks, but subject to technical difficulty.

In this example, I try to solve the aforementioned technical difficulty. I attempt to provide a trading framework for the aforementioned “group trading” by using the copula regression method to provide a benchmark for a specific stock using the information from several stocks that have historically moved together with it. The aim of this example is not to convince that the proposed method is better than the conventional method, but rather to motivate a new way of thinking pairs trading, in addition to the current literature.

5.4.1 Group Selection

For pairs trading, the basic assumption is that the two stocks within a pair should have similar risk factors, and the stock prices should move together. This inherent property ensures the market risk-free characteristic of pairs trading. This rule also applies to “group trading”, which requires all stocks in a particular group to have co-movements in their stock prices. Moreover, the reason to extend pairs trading to “group trading” is to utilize more dependency information by using multiple stocks to construct the benchmark for the target stock instead of solely using one. Therefore, this should be reflected in the group selection process as well. For simplicity, I use a three-dimensional setting to illustrate the idea for the rest of this example. However, it should be noted that all principles and strategies discussed should also be applicable to dimensions that are higher than three.

Let X and Y to be the two stocks selected for pairs trading and stock Z is the candidate which would join stocks X and Y to form a group. Assuming the stock prices are recorded as p_t^X , p_t^Y and p_t^Z and daily returns are r_t^X , r_t^Y and r_t^Z ,

then this group is accepted during the formation phase if the following criteria are satisfied:

1. p_t^X , p_t^Y and p_t^Z are co-integrated pairwise.
2. The alternative copula regression model $r_t^X | (r_t^Y, r_t^Z)$ should be significantly better than the null copula regression model $r_t^X | r_t^Y$ and the alternative copula regression model $r_t^Y | (r_t^X, r_t^Z)$ should be significantly better than the null copula regression model $r_t^Y | r_t^X$.

The first criterion is examined using the Engle-Granger co-integration test (Engle and Granger, 1987). This is to ensure that all three stocks within a group have co-movements in stock prices.

The second criterion is checked using the likelihood ratio test between the alternative model and the null model. According to Sklar's theorem (Sklar, 1959), there must exist a copula C linking the joint distribution function H of R_t^X , R_t^Y and R_t^Z to their separate cumulative densities F_X , F_Y and F_Z , i.e.

$$H(r_t^X, r_t^Y, r_t^Z) = C(F_X(r_t^X), F_Y(r_t^Y), F_Z(r_t^Z)). \quad (5.11)$$

If I denote $u_t^X = F_X(r_t^X)$, $u_t^Y = F_Y(r_t^Y)$, and $u_t^Z = F_Z(r_t^Z)$. Then according to Crane and Van Der Hoek (2008), the conditional densities of $r_t^X | (r_t^Y, r_t^Z)$, denoted as $F_{X|Y,Z}$, can be calculated as

$$F_{X|Y,Z} = [\partial C(u_t^X, u_t^Y, u_t^Z) / \partial u_t^Y \partial u_t^Z] / [\partial C(1, u_t^Y, u_t^Z) / \partial u_t^Y \partial u_t^Z]. \quad (5.12)$$

The log-likelihood value is then calculated for the alternative model and the null model. The test statistic, $D_{X|Y,Z}$, is defined as

$$D_{X|Y,Z} = 2 \sum \ln f_{X|Y,Z}(r_t^X | r_t^Y, r_t^Z) - 2 \sum \ln f_{X|Y}(r_t^X | r_t^Y). \quad (5.13)$$

Similarly, the test statistic $D_{Y|X,Z}$ used to examine between the alternative model $r_t^Y | (r_t^X, r_t^Z)$ and the null model $r_t^Y | r_t^X$ is defined as

$$D_{Y|X,Z} = 2 \sum \ln f_{Y|X,Z}(r_t^Y | r_t^X, r_t^Z) - 2 \sum \ln f_{Y|X}(r_t^Y | r_t^X). \quad (5.14)$$

If stock Z satisfies the above two criteria, forming a group with stocks X, Y and Z for group trading is seen as a better choice than simply performing pairs trading using stocks X and Y. This is because stock Z moves together with stocks X and Y, which is able to provide a better description of the variation of stocks X and Y through their joint dependence.

5.4.2 Group Trading Strategy

The trading strategy can be seen as a generalization of the two-dimension copula approach to multi-dimension. During the formation period, the marginal distributions are estimated as well as the copula parameter. Recall that according to the Sklar's theorem (Sklar, 1959), the estimated copula C can link the joint distribution function H of R_t^X , R_t^Y and R_t^Z to their separate cumulative densities F_X , F_Y and F_Z ,

During the trading period, I calculate the daily degrees of mispricing using the conditional densities similar to the two-dimension copula approach, i.e.

$$M_t^{X|Y,Z} = [\partial C(u_t^X, u_t^Y, u_t^Z) / \partial u_t^Y \partial u_t^Z] / [\partial C(1, u_t^Y, u_t^Z) / \partial u_t^Y \partial u_t^Z]. \quad (5.15)$$

$$M_t^{Y|X,Z} = [\partial C(u_t^X, u_t^Y, u_t^Z) / \partial u_t^X \partial u_t^Z] / [\partial C(u_t^X, 1, u_t^Z) / \partial u_t^X \partial u_t^Z]. \quad (5.16)$$

$$M_t^{Z|X,Y} = [\partial C(u_t^X, u_t^Y, u_t^Z) / \partial u_t^X \partial u_t^Y] / [\partial C(u_t^X, u_t^Y, 1) / \partial u_t^X \partial u_t^Y]. \quad (5.17)$$

The overall degrees of relative mispricing are determined by adding $(M_t^{X|Y,Z} - 0.5)$, $(M_t^{Y|X,Z} - 0.5)$ and $(M_t^{Z|X,Y} - 0.5)$ to MI_{3D}^X , MI_{3D}^Y and MI_{3D}^Z every day respectively. Note that MI_{3D}^X , MI_{3D}^Y and MI_{3D}^Z are set to zero at the beginning of the trading period. Assume that T_{MD} and SL_{MD} are the respective threshold levels of relative mispricing for opening trading positions, and closing trading positions to prevent substantial losses when stocks diverge. Furthermore, trading positions are closed when the corresponding mispricing index returns to 0. There are six different scenarios of opening and closing trading positions (assuming that no trades are open), which are summarized in Table 5.1.

Note that in each scenario, long and short positions are taken in equal weights to ensure that they are market risk-free. In the case that we long/short one stock and short/long the other basket of two stocks, we also take equal weights for the two stocks within the basket. At the end of the trading period, all trading positions are closed regardless of the values of MI_{3D}^X , MI_{3D}^Y and MI_{3D}^Z . It should be noted that T_{3D} and SL_{3D} are both pre-specified values. Hence, there are infinite combinations, and back-testing should be performed to determine the optimal values of T_{3D} and SL_{3D} .

5.4.3 Data Description and Settings

To perform my empirical analyses, I use stocks from the financial sectors of the major Asia Pacific markets, namely Australia, Hong Kong, Japan, Korea, and Singapore. I choose the financial sector as Gatev et al. (2006) tested pairs trading within four different industry groups, and found a higher profitability in financials and utilities. Do and Faff (2010) argued that this finding was within expectation because these companies have low differentiations and stable demand, as well as are subject to common macroeconomic shocks, resulting in strong co-movements in their stock prices.

To standardize the classification of financial companies across different markets, financial composites of indices from each market were compared. These include SP ASX200, Hang Seng Composite Index, KOSPI 200 Index, Nikkei 225 Index and Straight Times Index. The daily stock data are retrieved from Bloomberg database. The sample period starts from January 1, 2005 to December 31, 2014. To form the initial dataset, I first identify stocks classified under the financial industry in the stock exchanges of Australia, Hong Kong, Japan, Korea, and Singapore. Subsequently, in the stock exchanges of Australia and Hong Kong, I drop stocks with missing data in the sample period and penny stocks with stock prices that are less than 1. This gives the final sample which consists of 15 stocks from Australia, 7 stocks from Hong Kong, 15 stocks from Japan, 17 stocks from Korea, and 5 stocks from Singapore.

The formation period is set to be 12-month (252 trading days) and the trading period is set to be 6-month (126 trading days). During my sample period, I have 16 trading cycles in total, rolling on a half-year basis. In each trading cycle, the groups are selected according to the selection criteria mentioned in section 2.3.1. Accordingly, the stock pair portfolio and group portfolio are constructed.

To calculate the monthly excess returns, I adopt the measurement of return on committed capital used in Gatev et al. (2006) similar to chapter II. This is a more conservative measurement compared to the fully invested return, which only considers the money currently being traded as the principal amount. In contrast, return on committed capital also takes the amount set aside for potential trades into consideration. Adopting such a conservative measurement can increase the credibility of my results.

5.4.4 Main Results

In terms of mean excess return, the results summarized in Table 5.2 favour the multi-dimensional copula strategy for Australia and Singapore, and the 2-dimensional copula strategy for Hong Kong, Japan and Korea. However, the multi-dimensional copula strategy has the highest t-statistics for all markets except for Japan. This means that the mean excess returns of the multi-dimensional copula strategy are more significant than those of the other two strategies. Consistent with the cumulative returns, the mean excess returns of both the copula strategies are higher than those of the distance strategy except for

Japan, meaning that the copula strategies generally outperform the distance strategy in terms of returns.

Although high cumulative returns and mean excess returns may favour a particular strategy, the risks involved when achieving those returns should be taken into account. Although the 2-dimensional copula strategy outperforms the multi-dimensional copula strategy in 3 out of 5 markets in terms of cumulative returns, the cumulative returns of the multi-dimensional copula strategy are relatively less volatile. Furthermore, as shown in Table 5.2, the multi-dimensional copula strategy has the lowest standard deviation of returns for Hong Kong, Japan, and Korea, while the distance strategy is the least volatile for Australia and Singapore. Overall, the 2-dimensional copula strategy appears to be the most volatile strategy, as measured by standard deviation.

However, since upside volatility could be desirable, it is good to solely look at the downside volatility to reflect the “real risks” involved. As depicted in Table 5.2, similar to the inferences drawn from standard deviations, the multi-dimensional copula strategy has the least downside volatility for Hong Kong, Japan (very close to the lowest) and Korea, and the 2-dimensional copula strategy appears to have the most downside volatility in general. Overall, the multi-dimensional copula strategy emerges as the least volatile one, substantiating that it is more robust than the others.

Since both risk and return matter, the Sharpe ratio and the Sortino ratio are adopted to take into account the two factors at the same time (Sharpe, 1994).

From Table 5.2, we can see that the multi-dimensional copula strategy has the highest Sharpe and Sortino ratios for all markets except for Japan, which favours the distance strategy. This allows affirmative confirmation that the proposed strategy is able to better utilize the dependency information and measure relative pricing more comprehensively than the 2-dimensional strategies. Other than that, the 2-dimensional copula strategy also outperforms the distance strategy for all markets except for Japan and Korea. This is consistent with the fact that the copula framework may have the advantage of considering both linear and non-linear associations when measuring relative pricing, which may leads to higher returns and more trading opportunities. However, it is noted that though multi-dimensional copula method are better than other two methods in general, most of the t-statistics are still not significant for the mean excess returns. This is aligned with the fact that the profitability of the pairs trading strategy has decreased over time. The purpose of this study is to motivate a multi-dimensional copula framework for pairs trading, and it is believed that the profitability can be further improved with other technique in the future research.

Last but not least, Table 5.3 presents the overall comparison of trading statistics of the three different strategies. The multi-dimensional copula strategy has achieved the highest average number of trades per pair/group, followed closely by the two-dimensional copula strategy. This shows that the multi-dimensional copula strategy is able to identify more trading opportunities than the two-dimensional copula and the distance method, and these trading opportunities are generally profitable as the strategy has achieved higher Sharpe

ratios across different markets. However, trading positions are open for a longer period of time (in months) for the two-dimensional and multi-dimensional copula strategies. Having said that, there is greater certainty about the duration of an open position for the copula strategies as the standard deviations are significantly lower than that of the distance method.

5.5 REMARKS

In this chapter, I extend the current copula regression analysis to a multi-dimensional framework. To solve the issue of the limited availability of copulas at a multi-dimensional level, I adopt a non-parametric copula called the Bernstein copula. The advantages of it are two-fold. First, it uses the Bernstein polynomials to smooth the copula density, and thus can deal with any dependency structure, either symmetric or asymmetric. Second, it has the ease of implementation and can obtain closed-form solution of the conditional distribution of the dependent variable on its covariates. This is essential to ensure that the proposed model is computationally friendly in the multi-dimensional case.

I provide two examples of applications of this multi-dimensional framework to practical finance and insurance problems. The first example is a simulation analysis, which I conclude that the Bernstein copula can better describe the dependency structure than the Gaussian copula if the intrinsic structure is not normal. However, it should be noted that this depends on a relatively sufficient number of samples. When the intrinsic structure is indeed

normal, the performance of the Bernstein copula can still converge to the performance of Gaussian copula when the sample size is large enough. In the second application, I propose a “group trading” concept. This is only applicable when we can handle the multi-dimensional situation, which confirms the importance of the proposed multi-dimensional copula regression framework. I apply the proposed model to real financial data and observed consistent results with my hypothesis that adding more stocks can increase the robustness of the profitability, resulting in higher Sharpe ratios.

In conclusion, the proposed Bernstein copula regression analysis enables us to perform multi-dimensional analysis if necessary. This opens the door for applying the copula modelling technique to various problems in insurance and finance that previously cannot be done due to technical difficulties. I believe that the proposed method could have wide application in the future research.

Figure 5.1 Accuracies of the Benchmark with different sample size

(c) Gumbel Structure

(d) Gaussian Structure

Figure 5.2 Accuracies of the Bernstein Benchmark, Gaussian Benchmark and the Reported Loss Ratio (ABS: Absolute Error)

(a) Bernstein Benchmark

(b) Gaussian Benchmark

(c) Reported Loss Ratio

Figure 5.3 Loss Reserving Accuracies (ABS: Absolute Error)

(a) Bernstein Benchmark

(b) Gaussian Benchmark

(c) Reported Loss Ratio

Table 5.1 Trade Conditions and Positions for Multi-Dimensional Copula

Open Trade Condition	Trading Positions		Close Trade Condition	Stop Loss Condition
	Long	Short		
$MI_{3D}^X > T_{3D}$	Stock Y, Stock Z	Stock X	MI_{3D}^X falls to 0	MI_{3D}^X rises to SL_{3D}
$MI_{3D}^X < -T_{3D}$	Stock X	Stock Y, Stock Z	MI_{3D}^X rises to 0	MI_{3D}^X falls to $-SL_{3D}$
$MI_{3D}^Y > T_{3D}$	Stock X, Stock Z	Stock Y	MI_{3D}^Y falls to 0	MI_{3D}^Y rises to SL_{3D}
$MI_{3D}^Y < -T_{3D}$	Stock Y	Stock X, Stock Z	MI_{3D}^Y rises to 0	MI_{3D}^Y falls to $-SL_{3D}$
$MI_{3D}^Z > T_{3D}$	Stock X, Stock Y	Stock Z	MI_{3D}^Z falls to 0	MI_{3D}^Z rises to SL_{3D}
$MI_{3D}^Z < -T_{3D}$	Stock Z	Stock X, Stock Y	MI_{3D}^Z rises to 0	MI_{3D}^Z falls to $-SL_{3D}$

Table 5.2 Summary of Returns for Different Strategies and Different Markets

	Distance	2D Copula	MD Copula
Australia			
Mean Excess Return	0.0023	0.0061	0.0104
T-Statistic	0.6064	0.6900	1.5492
Sharpe Ratio	0.1471	0.1673	0.3757
Sortino Ratio	0.2482	0.2990	1.8396
Standard Deviation	0.0159	0.0363	0.0278
Downside Volatility	0.0094	0.0203	0.0057
Skewness	0.2473	0.4093	2.8386
Maximum	0.0371	0.0877	0.1071
Minimum	-0.0247	-0.0489	-0.0154
Hong Kong			
Mean Excess Return	-0.0003	0.0017	0.0013
T-Statistic	-0.0278	0.1852	0.2789
Sharpe Ratio	-0.0105	0.0700	0.1054
Sortino Ratio	-0.0164	0.1295	0.2142
Standard Deviation	0.0259	0.0239	0.0124
Downside Volatility	0.0166	0.0129	0.0061
Skewness	0.3259	0.6211	1.0760
Maximum	0.0416	0.0401	0.0248
Minimum	-0.0347	-0.0267	-0.0141
Japan			
Mean Excess Return	0.0086**	0.0109*	0.0057*
T-Statistic	2.6243	1.9970	1.7540
Sharpe Ratio	0.3956	0.3011	0.2644
Sortino Ratio	0.9043	0.6121	0.5457
Standard Deviation	0.0218	0.0361	0.0214
Downside Volatility	0.0095	0.0177	0.0104
Skewness	0.4843	0.5267	0.8353
Maximum	0.0690	0.1129	0.0719
Minimum	-0.0358	-0.0640	-0.0396
Korea			
Mean Excess Return	0.0046*	0.0085**	0.0066***
T-Statistic	1.9184	2.3325	3.6396
Sharpe Ratio	0.1968	0.2393	0.3734
Sortino Ratio	0.3645	0.3552	0.7440
Standard Deviation	0.0233	0.0356	0.0177
Downside Volatility	0.0126	0.0240	0.0089
Skewness	0.4262	-0.9678	0.0282
Maximum	0.0642	0.0857	0.0514

Minimum	−0.0415	−0.1300	−0.0382
Singapore			
Mean Excess Return	0.0026	0.0035	0.0105
T-Statistic	0.4359	0.8970	0.9780
Sharpe Ratio	0.1647	0.3390	0.3696
Sortino Ratio	0.3014	0.9065	0.9342
Standard Deviation	0.0155	0.0103	0.0283
Downside Volatility	0.0085	0.0039	0.0112
Skewness	0.1269	0.5192	0.5494
Median	0.0000	0.0000	0.0038
Maximum	0.0257	0.0196	0.0592
Minimum	−0.0201	−0.0075	−0.0290

*, **, ***, **** represent significances at 10%, 5%, 1% and 0.1% levels respectively.

Table 5.3 Overall Comparison of Trading Statistics of Different Strategies

	Distance	2D Copula	MD Copula
Average no. of trades per pair	1.0417	6.7917	8.0533
Standard Dev of number of round trips per pair	0.7506	2.1260	2.2316
Average time pairs are open in months	2.2321	4.2639	3.5548
Standard Deviation of time open, per pair, in months	1.8106	0.5384	0.5881

CHAPTER VI. CONCLUDING REMARKS

In this thesis, I mainly discuss the usage of the copula modelling technique in insurance and financial studies. In Chapter II and III, I first introduce two applications based on existing copula modelling technique to insurance and financial studies. Chapter II utilizes the copula modelling technique to deal with the associations between different stock returns, and Chapter III uses the copula modelling technique to model the relations between different insurance lines. Afterwards, I propose a copula framework with bootstrapping to deal with the data paucity problem in the context of cyber insurance in Chapter IV. Last but not least, in Chapter V, I try to extend the current analysis into multi-dimension using a non-parametric copula called Bernstein copula. Copula modelling technique can be seen as a generalization of the linear regression models, and it removes the assumptions such as normal marginal distribution imposed by the classical linear regression models or the assumption of marginal distributions from exponential families imposed by the generalized linear regression models. Moreover, it helps to investigate not only the linear associations but also the non-linear associations between random variables.

The technology of machine learning and development of big data are changing the prospect of insurance and financial industry. For example, instead of only observing driver's age, gender, educational level or occupation, there are more

tools to capture driving behaviours, personalities and even shopping behaviours to help decide on the risk level of a driver when he intends to buy auto insurance.

The characteristics of big data can be summarized as 3V, i.e., volume, velocity, and variety (Chen and Zhang, 2014). Volume means huge data size. Velocity describes that the growth of data is exponential. Last but not least, variety refers to all types of data that are available nowadays. With more information available, the most important task is to understand the associations between them, and the conventional linear regression models may not be able to handle this due to the complex associations.

Copula modelling technique is a good alternative when we investigate the associations between different variables. Conventional linear regression models have the advantages of easy implementation and interpretation, but may be subject to rigid assumptions on variables and inability to capture important non-linear relations such as tail dependence. Copula modelling technique can deal with these shortcomings and become more suitable when we model non-normal data and non-linear dependency structure.

However, there are also limitations associated with the copula modelling technique, especially the ability to handle multi-dimensional data. Though I propose an alternative multi-dimensional copula regression approach in the last chapter, the computational complexity remains a problem when the dimension is extremely high.

With big data becoming more popular in insurance and financial industry, the future research direction may focus on the development of copula modelling technique dealing with high-dimensional data.

REFERENCES

- Ane, T., and C. Kharoubi. 2003. *Dependence structure and risk measure. Journal of Business*, 76(3): 411-438.
- Andrade, S., V. D. Pietro, and M. Seasholes. 2005. Understanding the profitability of pairs trading. Working Paper, University of California, Berkeley.
- Antonio, K., and Beirlant, J. 2008. Issues in claims reserving and credibility: a semiparametric approach with mixed models. *Journal of Risk and Insurance*, 75(3): 643-676.
- Baer, W. S., and A. Parkinson, 2007. The role of cyber insurance in managing IT security. *IEEE Security and Privacy Magazine (May/June)*.
- Barracchini, C., and M. E. Addessi, 2014. Cyber risk and insurance coverage: An actuarial multistate approach, *Review of Economics & Finance*, 4: 57-69.
- Böhme, R., and G. Kataria, 2006. Models and measures for correlation in cyber-insurance, Workshop on the Economics of Information Security (WEIS). Cambridge University, UK.
- Braun, C. 2004. The prediction error of the chain ladder method applied to correlated run-off triangles. *ASTIN Bulletin*, 34(2): 399-423.

Brehm, P. 2002. Correlation and the aggregation of unpaid loss distributions. *CAS Forum*, 1-23.

Bürgi, R., M. M. Dacorogna, and R. Iles. 2008. Risk aggregation, dependence structure and diversification benefit. In: *Stress testing for financial institutions, Riskbooks*, Incisive Media, London, 265-306.

Carhart, M. M. 1997. On persistence in mutual fund performance. *The Journal of Finance*, 52: 57-82.

Chen, C. P., and C. Y. Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275: 314-347.

Chen, H., and R. D. MacMinn, and T. Sun. 2015. Multi-population mortality models: A factor copula approach. *Insurance: Mathematics and Economics*, 63: 135-146.

Conrad, J., and G. Kaul. 1989. Mean reversion in short-horizon expected returns. *Review of Financial Studies*, 2: 225-240.

Cont, R. 2001. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1: 223-236.

Cox, D. R. 1958. The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society: Series B*, 20: 215-242.

- Crane, G. J., and J. Van Der Hoek. 2008. Conditional expectation formulae for copulas. *Australian and New Zealand Journal of Statistics*, 50: 53-67.
- De Alba, E. 2002. Bayesian estimation of outstanding claims reserves. *North American Actuarial Journal*, 6(4): 1-20.
- De Alba, E. 2006. Claims reserving when there are negative values in the run-off triangle: Bayesian analysis using the three-parameter log-normal distribution. *North American Actuarial Journal*, 10(3): 45-49.
- De Jong, P. 2012. Modelling dependence between loss triangles. *North American Actuarial Journal*, 15(1): 74-86.
- De Long, J., A. Shleifer, L. Summers, and R. Waldmann. 1990. Noise trader risk in financial markets. *Journal of Political Economy*, 98(4): 703-738.
- Do, B., and R. Faff. 2010. Does simple pairs trading still work? *Financial Analysts Journal*, 66(4): 83-95.
- Efron, B., 1979. Bootstrap methods: Another look at the Jackknife, *Annals of Statistics*, 7: 1-26.
- Efron, B. and R. J. Tibshirani, 1994. *An introduction to the bootstrap*, 1st edition, Chapman and Hall/CRC.

- Embrechts, P., McNeil, A.J., and Straumann, D. 2002. Correlation and dependency in risk management: properties and pitfalls. *In Risk Management: Value at Risk and Beyond*, pp. 176-223. Cambridge University Press, Cambridge.
- Embrechts, P., and Puccetti, G. 2006. Aggregating risk capital, with an application to operational risk. *The Geneva Risk and Insurance Review*, 31: 71-90.
- England, P., and Verrall, R. 2002. Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8(3): 443-518.
- Engle, R. F., and Granger, W. J. 1987. Co-integration and error correction: representation, estimation and testing. *Econometrica*, 55: 251-276.
- Fama, E. F., and French, K. R. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33: 3-56.
- Ferreira, L. 2008. New tools for spread trading. *Futures*, 37(12): 38-41.
- Frees, E., and Valdez, E. 1998. Understanding relationships using copulas. *North American Actuarial Journal*, 2(1): 1-25.
- Frees, E., and Wang, P. 2005. Credibility using copulas. *North American Actuarial Journal*, 9(2): 31-48.
- Frees, E., and Wang, P. 2006. Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics*, 38(2): 360-373.

Gatev, E., W. N. Goetzmann, and K. G. Rouwenhorst. 2006. Pairs trading: Performance of a relative-value arbitrage rule. *Review of Financial Studies*, 19(3): 797-827.

Getczy, C., D. Musto, and A. Reed. 2002. Stocks are special too: An analysis of the equity lending market. *Journal of Financial Economics*, 66: 241-269.

Gordon, L.A., M.P. Loeb and T. Sohail, 2003. A framework for using insurance for cyber risk management, *Communications of the ACM*, 46: 81-85.

Herath, H.S.B and T.C. Herath, 2011. Copula-based actuarial model for pricing cyber-insurance policies. *Insurance Markets and Companies: Analysis and Actuarial Computations*, 2(1): 7-20.

Hess, L., Schmidt, K., and Zocher, M. 2006. Multivariate loss prediction in the multivariate additive model. *Insurance: Mathematics and Economics*, 39(2): 185-191.

Itzhak, B., F. Franzoni, and R. Moussawi. 2012. Hedge fund stock trading in the financial crisis of 2007–2009. *Review of Financial Studies*, 25(1): 1-54.

Jegadeesh, N. 1990. Evidence of predictable behavior of security returns. *The Journal of Finance*, 45: 881-898.

Jegadeesh, N., and S. Titman. 1995. Overreaction, delayed reaction, and contrarian profits. *Review of Financial Studies*, 8: 973-993.

- Keim, D., and A. Madhavan. 1997. Transaction costs and investment style: An inter-exchange analysis of institutional equity trades. *Journal of Financial Economics*, 46: 265-292.
- Kesar, S. and M. Xie, 2005. Bootstrap: a statistical method, retrieved from <http://www.stat.rutgers.edu/home/mxie/stat586/handout/Bootstrap1.pdf>.
- Li, F. 2014. Identifying asymmetric comovements of international stock market returns. *Journal of Financial Econometrics*, 12(3): 507-543.
- Liew, R. Q., and Y. Wu. Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19: 12-30.
- Luo, C, X. Chi, Y. Cong, and X. Yan. 2015. Measuring financial market risk contagion using dynamic MRS-Copula models: The case of Chinese and other international stock markets. *Economic Modelling*, 51: 657-671.
- Merz, M., and Wüthrich, M. 2008. Prediction error of the multivariate chain ladder reserving method. *North American Actuarial Journal*, 12(2): 175-197.
- Merz, M., and Wüthrich, M. 2009a. Combining chain-ladder and additive loss reserving methods for dependent lines of business. *Variance*, 3(2): 270-291.
- Merz, M., and Wüthrich, M. 2009b. Prediction error of the multivariate additive loss reserving method for dependent lines of business. *Variance*, 3(1): 131-151.

- Mukhopadhyay, A., D. Saha, B.B. Chakrabarti, A. Mahanti, A. Podder, 2005. Insurance for cyber-risk: A utility model. *Decision*, 32(1): 153-169.
- Nelson, R.B. 2006. *An Introduction to Copulas (2nd Ed.)*. New York: Springer.
- Nelson, R. B. 2006. *An Introduction to Copulas (2nd Ed.)*. New York: Springer.
- Ogut, H., S. Raghunathan, and N. Menon, 2011. Cyber security risk management: Public policy implications of correlated risk, imperfect ability to prove loss, and observability of self-protection. *Risk Analysis*, 31(3): 497-512.
- Oh, D. W., and A. J. Patton, 2015. Modelling dependence in high dimensions with factor copulas. Finance and Economics Discussion Series 2015-051. Washington: Board of Governors of the Federal Reserve System.
- Okhrin, O., Y. Okhrin, and W. Schmid. 2013. On the structure and estimation of hierarchical Archimedean copulas. *Journal of Econometrics*, 173: 189-204.
- Pardoe, I., 2000. An introduction to bootstrap methods using arc. University of Minnesota, St. Paul, MN 55108, Technical Report Number 631.
- Parsa, R. A., and S. A. Klugman. 2011. Copula regression. *Variance*, 5(1): 45-54.
- Perlin, M. S. 2009. Evaluation of pairs-trading strategy at the Brazilian financial market. *Journal of Derivatives & Hedge Funds*, 15(2): 122-136.

Pizzutilo, F. 2013. A note on the effectiveness of pairs trading for individual investors. *International Journal of Economics & Financial Issues*, 3(3): 764-771.

Sancetta, A., and S. Satchell, 2004. The Bernstein copula and its applications to modelling and approximations of multivariate distributions, *Econometric Theory*, 20: 1-38.

Sandoval, L., and I. Franca. 2012. Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, 391(1): 187-208.

Savu, C., and M. Trede, 2010. Hierarchical archimedean copulas. *Quantitative Finance*, 10: 295-304.

Schmidt, K. 2006. Optimal and additive loss reserving for dependent lines of business. *CAS Forum*: 319-351.

Shemyakin, A. and H. Youn. 2006. Copula models of joint last survivor analysis. *Applied Stochastic Models in Business and Industry*, 22(2): 211-224.

Shi, P., and Frees, E. 2010. Long-tail longitudinal modelling of insurance company expenses. *Insurance: Mathematics and Economics*, 47(3): 303-314.

Shi, P., and Frees, E. 2011. Dependent loss reserving using copulas. *ASTIN Bulletin*, 41(2): 449-486.

- Shi, P., and Valdez, E. 2014. Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, 55(1): 18-29.
- Sklar, A. 1959. Fonctions de répartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris*, 8: 229-231.
- Spreeuw, J. 2006. Types of dependence and time-dependent association between two lifetimes in single parameter copula models. *Scandinavian Actuarial Journal*, 2006(5): 286-309.
- Stander, Y., D. Marais, and I. Botha. Trading strategies with copulas. *Journal of Economic and Financial Sciences*, 6(1): 83-108.
- Sun, J., Frees, E., and Rosenberg, M. 2008. Heavy-tailed longitudinal data modelling using copulas. *Insurance: Mathematics and Economics*, 42(2): 817-830.
- Taylor, G. 2000. *Loss Reserving: An Actuarial Perspective*. Kluwer Academic Publishers.
- Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica*, 26 (1): 24–36.
- Varian, H., 2005. Bootstrap tutorial. *Mathematica Journal*, 9: 768-775.
- Verrall, R. J. 1996. Claims reserving and generalized additive models. *Insurance: Mathematics and Economics*, 19: 31-43.

Walsh, P. D., R. Biek, and L. A. Real. 2005. Wave-like spread of Ebola Zaire. *PLoS Biol*, 3(11), e371.

Wang, C.W., S. S. Yang, and H. C. Huang. 2015. Modelling multi-country mortality dependence and its application in pricing survivor index swaps - A dynamic copula approach. *Insurance: Mathematics and Economics*, 63: 30-39.

Wu, C., and Z. Lin. 2014. An economic evaluation of stock–bond return comovements with copula-based GARCH models. *Quantitative Finance*, 14: 1283-1296.

Wüthrich, M., and Merz, M. 2008. *Stochastic Claims Reserving Methods in Insurance*. John Wiley & Sons.

Yang, L., and S. Hamori. 2013. Dependence structure among international stock markets: A GARCH–copula analysis. *Applied Financial Economics*, 23(23): 1805-1817.

Zhang, Y. 2010. A general multivariate chain ladder model. *Insurance: Mathematics and Economics*, 46(3): 588-599.

Zhang, Y., Dukik, V., and Guszczka, J. 2012. A Bayesian nonlinear model for forecasting insurance loss payments. *Journal of the Royal Statistical Society*, 175: 637-656.

Zhu, W, K. S. Tan, and C. Wang, 2015. Modelling multi-population longevity risk with mortality dependence: A lévy subordinated hierarchical archimedean copulas (LSHAC) approach, working paper.

APPENDIX

Table A1. Claim Triangles for the Personal Auto Line (1989-2007)

Accident Year	Premium	0	1	2	3	4	5	6	7	8	9
1989	8,756,357	2,823,701	2,536,475	1,125,533	601,484	321,420	157,915	79,395	38,016	18,838	9,836
1990	9,784,849	3,181,381	2,723,043	1,226,151	633,645	321,623	154,692	74,083	39,773	17,525	11,638
1991	10,584,031	3,188,314	2,682,586	1,196,454	622,922	297,869	143,653	75,744	40,527	23,064	8,743
1992	11,443,613	3,557,266	2,907,921	1,288,319	640,096	311,520	159,719	82,672	38,593	18,236	9,626
1993	12,019,862	3,864,567	3,076,096	1,285,639	633,827	315,839	162,529	87,525	40,933	17,074	12,640
1994	12,906,846	4,254,977	3,184,599	1,282,801	673,810	358,060	171,278	77,177	39,540	19,901	13,663
1995	13,695,937	4,416,059	3,085,207	1,278,943	675,913	347,680	177,668	93,212	47,303	34,314	14,423
1996	14,187,681	4,364,315	2,930,960	1,284,297	677,882	374,280	179,656	93,677	50,548	24,770	13,121
1997	14,423,805	4,256,384	2,845,630	1,225,166	698,481	370,666	183,970	83,442	44,132	24,346	14,962
1998	13,986,682	4,208,129	2,696,292	1,247,432	711,825	363,532	169,271	84,741	45,850	20,908	16,492
1999	13,062,099	4,325,679	2,808,668	1,295,227	689,146	344,196	180,155	92,708	44,057	26,946	
2000	12,912,923	4,659,760	3,133,005	1,278,609	687,925	382,801	183,827	93,233	49,540		
2001	13,737,555	5,191,950	3,355,237	1,352,152	760,362	403,567	202,809	111,573			
2002	15,538,426	5,735,108	3,456,295	1,450,094	800,682	437,833	230,331				
2003	17,252,390	5,405,940	3,238,611	1,361,110	820,430	449,778					
2004	17,606,104	5,147,442	3,145,537	1,350,208	839,574						
2005	17,207,089	5,078,749	3,106,406	1,392,783							
2006	17,270,124	5,082,940	3,141,482								
2007	16,982,578	5,272,414									

Table A2. Claim Triangles for the Commercial Auto Line (1989-2007)

Accident Year	Premium	0	1	2	3	4	5	6	7	8	9
1989	308,625	60,022	59,233	31,770	23,503	10,801	6,744	3,939	1,861	1,984	483
1990	326,256	65,793	64,885	41,572	25,583	12,253	9,036	3,161	1,650	777	575
1991	332,437	61,901	59,135	37,744	23,797	12,557	6,607	3,058	624	515	669
1992	341,701	65,002	63,481	35,878	23,069	12,315	8,186	2,916	2,082	1,090	427
1993	353,859	72,021	71,960	40,640	22,557	10,700	6,561	4,012	1,553	1,146	622
1994	375,881	81,089	68,201	36,768	22,056	16,677	7,436	3,177	1,647	395	1,414
1995	393,215	82,401	67,921	37,313	21,220	12,467	7,146	3,057	673	1,846	588
1996	400,027	78,666	62,770	39,514	21,752	16,669	8,676	3,772	-498	98	512
1997	399,064	75,028	66,121	40,871	20,586	13,049	5,664	3,506	1,541	-720	472
1998	239,212	51,984	39,573	27,894	15,432	8,849	3,828	1,733	2,019	1,409	367
1999	225,695	54,084	38,906	29,282	15,152	8,077	3,435	1,686	1,963	657	
2000	216,044	54,176	46,846	29,851	17,657	10,182	4,907	2,941	541		
2001	226,678	54,579	38,797	23,295	13,914	8,434	5,057	1,994			
2002	251,753	58,602	45,844	27,878	20,270	11,653	5,272				
2003	273,891	60,152	45,950	28,146	22,175	13,497					
2004	294,545	55,946	46,755	29,210	23,914						
2005	314,875	63,715	54,244	46,795							
2006	347,173	74,354	50,243								
2007	362,794	73,776									