# RANK-BASED MULTIVARIATE SARMANOV FOR MODELING DEPENDENCE BETWEEN LOSS RESERVES

# RANK-BASED MULTIVARIATE SARMANOV FOR MODELING DEPENDENCE BETWEEN LOSS RESERVES

BY

Lan Wang, B.Sc.

A THESIS
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Master of Science (2023)
(Statistics)

McMaster University
Hamilton, Ontario, Canada

TITLE: Rank-Based Multivariate Sarmanov for Modeling Dependence Between Loss Reserves

AUTHOR: Lan Wang
B.Sc. (Mathematics and Statistics)
McMaster University,
Hamilton, Ontario, Canada

SUPERVISOR: Dr. Anas Abdallah

NUMBER OF PAGES: xi, 56

*To the spectacular world.*

# Abstract

The dependence between multiple lines of business has an important impact on determining loss reserves and risk capital, which are crucial elements of risk management for an insurance portfolio. In this work, we show that the Sarmanov family of multivariate distribution can be used for dependent lines of business using a rank-based method estimation. In fact, an inadequate choice of the dependence structure may negatively impact the estimation of the marginals, which might lead to an undesirable effect on reserve computation. Thus, we propose a two-stage inference strategy in this thesis. We show that this strategy leads to robust estimation and better capture the dependence between the risks. We also show that it leads to smaller risk capital and a better diversification benefit.

We introduce the two-stage inference using the Sarmanov distribution. First, we fit the marginals with generalized linear models (GLMs) and obtain the corresponding residuals. Secondly, the Sarmanov family of bivariate distributions links these marginals through the rank of residuals. We also show that this can be extended to a multivariate case.

To illustrate this method, we analyzed two sets of data. For the bivariate case, we considered an insurance portfolio consisting of personal and commercial auto lines provided by a major US property-casualty insurer. We also used the data from three lines of business of a large Canadian insurance company for the multivariate dependence case.

# Acknowledgements

First and most importantly, I would like to express my deepest gratitude to my supervisor Dr. Anas Abdallah, for his patience, helpful feedback, encouragement, and guidance throughout my master's study. I would also thank my defense committee, Dr. Pratheepa Jeganathan, and Dr. Traian Pirvu, for their time and support.

I am also grateful for the support of my friends and family, my mother, Lan Yu, my friend Runyue Wang, and Haoshuai Wang, who stayed by my side and supported me.

I would also like to thank my roommate's dog Bitcoin for all the entertainment and emotional support.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

For insurance companies, the production cycle is inverted because the insurer receives the price (premium) of the product before knowing the cost (claim). So the insurer needs to estimate the cost and ensure there is enough money aside to meet its commitments to its policyholders and claimants, which constitutes the reserve. Classical reserving methods are often determined under an independent assumption between the portfolio risk components. However, risks are related to each other in practice, and this dependence needs to be considered as a correlation exists between multiple lines of business. Therefore, it plays an important role in determining the reserve for the whole portfolio and, more importantly, calculating the risk capital. The risk capital is the amount that property & casualty insurers set aside as a buffer against potential losses from extreme and adverse events.

In order to get the loss reserves, we have the original loss triangles for each line of business with rows assigned as accident years and columns as development periods, which we can use to predict future claims and complete the lower part of the loss triangle.

To capture the dependencies between different loss triangles, the mainly used method involves the copula model. For example, Shi and Frees (2011) analyzed the dependent loss reserving between two lines of business using Gaussian and Frank copula. In this research we explore and study the Sarmanov family of distribution, which has also been used in literature. For example, Abdallah et al. (2016) used bivariate Sarmanov distributions with random effects taking into account the correlation between two lines of business. The original method using the Sarmanov distribution is to perform a one-stage inference, by simultaneously estimating the marginals and the dependence parameters. However, a change in the dependence structure would lead to different parameters estimation for the marginals, and thus, to a different total reserve. Consequently, this method has the undesirable effect of violating the linear property of the mean.

In this research, we propose to use a two-step inference method. In the first step, generalized linear models are fitted to the marginals, which will fix the parameters of the marginals and the estimations of the reserves. Then we link the dependence of GLMs using the rank-based method for bivariate and trivariate Sarmanov. This approach has been used in the copula model. For example, Côté et al. (2016) used the rank-based method in the Archimedean copula model with six lines of business. However, the rank-based method has never been introduced in the literature with the Sarmanov family of multivariate distributions, which is more flexible than copulas.

Some background knowledge about loss reserves, loss triangles and dependence between lines of business are reviewed in Chapter 2. Next, we introduce the data we used for this thesis in Chapter 3, which includes real data of an insurance portfolio consisting of personal and commercial auto lines provided by a major US property-casualty insurer, and three lines of business data from a large Canadian insurance company. The data will be analyzed in the next two chapters. Then we present the bivariate and multivariate Sarmanov distribution and introduce the rank-based method comparing it with the classical (one-stage inference) method in Chapter 4. In Chapter 5, we present the importance of risk capital and show a better diversification benefit with the two-stage inference method, using simulation and bootstrapping. Finally, Chapter 6 concludes and summarizes the comparison between the rank-based method and the one-stage inference method.

# Chapter 2

# Background

## 2.1 Loss Reserve in property and casualty insurance

A property and casualty insurance policy is a contract between two parties, the insurer and the insured, where the insurer is usually an insurance company and the insured purchase insurance product from the insurer. After receiving the premium paid from the insured, the insurer need to pay amounts of money to the insured once the accident or events mentioned in the agreement occurs. The amount of money the insurer needs to pay is called the claim amount. Therefore, the insurer needs to reserve amounts of money for the claim amount they need to pay in the future. The reserved amount of money is called the loss reserve, which is an estimation of an insurer's liability from future claims. The estimation procedure and technique are called loss reserving. Loss reserving is crucial for risk management, as it quantifies and predicts potential losses and controls the financial impact of risks.

Figure 2.1 shows the Lexis diagram for the lifetime of claims. The x-axis is the calendar time, while the y-axis is the years of development. The dot "•" in the figure shows the date that claims occur. The plus sign "+" gives the date that the claims have been declared to the insurer, and the claims are closed on date "×". The red vertical line is the current date, which in this figure is year 2011. The dotted blue line after the current date is the future claims the insurer will need to pay. So loss reserving is to predict the future claims that are needed to be paid by the insurer, given past open claims.

The loss reserving process normally involves analyzing historical claim data using statistical

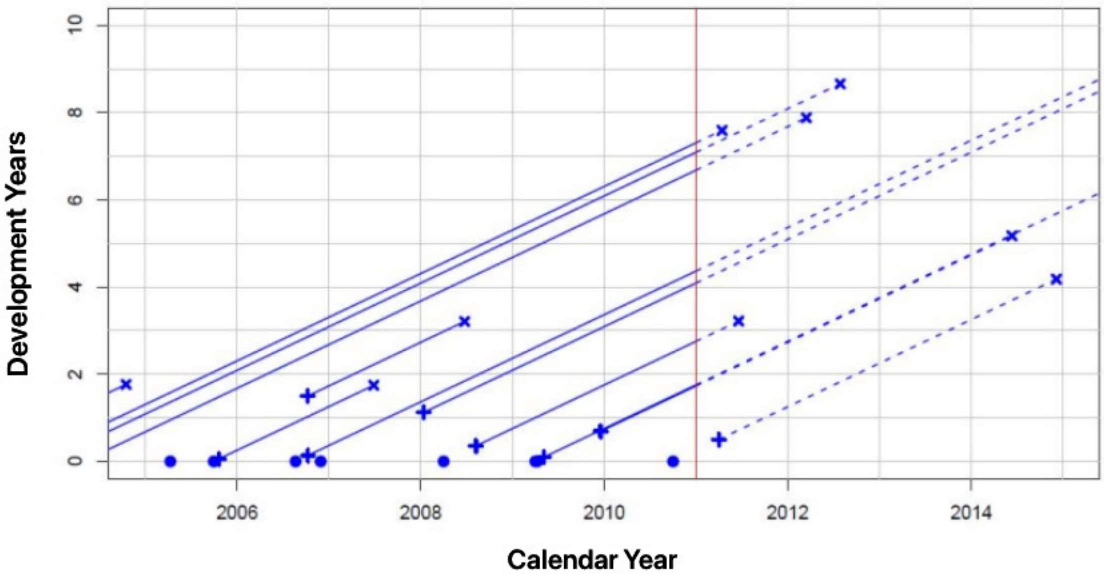



Figure 2.1: Lexis diagram for the lifetime of claims

models to predict future losses. The historical claim is often presented in the form of a loss triangle. It is a triangular-shaped table that shows the development of losses over a period of time for each accident year, which is the year that the accident occurs and the issuer requires to claim. The following shows an example of loss triangle:

$$\begin{array}{ccccc} X_{1,1} & X_{1,2} & \dots & X_{1,n-1} & X_{1,n} \\ X_{2,1} & X_{2,2} & \dots & X_{2,n-1} & \\ \vdots & \vdots & \iddots & & \\ X_{n-1,1} & X_{n-1,2} & & & \\ X_{n,n} & & & & \end{array}$$

where for $X_{i,j}$, $i$ indicates the accident year and $j$ indicates the development period. $X_{i,j}$ gives the loss incurred during the $i$th accident year and $j$th development period.

Loss triangles may show incremental claims or cumulative claims. The incremental claims represent additional claims that are needed to be paid during a given period of time, i.e., the net increase in claim over a development period, while the cumulative claims are the total amount of claims that have been reported or reserved over the period of time, i.e., the sum of current and previous incremental claims. In this thesis, we will use loss triangles with incremental claims.

In order to analyze the loss triangles and estimate loss reserves, we will need to complete the lower part of the loss triangle, i.e., predict the incremental claims for the future development period of a certain accident year. Several methods are used for loss reserving, for instance, the chain-ladder method and the Bornhuetter-Ferguson method, the most basic claims reserving and the most frequently used techniques in practice. The chain-ladder method predicts the future claims based on the pattern of historical claims and assumes the pattern will continue into the future, which might not always be accurate. The Bornhuetter-Ferguson method proposed by Bornhuetter and Ferguson (1972) is a hybrid approach that combines historical data with judgment and experience to produce more accurate estimations.

## 2.2 Dependence between lines of business

For now, we have been talking about loss reserving for one single line of business, but in practice, insurance company has multiple lines of business. Therefore, while estimating the loss reserve, we need to consider the dependence between different lines of business. This means that the loss reserves and risks of two or more loss triangles can be correlated with each other. To capture dependence between loss triangles, there exist two different approaches.

Various literature have been proceeding on studying distribution-free multivariate reserving methods. Braun (2004) showed the effectiveness of the multivariate chain-ladder method using simulated data and found it provides an accurate estimation of prediction error when taking the correlation between loss triangles into account. Merz and W$\ddot{u}$thrich (2008) also considered the prediction error of a modified multivariate chain-ladder model proposed by Schmidt (2006) and incorporated the dependence structure in to their model.

The other approach to modeling the dependence between lines of business is using parametric methods based on various distributional families, which involves assuming a specific distribution for the loss for each line of business, then modeling the dependence between the distributions. This approach allows greater flexibility in modeling the dependence and can provide more information to the actuaries by giving a reasonable range of loss reserves rather than only a mean square prediction error. Therefore, our thesis will focus on the parametric approach.

One commonly used method for parametric loss reserving is the copula model to capture dependence between lines of business. Copula can be used to describe the dependence structure between two or more random variables. Modelling using copula often starts by selecting appropriate copula functions and then estimating the dependence parameter based on the marginal distribution fitted into each line of business. Future claims can then be simulated using the copula model, and loss reserves can be estimated.

A lot of literature have studied on these reserving methods. Brehm (2002) proposed using a Gaussian copula to model the joint distribution of unpaid losses. Moreover, De Jong (2012) used a Gaussian copula correlation matrix to model the dependence between lines of business. Although, the Gaussian copula assumes the marginals to follow a normal distribution, both Brehm and De Jong assumed the loss distribution of each line of business follows a log-normal distribution, which is a commonly used distribution in modelling the losses in the insurance industry. Shi et al. (2012) and W*ü*thrich et al. (2013) used multivariate Gaussian copula to capture correlation due to accounting years using loss triangles, while W*ü*thrich et al. (2013) allowed the correlation matrix to vary over time and produces more accurate modeling of dependence.

Bootstrapping is also a popular parametric approach used for loss reserving, which involves resampling the historical data to simulate and generate new datasets (pseudo-responses). Bootstrapping is the method we will use to estimate the predictive distribution for unpaid losses. Kirschner et al. (2008) proposed the synchronized bootstrap, which aimed to estimate the prediction error of a multivariate dependence model. Taylor and McGuire (2007) modified their approach to account for the additional complexity introduced by the generalized linear model framework.

Shi and Frees (2011) used Frank and Gaussian copula to model the dependence between lines of business, and introduced a parametric bootstrapping method to estimate the prediction error. Frank Copula can capture both positive and negative dependence between the lines of business, as its parameter controls the strength and direction of the dependence. Shi and Frees (2011) used the bootstrap method to compare between Frank and Gaussian copula. Abdallah et al. (2015) used hierarchical Archimedean copulas (HAC) to model the dependence and correlation between loss triangles. They found it outperformed several other methods, including the Gaussian copula and multivariate chain-ladder method. Hierarchical Archimedean copulas can capture complex dependence structure by combining Archimedean copulas in a hierarchical manner. Futhermore, it can accommodate various types of marginal distribution and can capture both positive and negative dependence.

In this thesis, we will consider the Sarmanov family of a multivariate distribution. This family of distribution was first introduced in Sarmanov (1966) and Lee (1996) proposed using bivariate Sarmanov distribution to model dependence between bivariate random vectors. One of the reasons we selected this distribution instead of copula is that the Sarmanov distribution can easily provide closed-form expressions for loss reserves, while it is very complex for copula. The closed-form mean, variance, and covariance for loss reserves generated using the Sarmanov distribution is in Appendix 2.

Bahraoui et al. (2015) performed the bivariate Sarmanov distribution and copula, showing that the bivariate Sarmanov is more flexible than copulas in modeling dependence. In fact, in addition to the possibility of capturing both positive and negative dependencies, the bivariate Sarmanov model also provides more flexibility for tail dependence. As such, the paper showed bivariate Sarmanov distribution models skewed data, which is inappropriate for Gaussian copula. The author also proposed a method for estimating the dependence parameter for Sarmanov distribution based on maximum likelihood estimation. Furthermore, the applicability of Sarmanov's distribution results from its versatile structure that offers us flexibility in the choice of marginals and allows a closed form for the joint density. Abdallah et al. (2016) showed the potential of this family of distributions in a loss reserving context. The paper used random effects to accommodate the correlation between loss triangles. Ratovomirija et al. (2016) proposed a new method based on multivariate Sarmanov mixed Erlang distribution to model the joint distribution for lines of business. Bolancé and Vernic (2017) also provided three approaches based on multivariate Sarmanov distribution to model dependence loss reserving.

In this research, we propose to use a two-step inference method called the rank-based method. This approach uses the rank of the observations rather than the actual values of data in the analysis. We will link the dependence of GLMs using the rank-based method for Sarmanov distribution. This approach has been used in the copula model. Genest and Neschléhova (2012) discuss the rank-based methods for copula estimation. *Côté* et al. (2016) used the rank-based method that replaced the loss data with the rank of residuals in the Archimedean copula model with six lines of business. Residuals are the differences between the observed claim of the dependent variable and the predicted loss reserve. Further, the rank of residuals are the order of magnitude of the residuals. To our knowledge, the rank-based method has not been applied to Sarmanov distribution.

The rank-based method is robust to outliers and non-normality. Thus, we propose to use this method to accommodate the correlation between loss triangles using bivariate and trivariate Sarmanov distribution.

## 2.3 Modelling and Notation

In this research, we use the generalized linear model (GLM) as the marginals for each lines of business. GLM is a type of regression analysis that allows various distributions of the response variable, while there is linear or non-linear relationships between the response and predictor variables.

In our case, we will take the accident year and development period as the predictor variable when fitting the generalized linear model for a loss triangle, and for the response variable, we will use the loss ratio. The loss ratio is a key performance metric used to measure the profitability and loss for an insurance portfolio. It is used to represent the ratio of incurred losses to premiums earned, where the losses include paid insurance claims and adjustment expenses.

As mentioned above, in a loss triangle, the row would represent the year which an accident occurs, the column represents each year passed since the accident happened. We will use $i$ to indicate the accident year and $j$ as the development period, which are the row and column respectively. Let $\ell$ be the different lines of business, then we will denote $X_{ij}^{(\ell)}$ as the incremental payments in the loss triangle. Let $p_i^{(\ell)}$ be the premium for the $\ell$th lines of business and $i$th accident year, then $y_{ij}^{(\ell)} = X_{ij}^{(l)}/p_i^{(\ell)}$ is the loss ratio.

After calculating the loss ratio, we can fit the generalized linear model with accident year and development period as factor, using different types of independent distributions, to find out which distribution model fits the marginals well. The inference method used to choose the better model will be mentioned in the next chapter.

In order to fit the generalised linear model, we use the procedure shown in Abdallah et al. (2016), let $s_i^{(\ell)}$ be the effect of accident year, $t_j^{(\ell)}$ be the effect of development period, $i, j \in \{1, 2, ..., n\}$ then the systematic component for the $\ell$th line of business can be shown as:

$$\eta_{ij}^{(\ell)} = u^{(\ell)} + s_i^{(\ell)} + t_j^{(\ell)},$$

where $u^{(\ell)}$ is the intercept and for parameter identification, $s_i^{(\ell)}$ and $t_j^{(\ell)}$ are set to 0 for $i,\ j = 1$. Here we will give two examples of distributions. If we fit the log-normal distribution, then

$$a_{ij}^{(\ell)} = \eta_{ij}^{(\ell)}$$

where $a_{ij}^{(\ell)}$ is the mean of the log-normal distribution with standard deviation $b^{(\ell)}$. If we fit the Gamma distribution, we will have

$$\tau_{ij}^{(\ell)} = exp(\eta_{ij}^{(\ell)})/\alpha^{(\ell)}$$

where the non-zero $\alpha^{(\ell)}$ is the shape parameter and $\tau_{ij}^{(\ell)}$ is the scale parameter of the gamma distribution. We use maximum likelihood estimation for the parameter estimation of all the models.

With the estimated parameters, the total reserve can be estimated using

$$\sum_{\ell}\sum_{i}\sum_{j} p_i^{(\ell)} E(y_{ij}^{(\ell)})$$

where $E(y_{ij}^{(\ell)})$ is the mean of unpaid loss ratio. For log-normal distribution, we have

$$E(y_{ij}^{(\ell)}) = exp\Big[a_{ij}^{(\ell)} + \frac{(b^{(\ell)})^2}{2}\Big],$$

and for the gamma distribution, we have

$$E(y_{ij}^{(\ell)}) = \tau_{ij}^{(\ell)}\alpha^{(\ell)}$$

# Chapter 3

# Data

There are two sets of data we used in this thesis. Both of them come from real life data which have different lines of business that may be dependent with each other and can be analysed in this project.

## 3.1 Shi and Frees (2011) Data

The data we used for Bivariate case are the same as the ones used in Shi and Frees (2011) and Abdallah et al. (2016), which is an insurance portfolio consisting of two business lines personal and commercial automobile lines from a major US property casualty insurer. The data were collected from Schedule P of the National Association of Insurance Commissioners (NAIC) database. The NAIC is an organization created and governed by the head of insurance regulators from the whole US. It was created in 1871 to be used as a forum for information exchanging and is one of the largest insurance regulatory database. The Schedule P provides losses and aggregated claims within 10 years time, which can be arranged into loss triangles. It also gives the unpaid losses, premium earned for all lines of business.

Personal auto line is the insurance on personal vehicle, while the commercial automobile line is insurance for physical damage and liability coverages for the situation not covered by the personal auto line. The loss triangle of this dataset can be found in Appendix 1.

### 3.1.1 Inference for the marginals

Shi and Frees (2011) assumed that the personal auto line follows log-normal distribution and the commercial auto line follows gamma distribution. Here we will introduce some inference methods to check whether the data fits better with log-normal and gamma model.

**Akaike information criterion (AIC)**

The Akaike information criterion is an estimator of prediction error, which can be used to check the quality of statistical models, given a set of data and provide a means for model selection. It can also be used for non-nested model. When fitting models, adding parameters may cause increasing of the loglikelihood which would lead to overfitting, AIC adds a penalty term to resolve this problem.

Let $k$ be the number of estimated parameters in the statistic model, let $L$ be the maximum value of the likelihood function of the model, then the AIC can be expressed as

$$AIC = 2k - 2ln(\hat{L})$$

where $\hat{L}$ represents the estimated value of the maximum likelihood. The lower AIC value gives the better model.

We fitted log-normal and gamma distribution to both personal line and commercial auto line of business, and the corresponding Akaike information criterion (AIC) is in Table 3.1. The AIC results shows that personal line has lower AIC when fitted to log-normal model and therefore (i.e., is more fitted to a log-normal model) and commercial auto line is more fitted to gamma distribution.

Table 3.1: Fit statistics for marginals of personal and commercial Lines

| Lines of business/AIC | Lognormal | Gamma |
|---|---|---|
| Personal Line | -395.095 | -384.453 |
| Commercial Auto Line | -214.495 | -218.083 |

We can also use goodness-of-fit test to check if the data follows the distribution.

**Kolmogorov-Smirnov (KS) test**

The Kolmogorov-Smirnov (KS) test is a non-parametric test that can compare the observed data with a theoretical distribution for one-sample KS test. The two-sample KS test can compare two sets of observed data with each other.

The KS test produces a empirical cumulative distribution function for the non-parametric data, and measures the distance between the cumulative distribution function (cdf) of two distributions and provides whether they are from the same family of distribution.

The null hypothesis of KS test is that the data follows the specified distribution, or the two sets of data comes from the same distribution, while the alternative hypothesis is the opposite. The KS statistics with given cdf F(x) is calculated as:

$$D_n = sup_x|F_n(x) - F(x)|$$

where the $sup_x$ is the supremum.

If the p-value for the KS test is bigger than the significance level, we cannot reject the null hypothesis, and there is no enough evidence that the data do not come from the given distribution. If the p-value is very small, then we reject the null hypothesis and say that the data is not from the given distribution or the two sets of data does not come from the same distribution.

We do the KS test for the residuals of personal auto line with log-normal distribution and commercial auto line with gamma distribution. Table 3.2 shows that there is no strong evidence against saying personal auto line follows log-normal distribution and commercial auto line follows gamma distribution, although the fit of the Commercial auto is borderline.

Table 3.2: KS Test for marginals of personal and commercial Lines

| Lines of business/p-value | Personal Auto(Log-normal) | Commercial Auto(Gamma) |
|---|---|---|
| Kolmogorov-Smirnov (KS) test | 0.8732 | 0.077 |

## 3.2 Côté et al. (2016) Data

The data we used for both Bivariate and Trivariate case are the same ones used in Côté et al. (2016) which is real data from a large Canadian property and casualty insurance company. It includes the loss triangle, loss ratio, rank of residuals, gamma model with parameters of six lines of business: Atlantic Bodily injury, Ontario Bodily injury, West Bodily injury, Ontario Accident benefits excluding disability income, Ontario Accident benefits with disability income only and Country-wide Liability. For the trivariate case, we pick lines 2,4 and 5, which are Ontario Bodily injury, Ontario Accident benefits excluding disability income and Ontario Accident benefits with disability income only. We will also use line 2 and 4 for the bivariate case. This is because in all six lines of business, line 2, 4 and 5 are in the Ontairio Region and their products are auto insurance, which would lead to stronger dependence between these lines of business. A descriptive summary of the three lines of business is given in Table 3.3.

Table 3.3: Descriptive summary of three lines of business from a Canadian insurance company

| LOB | Region | Product | Coverage |
|---|---|---|---|
| 2 | Ontario | Auto | Bodily injury |
| 4 | Ontario | Auto | Accident benefits excluding disability income |
| 5 | Ontario | Auto | Accident benefits: disability income only |

Bodily injury coverage gives payments to the insured if they are injured or killed by an automobile accident which occurs through the fault of the vehicle owner who has no insurance, or by unidentified vehicles. The accident benefits coverage provides compensation for injury or death involved in a vehicle collision regardless of fault, including if the insured's role during the accident is the driver, passenger or a pedestrian. Disability income provides compensation if the accident results in a disability and the insured could not continue work at their regular employment because of this disability. The data is in Appendix 1.

### 3.2.1 Inference for the marginals

Côté et al. (2016) assumed that all three lines of business Ontario Bodily injury, Ontario Accident benefits excluding disability income and Ontario Accident benefits with disability income only follow gamma distribution. We check this using AIC comparing with log-normal distribution and KS test to see if they fit well enough for gamma distribution.

Table 3.4: Fit statistics for marginals of Line 2,4 and 5

| Lines of business/AIC | Lognormal | Gamma |
|---|---|---|
| 2 | -262.1514 | -270.1587 |
| 4 | -267.3952 | -276.1508 |
| 5 | -436.7875 | -443.9719 |

Table 3.5: KS Test for marginals of Line 2, 4 and 5

| Lines of business/p-value | 2 | 4 | 5 |
|---|---|---|---|
| Kolmogorov-Smirnov (KS) test (Gamma) | 0.6443 | 0.1356 | 0.4787 |

Table 3.4 and Table 3.5 shows that the three lines of business fit well for the gamma distribution.

# Chapter 4

# Sarmanov distribution and estimation

## 4.1 One-stage inference for the Dependence Structure

### 4.1.1 Bivariate distribution

Let $y_{ij}^{(\ell)}$ be the element from each line of business, where $\ell \in \{1, 2\}$, $f^{(\ell)}$ be the univariate probability density function, and $\psi^{(\ell)}(y_{ij}^{(\ell)})$ be nonconstant functions such that $\int_{-\infty}^{\infty} \psi^{(\ell)}(t) f^{(\ell)}(t)dt = 0$. If we use line 2 and 4 from Côté et al. (2016) data, $y_{ij}^{(\ell)}$ follow Gamma distribution, i.e., $y_{ij}^{(\ell)} \sim Gamma(\alpha^{(\ell)}, \tau_{ij}^{(\ell)})$. Here for convenience, we write $\alpha^{(\ell)}$ as $\alpha_\ell$, and $\tau_{ij}^{(\ell)}$ as $\tau_\ell$. Then the bivariate Sarmanov joint distribution can be expressed as

$$f^S(y_{ij}^{(1)}, y_{ij}^{(2)}) = f^{(1)}\big(y_{ij}^{(1)}; \alpha_1, \tau_1\big) f^{(2)}\big(y_{ij}^{(2)}; \alpha_2, \tau_2\big)\big(1 + \omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\big), \tag{4.1}$$

with the mixing function:

$$\psi^{(\ell)}(y_{ij}^{(\ell)}) = exp(-y_{ij}^{(\ell)}) - (1 + \tau_\ell)^{-\alpha_\ell}, \quad \ell = 1, 2. \tag{4.2}$$

This is because that Corollary 2 in Lee (1996) proposed that a mixing function can be defined as $\psi^{(\ell)}(y_{ij}^{(\ell)}) = exp(-y_{ij}^{(\ell)}) - L^{(\ell)}(1)$, where $L^{(\ell)}$ is the Laplace transform of $f^{(\ell)}$, evaluated at 1. Thus we get (4.2), as $y_{ij}^{(\ell)},\ \ell \in \{1, 2\}$, follow gamma distribution.

Similarly, if we use Personal and commercial auto lines from Shi and Frees (2011) data, our first lines of business follows normal distribution where the response variable is positive in order to acquire the logarithm, while the second lines of business follows gamma distribution, we will have:

$$f^S(y_{ij}^{(1)}, y_{ij}^{(2)}) = f^{(1)}\big(y_{ij}^{(1)}; a_1, b_1\big) f^{(2)}\big(y_{ij}^{(2)}; \alpha_2, \tau_2\big)\big(1 + \omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\big), \tag{4.3}$$

with the mixing function:

$$\psi^{(1)}(y_{ij}^{(1)}) = exp(-y_{ij}^{(1)}) - exp\Big(-a_1 + \frac{b_1^2}{2}\Big) \tag{4.4}$$

and

$$\psi^{(2)}(y_{ij}^{(2)}) = exp(-y_{ij}^{(2)}) - (1+\tau_2)^{-\alpha_2}.$$

The variable $\omega$ in (4.1) should be a real number, which requires the constraint

$$1 + \omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)}) \geq 0 \tag{4.5}$$

for all $y_{ij}^{(1)}, y_{ij}^{(2)}$. This is also a very important condition when coding the Sarmanov model.

As Theorem 2 in Lee (1996) mentioned, the correlation coefficient of $y_{ij}^{(1)}, y_{ij}^{(2)}$ is given as

$$\rho = \frac{\omega\nu_1\nu_2}{\sigma_1\sigma_2},$$

where

$$\mu_\ell = \int_{-\infty}^{\infty} t f^{(\ell)}(t)dt, \quad \sigma_\ell^2 = \int_{-\infty}^{\infty} (t-\mu_\ell)^2 f^{(\ell)}(t)dt, \quad \nu_\ell = \int_{-\infty}^{\infty} t\psi^{(\ell)}(t) f^{(\ell)}(t)dt,$$

therefore, if both lines of business follows gamma distribution, then

$$\sigma_\ell = \sqrt{\alpha_\ell}\tau_\ell, \quad \nu_\ell = \alpha_\ell\tau_\ell^2(1+\tau_\ell)^{-\alpha_\ell-1}.$$

As we know that $-1 \leq \rho \leq 1$, then we can obtain the lower and upper bound of $\omega$,

$$-\frac{1}{\sqrt{\alpha_1}\tau_1(1+\tau_1)^{-\alpha_1-1}\sqrt{\alpha_2}\tau_2(1+\tau_2)^{-\alpha_2-1}} \leq \omega \leq \frac{1}{\sqrt{\alpha_1}\tau_1(1+\tau_1)^{-\alpha_1-1}\sqrt{\alpha_2}\tau_2(1+\tau_2)^{-\alpha_2-1}}. \tag{4.6}$$

Similarly, if the two lines of business follows normal and gamma distribution, then

$$\sigma_1 = b_1, \quad \nu_1 = -b^2 exp\Big(-a_1 + \frac{b_1^2}{2}\Big).$$

Therefore the lower and upper bound of $\omega$ can be obtained as

$$-\frac{1}{b_1 exp(-a_1+b_1^2/2)\sqrt{\alpha_2}\tau_2(1+\tau_2)^{-\alpha_2-1}} \leq \omega \leq \frac{1}{b_1 exp(-a_1+b_1^2/2)\sqrt{\alpha_2}\tau_2(1+\tau_2)^{-\alpha_2-1}}. \tag{4.7}$$

The full proof of the bounds of $\omega$ can be found in Appendix 3.

### 4.1.2 Trivariate distribution

The trivariate distribution is similar to the bivariate distribution. We now have three lines of business, thus $y_{ij}^{(\ell)}$ with $\ell \in \{1,2,3\}$. Here we assume we use the three lines of business from Côté et al. (2016) data, then the distribution function is given as follows.

$$\begin{aligned} f^S(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}) = {} & f^{(1)}(y_{ij}^{(1)};\alpha_1,\tau_1) f^{(2)}(y_{ij}^{(2)};\alpha_2,\tau_2) f^{(3)}(y_{ij}^{(3)};\alpha_3,\tau_3) \\ & \times \big(1 + \omega_{12}\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)}) + \omega_{13}\psi^{(1)}(y_{ij}^{(1)})\psi^{(3)}(y_{ij}^{(3)}) \\ & \quad + \omega_{23}\psi^{(2)}(y_{ij}^{(2)})\psi^{(3)}(y_{ij}^{(3)}) + \omega_{123}\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\psi^{(3)}(y_{ij}^{(3)})\big). \end{aligned}$$

Here is a simpler version of the formula, where we write $\psi^{(i)}(y_{ij}^{(i)})$ as $\psi^{(i)}$ :

$$\begin{aligned} f^S(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}) = {} & f^{(1)}(y_{ij}^{(1)};\alpha_1,\tau_1) f^{(2)}(y_{ij}^{(2)};\alpha_2,\tau_2) f^{(3)}(y_{ij}^{(3)};\alpha_3,\tau_3) \\ & \times (1 + \omega_{12}\psi^{(1)}\psi^{(2)} + \omega_{13}\psi^{(1)}\psi^{(3)} + \omega_{23}\psi^{(2)}\psi^{(3)} + \omega_{123}\psi^{(1)}\psi^{(2)}\psi^{(3)}). \end{aligned} \tag{4.8}$$

However, as proposed in Ratovomirija et al. (2017), it is often assumed that $\omega_{i_1,\ldots,i_n} = 0$ for $n \geq 3$, so (4.8) can be written as follows.

$$\begin{aligned} f^S(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}) = {} & f^{(1)}(y_{ij}^{(1)};\alpha_1,\tau_1) f^{(2)}(y_{ij}^{(2)};\alpha_2,\tau_2) f^{(3)}(y_{ij}^{(3)};\alpha_3,\tau_3) \\ & \times (1 + \omega_{12}\psi^{(1)}\psi^{(2)} + \omega_{13}\psi^{(1)}\psi^{(3)} + \omega_{23}\psi^{(2)}\psi^{(3)}). \end{aligned} \tag{4.9}$$

Its mixing function $\psi^{(\ell)}(y_{ij}^{(\ell)})$ is the same as the bivariate case, so for gamma distribution, the mixing function can be written as follows.

$$\psi^{(\ell)}(y_{ij}^{(\ell)}) = exp(-y_{ij}^{(\ell)}) - (1+\tau_{ij}^{(\ell)})^{-\alpha^{(\ell)}}, \quad \ell = 1,2,3.$$

The four variables $\omega_{12}, \omega_{13}, \omega_{23}$, and $\omega_{123}$ in (4.8) should be a real number, which requires the condition

$$1 + \omega_{12}\psi^{(1)}\psi^{(2)} + \omega_{13}\psi^{(1)}\psi^{(3)} + \omega_{23}\psi^{(2)}\psi^{(3)} + \omega_{123}\psi^{(1)}\psi^{(2)}\psi^{(3)} \geq 0, \tag{4.10}$$

for all $y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}$. After setting $\omega_{123} = 0$, we have the following condition.

$$1 + \omega_{12}\psi^{(1)}\psi^{(2)} + \omega_{13}\psi^{(1)}\psi^{(3)} + \omega_{23}\psi^{(2)}\psi^{(3)} \geq 0. \tag{4.11}$$

Bolancé and Vernic (2017) showed that the conditions in bivariate case still need to be applied, so we add the following restrictions for trivariate distribution.

$$1 + \omega_{cd}\psi^{(c)}(y_{ij}^{(c)})\psi^{(d)}(y_{ij}^{(d)}) \geq 0, \quad 1 \leq c < d \leq 3. \tag{4.12}$$

Similarly, if the lines of business follow gamma distribution, the bound of the correlation coefficient of each $\omega_{cd}$, $1 \leq c < d \leq 3$ need to be considered. We should also apply the following condition.

$$-\frac{1}{\sqrt{\alpha_c}\tau_c(1+\tau_c)^{-\alpha_c-1}\sqrt{\alpha_d}\tau_d(1+\tau_d)^{-\alpha_d-1}} \leq \omega \leq \frac{1}{\sqrt{\alpha_c}\tau_c(1+\tau_c)^{-\alpha_c-1}\sqrt{\alpha_d}\tau_d(1+\tau_d)^{-\alpha_d-1}} \tag{4.13}$$

for $1 \leq j < k \leq 3$.

### 4.1.3 One-stage inference

We use one-step inference method for the estimation, which estimates the marginals and the $\omega$ simultaneously using maximum likelihood estimation.

The loglikelihood of the bivariate Sarmanov distribution when using loss ratio as $y_{ij}^{(1)}, y_{ij}^{(2)}$ is given below.

$$\ell = \sum_{i=1}^{n}\sum_{j=1}^{n+1-i} log\ f^{(1)}(y_{ij}^{(1)}, \alpha_1, \tau_1)f^{(2)}(y_{ij}^{(2)}, \alpha_2, \tau_2) + \sum_{i=1}^{n}\sum_{j=1}^{n+1-i} log\ h(y_{ij}^{(1)}, y_{ij}^{(2)}, \omega), \tag{4.14}$$

where $h(y_{ij}^{(1)}, y_{ij}^{(2)}, \omega) = 1 + \omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})$ is the density of the Sarmanov distribution.

Similarly, we can also calculate the loglikelihood of the trivariate Sarmanov distribution when using loss ratio $y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}$ by the formula below. Here $\vec{\omega}$ includes $\{\omega_{12}, \omega_{13}, \omega_{23}\}$.

$$\ell = \sum_{i=1}^{n}\sum_{j=1}^{n+1-i} log\ f^{(1)}(y_{ij}^{(1)}, \alpha_1, \tau_1)f^{(2)}(y_{ij}^{(2)}, \alpha_2, \tau_2)f^{(3)}(y_{ij}^{(3)}, \alpha_3, \tau_3) + \sum_{i=1}^{n}\sum_{j=1}^{n+1-i} log\ h(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}, \vec{\omega}), \tag{4.15}$$

where

$$h(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}, \vec{\omega}) = 1+\omega_{12}\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})+\omega_{13}\psi^{(1)}(y_{ij}^{(1)})\psi^{(3)}(y_{ij}^{(3)})+\omega_{23}\psi^{(2)}(y_{ij}^{(2)})\psi^{(3)}(y_{ij}^{(3)})$$

is the density of the Sarmanov distribution.
We optimize the loglikelihood function to estimate $\vec{\omega}$, $\alpha_i$, and $\tau_i$ $(i = 1, 2, 3)$ in the Sarmanov distribution.

Here we can use this one-stage estimation method to estimate the dependence parameters $\omega$ for the bivariate Sarmanov model for the Personal and Commercial Auto lines from Shi and Frees data (2011). The results are shown in Table 4.1.

Table 4.1: Estimated omega for bivariate Sarmanov model with Personal and Commercial lines using one-step inference method

| Lines | Estimated omega | Loglikelihood | Standard error |
|---|---|---|---|
| Personal and Commercial | -0.0000837 | 346.5932 | 0.6403 |

As shown in Table 4.1, the estimated omega is smaller than the standard error, which means the estimated omega is unlikely to be significant, but we still need to test whether the Sarmanov model is better than the independent model. Apart from the AIC methods we mentioned above, there are some other inference methods we used in this thesis to check the significance of the dependence parameter.

**Bayesian information criterion (BIC)**

The Bayesian information criterion is one of the methods for model selection. BIC has larger penalty term than AIC, and it cannot deal with overfitting.

Let $k$ be the number of estimated parameters in the statistic model, $n$ be the number of data points in the model, and $\hat{L}$ be the estimated likelihood. Then, the BIC can be expressed as follows.

$$BIC = k \times ln(n) - 2ln(\hat{L}).$$

The lower BIC value gives the better model.

**Likelihood-ratio test**

The likelihood-ratio test, also known as likelihood-ratio chi-squared test, evaluates the goodness of fit between two nested statistical models using the ratio of their likelihood, where the nested model means that one of the models is the special case of the other.

The null hypothesis is that the model with fewer parameters is the better model, and the alternative hypothesis is that the full model is a good fit. The test statistic is expressed as follows.

$$LRT = -2ln\frac{L_s(\hat{\theta})}{L_g(\hat{\theta})},$$

where $L_s(\hat{\theta})$ is the likelihood of the model with fewer parameter and $L_g(\hat{\theta})$ is the likelihood of model with more parameters. The likelihood ratio can also be shown as a difference between the log-likelihoods.

$$LRT = -2\Big[\ell_s(\hat{\theta} - \ell_g(\hat{\theta}))\Big].$$

We set the significance level and the difference in the degree of freedom between the two models, check using the chi-square table and compare the result with our calculated test statistics. If the test statistic is larger, then we reject the null hypothesis. We also can calculate the p-value and compare it with the significance level $\alpha$. If the p-value is smaller than $\alpha$, then we reject the null hypothesis and say the model with more parameters has a significant improvement over the simpler model.

The likelihood-ratio test can also be used to determine the significance of the parameter. For example, for the trivariate model, in (4.9), we know that there are three parameters $\omega_{12}, \omega_{13}, \omega_{23}$ in the model. We can test the significance of a certain parameter by setting a new model with the certain parameter equal to 0 and compare with the original full model. If the p-value is small, we would conclude the parameter is significant for the model. If the p-value is large, then the parameter is not significant, and we can consider it as 0 in the model.

**Wald Test**

The Wald test is also one of the hypothesis tests used to determine whether the estimated parameters in a model are significant. Unlike the likelihood-ratio test, it only requires the estimation of the model and has a shorter computational time. The Wald statistic can be written as follows.

$$W^2 = \frac{(\hat{\beta} - \beta_0)^2}{Var(\hat{\beta})} \sim \chi_1^2,$$

where $\hat{\beta}$ is the maximum-likelihood estimation of the parameter, and $\beta_0$ is usually set to 0 because we want to test whether the parameter is significant or not. If the p-value of the Wald test is small or equal 0, we can reject the null hypothesis and say the parameter is significant.

We can use these statistical inference tests to determine whether bivariate Sarmanov distribution with one-stage estimation improves the independent model of personal and commercial auto lines.

Table 4.2: AIC and BIC for bivariate Sarmanov model with Personal and Commercial lines using one-step inference method

| Model | AIC | BIC |
|---|---|---|
| Independent | -613.1788 | -532.8931 |
| Bivariate Sarmanov with one-step inference | -611.1864 | -500.5932 |

From the AIC and BIC result in Table 4.2, we can see that the bivariate Sarmanov model using one-step inference method is not better than the independent case. We can also use likelihood-ratio test to check whether it is useful to add the dependence parameter in this model, and use Wald test to check whether the dependence parameter $\omega$ is significant:

Table 4.3: Significant tests for bivariate Sarmanov model with Personal and Commercial lines using one-step inference method

| Significant tests | likelihood-ratio test | Wald test |
|---|---|---|
| Test statistic | $7.1009 \cdot 1e - 06$ | $1.1 \cdot 1e - 08$ |
| p-value | 0.9979 | 1.0 |

In Table 4.3, we see that the test statistic for both test is small with large p-values, which indicates that we cannot reject the null hypothesis in both cases, meaning the independent model is better than the bivariate Sarmanov model using loss ratio.

However we have other sets of data, we can check if the Sarmanov model with one-step inference method captures the dependence and improves other independent models. Table 4.4 gives the dependence paarameter $\omega$ estimation for line 2 & 4 from Côté et al. (2016) data.

| Lines | Estimated omega | Loglikelihood |
|---|---|---|
| 2&4 | 436.9040 | 315.1206 |

Table 4.4: Estimated omega for bivariate Sarmanov model with Line 2 & 4 using one-step inference method

In this case, the standard error of $\hat{\omega}$ is not computable, therefore we cannot use the Wald test to check the significance. We can still use AIC, BIC and likelihood-ratio test to see if the bivariate Sarmanov model with one-step inference method is better than the independent model.

Table 4.5: AIC and BIC for bivariate Sarmanov model with line 2&4 using one-step inference method

| Model for line 2&4 | AIC | BIC |
|---|---|---|
| Independent | -546.3281 | -438.3089 |
| Bivariate Sarmanov with one-step inference | -548.2413 | -437.5216 |

We see that as BIC has larger penalty term than AIC. It shows that the bivariate Sarmanov model with one-inference method is not much better than the independent model, while AIC shows it provides better fit than the independent case. We use the likelihood-ratio test to check whether it is a better model.

Table 4.6: Significant tests for bivariate Sarmanov model with line 2&4 using one-step inference method

| Significant tests | likelihood-ratio test |
|---|---|
| Test statistic | 3.91314 |
| p-value | 0.04791 |

Table 4.6 shows the null hypothesis of independence is rejected at the 5% level. We conclude that the bivariate Sarmanov model with one-step inference provides a better fit than the independent model for line 2&4.

For the trivariate case, we use line 2, 4 and 5 from the *Côté* et al.(2016) data. We need to estimate the three $\omega$'s, $\omega_{12}, \omega_{13}, \omega_{23}$ in (4.9) using the one-step inference method. We use the maximum log-likelihood estimation where the log-likelihood function is given in (4.15).

We first use Wald test to check if the three parameters are significant.

Table 4.7: Estimated omega for trivariate Sarmanov model with Line 2 & 4 & 5 using one-step inference method

| Lines 2&4&5 | $\omega_{24}$ | $\omega_{25}$ | $\omega_{45}$ |
|---|---|---|---|
| Estimated omega | 374.7942 | -110.3272 | -165.7813 |
| Log-likelihood | 556.4291 | | |

Table 4.8: Significant tests for trivariate Sarmanov model with line 2&4&5 using one-step inference method

| likelihood-ratio test | $\omega_{24}$ | $\omega_{25}$ | $\omega_{45}$ |
|---|---|---|---|
| Test statistic | 2.2803 | -0.1351384 | 1.061 |
| p-value | 0.1310 | 1.00 | 0.3030 |

Table 4.8 shows that all three parameters are not significant. Next, we compare the AIC and BIC to see if the trivariate Sarmanov model using one-step inference is better than the independent model.

Table 4.9: AIC and BIC for trivariate Sarmanov model with line 2&4&5 using one-step inference method

| Model for line 2&4&5 | AIC | BIC |
|---|---|---|
| Independent | -1026.6996 | -837.2370 |
| Trivariate Sarmanov with one-step inference | -986.8582 | -791.1837 |

Result from Table 4.9 shows that the trivariate Sarmanov model using one-step inference method is not better than the independent model for line 2, 4 and 5, i.e. it does not provides better fit in capturing the dependence.

After obtaining the estimated parameters, we use them to calculate reserve as follows.

$$\sum_{\ell}\sum_{i}\sum_{j} p_i^{(\ell)} E[y_{ij}^{(\ell)}]$$

which is mentioned in Section 2.3. For the one-step inference method, we get the estimated reserves for the bivariate and trivariate cases.

As the dependence parameter of bivariate Sarmanov for Personal Commercial Line and trivariate Sarmanov for line 2, 4 and 5 are not significant, the reserve is close to the independent reserve. Once the dependence becomes significant, such as line 2&4 bivariate Sarmanov, the total reserve differs more from the reserve obtained in the independent case.

Table 4.10: Reserve calculation of one-method inference vs. independent

| Models | Reserve 1st line | Reserve for 2nd line | Reserve for 3rd line | Total Reserve |
|---|---|---|---|---|
| Independent P&C | 6,464,075 | 490,652 | - | 6,954,727 |
| Bivariate P&C | 6,464,318 | 490,702 | - | 6,955,020 |
| Independent 2&4&5 | 132,919 | 73,220 | 18,288 | 224,426 |
| Bivariate 2&4 | 129,397 | 71,457 | - | 219,144 |
| Trivariate 2&4&5 | 135,061 | 70,857 | 18,752.67 | 224,671 |

## 4.2 Two-stage rank-based inference for the Dependence Structure

Rank-based methods replace the actual value with the ranks of observation in the dependence structure. This method is often used when the distribution of data is not normal or unknown, or the data have outliers that may affect the results. It is more robust for the non-normal distributions, provides more accurate and reliable results in such circumstances.

Rank-based methods do not need to re-estimate the marginals, which is required for the one-step inference method. Re-estimating the marginals may cause big effect to the loss reserves. First of all, it might cause violation for linear property of the mean. As mentioned above, with the estimated parameters, the total reserve can be estimated using

$$\sum_{\ell}\sum_{i}\sum_{j} p_i^{(\ell)} E[y_{ij}^{(\ell)}].$$

But re-estimating the marginals could cause the parameters to deviate from the original parameters, then the new $E[\sum_\ell \sum_i \sum_j y_{ij}^{(\ell)}]$ produced using dependence model will not be equal to the original $\sum_\ell \sum_i \sum_j E[y_{ij}^{(\ell)}]$. This violates the linear property of the mean. As shown in Table 4.10, the estimation of the reserve changed for the one-step inference method, but it stays the same if we use rank-based method.

Also, while using the dependence model with distribution, if we re-estimate the marginals not knowing if we chose the correct distribution or correct dependence model, it will cause much bigger error, and hard to check whether the error comes from the incorrect dependence structure or the marginal distribution.

Rank-based method avoids re-estimating the parameters, separate the marginals from the dependence structure and directly estimate the dependence parameter. This method is more robust than the one-step inference method.

In loss reserving, rank-based method involves using the rank of residuals to perform statistical inference or create statistical models, rather than using the loss data, such as loss ratio.

### 4.2.1 Bivariate and Trivariate distribution

For rank-based method, we use rank of residuals $R_{ij}$ instead of loss ratio $y_{ij}$ in the dependence structure, to separate it from the marginals. The residual for each observation is the difference between predicted values of dependent variable and observed values of it. If we use the data from Shi and Frees (2011) with log-normal and gamma distribution, the bivariate Sarmanov distribution will be written as

$$f^S(y_{ij}^{(1)}, y_{ij}^{(2)}) = f^{(1)}\big(y_{ij}^{(1)}; a_1, b_1\big) f^{(2)}\big(y_{ij}^{(2)}; \alpha_2, \tau_2\big)\big(1 + \omega\psi^{(1)}(R_{ij}^{(1)})\psi^{(2)}(R_{ij}^{(2)})\big), \tag{4.16}$$

where the loss ratio $y_{ij}$ in the mixing function is changed to the rank of residuals $R_{ij}$ as below.

$$\psi^{(1)}(R_{ij}^{(1)}) = exp(-R_{ij}^{(1)}) - exp\Big(-a_1 + \frac{b_1^2}{2}\Big),$$
$$\psi^{(2)}(R_{ij}^{(2)}) = exp(-R_{ij}^{(2)}) - (1+\tau_2)^{-\alpha_2}.$$

Therefore the bound of the parameter $\omega$ will become:

$$1 + \omega\psi^{(1)}(R_{ij}^{(1)})\psi^{(2)}(R_{ij}^{(2)}) \geq 0 \tag{4.17}$$

The lower and upper bound given in (4.6) and (4.7) still remains the same for rank-based method.

For the trivariate Sarmanov distribution, the distribution function using rank-based method follows.

$$\begin{aligned} f^S(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)}) = \; & f^{(1)}(y_{ij}^{(1)}; \alpha_1, \tau_1) f^{(2)}(y_{ij}^{(2)}; \alpha_2, \tau_2) f^{(3)}(y_{ij}^{(3)}; \alpha_3, \tau_3) \\ & * \big(1 + \omega_{12}\psi^{(1)}(R_{ij}^{(1)})\psi^{(2)}(R_{ij}^{(2)}) + \omega_{13}\psi^{(1)}(R_{ij}^{(1)})\psi^{(3)}(R_{ij}^{(3)}) \\ & \quad + \omega_{23}\psi^{(2)}(R_{ij}^{(2)})\psi^{(3)}(R_{ij}^{(3)})\big), \end{aligned}$$

where we have the mixing function

$$\psi^{(\ell)} = \psi^{(\ell)}(R_{ij}^{(\ell)}) = exp(-R_{ij}^{(\ell)}) - (1+\tau_{ij}^{(\ell)})^{-\alpha^{(\ell)}}, \quad \ell = 1,2,3. \tag{4.18}$$

Also, the bound of $\omega$ should include (4.13) for each $\omega_{cd}$, $1 \le c < d \le 3$, and the following constraint need to be satisfied.

$$1 + \omega_{12}\psi^{(1)}(R_{ij}^{(1)})\psi^{(2)}(R_{ij}^{(2)}) + \omega_{13}\psi^{(1)}(R_{ij}^{(1)})\psi^{(3)}(R_{ij}^{(3)}) + \omega_{23}\psi^{(2)}(R_{ij}^{(2)})\psi^{(3)}(R_{ij}^{(3)}) >= 0,$$

and

$$1 + \omega_{cd}\psi^{(c)}(R_{ij}^{(c)})\psi^{(d)}(R_{ij}^{(d)}) \ge 0, \qquad 1 \le c < d \le 3. \tag{4.19}$$

### 4.2.2 Rank-based Estimation method (Two-stage inference)

When using Rank-based method, two-step inference method is used. Here we assume line 1 follows normal distribution and line 2 follows gamma distribution. We first estimate the parameters of the marginals $a_1, b_1, \alpha_2, \tau_2$, and use the estimated marginals to calculate the rank of residuals. Then $\omega$ can be estimated. We use maximum likelihood estimation method in both stages.

The loglikelihood of the marginals of the bivariate Sarmanov distribution is written as below, where $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$ are the loss ratios of line of business (1) and (2).

$$\ell_{marginals} = \sum_{i=1}^{n} \sum_{j=1}^{n+1-i} log\ f^{(1)}(y_{ij}^{(1)}, a_1, b_1) f^{(2)}(y_{ij}^{(2)}, \alpha_2, \tau_2) \tag{4.20}$$

Then we calculate the residuals for line $\ell = 1, 2$ as follows.

$$r_{ij}^{(1)} = \frac{ln(y_{ij}^{(1)}) - a_{ij}^{(1)}}{b^{(1)}},$$

$$r_{ij}^{(2)} = \frac{y_{ij}^{(2)}}{\tau_{ij}^{(2)}}.$$

From the residuals, the rank of residuals is obtained as follows.

$$R_{ij}^{(\ell)} = \frac{1}{55+1} \sum_{i*=1}^{10} \sum_{j*=1}^{11-i*} \vec{1}(r_{i*j*}^{(\ell)} \le r_{ij}^{(\ell)}). \tag{4.21}$$

Here $\vec{1}(A)$ is the indicator function.

Then, we optimize the pseudo-likelihood with new obtained Rank of residuals, where the pseudo-likelihood is given as below.

$$\ell = \sum \sum log\ h(R_{ij}^{(1)}, R_{ij}^{(2)}, \omega). \tag{4.22}$$

(4.22) gives the loglikelihood of Sarmanov distribution, where

$$h(R_{ij}^{(1)}, R_{ij}^{(2)}, \omega) = f^{(1)}(y_{ij}^{(1)}, a_1, b_1) f^{(2)}(y_{ij}^{(2)}, \alpha_2, \tau_2)(1 + \omega \psi^{(1)}(R_{ij}^{(1)}) \psi^{(2)}(R_{ij}^{(2)})).$$

But in this case we do not re-estimated the marginals. The mixing (4.4) and (4.2) are used. The estimated $\omega$ can be obtained by optimizing the loglikelihood function.

Similarly, for the trivariate case, if we have all three lines of business follow gamma distribution, then we estimate the parameters using maximum loglikelihood.

$$\ell_{marginals} = \sum_{i=1}^{n} \sum_{j=1}^{n+1-i} log\ f^{(1)}(y_{ij}^{(1)}, \alpha_1, \tau_1) f^{(2)}(y_{ij}^{(2)}, \alpha_2, \tau_2) f^{(3)}(y_{ij}^{(3)}, \alpha_3, \tau_3). \tag{4.23}$$

Then, we calculate the rank of residual from the estimated parameters, optimize the pseudo-likelihood of trivariate Sarmanov distribution and obtain the estimation of $\omega_{12}, \omega_{13}, \omega_{23}$:

$$\ell = \sum \sum log\ h(R_{ij}^{(1)}, R_{ij}^{(2)}, R_{ij}^{(3)}, \omega_{12}, \omega_{13}, \omega_{23}). \tag{4.24}$$

For the rank-based method, we first use Kendall's tau test to check the dependence between the residuals of two lines of business.

**Kendall's $\tau$ Test**

Kendall's $\tau$ coefficient is used to measure the dependence between two sets of data. Kendall's $\tau$ test is a non-parametric hypothesis test for the dependence based on the $\tau$ coefficient. As shown in Genest and Neschléhova (2011) and summarized in Côté et al. (2016), the formula used to calculate Kendall's $\tau$ for multiple sets of data, such as residuals of multiple lines of business is given as below.

$$\tau_{d,n} = \frac{1}{2^{d-1} - 1} \left[ -1 + \frac{2^d}{n(n-1)} \sum_{(i,j) \neq (i*,j*)} 1 \Big( r_{i*j*}^{(1)} \leq r_{ij}^{(1)}, ..., r_{i*j*}^{(d)} \leq r_{ij}^{(d)} \Big) \right], \tag{4.25}$$

where $d$ is the number of sets of data and $n$ is the number of data in each set. The variance of $\tau$ given $d$ and $n$ can be calculated by:

$$Var(\tau_{d,n}) = \frac{n(2^{2d+1} + 2^{d+1} - 4 * 3^d) + 3^d(2^d + 6) - 2^{d+2}(2^d + 1)}{3^d(2^{d-1} - 1)^2 n(n-1)},$$

as shown in Section 2.2 of Côté et al.(2016). As the Kendall's test use chi-square test to determine the p-value, we calculate the p-value of kendall's test by:

$$p = 2 * \left(1 - cdf_{normal}\left(\left|\tau_{d,n}/\sqrt{Var(\tau_{d,n})}\right|\right)\right).$$

Here, we first check the dependence between the residuals of personal and commercial auto line from Shi and Frees (2011) data.

Table 4.11: Kendall tau for Personal and Commercial lines

| LOB | Personal&Commercial Auto Line |
|---|---|
| Kendall tau | -0.1556 |
| Kendall test p-value | 0.09355 |

Based on the p-value of the Kendall's test given in Table 4.11, we conclude that the null hypothesis of independence is rejected at the 10% level. Therefore, we can say that there exists a significant but small dependency between the two lines of business. However, as the Kendall's $\tau$ statistic in Table 4.11 is negative, which indicates a negative association between the two lines of business, we need to use the negative of rank of residuals for the second line of business when estimating $\omega$. Thus, we optimize the following pseudo-likelihood in this case:

$$\ell = \sum\sum log\ h(R_{ij}^{(1)}, -R_{ij}^{(2)}, \omega).$$

This allows us to obtain the estimated $\omega$ in Table 4.12.

Table 4.12: Estimated omega for bivariate Sarmanov model with Personal and Commercial lines using rank-based method

| Lines | Estimated omega | Pseudo-likelihood | Standard error |
|---|---|---|---|
| Personal and Commercial | -10.14954 | 609.7023 | 1.3985 |

For the rank-based method, we maximize the pseudo-likelihood instead of log-likelihood function while estimating $\omega$. We cannot use the AIC, BIC and likelihood-ratio test as we do not use the same data as the independent case and we do not re-estimate the marginals.

But we can still use Wald Test to check the significance of the dependence parameter $\omega$ in this case.

Table 4.13: Wald Test for bivariate Sarmanov model with Personal and Commercial lines using rank-based method

| Significant tests | Wald test |
|---|---|
| Test statistic | 73.7 |
| p-value | 0.0 |

From Table 4.13, we see that Wald test shows the estimated $\omega$ is significant.

Bootstrap method can also be used to check whether a parameter is significant as pointed out in Côté et al. (2016). If we simulate and estimate the parameter 5,000 times, then we can check if the 95% confidence interval of the 5,000 estimation includes 0. If it does not include 0, then the estimated parameter is significant.

We can also use the bootstrapping method to check whether the dependence parameter $\omega$ is significant. We simulate the loss ratio using bivariate Sarmanov with $\omega$ estimated using rank-based method for 5,000 times, and estimate the new dependence parameter $\omega^*$ each time. The simulation and bootstrapping procedure will be illustrate throughly in the next chapter.

Figure 4.1 shows the distribution of the 5,000 $\omega$'s, the blue line gives the 95% confidence interval and we can see that the confidence interval does not include 0. This indicates that the estimation of $\omega$ is significant in the bivariate Sarmanov model using rank-based method for personal and commercial auto line.

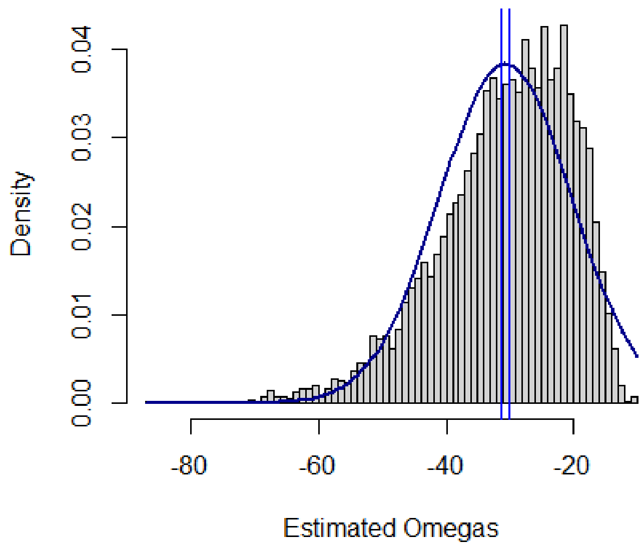
We can also work the similar procedure with line 2, 4 and 5 from the Côté et al.(2016) data. Table 4.14 gives the Kendall's $\tau$ test between all three lines of business.

Table 4.14: Kendall tau for line 2&4&5

| LOB | Line 2&4 | Line 2&5 | Line 4&5 | Line 2&4&5 |
|---|---|---|---|---|
| Kendall tau | 0.2444 | 0.2094 | 0.2000 | 0.2180 |
| Kendall test p-value | 0.0084 | 0.0240 | 0.0311 | 4.7064e-05 |

Table 4.14 shows that the three lines of business are positively correlated. The fact the three lines of business are positively correlated is due in part to exogenous common factors such

Figure 4.1: 5,000 $\omega^*$ estimations using bootstrap for bivariate Sarmanov Personal & Commercial lines with rank-based method




as inflation and interest rates. Furthermore, strategic decisions can impact several portfolios, e.g., the acceleration of payments on all lines of the insurance sector could induce some positive dependence. At a granular level, the positive association between Ontario AB and BI can be explained by the fact that the same accident will often arise in both coverage.

We will still take line 2 & 4 as another demonstration for bivariate case and then talk about the trivariate case.

In this bivariate case, we use (4.20) and (4.21) in the gamma-gamma version to compute the rank of residuals which we plug in (4.22) to estimate the omega. Table 4.15 gives the result of $\omega$ estimation using rank-based method for line 2 & 4.

Table 4.15: Estimated omega for bivariate Sarmanov model with line 2&4 using rank-based method

| Lines | Estimated omega | Pseudo-likelihood | Standard error |
|---|---|---|---|
| 2&4 | 24.5244 | 369.2047 | 0.7632144 |

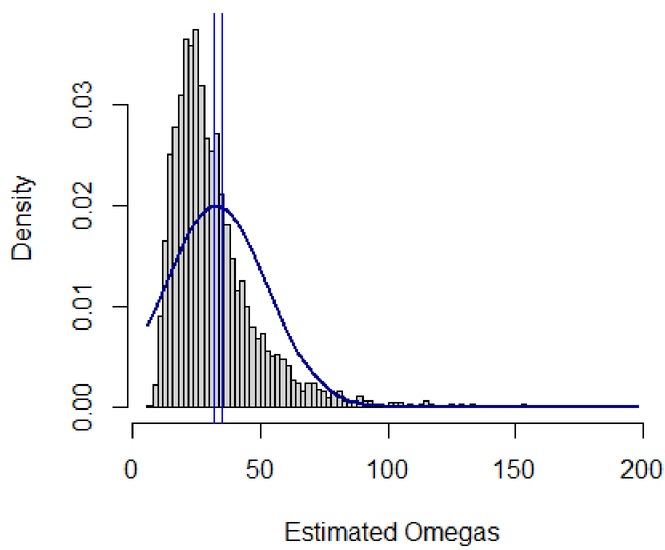
We can then use Wald test to check the significance of the dependence parameter $\omega$.

Table 4.16: Wald Test for bivariate Sarmanov model with line 2&4 using rank-based method

| Significant tests | Wald test |
|---|---|
| Test statistic | 788.0 |
| p-value | 0.0 |

The Wald test result in Table 4.16 shows strong significance of omega. We can also use the bootstrapping method to check the significance of $\omega$. Similarly, we simulate the loss data using bivariate Sarmanov with $\omega$ estimated using rank-based method for 5,000 times, and estimate the new $\omega^*$ each time.

Figure 4.2: 5,000 $\omega^*$ estimations using bootstrap for bivariate Sarmanov line 2&4 with rank-based method




In figure 4.2, the blue lines give the 95% confidence interval, and we can see from the figure that it does not include 0. This means that the estimation of $\omega$ is significant in the bivariate Sarmanov model using rank-based method for line 2&4.

For the trivariate case, we estimate the $\omega_{12}, \omega_{13}, \omega_{23}$ from (4.24) after calculating the rank of residuals using (4.23) and (4.21). Table 4.17 gives the result of estimated omegas.

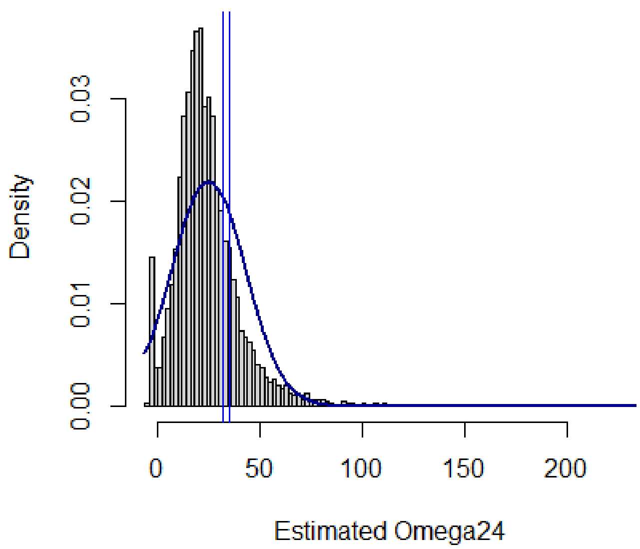
As shown in Table 4.14, Kendall tau test shows the three lines of business are dependent with each other, we will also use bootstrapping method directly to check the significance of

Table 4.17: Estimated omega for trivariate Sarmanov model with Line 2 & 4 & 5 using rank-based method

| Lines 2&4&5 | $\omega_{24}$ | $\omega_{25}$ | $\omega_{45}$ |
|---|---|---|---|
| Estimated omega | 25.2962 | 30.4092 | 61.4528 |
| Pseudo-likelihood | 678.9434 | | |

the three dependence parameters.

Figure 4.3: 5,000 $\omega^*_{24}$ estimations using bootstrap for trivariate Sarmanov line 2&4&5 with rank-based method




From the bootstrap result given in Figure 4.3, 4.4, 4.5, we can conclude that the dependence parameters are all significant for the trivariate Sarmanov distribution using rank-based method.

Figure 4.4: 5,000 $\omega_{25}^{*}$ estimations using bootstrap for trivariate Sarmanov line 2&4&5 with rank-based method

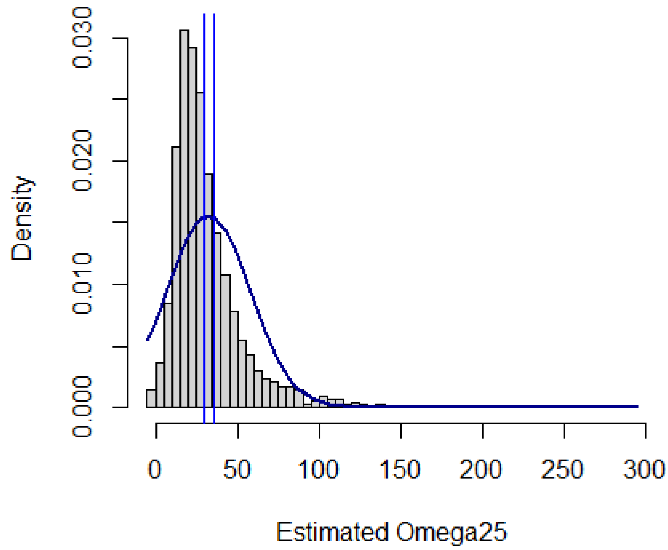
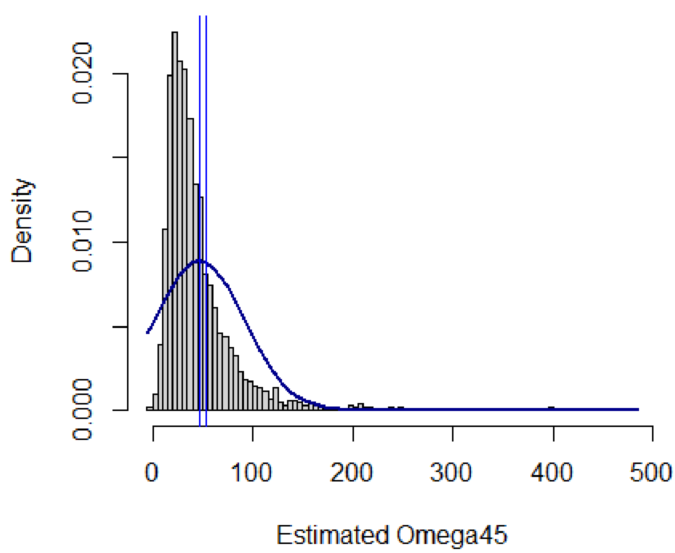



Figure 4.5: 5,000 $\omega_{45}^{*}$ estimations using bootstrap for trivariate Sarmanov line 2&4&5 with rank-based method



# Chapter 5

# Risk Capital

In addition to reserves, companies also set aside some amount of fund as a buffer, in case of potential losses caused by extreme events, this is the risk capital. It represents the amount of money that the company can lose without causing significant harm for the financial situation. If a dependence model provides lower risk capital, then it means using this model produces lower risks, i.e. the better model will have lower risk capital. In order to measure risk capital, we use two numerical measurement, value at risk (VaR) and tail value at risk (TVaR).

The $VaR_k$ is calculated as the 100(1-$k$) percentile of the loss distribution, where $k \in (0, 1)$ is the risk tolerance. The risk of potential losses can be quantified by this statistic.

The $TVaR$ is also known as tail conditional expectation, which calculates the expectation of potential loss when an event outside of certain probability occurs. In order to calculate the tail value at risk, we have:

$$TVaR_k(X) = E[X|X > VaR_k(X)].$$

Capital allocation is the share of the risk capital to be allocated to each line of business, which was first introduced in Tasche (1999), and also summarized in Bargès et al. (2009).

After $n$ simulations, we get $n$ sets of simulated data, then given

$$y^{(\ell)} = \sum_i \sum_j p_i^{(\ell)} y_{ij}^{(\ell)}$$

is the unpaid loss for $\ell$th line of business, $S = \sum_\ell y^{(\ell)}$ is the total unpaid loss, TVaR- based capital allocation can be written as

$$
\begin{aligned}
&TVaR_k(y^{(\ell)}; S) \\
&= \frac{1}{n(1-k)} \left[ \sum_{j=1}^{n} y_j^{(\ell)} 1(S_j > VaR_k(X)) + \frac{F_n(VaR_k(X)) - k}{\frac{1}{n}\sum_{i=1}^{n} 1(S_i = VaR_k(X))} \sum_{j=1}^{n} y_j^{(\ell)} 1(S_j = VaR_k(X)) \right],
\end{aligned}
$$

where $F_n$ is the empirical cumulative distribution function of $S$.

In order to calculate the risk capital of independent case, we obtain the risk capital separately, "Silo" method introduced in Ajne (1994) is put to use:

- Calculate the risk measure $VaR^{(silo)}, TVaR^{(silo)}$ for each line of business.
- Obtain the sum: $\sum_i VaR^{(i)} = VaR^{(silo)}$ where $i \in \{1, ..., n\}$.
- Then obtain the Risk Capital using $RC^{(i)} = TVaR^{(i)}_{99\%} - TVaR^{(i)}_{60\%}$, where $i = 1, ..., n$.
- Obtain the sum of the risk capitals:$RC^{silo} = \sum_i RC^{(i)}$, where $i \in \{1, ..., n\}$.

For Sarmanov method, we obtain the risk capital simultaneously:

- Calculate the risk measure $VaR^{(Sarmanov)}, TVaR^{(Sarmanov)}$ for the dependent model.
- Obtain the risk capital: $RC^{(Sarmanov)} = TVaR^{(Sarmanov)}_{99\%} - TVaR^{(Sarmanov)}_{60\%}$.

If we have $RC^{silo} - RC^{Sarmanov} > 0$, then this means the risk capital of the dependent case is smaller than the independent case, so using Sarmanov distribution decreases the risk capital of the lines of business and increases diversification benefit.

## 5.1 Simulation procedure

The simulation procedures are the same for both one-step inference method and rank-based method. We only use the rank-based method to estimate the dependence parameter $\omega$.

To geneerate realizaions from the multivariate Sarmanov distrubtion, we use the inversion method, based on the conditional cumulative distrbution function, as described in Pelican and Vernic (2013).

### 5.1.1 Bivariate Sarmanov simulation (one-step inference and rank-based method)

We generate the bivariate Sarmanov distribution using the conditional simulation method which has the following steps:

- Generate a set of observed values $y_{ij}^{(1)}$ from a random variable that follows Gamma distribution $y_{ij}^{(1)} \sim Gamma(\alpha_1, \tau_1)$.
- Calculate the cumulative distribution function of the conditional distribution $F_{CDF}(y_{ij}^{(2)}|y_{ij}^{(1)})$

  The density function of the conditional distribution can be written as:

$$\begin{aligned} f(y_{ij}^{(2)}|y_{ij}^{(1)}) &= \frac{f^{(1)}\big(y_{ij}^{(1)};\alpha_1,\tau_1\big)f^{(2)}\big(y_{ij}^{(2)};\alpha_2,\tau_2\big)\big(1+\omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\big)}{f^{(1)}\big(y_{ij}^{(1)};\alpha_1,\tau_1\big)} \\ &= f^{(2)}\big(y_{ij}^{(2)};\alpha_2,\tau_2\big)\big(1+\omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\big) \\ &= f^{(2)}\big(y_{ij}^{(2)};\alpha_2,\tau_2\big)+\omega f^{(2)}\big(y_{ij}^{(2)};\alpha_2,\tau_2\big)\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)}). \end{aligned}$$

  Therefore the cumulative distribution can be calculated as:

$$F(y_{ij}^{(2)}|y_{ij}^{(1)}) = F\big(y_{ij}^{(2)};\alpha_2,\tau_2\big)+\omega\psi^{(1)}(y_{ij}^{(1)})\int f^{(2)}\big(y_{ij}^{(2)};\alpha_2,\tau_2\big)\psi^{(2)}(y_{ij}^{(2)})\ dy_{ij}^{(2)}. \quad (5.1)$$

- Generate a set of observed values $y_{ij}^{(2)}$ from the conditional distribution of a random variable $(y_{ij}^{(2)}|y_{ij}^{(1)} = y_{ij}^{(1)})$.

### 5.1.2 Trivariate Sarmanov simulation (one-step inference and rank-based method)

For the texts below, $\psi^{(k)}(y_{ij}^{(k)})$ will be writen as $\psi_k$ for short, $f^{(k)}\big(y_{ij}^{(k)};\alpha_k,\tau_k\big)$ will be written as $f(y_{ij}^{(k)})$.

- Generate a set of observed values $y_{ij}^{(1)}$ from a random variable that follows Gamma distribution $y_{ij}^{(1)} \sim Gamma(\alpha_1, \tau_1)$.

- Calculate the cumulative distribution function of the conditional distribution

$$F_{CDF}(y_{ij}^{(2)}|y_{ij}^{(1)}) = F(y_{ij}^{(2)}) + \omega_{12}\psi_1 \int f(y_{ij}^{(2)})\psi_2 \; dy_{ij}^{(2)},$$

where $\omega_{12}$ is from the trivariate Sarmanov distribution between $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$, as the joint distribution $f(y_{ij}^{(1)}, y_{ij}^{(2)})$ is from the trivariate model.

- Generate a set of observed values $y_{ij}^{(2)}$ from the conditional distribution of a random variable $(y_{ij}^{(2)}|y_{ij}^{(1)} = y_{ij}^{(1)})$.
- Calculate the cumulative distribution function of the conditional distribution $F_{CDF}(y_{ij}^{(3)}|y_{ij}^{(1)}, y_{ij}^{(2)})$,

$$\begin{aligned}
f(y_{ij}^{(3)}|y_{ij}^{(1)}, y_{ij}^{(2)}) &= \frac{f(y_{ij}^{(1)}, y_{ij}^{(2)}, y_{ij}^{(3)})}{f(y_{ij}^{(1)}, y_{ij}^{(2)})} \\
&= \frac{f(y_{ij}^{(1)})f(y_{ij}^{(2)})f(y_{ij}^{(3)})(1 + \omega_{12}\psi_1\psi_2 + \omega_{13}\psi_1\psi_3 + \omega_{23}\psi_2\psi_3)}{f(y_{ij}^{(1)})f(y_{ij}^{(2)})(1 + \omega_{12}\psi_1\psi_2)} \\
&= \frac{f(y_{ij}^{(3)})(1 + \omega_{12}\psi_1\psi_2)}{1 + \omega_{12}\psi_1\psi_2} + \frac{f(y_{ij}^{(3)})\omega_{13}\psi_1\psi_3}{1 + \omega_{12}\psi_1\psi_2} + \frac{f(y_{ij}^{(3)})\omega_{23}\psi_2\psi_3}{1 + \omega_{12}\psi_1\psi_2} \\
&= f(y_{ij}^{(3)}) + \frac{f(y_{ij}^{(3)})\omega_{13}\psi_1\psi_3}{1 + \omega_{12}\psi_1\psi_2} + \frac{f(y_{ij}^{(3)})\omega_{23}\psi_2\psi_3}{1 + \omega_{12}\psi_1\psi_2}.
\end{aligned}$$

Therefore the cumulative distribution can be calculated as:

$$\begin{aligned}
F_{CDF}(y_{ij}^{(3)}|y_{ij}^{(1)}, y_{ij}^{(2)}) &= \int_{-\infty}^{y_{ij}^{(3)}} f(y_{ij}^{(3)}|y_{ij}^{(1)}, y_{ij}^{(2)}) \; dy_{ij}^{(3)} \\
&= F(y_{ij}^{(3)}) + \frac{\omega_{13}\psi_1 \int f(y_{ij}^{(3)})\psi_3 \; dy_{ij}^{(3)}}{1 + \omega_{12}\psi_1\psi_2} + \frac{\omega_{23}\psi_2 \int f(y_{ij}^{(3)})\psi_3 \; dy_{ij}^{(3)}}{1 + \omega_{12}\psi_1\psi_2}.
\end{aligned}$$

- Generate a set of observed values $y_{ij}^{(2)}$ from the conditional distribution of a random variable $(y_{ij}^{(3)}|y_{ij}^{(1)} = y_{ij}^{(1)}, y_{ij}^{(2)} = y_{ij}^{(2)})$.

The following are the steps of the simulation procedure for bivariate or multivariate case.

- The original loss ratio can be written in the following loss triangle:

$$
\begin{matrix}
y_{1,1}^{(1)} & \cdots & y_{1,10}^{(1)} & & y_{1,1}^{(2)} & \cdots & y_{1,10}^{(2)} & \\
\vdots & \iddots & & & \vdots & \iddots & & \cdots \\
y_{10,1}^{(1)} & & & & y_{10,1}^{(2)} & & &
\end{matrix}
$$

- Estimation of $\vec{\omega}$, $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$, $k \geq 2$ (or $a_1, b_1$ based on the distribution) where $\omega$ is from the Sarmanov distribution and $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ are the marginals for two lines of business. Bounds of $\vec{\omega}$ are based on estimated marginals $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$.

  For one-step inference method, we estimate $\vec{\omega}$, $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ simultaneously.

  For rank-based method, we use two-step inference method, which $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ are estimated first, then we estimate $\vec{\omega}$.

- Simulate the lower part (45 observations) of the triangle with the estimated parameters $\vec{\omega}, \alpha_1, \tau_1, ..., \alpha_k, \tau_k$ obatianed above.

$$
\begin{matrix}
 & & y_{2,10}^{(1)} & & & & y_{2,10}^{(2)} & \\
 & \iddots & \vdots & & & \iddots & \vdots & \cdots \\
y_{10,2}^{(1)} & \cdots & y_{10,10}^{(1)} & & y_{10,2}^{(2)} & \cdots & y_{10,10}^{(2)} &
\end{matrix}
$$

- Calculate the reserve from the simulated lower part of the triangle.

Now we use the simulation method to simulate the lower part of the loss triangle and compute the risk capital. For the Personal and commercial auto lines in Shi and Frees (2011), we calculate the TVaR99% and risk capital for using rank of residuals, comparing it with Silo method and other used Copula model in Shi and Frees (2011). Table 5.1 and 5.2 gives the TVaR 99% and risk capital after 50,000 simulations for the lower part of the loss triangle.

Table 5.1: 50,000 Simulations TVaR 99% comparison Personal Commercial

| Model | TVaR 99% |
|---|---|
| Silo Method | 7,594,465 |
| Sarmanov with one-step inference method | 7,526,434 |
| Sarmanov with rank-based method | **7,279,902** |
| Gausian Copula (From Shi & Frees (2011)) | 7,453,552 |

The comparison shows that bivariate Sarmanov model using rank of residuals produces lower risk measures than the silo method and Gaussian Copula model, which means it out-

Table 5.2: 50,000 Simulations risk capital comparison Personal Commercial

| Model | Risk Capital | Gain |
|---|---|---|
| Silo Method | 436,433 | - |
| Sarmanov with one-step inference method | 375,361 | 13.99% |
| Sarmanov with rank-based method | **367,510** | **15.79%** |

performed the other two method.

We then compare the Risk Capital for trivariate case line 2&4&5 and line 2&4 bivariate case for both one-step inference method and rank-based method.

Table 5.3: 50,000 Simulations Risk Capital comparison 245

| Model | Line 2 | Line 4 | Line 5 | Total | Gain |
|---|---|---|---|---|---|
| Silo | 16,163 | 11,301 | 2,455 | 29,920 | - |
| Biv 24 one-step inference | 13,474 | 5,972 | - | 21,902 | 26.80% |
| Biv 24 rank-based method | 13,549 | 5,820 | - | 21,825 | 27.06% |
| Triv 245 one-step inference method | 14,034 | 6,144 | 232 | 20,411 | 31.78% |
| Triv 245 rank-based method | 13,458 | 5,800 | 246 | **19,505** | **34.81%** |

Table 5.3 shows that the bivariate Sarmanov with rank-based method is better than the silo method and one-step inference method, the trivariate Sarmanov shows lower risks than the bivariate case, with low risk capital in total and higher gain.

## 5.2 The Bootstrap procedure

In order to calculate reserves and risk capital, we use bootstrapping method to generate sample data and estimate the parameters. We use the same bootrstrap algorithm as Taylor and McGuire (2007), which is also shown in Shi and Frees (2011) and Abdallah et al. (2016). The following are the steps included in the bootstrapping method for bivariate or multivariate case.

- The original loss ratio can be written in the following loss triangle:

$$\begin{array}{ccccccc} y_{1,1}^{(1)} & \cdots & y_{1,10}^{(1)} & & y_{1,1}^{(2)} & \cdots & y_{1,10}^{(2)} \\ \vdots & \iddots & & & \vdots & \iddots & & \cdots \\ y_{10,1}^{(1)} & & & & y_{10,1}^{(2)} & & \end{array}$$

- Estimation of $\vec{\omega}$, $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$, $k \geq 2$ (or $a_1, b_1$ based on the distribution) where $\omega$ is from the Sarmanov distribution and $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ are the marginals for two lines of business. Bounds of $\vec{\omega}$ are based on estimated marginals $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$.

  For one-step inference method, we estimate $\vec{\omega}$, $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ simultaneously.

  For rank-based method, we use two-step inference method, which $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ are estimated first, then we estimate $\vec{\omega}$.

- Simulate a sample (of 55 observations) from the Sarmanov distribution using the parameters $\vec{\omega}$, $\alpha_1, \tau_1, ..., \alpha_k, \tau_k$ estimated above.

  Then we have simulated data (pseudo-response): $\begin{matrix} y^{*(1)}_{1,1} & \cdots & y^{*(1)}_{1,10} & & y^{*(2)}_{1,1} & \cdots & y^{*(2)}_{1,10} & \\ \vdots & \iddots & & & \vdots & \iddots & & \cdots \\ y^{*(1)}_{10,1} & & & & y^{*(2)}_{10,1} & & & \end{matrix}$

- Estimate parameters $\vec{\omega}^*, \alpha^*_1, \tau^*_1, ..., \alpha^*_k, \tau^*_k$ from the new simulated data (Different calculation for different method).

- Simulate the lower part (45 observations) of the triangle with the new estimated parameters $\vec{\omega}^*, \alpha^*_1, \tau^*_1, ..., \alpha^*_k, \tau^*_k$ obatianed above.

$$\begin{matrix} & & y^{*(1)}_{2,10} & & & & y^{*(2)}_{2,10} & \\ & \iddots & \vdots & & & \iddots & \vdots & \cdots \\ y^{*(1)}_{10,2} & \cdots & y^{*(1)}_{10,10} & & y^{*(2)}_{10,2} & \cdots & y^{*(2)}_{10,10} & \end{matrix}$$

- Calculate the reserve from the simulated lower part of the triangle.

Now we can use the bootstrap method with estimation and simulations which provides us the lower part of the simulated loss triangles and compute the risk capital.

We use the Kolmogorov-Smirnov test to check whether simulation procedure produces adequate datasets, as shown in Table 5.4. We observe that the null hypothesis is not rejected, except for the Commercial line of business from Shi and Frees (2011) data. This is not surprising, as both the goodness of fit test (Gamma for Commerical line of business) and dependence (Kendall tau test) were borderline for this dataset.

For Personal and commercial auto lines, Table 5.5 and Table 5.6 gives the TVaR 99% and risk capital after 5,000 times of bootstrap, which including simulation of the upper part of

Table 5.4: ks test for simulated vs original loss ratios

| Model/ p-value | 1st line | 2nd line | 3rd line |
|---|---|---|---|
| Bivariate Personal & Commercial | 0.9989 | 0.0005792 | - |
| Bivariate 2 & 4 | 0.9031 | 0.9789 | - |
| Trivariate 2 & 4 & 5 | 0.9031 | 0.9789 | 0.9789 |

the loss triangle, estimating $\omega^*$ for all the simulations and then using the new $\omega^*$ in the bivariate Sarmanov distribution to simulate the lower part of the loss triangle. As bootstrap is more computationally intensive, we use less simulations for this part.

Table 5.5: 5,000 Bootstrap TVaR 99% comparison Personal Commercial

| Model | TVaR 99% |
|---|---|
| Silo Method | 8,399,543 |
| Sarmanov with rank-based method | **7,913,426** |
| Gausian Copula (From Shi & Frees (2011)) | 7,923,715 |

Table 5.6: 5,000 Bootstrap risk capital comparison Personal Commercial

| Model | Risk Capital | Gain |
|---|---|---|
| Silo Method | 962,251 | - |
| Sarmanov with rank-based method | **790,212** | **17.88%** |

The comparison shows that bivariate Sarmanov model using rank of residuals produces lower risk measures than the silo method and Gaussian Copula model, which leads to the conclusion that it outperform the other two methods for this dataset.

We then use the bootstrap method to compare the Risk Capital for trivariate case line 2&4&5 and line 2&4 bivariate case for both one-step inference method and rank-based method.

Table 5.7: 5,000 Bootstrap Risk Capital comparison 245

| Model | Line 2 | Line 4 | Line 5 | Total | Gain |
|---|---|---|---|---|---|
| Silo | 35,471 | 26,899 | 5,563 | 67,934 | - |
| Biv 24 one-step inference | 28,233 | 18,320 | - | 52,117 | 23.28% |
| Biv 24 rank-based method | 24,717 | 17,978 | - | 48,258 | 28.96% |
| Triv 245 rank-based method | 24,548 | 17,970 | 1,591 | **44,110** | **35.07%** |

The bootstrap result also confirms the result we get from the simulation only method, that the trivariate Sarmanov distribution with rank-based method provides a better fit than the silo and bivariate Sarmanov model.

# Chapter 6

# Disscussion and Conclusion

We explored the use of bivariate and trivariate Sarmanov distribution with original one-step inference method, introduced a new rank-based method for Sarmanov distribution, showed the difference of both estimation procedure and analyzed two sets of data using such methods. We also provided and used the method to simulate the data using Sarmanov model, gave the bootstrap method and used it to calculate the risk capital.

The two sets of data we used are the real-life data from credible resources, the first set of data is from a major US property casualty insurer which provides personal and commercial auto lines of business. This data set has been widely used in the reserving literature, and we have checked that the personal auto line follows log-normal distribution, while the commercial auto line follows the gamma distribution. The second set of data is provided by a large Canadian property and casualty insurance company, where we chose 3 lines of business, which are Ontario Bodily injury, Ontario Accident benefits excluding disability income and Ontario Accident benefits with disability income only, and showed that all three lines of business follows gamma distribution.

Then we introduced the Sarmanov distribution, and showed using one-step inference method, the bivariate Sarmanov distribution could not capture the dependence between personal and commercial auto line. Although it can be used for line 2 and 4 for the Ontario Auto insurance, the trivariate Sarmanov distribution also does not work better than the independent case for line 2, 4 and 5.

However, when we used the rank-based method for Sarmanov distribution, it can capture the dependence between personal and commercial auto line, and also shows significance when dealing with line 2, 4 or line 2, 4 and 5 of Ontario Auto insurance for bivariate and trivariate Sarmanov distribution.

We also provided the simulation and bootstrap method for Sarmanov model, and by calculating and comparing the risk capitals, we can conclude that the model using rank-based method provides a better fit than the silo method and the model using one-step inference method. We also compared the TVaR 99% with the data from Shi and Frees (2011) and found it works better than the Gaussian copula model. From line 2, 4 and 5 of Ontario auto insurance data, we can see that trivariate Sarmanov model with rank-based method provides a better fit than the bivariate Sarmanov model, this could lead to an extension to further discussion about Sarmanov mdoel with more lines of business included.

Above all, Sarmanov distribution can capture the dependence between distributions and is easy to comprehend. Rank-based method provides a more robust estimation for the dependence parameters. There could be possibility to explore using Sarmanov distribution but with marginals GLMs which includes factors other than just accident year and development period, such as other aspects of the company or geographic locations, i.e. adding variables to the generalized linear model. Sarmanov model could also be put into use for areas beyond insurance, such as analyzing the dependence between measurable air pollution and water pollution, or be used in biological system, etc. There exist dependencies in all kinds of areas, and Sarmanov distribution can be used wherever there is data that can be ranked and follows certain distribution.

# Reference

Abdallah, A., J.-P. Boucher, and H. Cossette. 2015. Modeling Dependence between Loss Triangles with Hierarchical Archimedean Copulas. ASTIN Bulletin 45(3): 577-599.

Abdallah, Anas, et al. 2016. “Sarmanov Family of Bivariate Distributions for Multivariate Loss Reserving Analysis.” North American Actuarial Journal, vol. 20, no. 2.

Ajne B, 1994. Additivity of chain-ladder projections. ASTIN Bull 24:311-318

Bahraoui, Zuhair, et al. 2015. “On the bivariate Sarmanov distribution and copula. An application on insurance data using truncated marginal distributions” July-December.

Bargès M, Cossette H, Marceau E, 2009. TVaR-based capital allocation with copulas. Insur Math Econ 45:348-361

Bolance, Catalina, and Raluca Vernic. 2017. “Multivariate Count Data Generalized Linear Models: Three Approaches Based on the Sarmanov Distribution.” SSRN Electronic Journal.

Bornhuetter and Ferguson, 1972,”The Actuary and IBNR”.

Braun, C. 2004. The Prediction Error of the Chain Ladder Method Applied to Correlated Run-off Triangles. ASTIN Bulletin 34(2): 399–434.

Brehm, P. 2002. Correlation and the Aggregation of Unpaid Loss Distributions. Casualty Actuarial Society Forum 2(Fall): 1–23.

Côté, Marie-Pier, et al. “Rank-Based Methods for Modeling Dependence between Loss Triangles.” European Actuarial Journal, vol. 6, no. 2, 2016.

De Jong, P. 2012. Modeling Dependence between Loss Triangles. North American Actuarial Journal 16(1): 74–86.

Genest C, Neslehova J, 2012. Copulas and copula models. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics, 2nd edn. Wiley, Chichester

Kirschner, G., C. Kerley, and B. Isaacs. 2008. Two Approaches to Calculating Correlated Reserve Indications Across Multiple Lines of Business. Variance 2(1): 15-38.

Lee, M.-L. T. 1996. Properties and Applications of the Sarmanov Family of Bivariate Distributions. Comm. Statist. Theory Methods, 25(6):1207–1222.

Merz, M. and Wüthrich, M. 2008. Prediction error of the multivariate chain ladder reserving method. North American Actuarial Journal 12(2), 175-197.

Pelican, E., & Vernic, R. (2013). Maximum-likelihood estimation for the multivariate Sarmanov distribution: simulation study. International Journal of Computer Mathematics, 90(9), 1958-1970.

Ratovomirija, Gildas, et al. 2016. "On Some Multivariate Sarmanov Mixed Erlang Reinsurance Risks: Aggregation and Capital Allocation." SSRN Electronic Journal.

Schmidt, K. 2006. Optimal and Additive Loss Reserving for Dependent Lines of Business. Casualty Actuarial Society Forum 2006a(fall): 319–351.

Shi, P., S. Basu, and G. Meyers. 2012. A Bayesian Log-Normal Model for Multivariate Loss Reserving. North American Actuarial Journal 16(1): 29–51.

Shi P, Frees E 2011. Dependent loss reserving using copulas. ASTIN Bull 41:449–486

Tasche D, 1999. Risk contributions and performance measurement. Working paper, Technische Universität München, Germany

Taylor, G., and G. McGuire. 2007. A Synchronous Bootstrap to Account for Dependencies between Lines of Business in the Estimation of Loss Reserve Prediction Error. North American Actuarial Journal 11(3): 70–88.

Wüthrich, M., M. Merz, and E. Hashorva. 2013. Dependence Modelling in Multivariate Claims Run-off Triangles. Annals of Actuarial Science 7(1): 3–25.

# Appendix 1 Data

Table 6.1: Incremental paid losses for personal auto line

| year | premium | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1988 | 4 711 333 | 1 376 384 | 1 211 168 | 535 883 | 313 790 | 168 142 | 79 972 | 39 235 | 15 030 | 10 865 | 4 086 |
| 1989 | 5 335 525 | 1 576 278 | 1 437 150 | 652 445 | 342 694 | 188 799 | 76 956 | 35 042 | 17 089 | 12 507 | |
| 1990 | 5 947 504 | 1 763 277 | 1 540 231 | 678 959 | 364 199 | 177 108 | 78 169 | 47 391 | 25 288 | | |
| 1991 | 6 354 197 | 1 779 698 | 1 498 531 | 661 401 | 321 434 | 162 578 | 84 581 | 53 449 | | | |
| 1992 | 6 738 172 | 1 843 224 | 1 573 604 | 613 095 | 299 473 | 176 842 | 106 296 | | | | |
| 1993 | 7 079 444 | 1 962 385 | 1 520 298 | 581 932 | 347 434 | 238 375 | | | | | |
| 1994 | 7 254 832 | 2 033 371 | 1 430 541 | 633 500 | 432 257 | | | | | | |
| 1995 | 7 739 379 | 2 072 061 | 1 458 541 | 727 098 | | | | | | | |
| 1996 | 8 154 065 | 2 210 754 | 1 517 501 | | | | | | | | |
| 1997 | 8 435 918 | 2 206 886 | | | | | | | | | |

Table 6.2: Incremental paid losses for commercial auto line

| year | premium | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1988 | 267 666 | 33 810 | 45 318 | 46 549 | 35 206 | 23 360 | 12 502 | 6 602 | 3 373 | 2 373 | 778 |
| 1989 | 274 526 | 37 663 | 51 771 | 40 998 | 29 496 | 12 669 | 11 204 | 5 785 | 4 220 | 1 910 | |
| 1990 | 268 161 | 40 630 | 56 318 | 56 182 | 32 473 | 15 828 | 8 409 | 7 120 | 1 125 | | |
| 1991 | 276 821 | 40 475 | 49 697 | 39 313 | 24 044 | 13 156 | 12 595 | 2 908 | | | |
| 1992 | 270 214 | 37 127 | 50 983 | 34 154 | 25 455 | 19 421 | 5 728 | | | | |
| 1993 | 280 568 | 41 125 | 53 302 | 40 289 | 39 912 | 6 650 | | | | | |
| 1994 | 344 915 | 57 515 | 67 881 | 86 734 | 18 109 | | | | | | |
| 1995 | 371 139 | 61 553 | 132 208 | 20 923 | | | | | | | |
| 1996 | 323 753 | 112 103 | 33 250 | | | | | | | | |
| 1997 | 221 448 | 37 554 | | | | | | | | | |

Tables 6.3–6.5 provide the net earned premiums and the cumulative paid losses for accident years 2003–12 inclusively for each of LOBs 2, 4, 5 developed over at most ten years. To preserve confidentiality, all figures were multiplied by a constant. Table 6.6 provides the parameters for the GLMs of three LOBs.

Table 6.3: Cumulative paid losses for LOB 2.

| Accident | Development Lag (in months) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 | Premiums |
| 2003 | 3488 | 14559 | 27249 | 37979 | 49561 | 55957 | 58406 | 60862 | 63280 | 63864 | 85421 |
| 2004 | 1169 | 12781 | 20550 | 31547 | 42808 | 47385 | 50251 | 50978 | 51272 | | 98579 |
| 2005 | 1478 | 10788 | 25499 | 34279 | 43057 | 49360 | 52329 | 52544 | | | 103062 |
| 2006 | 1186 | 11852 | 22913 | 32537 | 41824 | 48005 | 52542 | | | | 108412 |
| 2007 | 1737 | 13881 | 25521 | 38037 | 43684 | 47755 | | | | | 111176 |
| 2008 | 1571 | 12153 | 27329 | 41832 | 51779 | | | | | | 112050 |
| 2009 | 1199 | 17077 | 29876 | 44149 | | | | | | | 112577 |
| 2010 | 1263 | 16073 | 28249 | | | | | | | | 113707 |
| 2011 | 986 | 10003 | | | | | | | | | 126442 |
| 2012 | 683 | | | | | | | | | | 130484 |

Table 6.4: Cumulative paid losses for LOB 4.

| Accident | Development Lag (in months) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 | Premiums |
| 2003 | 13714 | 24996 | 31253 | 38352 | 44185 | 46258 | 47019 | 47894 | 48334 | 48902 | 116491 |
| 2004 | 6883 | 16525 | 24796 | 29263 | 32619 | 33383 | 34815 | 35569 | 35612 | | 111467 |
| 2005 | 7933 | 22067 | 32801 | 38028 | 44274 | 44948 | 46507 | 46665 | | | 107241 |
| 2006 | 7052 | 18166 | 25589 | 31976 | 36092 | 38720 | 39914 | | | | 105687 |
| 2007 | 10463 | 23982 | 31621 | 36039 | 38070 | 41260 | | | | | 105923 |
| 2008 | 9697 | 28878 | 41678 | 47135 | 50788 | | | | | | 111487 |
| 2009 | 11387 | 37333 | 48452 | 55757 | | | | | | | 113268 |
| 2010 | 12150 | 32250 | 40677 | | | | | | | | 121606 |
| 2011 | 5348 | 14357 | | | | | | | | | 110610 |
| 2012 | 4612 | | | | | | | | | | 104304 |

Table 6.5: Cumulative paid losses for LOB 5.

| Accident | Development Lag (in months) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 | Premiums |
| 2003 | 3043 | 5656 | 7505 | 8593 | 9403 | 10380 | 10450 | 10812 | 10856 | 10860 | 116491 |
| 2004 | 2070 | 4662 | 6690 | 8253 | 9286 | 9724 | 9942 | 10086 | 10121 | | 111467 |
| 2005 | 2001 | 4825 | 7344 | 8918 | 9824 | 10274 | 10934 | 11155 | | | 107241 |
| 2006 | 1833 | 4953 | 7737 | 9524 | 10986 | 11267 | 11579 | | | | 105687 |
| 2007 | 2217 | 5570 | 7898 | 8885 | 9424 | 10402 | | | | | 105923 |
| 2008 | 2076 | 5681 | 8577 | 10237 | 12934 | | | | | | 111487 |
| 2009 | 2025 | 6225 | 9027 | 10945 | | | | | | | 113268 |
| 2010 | 2024 | 5888 | 8196 | | | | | | | | 121606 |
| 2011 | 1311 | 3780 | | | | | | | | | 110610 |
| 2012 | 912 | | | | | | | | | | 104304 |

Table 6.6: Parameter and Reserve Estimations.

| LOB $\ell$ | | 2 | 4 | 5 |
|---|---|---|---|---|
| GLM | | Gamma | Gamma | Gamma |
| $u^{(\ell)}$ | | $-3.628$ (0.148) | $-2.365$ (0.173) | $-4.064$ (0.148) |
| Accident Year | 2 | $-0.750$ (0.151) | $-0.413$ (0.174) | $-0.121$ (0.151) |
| | 3 | $-0.729$ (0.160) | $-0.196$ (0.183) | 0.171 (0.161) |
| | 4 | $-0.651$ (0.168) | $-0.112$ (0.190) | 0.129 (0.168) |
| | 5 | $-0.741$ (0.174) | $-0.095$ (0.199) | 0.092 (0.173) |
| | 6 | $-0.574$ (0.185) | $-0.001$ (0.210) | 0.396 (0.187) |
| | 7 | $-0.574$ (0.200) | 0.197 (0.227) | 0.254 (0.200) |
| | 8 | $-0.658$ (0.220) | $-0.012$ (0.253) | 0.055 (0.222) |
| | 9 | $-1.147$ (0.255) | $-0.628$ (0.295) | $-0.259$ (0.260) |
| | 10 | $-1.625$ (0.340) | $-0.754$ (0.393) | $-0.676$ (0.348) |
| Dev. Lag | 2 | 2.061 (0.145) | 0.450 (0.167) | 0.419 (0.149) |
| | 3 | 2.065 (0.151) | $-0.055$ (0.175) | 0.114 (0.155) |
| | 4 | 2.018 (0.158) | $-0.507$ (0.183) | $-0.358$ (0.163) |
| | 5 | 1.818 (0.166) | $-0.759$ (0.193) | $-0.582$ (0.173) |
| | 6 | 1.297 (0.176) | $-1.580$ (0.207) | $-1.154$ (0.182) |
| | 7 | 0.773 (0.193) | $-1.899$ (0.223) | $-1.870$ (0.201) |
| | 8 | $-0.493$ (0.216) | $-2.670$ (0.250) | $-2.103$ (0.219) |
| | 9 | $-0.429$ (0.255) | $-3.762$ (0.298) | $-3.849$ (0.257) |
| | 10 | $-1.358$ (0.340) | $-2.960$ (0.393) | $-6.248$ (0.348) |
| sd or scale | | 10.700 (2.009) | 8.038 (1.502) | 10.078 (1.891) |
| Reserve | | 132,919 | 73,220 | 18,288 |
| C-L Reserve | | 146,794 | 75,551 | 18,726 |

# Appendix 2 Closed-form expression

This appendix will provide the closed-form expression for expectation, variance and covariance for loss reserve.

$$E[R_{tot}] = E[R^{(1)} + R^{(2)}] = E\left[\sum_{l=1}^{2}\sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(l)} Y_{i,j}^{(l)}\right] = \sum_{l=1}^{2}\sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(l)} E[Y_{i,j}^{(l)}]$$

$$E[Y_{i,j}^{(1)}] = e^{\mu_{i,j}^{(1)}+\sigma^2/2},\ E[Y_{i,j}^{(2)}] = \alpha * \tau_{i,j}^{(2)}$$

$$E[R_{tot}] = \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)} e^{\mu_{i,j}^{(1)}+\sigma^2/2} + \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(2)} \alpha * \tau_{i,j}^{(2)}$$

$$Var(R_{tot}) = Var(R^{(1)} + R^{(2)}) = Var(R^{(1)}) + Var(R^{(2)}) + 2Cov(R^{(1)}, R^{(2)})$$

$$Var(R^{(1)}) = Var(\sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)} Y_{i,j}^{(1)}) = \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)2} Var(Y_{i,j}^{(1)})$$
$$= \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)2} (e^{\sigma^2} - 1) e^{2\mu_{i,j}^{(1)}+\sigma^2}$$

$$Var(R^{(2)}) = Var(\sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(2)} Y_{i,j}^{(2)}) = \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(2)2} Var(Y_{i,j}^{(2)})$$
$$= \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(2)2} \alpha * \tau_{i,j}^{(2)2}$$

$$Cov(R^{(1)}, R^{(2)}) = \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)} p_i^{(2)} \Big( E[Y_{i,j}^{(1)} Y_{i,j}^{(2)}] - E[Y_{i,j}^{(1)}]E[Y_{i,j}^{(2)}] \Big)$$

$$E[Y_{i,j}^{(1)}]E[Y_{i,j}^{(2)}] = (e^{\mu_{i,j}^{(1)} + \sigma^2/2})(\alpha * \tau_{i,j}^{(2)})$$

$$f(y_{ij}^{(1)}, y_{ij}^{(2)}) = f^{(1)}\big(y_{ij}^{(1)}; \alpha_1, \tau_1\big) f^{(2)}\big(y_{ij}^{(2)}; \alpha_2, \tau_2\big)\big(1 + \omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\big)$$

$$E[Y_{i,j}^{(1)} Y_{i,j}^{(2)}] = (e^{\mu_{i,j}^{(1)} + \sigma^2/2})(\alpha * \tau_{i,j}^{(2)}) + \Big( \int_0^{\infty}\int_{-\infty}^{\infty} y_{ij}^{(1)} y_{ij}^{(2)} f(y_{ij}^{(1)}) f(y_{ij}^{(2)}) \big(\omega\psi^{(1)}(y_{ij}^{(1)})\psi^{(2)}(y_{ij}^{(2)})\big) dy_{ij}^{(1)} dy_{ij}^{(2)} \Big)$$
$$= (e^{\mu_{i,j}^{(1)} + \sigma^2/2})(\alpha * \tau_{i,j}^{(2)}) + \int y_{ij}^{(1)} f(y_{ij}^{(1)})(exp(-y_{ij}^{(1)}) - exp(-\mu + \sigma^2/2)) dy_{ij}^{(1)}$$
$$* \int y_{ij}^{(2)} f(y_{ij}^{(2)})(exp(-y_{ij}^{(2)}) - (1+\tau)^{-\alpha} dy_{ij}^{(2)}$$
$$= (e^{\mu_{i,j}^{(1)} + \sigma^2/2})(\alpha * \tau_{i,j}^{(2)}) + (-exp(\sigma^2) + \int y_{ij}^{(1)} exp(-y_{ij}^{(1)}) f(y_{ij}^{(1)}) dy_{ij}^{(1)})$$
$$* (-\alpha_2\tau_2 - \alpha_2\tau_2^{-\alpha_2+1} + \int y_{ij}^{(2)} exp(-y_{ij}^{(2)}) f(y_{ij}^{(2)}) dy_{ij}^{(2)})$$

$$
\begin{aligned}
Var(R_{tot}) &= \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)2}(e^{\sigma^2}-1)e^{2\mu_{i,j}^{(1)}+\sigma^2} + \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(2)2}\alpha * \tau_{i,j}^{(2)2} \\
&+ 2 * \sum_{i=2}^{n}\sum_{j=n-i+2}^{n} p_i^{(1)}p_i^{(2)} \\
&* \Bigg( (e^{\mu_{i,j}^{(1)}+\sigma^2/2})(\alpha * \tau_{i,j}^{(2)}) + \Big( \int_0^{\infty}\int_{-\infty}^{\infty} \big(\omega y_{ij}^{(1)} y_{ij}^{(2)} f(y_{ij}^{(1)}) f(y_{ij}^{(2)}) \psi^{(1)}(y_{ij}^{(1)}) \psi^{(2)}(y_{ij}^{(2)})\big) dy_{ij}^{(1)} dy_{ij}^{(2)} \Big) \\
&- (e^{\mu_{i,j}^{(1)}+\sigma^2/2})(\alpha * \tau_{i,j}^{(2)}) \Bigg)
\end{aligned}
$$

# Appendix 3 Proof for omega bounds

$$
\begin{aligned}
\rho &= \frac{E[X_1X_2] - E[X_1]E[X_2]}{\sigma_1\sigma_2} \\
&= \frac{\int x_1x_2(f_1(x_1)f_2(x_2)(1+\omega\psi_1(x_1)\psi_2(x_2)))dx_1dx_2 - \int x_1f_1(x_1)dx_1 \int x_2f_s(x_2)dx_s}{\sigma_1\sigma_2} \\
&= \frac{\omega \int x_1f_1(x_1)\psi_1(x_1)dx_1 \int x_2f_2(x_2)\psi_2(x_2)dx_2}{\sigma_1\sigma_2} \\
&= \frac{\omega\nu_1\nu_2}{\sigma_1\sigma_2}
\end{aligned}
$$

with $\nu_i = \int x_if_i(x_i)\psi_i(x_i)dx_i$ and $\psi_i(x_i) = exp(-x_i) - L_i(1)$ where $L_i(1)$ is the Laplace transform evaluated at 1.

Let $X_1 \sim Normal(a, b^2)$, $X_2 \sim Gamma(\alpha, \tau)$, from Pelican and Vernic (2013) and Lee (1996), we have

$$L_1(1) = exp(-a + b^2/2)$$

$$L_2(1) = (1+\tau)^{-\alpha}$$

Given notation:

$$f_1(x_1) = n(x_1; a, b^2) = \frac{1}{b\sqrt{2\pi}} exp\left( - \frac{(x_1 - a)^2}{2b^2} \right)$$

$$f_2(x_2) = h(x_2; \alpha, \tau) = \frac{\theta_t^{(2)\alpha-1}}{\Gamma(\alpha)\tau^\alpha} exp(-x_2/\tau)$$

Then we can get

$$\begin{aligned}
\nu_1 &= \int x_1 f_1(x_1)\psi_1(x_1)dx_1 \\
&= \int x_1(exp(-x_1) - L_1(1))n(x_1; a, b^2)dx_1 \\
&= \int (x_1 exp(-x_1) - x_1 exp(-a + b^2/2))n(x_1; a, b^2)dx_1 \\
&= \int x_1 exp(-x_1)n(x_1; a, b^2)dx_1 - exp(-a + b^2/2)\int x_1 n(x_1; a, b^2)dx_1 \\
&= \int x_1 n(x_1; a - b^2, b^2)exp(-a + b^2/2)dx_1 - exp(-a + b^2/2)a \\
&= exp(-a + b^2/2)(a - b^2) - exp(-a + b^2/2)a \\
&= -b^2 exp(-a + b^2/2)
\end{aligned}$$

$$\begin{aligned}
\nu_2 &= \int x_2 f_2(x_2)\psi_2(x_2)dx_2 \\
&= \int x_2(exp(-x_2) - L_2(1))h(x_2; \alpha, \tau)dx_2 \\
&= \int x_2(exp(-x_2) - (1 + \tau)^{-\alpha})h(x_2; \alpha, \tau)dx_2 \\
&= \int x_2 exp(-x_2)h(x_2; \alpha, \tau)dx_2 - (1 + \tau)^{-\alpha}\int x_2 h(x_2; \alpha, \tau)dx_2 \\
&= \int x_2 h(x_2; \alpha, \frac{\tau}{1 + \tau})(1 + \tau)^{-\alpha}dx_2 - (1 + \tau)^{-\alpha} * \alpha\tau \\
&= (1 + \tau)^{-\alpha}(\frac{\alpha\tau}{1 + \tau}) - (1 + \tau)^{-\alpha} * \alpha\tau \\
&= \alpha(1 + \tau)^{-\alpha}(\frac{\tau}{1 + \tau} - \tau) \\
&= -\alpha\tau^2(1 + \tau)^{-\alpha-1}
\end{aligned}$$

As $\sigma_1 = b,\ \sigma_2 = \sqrt{\alpha}\tau$, then

$$\rho = \frac{\omega\nu_1\nu_2}{\sigma_1\sigma_2} = \frac{\omega b^2 exp(-a + b^2/2)\alpha\tau^2(1 + \tau)^{-\alpha-1}}{b\sqrt{\alpha}\tau} = \omega b exp(-a + b^2/2)\sqrt{\alpha}\tau(1 + \tau)^{-\alpha-1}$$

Because $-1 \leq \rho \leq 1$, we have $-1 \leq \omega b exp(-a + b^2/2)\sqrt{\alpha}\tau(1 + \tau)^{-\alpha-1} \leq 1$, and since $b exp(-a + b^2/2)\sqrt{\alpha}\tau(1 + \tau)^{-\alpha-1} \geq 0$, we have the omega bound

$$-\frac{1}{b exp(-a + b^2/2)\sqrt{\alpha}\tau(1 + \tau)^{-\alpha-1}} \leq \omega \leq \frac{1}{b exp(-a + b^2/2)\sqrt{\alpha}\tau(1 + \tau)^{-\alpha-1}}$$

Similarly, we can get the omega bound for margnals as two gamma distributions:

$$-\frac{1}{\sqrt{\alpha_1}\tau_1(1+\tau_1)^{-\alpha_1-1}\sqrt{\alpha_2}\tau_2(1+\tau_2)^{-\alpha_2-1}} \leq \omega \leq \frac{1}{\sqrt{\alpha_1}\tau_1(1+\tau_1)^{-\alpha_1-1}\sqrt{\alpha_2}\tau_2(1+\tau_2)^{-\alpha_2-1}}$$