Individual Loss Reserving for Multi-coverage Insurance

A Preprint

Roxane Turcotte
Université du Québec à Montréal
Département de mathématiques
Email : turcotte.roxane@uqam.ca
Peng Shi
Department of Risk and Insurance
Wisconsin School of Business
University of Wisconsin-Madison
Email: pshi@bus.wisc.edu

Abstract

Individual loss reserving methods have undergone substantial development in the past decade, driven by increased accessibility to granular-level insurance claims data. This paper presents a micro loss reserving model tailored for multi-coverage insurance policies, where a single insurance claim might trigger payments from multiple coverage types. We employ a copula-based multivariate regression approach to jointly model the settlement time and loss amount, effectively capturing the dependence among various types of loss amounts and their correlation with the settlement time. We stress the importance of considering both types of dependence for accurate reserving prediction and uncertainty quantification. Furthermore, we propose computationally efficient algorithms for parameter estimation and dynamic prediction. Through numerical experiments and real data analysis, we demonstrate the effectiveness of our proposed multivariate predictive model in loss reserving applications.

Keywords: copula regression, dependent risks, dynamic prediction, loss reserving, multi-coverage insurance

1 Introduction

Loss reserving, the process of predicting the insurer's outstanding liability, is a quintessential actuarial function in the insurance industry. Accurate estimation of loss reserves is crucial to several key operations within an insurance company, including claims management, ratemaking, and financial reporting (Frees (2015)). Reserving methods can be broadly categorized into aggregate (macro) and individual (micro) approach, depending on the type of claims data utilized for model development. This work specifically concentrates on individual loss reserving methods tailored for multi-coverage insurance products.

Individual loss reserving methods were initially formulated within a marked Poisson process framework to accommodate granular claims data (see Arjas (1989), Norberg (1993, 1999)). The first implementation of this model in empirical studies was due to Antonio and Plat (2014). The point process approach is further extended to the marked Cox process to account for overdispersion (see Avanzi et al. (2016), Badescu et al.(2016, 2019)), and to a copula-based point process to accommodate informative terminal events (see Yang et al. (2024)). The core advantage of granular models lies in their ability to integrate detailed information into the forecasting process, including but not limited to case reserves (Taylor et al. (2008)), claim markers (i.e. claim-specific characteristics) (Godecharle and Antonio (2015)), and environmental changes (Okine (2023a)). As claims data becomes more complex and voluminous, recent literature has witnessed the application of various machine learning methods in the context of individual loss reserving. Examples include Wüthrich (2018), Lopez et al. (2019), Duval and Pigeon (2019), and Delong and Wüthrich (2020). We refer readers to the recent survey of Taylor (2019) for detailed discussion on these developments.

Departing from the aforementioned literature, our work aims to develop an individual loss reserving method specifically designed for insurance policies providing multiple types of coverage. Multi-coverage contracts are prevalent in nonlife insurance, with examples such as automobile insurance indemnifying financial losses due to collision and third party liability (both property damage and bodily injury), homeowner insurance providing protection for building, contents, as well as liability, and workers compensation offering benefits for wage replacement, medical treatment, and vocational rehabilitation. A distinctive aspect of claims data from multi-coverage policies is the interrelation among claims from different coverage types. This unique feature demands careful

consideration in the development of methodologies for individual loss reserving for multi-coverage policies. Despite the rapid expansion of the individual loss reserving literature, methods adept at accommodating dependence among multiple types of claims remain scarce. A recent example is Michaelides et al. (2023) where the across-coverage dependence is introduced via activation patterns modeled with a multinomial logit regression while the loss amounts from different coverage types are treated as independent. Our research addresses this gap by proposing an individual loss reserving model that explicitly captures and quantifies the dependence among different types of claims originating from a multi-coverage insurance policy.

It is noteworthy that the underdevelopment of individual loss reserving methods for multi-coverage policies is in contrast to the extensive research on multivariate loss reserving methods in the aggregate context using loss triangle data. In the literature of aggregate loss reserving, various multivariate methods have been proposed to account for the dependence between different lines of business. Notable examples include nonparametric approaches such as those proposed by Merz and Wüthrich (2008), Merz and Wüthrich (2009), and Zhang (2010), and parametric approaches as evidenced by Shi and Frees (2011), Shi et al. (2012), and De Jong (2012). Given that dependence is a shared consideration for model development in both multi-coverage policies and multi-line insurance, one would naturally anticipate a parallel advancement in multivariate methods for individual reserving. Compared to the advanced state of multivariate aggregate loss reserving methods, the underdevelopment of individual loss reserving methods for multi-coverage insurance can be attributed to two primary reasons. First, from a methodology standpoint, the concept of dependence is well understood as an element-wise relationship in the context of multiple loss triangles. However, the notion of pairwise association becomes less clear when dealing with claim-level losses from multiple coverage types. Second, from a data perspective, access to granular loss data is more limited due to the constraints on data collection and data quality, and this issue is further compounded when considering claim data categorized by coverage type.

Furthermore, the development of multivariate granular reserving methods falls behind their ratemaking counterpart. It is widely recognized that pricing models in non-life insurance heavily rely on granular data driven by the competitive nature of the market. In particular, numerous multivariate pricing methods have been proposed for multi-coverage policies, such as automobile insurance (Frees and Valdez (2008), Frees et al. (2009), and Shi et al. (2016)) and commercial

property insurance (Frees et al. (2016)). Additionally, these methods have been expanded for pricing multivariate insurance risks in a much broader sense, including multi-peril risks (Frees et al. (2010), Shi and Yang (2018), and Yang and Shi (2019)) and spatially correlated risks (Zhao et al. (2021) and Huang et al. (2023)). Beyond their application in reserving, individual loss reserving methods arguably play a more vital role in ratemaking (Crevecoeur et al. (2023) and Okine (2023b)). Therefore, it is imperative to develop a multivariate individual loss reserving method to align with the demands of the ratemaking task.

In this paper, we present a copula-based granular reserving model designed for estimating loss reserves for multi-coverage insurance. Specifically, we employ a multivariate copula to jointly model the ultimate losses from different coverage types within a given claim, as well as the settlement time of the claim. Consequently, our proposed approach not only enables the quantification of dependence among losses of multiple coverage types but also captures the association between the size of the claim and its settlement time - a notable relationship identified in recent literature (see Okine et al. (2022) and Yang et al. (2024)). The resulting model facilitates dynamic predictions for an insurer's outstanding liability by leveraging information across coverage types and settlement delay. To tackle the estimation challenges due to imbalanced and censored observations in the data, we further introduce a stage-wise approach to estimate parameters in the proposed model. We investigate the efficacy of the stage-wise estimation using simulation studies, and demonstrate dynamic prediction in a loss reserving application using a large portfolio of automobile insurance claims from a Canadian insurance company.

Lastly, it is worth commenting on the significant role of dependence in the individual reserving context. While the consideration of dependence among various types of losses is a shared aspect in both aggregate and individual loss reserving methods, its implications on reserve prediction are different. In the aggregate reserving context, neglecting dependence is detrimental to accurately quantifying reserving variability. In simpler terms, uncertainty is underestimated when losses are positively correlated and overestimated when losses are negatively correlated. In the proposed individual loss reserving model, dependence affects prediction uncertainty similarly. Moreover and of greater importance, accounting for dependence enables dynamic updating of predictions for the loss from one coverage type based on the loss development from other coverage types.

The rest of the article is organized as follows: Section 2 describes the granular claims data in

automobile insurance that motivate our work. Section 3 introduces the copula-based individual loss reserving method for multi-coverage insurance. Section 4 delves into the two-stage estimation method through simulation studies. Section 5 applies the proposed method to the automobile insurance claims data and demonstrates dynamic prediction of outstanding claim payments. Section 6 concludes the paper.

2 Data

We consider a portfolio of private automobile insurance claims obtained from a Canadian insurance company, focusing on regions such as Ontario, Alberta, and the Atlantic provinces where mandatory coverages aren't provided by public insurance companies. The portfolio consists of 563,426 insurance claims recorded between January 1st, 2015 and July 31st, 2021. The data contain detailed information for these claims collected by the insurer up to July 31st, 2021. It is important to note that not all claims have been settled by the end of our observation period.

Each insurance claim involves up to three types of coverage, Automobile Physical Damage (APD), Loss of Use (LU) and Bodily Injury (BI). APD and LU cover the policyholder's losses related to car damage and the need for a replacement car during repairs, respectively. BI compensates for third-party medical expenses when the insured is at fault. Notably, Canada's public health insurance system limits medical expenses to what isn't already covered, such as compensation for severe physical impairment or loss of income resulting from a car accident.

The dataset exhibits an imbalance wherein not all claims activate all three types of coverage. Table 1 provides a comprehensive overview of the combinations of coverage types within the insurance portfolio. It is evident from the table that the predominant portion of claims emanates from the APD coverage, with the majority of multi-coverage claims also being triggered by the APD coverage. This observation is expected, as LU and BI losses frequently follow APD losses in many cases.

Table 1: Number and percentage of claims by coverage type

Coverage	Number	Percentage
APD	215,971	38.33
LU	304	0.05
BI	6,260	1.11
(APD, LU)	307,460	54.57
(APD, BI)	11,110	1.97
(LU,BI)	12	0.002
(APD, LU,BI)	22,309	3.96
Total	563,426	100.00%

The central outcome variables underpinning our analysis are the settlement time and loss amount. Within our context, settlement time refers to the period an insurer takes to process and settle an insurance claim subsequent to its reporting to the insurer, while loss amount indicates the ultimate losses delineated by coverage type. Table 2 offers descriptive statistics for these critical outcome variables. It is notable that both settlement time (measured in days) and loss amount (measured in CAD) exhibit right skewness and heavy-tailed distributions. Moreover, we observe that APD and LU claims are relatively high in frequency, although APD claims tend to have larger severity compared to LU claims. Additionally, BI claims occur less frequently, yet their loss severity surpasses that of the other two types of claims.

Table 2: Descriptive statistics for settlement time and ultimate loss amount by coverage

Variable	NumOfObs	Mean	SD	Min	25th	50th	75th	Max
Settlement Time	563,426	87	168	0	22	44	83	2,389
Loss - APD	556,850	4,756	8,381	0	524	2,313	5,573	533,574
Loss - LU	330,085	419	544	0	83	307	602	52,777
Loss - BI	39,691	15,331	66,334	0	0	1,139	5,783	2,244,794

To investigate the relationship among these outcome variables, we present in Table 3 their pairwise associations using Kendall's and Spearman's rank correlation coefficients. Two types

of associations are particularly pertinent to our study: the dependence among different types of coverage amounts and the dependence between settlement time and loss amount. The table reveals a positive correlation among losses of different coverage types. For instance, APD and LU exhibit a strong correlation, which can be reasonably explained by the fact that severe damage to the car necessitates a longer time for repairs. Moreover, the table illustrates a positive relationship between settlement time and loss amount. This positive correlation is also expected because larger claims typically take longer to settle due to the expertise and resources involved in the settlement process.

Table 3: Pairwise rank correlation between settlement time and ultimate losses by coverage

		Spearman's ρ			
		Y_{APD}	Y_{LU}	Y_{BI}	T
Kendall's τ	Y_{APD}	1	*0.5516*	*0.0722*	*0.4180*
	Y_{LU}	0.4112	1	*0.0057*	*0.1322*
	Y_{BI}	0.0517	0.0042	1	*0.6452*
	T	0.2982	0.0944	0.4838	1

Furthermore, our dataset encompasses a rich array of both policy-level and claim-level information, serving as covariates in our analytical model. Table 4 summarizes the descriptive statistics of these predictors. All the covariates are presented as categorical variables with the percentage reported for each level. The majority of the covariates are policy-level attributes which contain driver characteristics such as age and gender, and metrics pertaining to the vehicle's usage, such as its purpose and the distance driven. In addition, amidst these covariates, there is one claim-specific variable describing the degree of responsibility assigned for the accident.

Table 4: Descriptive statistics for the covariates (percentage by level of each variable)

Covariate	Categories			
Level of responsibility	At fault & NA	Partly at fault	Not at fault	Not applicable
	28.86	2.24	38.38	30.52
Province	Ontario	Alberta	Atlantic	
	60.10	24.42	15.48	
Gender	Female	Male & NA		
	32.83	67.17		
Decade of birth	40s & before	50s	60s	70s
	8.9	11.13	15.22	13.85
	80s	90s	00s & after	Not applicable
	13.50	8.87	0.66	27.87
Main purpose usage	Pleasure	Business & Commercial	Commute & NA	
	35.08	64.93		
Annual kilometers	$[0, 14999)$	$[15000, 19999)$	$[20000, 24999)$	$[25000, 29999)$
	40.43	17.24	30.08	4.07
	$[30000, 34999)$	[35000+ & NA		
	3.92	4.25		

3 Methodology

3.1 Some Notations

Consider an insurance policy offering k types of coverage. For a given insurance claim indexed by i, T_i represents the settlement time, indicating the duration from the claim's reporting to the insurer to its closure. Let $\eta_i^{(l)}$, for $l = 1, \ldots, k$, be a binary variable indicating whether the lth type of coverage is activated for the claim. At any given time τ during a claim's progression, i.e. $\tau \in [0, T_i]$, let $Y_i^{(l)}(\tau)$ represent the cumulative paid losses from coverage type l up to time τ. Furthermore, we define the corresponding ultimate losses as $Y_i^{(l)} = Y_i^{(l)}(T_i)$. In addition, denote baseline covariates as $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{ip})'$.

In our study, we operate under the assumption that the coverage information $\eta_i^{(l)}$ is available at the time of claim submission. That is, the insurer is able to identify coverage of all types when the claim is reported. It's important to note that this assumption should not be perceived as a limitation of our work. Firstly, it's a reasonable assumption for personal automobile insurance and is corroborated by our dataset. Secondly, even if this assumption is not applicable in certain contexts, our proposed method remains valuable. In such cases, the method can be complemented by a separate model focused specifically on whether coverage of each type is triggered. In this work, we focus on reported claims. If one would like to model incurred but not reported (IBRN) losses, they could use our model in combination with a model predicting the number of not reported claims and their triggered coverage(s).

3.2 Joint Model

Considering the interconnected nature of settlement time and loss amounts, we employ a multivariate regression framework to describe their relationship. Specifically, given covariates $\boldsymbol{X}_i$, the (conditional) joint distribution function of $(Y_i^{(1)}, \ldots, Y_i^{(k)}, T_i)$, denoted by F, can be represented via a $(k+1)$-variate copula as:

$$\begin{aligned} & F(y_1, \ldots, y_k, t|\boldsymbol{X}_i) \\ =&\Pr(Y_i^{(1)} \leq y_1, \ldots, Y_i^{(k)} \leq y_k, T_i < t|\boldsymbol{X}_i) \\ =&H(F_1(y_1|\boldsymbol{X}_i), \ldots, F_k(y_k|\boldsymbol{X}_i), F_T(t|\boldsymbol{X}_i)) \end{aligned} \tag{1}$$

where F_l is the distribution function of $Y_i^{(l)}$ for $l \in \{1, \ldots, k\}$, , F_T is the distribution of T_i, and H is a $(k+1)$-variate copula. Let $\bar{H}$ denote the survival copula associated with H. We can express the joint survial function of $(Y_i^{(1)}, \ldots, Y_i^{(k)}, T_i)$, denoted by S, by:

$$\begin{aligned} & S(y_1, \ldots, y_k, t|\boldsymbol{X}_i) \\ =&\Pr(Y_i^{(1)} > y_1, \ldots, Y_i^{(k)} > y_k, T_i > t|\boldsymbol{X}_i) \\ =&\bar{H}(1 - F_1(y_1|\boldsymbol{X}_i), \ldots, 1 - F_k(y_k|\boldsymbol{X}_i), 1 - F_T(t|\boldsymbol{X}_i)) \end{aligned} \tag{2}$$

Next we delineate the marginal models for $Y_i^{(l)}$ and T_i. We employ the generalized linear models (GLMs) for the specification of the ultimate payment $Y_i^{(l)}$, for $l \in \{1, \ldots, k\}$, and an accelerated

failure time model (AFT) for the settlement time T_i. Specifically, the marginal model for the loss amount is expressed as:

$$\begin{aligned} Y_i^{(l)}|\boldsymbol{X}_i &\sim \text{ED}(\mu_i^{(l)}, \phi) \\ \text{E}(Y_i^{(l)}|\boldsymbol{X}_i) &= \mu_i = s_l(\boldsymbol{X}_i; \boldsymbol{\alpha}^{(l)}) \\ \text{Var}(Y_i^{(l)}|\boldsymbol{X}_i) &= \phi_l V(\mu_i^{(l)}) \end{aligned}$$

where ED denotes the exponential dispersion family parameterized by the mean μ and dispersion ϕ. Furthermore, the mean is modeled through a smooth function $s(\cdot; \boldsymbol{\alpha})$ with parameters $\boldsymbol{\alpha}$, while the dispersion is considered constant. The marginal model for the settlement time is given by:

$$\begin{aligned} \log(T_i) &= \zeta(\boldsymbol{X}_i; \boldsymbol{\beta}) + \sigma W_i \\ W_i|\boldsymbol{X}_i &\overset{i.i.d.}{\sim} F_0 \end{aligned}$$

where F_0 belongs to a log-location-scale family. Under the AFT model, we have

$$F_T(t|\boldsymbol{X}_i) = F_0\left(\frac{\log(t) - \zeta(\boldsymbol{X}_i; \boldsymbol{\beta})}{\sigma}\right)$$

In above, $\zeta(\cdot; \boldsymbol{\beta})$ is a smooth function parameterized by $\boldsymbol{\beta}$. In both the GLM and AFT, we consider the special case of a linear model, i.e., $s_l(\boldsymbol{X}_i; \boldsymbol{\alpha}^{(l)}) = \boldsymbol{X}_i'\boldsymbol{\alpha}^{(l)}$ and $\zeta(\boldsymbol{X}_i; \boldsymbol{\beta}) = \boldsymbol{X}_i'\boldsymbol{\beta}$. Note that an intercept should be included in the linear model.

Lastly, we utilizes a $(k+1)$-variate Gaussian copula with an unstructured correlation matrix. We denote the Gaussian copula as:

$$H(u_1, \ldots, u_k, u_{k+1}) = \Phi_{k+1}(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_k), \Phi^{-1}(u_{k+1}); \boldsymbol{\Sigma}) \tag{3}$$

where $\boldsymbol{\Sigma}$ is the $(k+1) \times (k+1)$ correlation matrix. The Gaussian copula is most commonly used in multivariate analysis. The selection of the copula is a balance of interpretability, model complexity, and computational efficiency. In particular, the following two properties of the Gaussian copula are desirable for our method. First, the copula is symmetric such that $\bar{H} = H$. Second, consider the partial derivatives for $l = 1, \ldots, k-1$:

$$\partial_{l:T}\bar{H}(u_1, \ldots, u_k, u_{k+1}) = \frac{\partial^{l+1}}{\partial u_1 \cdots \partial u_l \partial u_{k+1}}\bar{H}(u_1, \ldots, u_k, u_{k+1})$$

Straightforward calculations show:

$$\begin{aligned}
&\partial_{l:T}\bar{H}(u_1,\ldots,u_k,u_{k+1}) \\
=&\Phi_{k-l}((\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{\mathbf{11}}^{-1}\boldsymbol{\Sigma}_{21}')^{-1/2}(\boldsymbol{z}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z}_1))\frac{\phi_{l+1}(\boldsymbol{z}_1)}{\phi(\Phi^{-1}(u_1))\cdots\phi(\Phi^{-1}(u_l))\phi(\Phi^{-1}(u_{k+1}))}
\end{aligned}$$

where for $l=1,\ldots,k-1$:

$$\begin{aligned}
\boldsymbol{z}_1&=(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_l),\Phi^{-1}(u_{k+1})) \\
\boldsymbol{z}_2&=(\Phi^{-1}(u_{l+1}),\ldots,\Phi^{-1}(u_k))
\end{aligned}$$

and

$$\boldsymbol{\Sigma}_{11}=\begin{pmatrix}
1 & \rho_{12} & \cdots & \rho_{1l} & \rho_{1,k+1} \\
\rho_{12} & 1 & \cdots & \rho_{2l} & \rho_{2,k+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\rho_{1l} & \rho_{2l} & \cdots & 1 & \rho_{l,k+1} \\
\rho_{1,k+1} & \rho_{2,k+1} & \cdots & 1 & 1
\end{pmatrix},$$

$$\boldsymbol{\Sigma}_{21}=\begin{pmatrix}
\rho_{1,l+1} & \cdots & \rho_{l,l+1} & \rho_{l+1,k+1} \\
\vdots & \ddots & \vdots & \vdots \\
\rho_{1,k} & \cdots & \rho_{l,k} & \rho_{k,k+1}
\end{pmatrix},\ \boldsymbol{\Sigma}_{22}=\begin{pmatrix}
1 & \rho_{l+1,l+2} & \cdots & \rho_{l+1,k} \\
\rho_{l+1,l+2} & 1 & \cdots & \rho_{l+2,k} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{l+1,k} & \rho_{l+2,k} & \cdots & 1
\end{pmatrix}.$$

3.3 Estimation

In practice, it is common for reserving models to be trained prior to the valuation date, where the insurer's book of claims typically comprises both open and closed claims. Let's consider a portfolio of n insurance claims. We denote C_i as the valuation time (from the reporting to the valuation) associated with the ith claim ($i=1,\ldots,n$). It's worth noting that for a common valuation time for the portfolio, C_i could be claim-specific due to different reporting times. We define $\Delta_i=\mathbf{1}(T_i\leq C_i)$. Furthermore, we define $\tilde{T}_i=\min\{T_i,C_i\}$ and $\tilde{Y}_i^{(l)}=Y_i^{(l)}(\tilde{T}_i)=\min\{Y_i^{(l)}(T_i),Y_i^{(l)}(C_i)\}$. Denote $\tilde{\boldsymbol{Y}}_i=(\tilde{Y}_i^{(1)},\ldots,\tilde{Y}_i^{(k)})'$ and $\boldsymbol{\eta}_i=(\eta_i^{(1)},\ldots,\eta_i^{(k)})$. To facilitate presentation of the density and survival functions for different coverage combination in a general form, we define $\boldsymbol{v}=(v_1,\ldots,v_k)$, where

$v_l \in 0, 1$ for $l = 1, \ldots, k$ (analogous to $\boldsymbol{\eta}_i$), and introduce :

$$g(u_1, \ldots, u_k, u_{k+1}|\boldsymbol{v}) = \frac{\partial^{\sum_{l=1}^{k} v_l + 1}}{\partial u_1^{v_1} \cdots, \partial u_k^{v_k} \partial u_{k+1}} H(u_1^{v_1}, \ldots, u_k^{v_k}, u_{k+1}) \tag{4}$$

$$\bar{G}(u_1, \ldots, u_k, u_{k+1}|\boldsymbol{v}) = \bar{H}(u_1^{v_1}, \ldots, u_k^{v_k}, u_{k+1}). \tag{5}$$

Denote the vector of model parameters as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \boldsymbol{\theta}_T, \boldsymbol{\theta}_\Sigma)$, where $\boldsymbol{\theta}_l$, $l = 1, \ldots, k$, represents the parameter vector in the marginal model of $Y_i^{(l)}$, $\boldsymbol{\theta}_T$ is the parameter vector in the marginal model of T_i, and $\boldsymbol{\theta}_\Sigma$ denotes the parameter vector in the copula. Let $\mathcal{D}_n = \{\boldsymbol{X}_i, \delta_i, \boldsymbol{\eta}_i, \tilde{t}_i, \tilde{\boldsymbol{y}}_i : i = 1, \ldots, n\}$ be the data available by the valuation, which contains information on valuation time, claim status, and settlement time for closed claims, along with baseline covariates. The overall loglikelihood function for the data can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{ \delta_i \log f(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(k)}, \tilde{t}_i|\boldsymbol{X}_i, \boldsymbol{\eta}_i) + (1 - \delta_i) S(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(k)}, \tilde{t}_i|\boldsymbol{X}_i, \boldsymbol{\eta}_i) \right\}, \tag{6}$$

where

$$\begin{aligned} f(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(k)}, \tilde{t}_i|\boldsymbol{X}_i, \boldsymbol{\eta}_i) &= f_T(\tilde{t}_i|\boldsymbol{X}_i) \prod_{l=1}^{k} \left\{ f_l(\tilde{y}_i^{(l)}|\boldsymbol{X}_i) \right\}^{\eta_i^{(l)}} \\ &\quad \times g(F_1(\tilde{y}_i^{(1)}|\boldsymbol{X}_i), \ldots, F_k(\tilde{y}_i^{(k)}|\boldsymbol{X}_i), F_T(\tilde{t}_i|\boldsymbol{X}_i)|\boldsymbol{\eta}_i), \\ S(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(k)}, \tilde{t}_i|\boldsymbol{X}_i, \boldsymbol{\eta}_i) &= \bar{G}(1 - F_1(\tilde{y}_i^{(1)}|\boldsymbol{X}_i), \ldots, 1 - F_k(\tilde{y}_i^{(k)}|\boldsymbol{X}_i), 1 - F_T(\tilde{t}_i|\boldsymbol{X}_i)|\boldsymbol{\eta}_i). \end{aligned}$$

The model parameters can be estimated using the full information likelihood approach by directly maximizing the overall log-likelihood function defined by (6). However, the maximum likelihood estimation could be computationally inefficient and impractical. To address this issue, we propose a stage-wise estimation method to implement the full information likelihood in a computationally efficient manner. Let's denote $\boldsymbol{\theta}_{\Sigma_{YT}} = \{\theta_{lT} : l = 1, \ldots, k\}$ where θ_{lT} describes the dependence between $Y_i^{(l)}$ and T_i, and $\boldsymbol{\theta}_{\Sigma_{YY}} = \{\theta_{ll'} : l, l' = 1, \ldots, k \text{ and } l < l'\}$ where $\theta_{ll'}$ describes the dependence between $Y_i^{(l)}$ and $Y_i^{(l')}$. Further denote $\boldsymbol{\theta}_\Sigma = (\boldsymbol{\theta}_{\Sigma_{YT}}, \boldsymbol{\theta}_{\Sigma_{YY}})$. Let $H_{l:T}$ be the bivariate copula associated with $(Y_i^{(l)}, T_i)$, i.e.

$$H_{l:T}(u_l, u_{k+1}) = H(1, \ldots, 1, u_l, 1, \ldots, 1, u_{k+1}), \text{ for } l = 1, \ldots, k.$$

And let $h_{l:T}$ and $\bar{H}_{l:T}$ represent the corresponding density and survival copula of $H_{l:T}$. In a similar manner, we define the trivariate copula $H_{(l,l'):T}$, copula density $h_{(l,l'):T}$, and survival copula $\bar{H}_{(l,l'):T}$

for a triplet $(Y_i^{(l)}, Y_i^{(l')}, T_i)$. The stage-wise estimation procedure is summarized as follows:

1. Estimate parameter $\boldsymbol{\theta}_T$ in the marginal model for the settlement time T_i by

$$\hat{\boldsymbol{\theta}}_T = \arg\max \mathcal{L}_T(\boldsymbol{\theta}_T)$$

where

$$\mathcal{L}_T(\theta_T) = \sum_{i=1}^{n} \left\{ \delta_i \log f_T(\tilde{t}_i|\boldsymbol{X_i}) + (1-\delta_i) \log(1 - F_T(\tilde{t}_i|\boldsymbol{X_i})) \right\}.$$

2. For $l = 1, ..., k$, and $\eta_i^{(l)} = 1$, estimate parameter $\boldsymbol{\theta}_l$ in the marginal model for $Y_i^{(l)}$ and the dependence parameter θ_{lT} in the bivariate copula for pair $(Y_i^{(l)}, T_i)$ simultaneously. Specifically, estimates are obtained by maximizing the likelihood function based on the conditional distribution of $Y_i^{(l)}|T_i$ while fixing the estimates $\hat{\boldsymbol{\theta}}_T$ from the first stage. That is,

$$\hat{\boldsymbol{\theta}}_l, \hat{\theta}_{lT} = \arg\max \mathcal{L}_l(\boldsymbol{\theta}_l; \hat{\boldsymbol{\theta}}_T),$$

where

$$\begin{aligned}\mathcal{L}_l(\boldsymbol{\theta}_l, \theta_{lT}; \hat{\boldsymbol{\theta}}_T) \propto & \sum_{i=1}^{n} \delta_i \left\{ \log f_l(\tilde{y}_i^{(l)}|\boldsymbol{X_i}) + \log h_{l:T}(F_l(\tilde{y}_i^{(l)}|\boldsymbol{X_i}), F_T(\tilde{t}_i|\boldsymbol{X_i})) \right\} \\ & + \sum_{i=1}^{n} (1-\delta_i) \log \bar{H}_{l:T}(1 - F_l(\tilde{y}_i^{(l)}|\boldsymbol{X_i}), 1 - F_T(\tilde{t}_i|\boldsymbol{X_i})).\end{aligned}$$

3. Fixing the estimated parameters from the previous two stages, estimate dependence parameters $\boldsymbol{\theta}_{\Sigma_{YY}}$ based on the full likelihood by

$$\hat{\boldsymbol{\theta}}_{\Sigma_{YY}} = \arg\max \mathcal{L}(\boldsymbol{\theta}_{\Sigma_{YY}}; \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\theta}}_{\Sigma_{YT}}),$$

where

$$\begin{aligned}& \mathcal{L}(\boldsymbol{\theta}_{\Sigma_{YY}}; \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\theta}}_T, \hat{\boldsymbol{\theta}}_{\Sigma_{YT}}) \\ \propto & \sum_{i=1}^{n} \delta_i \log g(F_1(\tilde{y}_i^{(1)}|\boldsymbol{X}_i), \ldots, F_k(\tilde{y}_i^{(k)}|\boldsymbol{X}_i), F_T(\tilde{t}_i|\boldsymbol{X}_i)|\boldsymbol{\eta}_i) \\ & + \sum_{i=1}^{n} (1-\delta_i) \log \bar{G}(1 - F_1(\tilde{y}_i^{(1)}|\boldsymbol{X}_i), \ldots, 1 - F_k(\tilde{y}_i^{(k)}|\boldsymbol{X}_i), 1 - F_T(\tilde{t}_i|\boldsymbol{X}_i)|\boldsymbol{\eta}_i).\end{aligned}$$

To further improve computational efficiency, dependence parameters $\boldsymbol{\theta}_{\Sigma_{YY}}$ in the last stage can be estimated pairwisely based on the conditional distribution of $(Y_i^{(l)}, Y_i^{(l')})|T_i$. That is,

for $l, l' = 1, \ldots, k$, $l < l'$, and $\eta_i^{(l)} = \eta_i^{(l')} = 1$, we have:

$$\hat{\theta}_{ll'} = \arg\max \mathcal{L}_{ll'}(\theta_{ll'}; \hat{\boldsymbol{\theta}}_l, \hat{\boldsymbol{\theta}}_{l'}, \hat{\boldsymbol{\theta}}_T, \hat{\theta}_{lT}, \hat{\theta}_{l'T}),$$

where

$$\begin{aligned}
&\mathcal{L}_{ll'}(\theta_{ll'}; \hat{\boldsymbol{\theta}}_l, \hat{\boldsymbol{\theta}}_{l'}, \hat{\boldsymbol{\theta}}_T, \hat{\theta}_{lT}, \hat{\theta}_{l'T}) \\
\propto &\sum_{i=1}^{n} \delta_i \log h_{(l,l'):T}(F_l(\tilde{y}_i^{(l)}|\boldsymbol{X}_i), F_{l'}(\tilde{y}_i^{(l')}|\boldsymbol{X}_i), F_T(\tilde{t}_i|\boldsymbol{X}_i)) \\
&+ \sum_{i=1}^{n} (1-\delta_i) \log \bar{H}_{(l,l'):T}(1 - F_l(\tilde{y}_i^{(l)}|\boldsymbol{X}_i), 1 - F_{l'}(\tilde{y}_i^{(l')}|\boldsymbol{X}_i), 1 - F_T(\tilde{t}_i|\boldsymbol{X}_i)).
\end{aligned}$$

It's worth noting that the stage-wise estimation strikes a balance between statistical efficiency and computational efficiency. To further enhance efficiency, one could either iterate the stage-wise procedure multiple times or utilize the stage-wise estimates as initial values in the full likelihood approach. This approach allows for flexibility in optimizing computational resources while still achieving reliable parameter estimates.

3.4 Dynamic Prediction

In the context of loss reserving, the objective is to predict the ultimate loss for each open claim as well as for the entire portfolio. To set appropriate reserves, reserving actuaries aim not only to provide a point prediction of outstanding payments but also to quantify the uncertainty surrounding reserves accurately. With this goal in mind, we delve into the predictive distribution of outstanding payments. We introduce a simulation-based algorithm enabling analysts to derive the predictive loss distribution for individual claims using the proposed multivariate reserving model. This approach equips actuaries with the tools needed to make informed decisions while accounting for the inherent uncertainty in loss reserving.

Let τ be the valuation time for an open claim. Define $y_\tau^{(l)} = Y_i^{(l)}(\tau)$, $l = 1, \ldots, k$, to be the cumulative paid losses from coverage type l by time τ. Without loss of generality, assume $\eta^{(l)} = 1$ for $l \in \{1, \ldots, k\}$. Our prediction relies on the conditional multivariate predictive distribution of

$Y_i^{(1)}, \ldots, Y_i^{(k)}, T_i$ given $(Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i > \tau, ; \boldsymbol{X}_i)$, which can be derived as:

$$
\begin{aligned}
&S_\tau(y_1 + y_\tau^{(1)}, \ldots, y_k + y_\tau^{(k)}, t + \tau | y_\tau^{(1)}, \ldots, y_\tau^{(k)}, \tau; \boldsymbol{X}_i) \\
=&\Pr(Y_i^{(1)} > y_1 + y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_k + y_\tau^{(k)}, T_i > t + \tau | Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i > \tau; \boldsymbol{X}_i) \\
=&\frac{\bar{H}(1 - F_1(y_1 + y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, 1 - F_k(y_k + y_\tau^{(k)} | \boldsymbol{X}_i), 1 - F_T(t + \tau | \boldsymbol{X}_i))}{\bar{H}(1 - F_1(y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, 1 - F_k(y_\tau^{(k)} | \boldsymbol{X}_i), 1 - F_T(\tau | \boldsymbol{X}_i))}
\end{aligned} \tag{7}
$$

for $y_1 \geq 0, \ldots, y_k \geq 0, t \geq 0$. In principle, the predictive distribution can be derived for the settlement time and ultimate losses for each coverage type from (7). However, it is mathematically involved and the explicit form is not readily available, which further complicates deriving the predictive distribution of outstanding payments for the entire portfolio. Consequently, we obtain the predictive distribution of these outcomes using Monte Carlo simulation. We employ a sequential approach based on the following decomposition of the conditional joint distribution of $Y_i^{(1)}, \ldots, Y_i^{(k)}, T_i$ given $(Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i > \tau, ; \boldsymbol{X}_i)$ as below:

$$
\begin{aligned}
&f_\tau(y_1 + y_\tau^{(1)}, \ldots, y_k + y_\tau^{(k)}, t + \tau | Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i > \tau; \boldsymbol{X}_i) \\
=&f(t + \tau | Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i > \tau; \boldsymbol{X}_i) \\
&\times f(y_1 + y_\tau^{(1)} | Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i = t + \tau; \boldsymbol{X}_i) \\
&\times f(y_2 + y_\tau^{(2)} | Y_i^{(1)} = y_1 + y_\tau^{(1)}, Y_i^{(2)} > y_\tau^{(2)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i = t + \tau; \boldsymbol{X}_i) \\
&\vdots \\
&\times f(y_k + y_\tau^{(k)} | Y_i^{(1)} = y_1 + y_\tau^{(1)}, \ldots, Y_i^{(k-1)} = y_{k-1} + y_\tau^{(k-1)}, Y_i^{(k)} > y_\tau^{(k)}, T_i = t + \tau; \boldsymbol{X}_i).
\end{aligned}
$$

We summarize the procedure for generating realizations of $(Y_i^{(1)}, \ldots, Y_i^{(k)}, T_i)$ from the predictive distribution (7) in the following algorithmic manner:

(1) Generate $T_i = t + \tau$ from distribution

$$
\begin{aligned}
&S(t + \tau | Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i > \tau; \boldsymbol{X}_i) \\
=&\frac{\bar{C}(\bar{F}_1(y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, \bar{F}_k(y_\tau^{(k)} | \boldsymbol{X}_i), \bar{F}_T(t + \tau | \boldsymbol{X}_i))}{\bar{C}(\bar{F}_1(y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, \bar{F}_k(y_\tau^{(k)} | \boldsymbol{X}_i), \bar{F}_T(\tau | \boldsymbol{X}_i))}.
\end{aligned}
$$

(2) Generate $Y_i^{(1)} = y_1 + y_\tau^{(1)}$ given $T_i = t + \tau$ from distribution

$$\begin{aligned}
&S(y_1 + y_\tau^{(1)} | Y_i^{(1)} > y_\tau^{(1)}, \ldots, Y_i^{(k)} > y_\tau^{(k)}, T_i = t + \tau; \boldsymbol{X}_i) \\
=&\frac{\partial_T \bar{C}(\bar{F}_1(y_1 + y_\tau^{(1)} | \boldsymbol{X}_i), \bar{F}_2(y_\tau^{(2)} | \boldsymbol{X}_i) \ldots, \bar{F}_k(y_\tau^{(k)} | \boldsymbol{X}_i), \bar{F}_T(t + \tau | \boldsymbol{X}_i))}{\partial_T \bar{C}(\bar{F}_1(y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, \bar{F}_k(y_\tau^{(k)} | \boldsymbol{X}_i), \bar{F}_T(t + \tau | \boldsymbol{X}_i))}.
\end{aligned}$$

(3) For $l = 1$, generate $Y_i^{(l+1)} = y_{l+1} + y_\tau^{(l+1)}$ given $Y_i^{(1)} = y_1 + y_\tau^{(1)}, \ldots, Y_i^{(l)} = y_l + y_\tau^{(l)}$ and $T_i = t + \tau$ from distribution

$$\begin{aligned}
&S(y_{l+1} + y_\tau^{(l+1)} | Y_i^{(1)} = y_1 + y_\tau^{(1)}, \ldots, Y_i^{(l)} = y_l + y_\tau^{(l)}, Y_i^{(l+1)} > y_\tau^{(l+1)}, \cdots, Y_i^{(k)} > y_\tau^{(k)}, T_i = t + \tau; \boldsymbol{X}_i) \\
=&\frac{\partial_{l:T} \bar{C}(\bar{F}_1(y_1 + y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, \bar{F}_{l+1}(y_{l+1} + y_\tau^{(l+1)} | \boldsymbol{X}_i), \bar{F}_{l+2}(y_\tau^{(l+2)} | \boldsymbol{X}_i), \cdots, \bar{F}_k(y_\tau^{(k)} | \boldsymbol{X}_i), \bar{F}_T(t + \tau | \boldsymbol{X}_i))}{\partial_{l:T} \bar{C}(\bar{F}_1(y_1 + y_\tau^{(1)} | \boldsymbol{X}_i), \ldots, \bar{F}_l(y_l + y_\tau^{(l)} | \boldsymbol{X}_i), \bar{F}_{l+1}(y_\tau^{(l+1)} | \boldsymbol{X}_i), \ldots, \bar{F}_k(y_\tau^{(k)} | \boldsymbol{X}_i), \bar{F}_T(t + \tau | \boldsymbol{X}_i))}
\end{aligned}$$

(4) Set $l \leftarrow l + 1$, go to step (3). Stop when $l = k$. In the questions provided earlier, we define $\bar{F} = 1 - F$. This algorithm enables the generation of a random sample of settlement time and ultimate loss amount by coverage type for each individual open claim. From these samples, we can derive the predictive distribution for these outcomes using non-parametric methods, and similarly the predictive distribution of outstanding payments for the entire insurance portfolio.

4 Numerical Experiments

We conduct two sets of numerical experiments to investigate the operating characteristics of the proposed methodology. Within the context of loss reserving applications, the first set focuses on the finite sample performance of the proposed stage-wise estimation technique, and the second set aims to assess the prediction performance of the copula-based multivariate regression.

In the data generating process, we assume there are $k = 3$ types of coverage for each insurance claim. We consider a shared vector of predictors for both settlement time and loss amount. Let $\boldsymbol{X}_i = (X_{i1}, X_{i2})$ be the predictors and assume $X_{i1} \sim Bernoulli(0.4)$ and $X_{i2} \sim Normal(0, 1)$.

We use a Gamma GLM for the ultimate claim amount of each type. Specifically, the marginal models for $Y_i^{(l)}$, $l = 1, 2, 3$, are specified as below:

$$\begin{aligned}
Y_i^{(l)} | \boldsymbol{X}_i &\sim Gamma\left(\mu_i^{(l)}, \phi^{(l)}\right) \\
\mu_i^{(l)} &= \exp\left\{\alpha_0^{(l)} + \alpha_1^{(l)} X_{i1} + \alpha_2^{(l)} X_{i2}\right\}
\end{aligned}$$

We set the mean parameters as $(\alpha_0^{(1)}, \alpha_1^{(1)}, \alpha_2^{(1)}) = (5.00, 2.50, 0.50)$, $(\alpha_0^{(2)}, \alpha_1^{(2)}, \alpha_2^{(2)}) = (4.50, 1.50, 0.05)$, and $(\alpha_0^{(3)}, \alpha_1^{(3)}, \alpha_2^{(3)}) = (6.00, 2.00, 0.50)$. The dispersion parameters are set to be $(\phi^{(1)}, \phi^{(2)}, \phi^{(3)}) = (0.20, 0.10, 0.15)$.

We use Weibull AFT for the settlement time. The marginal model for T_i is specified as:

$$\log(T_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \sigma W_i$$
$$W_i | \boldsymbol{X}_i \sim F_0(w) = 1 - \exp\{-\exp(w)\}$$

Here W_i is the Gumbel (Extreme Value) distribution. Under this formulation, T_i is Weibull distribution with:

$$F_T(t|\boldsymbol{X}_i) = 1 - \exp\left\{-\left(\frac{t}{\exp(\boldsymbol{X}_i'\boldsymbol{\beta})}\right)^{\frac{1}{\sigma}}\right\}$$
$$h_T(t|\boldsymbol{X}_i) = h_0(t)\exp\left\{-\boldsymbol{X}_i'\boldsymbol{\beta}/\sigma\right\} = \frac{1}{\sigma}t^{\frac{1}{\sigma}-1}\exp\left\{-\boldsymbol{X}_i'\boldsymbol{\beta}/\sigma\right\}$$

Furthermore, the AFT model is a proportional hazard regression. We set location parameters $(\beta_0, \beta_1, \beta_2) = (4.50, 0.25, 0.05)$ and scale parameter $\sigma = 2$.

We consider a four-dimensional Gaussian copula to model the (conditional) joint distribution of $\left(Y_i^{(1)}, Y_i^{(2)}, Y_i^{(3)}, T_i\right)$. We assume $\theta_{1T} = \theta_{2T} = \theta_{3T} := \rho_{YT}$ and $\theta_{12} = \theta_{13} = \theta_{23} := \rho_{YY}$. The correlation matrix $\boldsymbol{\Sigma}$ of the copula is expressed as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \boldsymbol{\Sigma}_{YT}' \\ \boldsymbol{\Sigma}_{YT} & \boldsymbol{\Sigma}_{YY} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{YT} & \rho_{YT} & \rho_{YT} \\ \rho_{YT} & 1 & \rho_{YY} & \rho_{YY} \\ \rho_{YT} & \rho_{YY} & 1 & \rho_{YY} \\ \rho_{YT} & \rho_{YY} & \rho_{YY} & 1 \end{pmatrix} \tag{8}$$

We consider three levels of dependence in the simulation, varying from low ($\rho_{YT} = -0.15, \rho_{YY} = 0.15$), medium ($\rho_{YT} = -0.5, \rho_{YY} = 0.5$) to high ($\rho_{YT} = -0.85, \rho_{YY} = 0.85$). We choose the above dependence parameters for illustration purposes. Nonetheless, our model can readily accommodate more flexible association such as unstructured dependence.

The described data generating process enables us to generate a data set of n claims with settlement time and ultimate loss amount from multiple coverage types. In the context of loss reserving, the copula model is formulated at the valuation time with both open and closed claims. To accommodate this scenario, we adopt a three-year (1095 days) window. For each simulated claim,

we generate an occurrence date uniformly between 0 and 1095. The claim is deemed closed if the sum of the occurrence date and the settlement time falls within three years; otherwise it remains open. This procedure essentially assumes the valuation time to be $C_i \sim Uniform(0, 1095)$. For open claims, we further generate the paid loss as of the valuation date from $Y_i(C_i)^{(l)} \sim Uniform(0, Y_i^{(l)})$ for $l = 1, 2, 3$. We use data on both open and closed claims as of the valuation date to develop the copula model and to predict outstanding payments for open claims.

4.1 Finite Sample Performance

In this experiment, we let the sample size (number of claims m) be 500 and 1000, and the association parameters to be low, medium, and high. For each scenario, we replicate the simulation 100 times, and we estimate the model parameters using the stage-wise procedure described in Section 3.3. The estimation results are summarized in Table 5 and Table 6 for different scenarios of sample size. Specifically, we report the bias and the associated standard error for each parameter. There are several observations to highlight. First, across all settings with different sample sizes and dependence levels, model parameters are estimated with a negligible bias and a small standard error. Second, as sample size increases, both bias and standard error decrease as anticipated.

Table 5: Performance of stage-wise estimation: Sample size = 500

Correlation	Low		Medium		High	
Parameters	Bias	SE	Bias	SE	Bias	SE
$\beta_0^{(T)} = 4.50$	0.0012	0.0013	0.0023	0.0013	0.0029	0.0015
$\beta_1^{(T)} = 0.25$	0.0056	0.0020	0.0049	0.0024	0.0039	0.0021
$\beta_2^{(T)} = 0.05$	0.0022	0.0010	0.0029	0.0011	0.0015	0.0010
$\sigma^{(T)} = 2.00$	0.0120	0.0035	0.0076	0.0037	0.0148	0.0036
$\alpha_0^{(1)} = 5.00$	0.0009	0.0005	0.0024	0.0006	0.0002	0.0005
$\alpha_1^{(1)} = 2.50$	0.0043	0.0008	0.0012	0.0008	0.0032	0.0008
$\alpha_2^{(1)} = 0.50$	0.0003	0.0004	0.0019	0.0005	0.0001	0.0004
$\phi^{(1)} = 0.20$	0.0013	0.0003	0.0006	0.0003	0.0014	0.0003
$\alpha_0^{(2)} = 4.50$	0.0005	0.0003	0.0007	0.0003	0.0003	0.0002
$\alpha_1^{(2)} = 1.50$	0.0005	0.0004	0.0005	0.0004	0.0015	0.0004
$\alpha_2^{(2)} = 0.50$	0.0005	0.0002	0.0007	0.0002	0.0002	0.0002
$\phi^{(2)} = 0.10$	0.0006	0.0001	0.0002	0.0002	0.0006	0.0002
$\alpha_0^{(3)} = 6.00$	0.0018	0.0004	0.0018	0.0004	0.0003	0.0004
$\alpha_1^{(3)} = 2.00$	0.0001	0.0006	0.0015	0.0007	0.0011	0.0006
$\alpha_2^{(3)} = 0.50$	0.0001	0.0003	0.0009	0.0003	0.0005	0.0003
$\phi^{(3)} = 0.15$	0.0007	0.0002	0.0009	0.0002	0.0009	0.0002
$\theta_{1,T}$	0.0012	0.0022	0.0057	0.0016	0.0031	0.0007
$\theta_{2,T}$	0.0077	0.0020	0.0024	0.0016	0.0016	0.0006
$\theta_{3,T}$	0.0076	0.0021	0.0024	0.0017	0.0007	0.0005
$\theta_{1,2}$	0.0001	0.0021	0.0026	0.0017	0.0011	0.0006
$\theta_{1,3}$	0.0035	0.0018	0.0047	0.0016	0.0012	0.0006
$\theta_{2,3}$	0.0040	0.0019	0.0024	0.0017	0.0015	0.0006

Table 6: Performance of stage-wise estimation: Sample size = 1000

Correlation	Low		Medium		High	
Parameters	Bias	SE	Bias	SE	Bias	SE
$\beta_0^{(T)} = 4.50$	0.0006	0.0007	0.0004	0.0007	0.0001	0.0007
$\beta_1^{(T)} = 0.25$	0.0017	0.0011	0.0014	0.0010	0.0021	0.0011
$\beta_2^{(T)} = 0.05$	0.0011	0.0005	0.0022	0.0005	0.0006	0.0005
$\sigma^{(T)} = 2.00$	0.0004	0.0015	0.0078	0.0014	0.0152	0.0018
$\alpha_0^{(1)} = 5.00$	0.0021	0.0003	0.0031	0.0003	0.0003	0.0003
$\alpha_1^{(1)} = 2.50$	0.0015	0.0005	0.0015	0.0004	0.0008	0.0004
$\alpha_2^{(1)} = 0.50$	0.0003	0.0002	0.0001	0.0002	0.0001	0.0002
$\phi^{(1)} = 0.20$	0.0010	0.0002	0.0013	0.0001	0.0012	0.0001
$\alpha_0^{(2)} = 4.50$	0.0003	0.0001	0.0013	0.0001	0.0002	0.0001
$\alpha_1^{(2)} = 1.50$	0.0000	0.0002	0.0002	0.0002	0.0003	0.0002
$\alpha_2^{(2)} = 0.50$	0.0003	0.0001	0.0006	0.0001	0.0000	0.0001
$\phi^{(2)} = 0.10$	0.0002	0.0001	0.0003	0.0001	0.0008	0.0001
$\alpha_0^{(3)} = 6.00$	0.0020	0.0002	0.0021	0.0002	0.0003	0.0002
$\alpha_1^{(3)} = 2.00$	0.0008	0.0003	0.0004	0.0003	0.0003	0.0003
$\alpha_2^{(3)} = 0.50$	0.0004	0.0002	0.0002	0.0001	0.0003	0.0001
$\phi^{(3)} = 0.15$	0.0002	0.0001	0.0004	0.0001	0.0012	0.0001
$\theta_{1,T}$	0.0093	0.0010	0.0057	0.0008	0.0031	0.0003
$\theta_{2,T}$	0.0060	0.0009	0.0061	0.0008	0.0025	0.0003
$\theta_{3,T}$	0.0065	0.0012	0.0045	0.0008	0.0028	0.0003
$\theta_{1,2}$	0.0015	0.0011	0.0023	0.0008	0.0017	0.0003
$\theta_{1,3}$	0.0006	0.0010	0.0015	0.0007	0.0020	0.0003
$\theta_{2,3}$	0.0009	0.0009	0.0024	0.0008	0.0017	0.0003

4.2 Prediction Accuracy

In our prediction analysis, we focus on two key quantities of interests at the valuation time τ for each claim. The first is the settlement time T_i, and the second is the ultimate losses denoted by

$W_i = Y_i^{(1)} + Y_i^{(2)} + Y_i^{(3)}$. We consider a portfolio of 2,500 simulated insurance claims. We then derive the predictive distributions of the outcomes of interest at both claim and portfolio level using the simulation method detailed in Section 3.4.

We first examine the predictive distribution of outcomes of interest at the individual claim level. Specifically, for each open claim, we obtain the predictive distribution for T_i, denoted by $\hat{F}_{T_i}(\cdot|H_i(\tau))$, where $H_i(\tau)$ denotes the historical data up to the valuation time. We employ probability integral transformation (PIT) to assess the probabilistic calibration of the predictive distribution. This entails the calculation of the PITs using the actual values of t_i as $\hat{F}_{T_i}(t_i|H_i(\tau))$. A well-calibrated predictive distribution would exhibit PITs conforming to a uniform distribution over $[0, 1]$. For visualization purposes, we compute the normal scores of the PITs using $\Phi^{-1}(\hat{F}_{T_i}(t_i|H_i(\tau)))$, with normality as the null pattern. Figure 1 exhibits the QQ plot of normal scores and the histogram of PITs. The results suggest that the predictive distribution of settlement time is probabilistically calibrated.

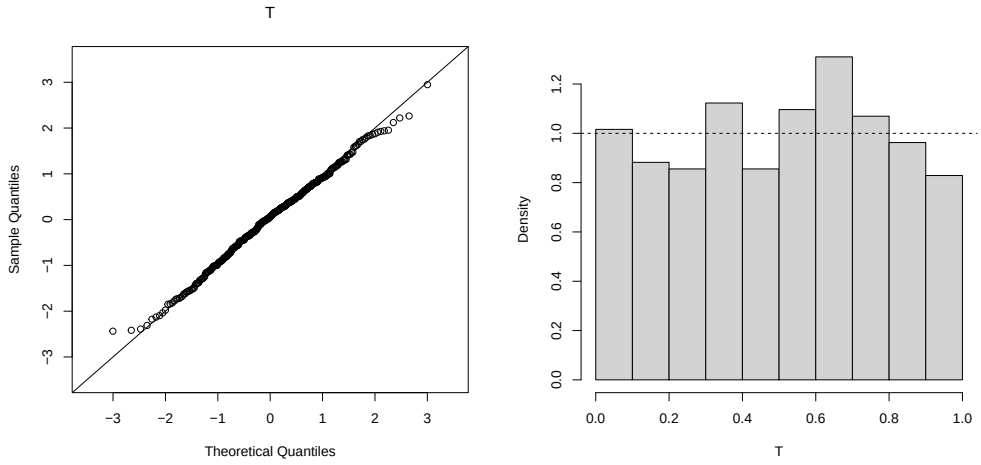

Figure 1: QQ plot of normal scores and histogram of PITs for settlement time.

We apply the same procedure to $Y_i^{(l)}$ for $l = 1, 2, 3$, and display the corresponding results in Figure 2. Similarly, the normality of the normal scores and the uniform distributions of PITs signify accurate predictions for the loss amount by coverage type. We replicate this analysis across various levels of dependence, and the findings remain consistent. To maintain brevity, we solely present the results corresponding to medium dependence.

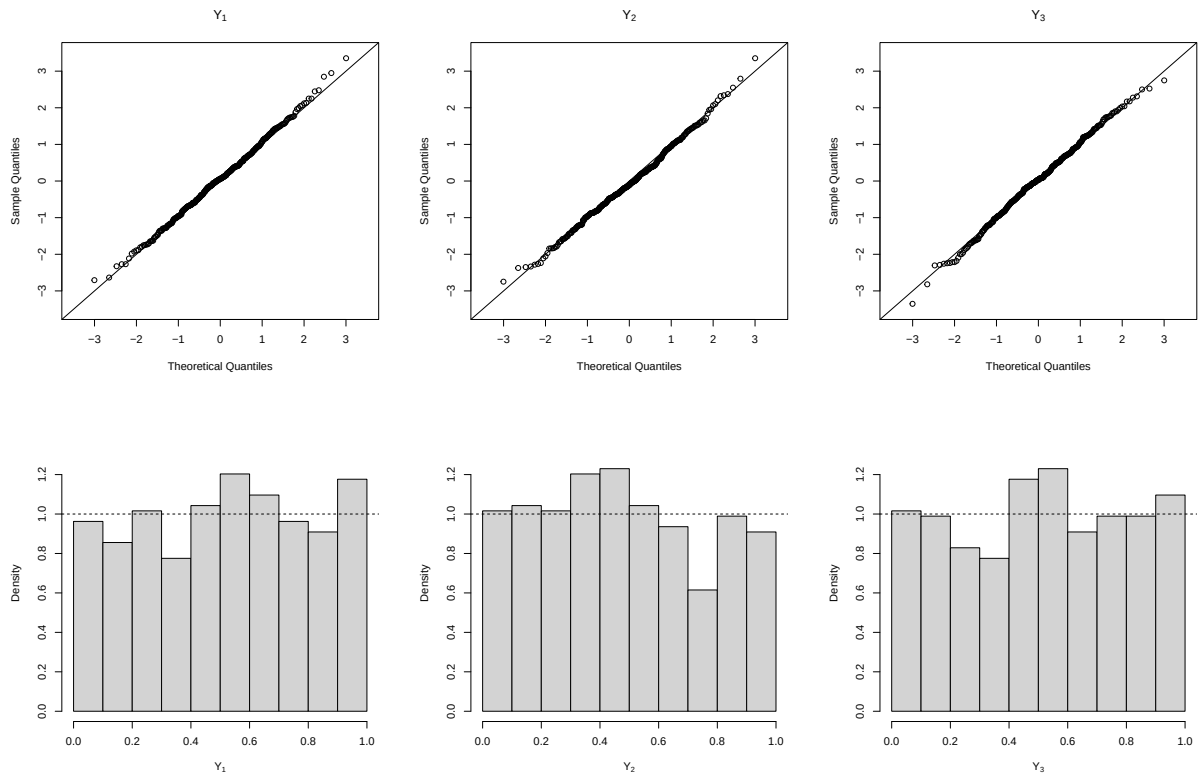

Figure 2: QQ plots of normal scores and histograms of PITs for loss amount by coverage type.

Next we examine the predictive distribution of the total losses for the entire insurance portfolio. To this end, we perform two sets of analysis to explore the effect of dependence on the predictive distribution of total losses. Firstly, we investigate the effect of the correlation between settlement time and loss amount. Specifically, we consider two scenarios corresponding to either a positive or a negative correlation, i.e. $\rho_{TY} \in \{-0.5, 0.5\}$. To maintain clarity, we set the correlation among loss amounts of different coverage types as zero, i.e. $\rho_{YY} = 0$. Figure 3 shows the predictive distribution of the portfolio losses. As a reference, the actual realized losses are displayed as the vertical dotted line. For comparison, we also present the predictive distributions obtained under the incorrect assumption of independence. It becomes evident that disregarding the correlation between settlement time and loss amount significantly biases the prediction of portfolio losses. Specifically, one might overestimate (underestimate) the outstanding liability when the settlement time is negatively (positively) related to the loss amount.

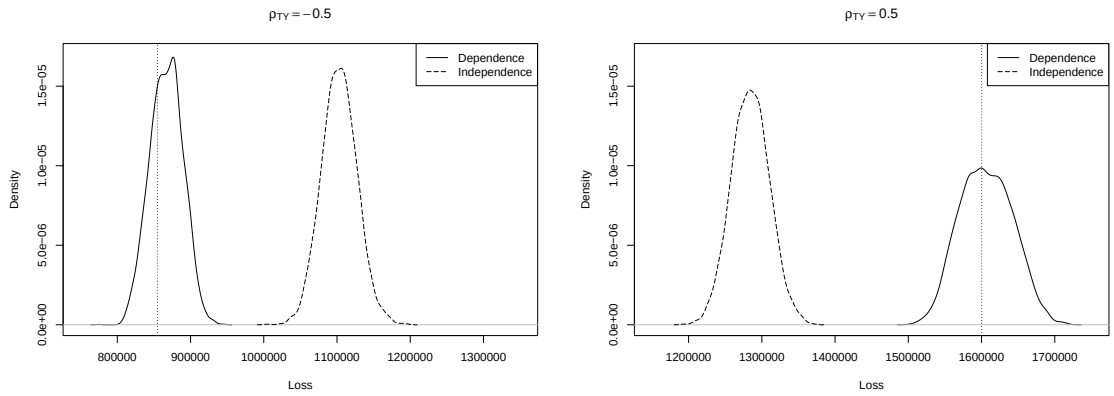

Figure 3: Predictive distributions of portfolio losses under correct dependence and incorrect independence specification between settlement time and loss amount. The left and right panels correspond to the negative and positive dependence respectively.

Secondly, we delve into the effect of the correlation among different types of losses. Similar to the previous analysis, we set $\rho_{TY} = 0$ and consider two dependence cases $\rho_{YY} \in \{-0.5, 0.5\}$ in the data generating process. For each scenario, we compare the predictive distributions of portfolio losses derived under the true dependence with those derived when incorrectly assuming independence. The outcomes are presented in Figure 4. A comparison between Figure 3 and Figure 4 suggests that the dependence among different types of losses plays a distinct role compared to the dependence between settlement time and loss amount in terms of their effect on the predictive distribution of portfolio losses. In particular, the dependence across loss types has minimal impact on the center of the predictive distribution; instead, its effect is more pronounced in terms of reserving uncertainty. That is, a negative (positive) correlation leads to smaller (larger) variation, and thus the misspecified independence model significantly overestimates (underestimates) the reserving uncertainty. In contrast, the dependence between settlement time and loss amount could substantially shift the predictive distribution while having little effect on the prediction uncertainty.

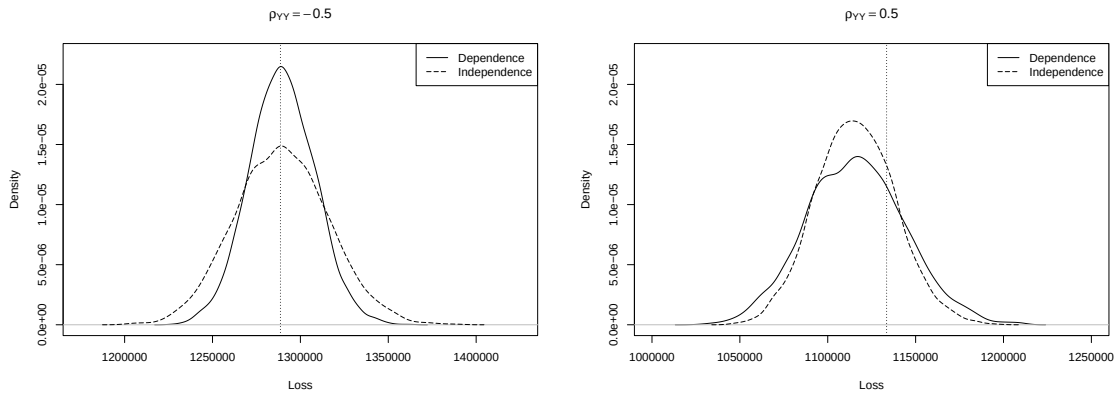

Figure 4: Predictive distributions of portfolio losses under correct dependence and incorrect independence specification among loss amount of different coverage types. The left and right panels correspond to the negative and positive dependence respectively.

5 Application

The proposed approach is calibrated on the Canadian insurance dataset described in Section 2. In the reserving context, actuaries are tasked with evaluating the outstanding liabilities as of the valuation date. To illustrate the process, we designate December 31th, 2016, as the valuation date. Utilizing the information accessible at this time, we train our model. Subsequently, we employ the outstanding payments on the open claims as test data for an out-of-sample analysis.

Specifically, there are in total 161,156 claims reported by the valuation time, out of which, 140,427 claims are closed. The model is estimated using data on both closed and open claims as of December 31th, 2016, and is employed to predict the outstanding payments of the 20,729 open claims. The actual paid losses between January 1st, 2017 and the end of our observation period July 31st, 2021 are summarized in Table 7 and will be used to assess the predictive performance of our model. The information in the "%CompDev" row represents the percentage of claims settled by the end of the observation period, indicating those for which we observed complete development within this timeframe. It is important to note that 630 claims remain open as of the end of our data collection period, July 31st, 2021. As a result, the information regarding settlement amounts for these claims is incomplete. Despite this, these files are kept in the analysis because they represent complex cases requiring a longer settlement time, which is precisely the problem tackled by this study. Therefore, the observed realization of the outstanding liability should be understood as an

approximation and the minimum amount necessary to settle all open claims.

Table 7: Summary statistics of loss development since evaluation

	Loss - APD	Loss - LU	Loss - BI	Total
Num Of Claims	19,753	13,072	6,171	20,729
%CompDev (percentage)	97.31	97.31	90.37	96.96
TotalObsAmount (in CAD)	58,365,065	2,885,441	286,882,559	348,133,065

The most predictive covariates for each coverage and the time to settlement were the province of loss and the level of responsibility in the accident. These two components have a direct influence on the level of compensation because of contract terms like the province regulation or deductibles that depend on whether the policyholder was or was not at-fault. In the case of the LU and APD coverages, there is one additional predictive covariate which is the main use of the vehicle. It is intuitive to think that how much the vehicle is used and what for have an influence over the need to get a replacement car. In the absence of information on the value of each vehicle, these predictor act as a proxy in the prediction of the material damage loss to the vehicle : business cars and vehicles that travel a great deal tend to be newer, more expensive cars. To model the time to settlement, the vehicle use is included in addition to the province and the responsibility because this factor influences how proactive the policyholder is to settle the claim.

The coverage losses and the time to settlement are modeled with Gamma and Weibull distributions. The heaviness of each coverage distribution tail is different. While BI claims are half as numerous as the other two, the outstanding liabilities associated with this coverage are more than four times higher than the combined amount of APD and LU coverages. Additionally, more BI claims are still open by the end of the observation period (90.37% vs 97.31%). The Generalized Beta distribution of the second kind was employed to model the marginal distribution of each coverage. This distribution is known to be able to accommodate heavier tails compared to the Gamma distribution and offers greater flexibility in shape. It is characterized by three shape parameters and a scale parameter, ν. Covariates are incorporated into the first shape parameter of each marginal distribution. The estimated parameters can be found in Table 8. The copula estimated parameters are presented in 9.

Table 8: Estimated parameters in the marginal models for the settlement time and loss amount

	Loss - APD	Loss - LU	Loss - BI	Settlement Time
Parameter	GB II	GB II	GB II	Weibull
β_0	-1.5646	-0.2707	-1.5325	0.5620
$\beta_{Resp.(NotFault)}$	0.0028	-0.1330	-0.3292	0.2024
$\beta_{Resp.(PartlyFault)}$	0.0371	-0.2583	0.1963	0.3055
$\beta_{Resp.(NotApp.)}$	0.1042	0.0370	0.1847	0.0738
$\beta_{Prov.(Atlantic)}$	0.0464	-0.0261	-0.2469	-0.0243
$\beta_{Prov.(Ontario)}$	0.0293	0.0797	-0.5115	-0.0001
$\beta_{Use(Pleasure)}$	-0.0087	-0.0101		0.2485
$\log(\nu)$	-4.5049	7.2740	8.9714	-1.3206
$\log(shape\,2)$	4.6028	0.4789	1.5717	
$\log(shape\,3)$	2.2963	2.1408	2.5330	

Table 9: Estimated association parameters in the Gaussian copula

Parameter	Estimate	Parameter	Estimate
$\theta_{T,Y_{APD}}$	0.7515	$\theta_{Y_{APD},Y_{LU}}$	0.7640
$\theta_{T,Y_{LU}}$	0.7496	$\theta_{Y_{APD},Y_{BI}}$	0.7063
$\theta_{T,Y_{BI}}$	0.9035	$\theta_{Y_{LU},Y_{BI}}$	0.7010

With these estimated parameters, a distribution of the outstanding liabilities for each coverage can be generated. Figure 5 displays the predictive distribution based on 1,000 replicates of the reserve by coverage and for the entire portfolio. The solid black lines represent the median of the distributions and the dotted lines the 95^{th} and 99^{th} quantiles of the distributions. It provides an estimate of the proportion of the reserve that should be asigned to each coverage. As expected, the predictive distributions tend to be higher than the censored approximations of the reserves represented by the red lines. The censored approximations of the reserves are the payments made between evaluation date (December 31th, 2016) and the end of observation period (July 31st, 2021), that is the observed loss amount in the test set. This means that the red lines in Figure 5

will move to the right, closer to the medians of the distributions once all claims are closed. The approximations from the test set (red lines) respectively stand at the 23.7^{th} quantile of the simulated distribution for the APD coverage, 45.6^{th} for LU, 39.1^{th} for BI and 38.2^{th} for all coverages. Besides, no matter the number of coverages triggered by a claim, there is only one settlement time for the whole claim. APD and LU coverages are short tail businesses. However, when a claim triggers the BI coverage in addition to these coverages, the claim might remained open because the BI component is not settled, while the APD and LU parts are likely to be paid entirely after a while. To avoid overestimating the reserve, a prediction is made only for the coverages that are going to generate a future payment recorded in the test set. It seems reasonable to believe the coverage is settled if nothing is recorded within 5.5 years after the evaluation date. This is in line with our working assumption that the coverage information is available, i.e. the insurer is able to identify coverage of all types requiring a future payment. A specific model could be used to complement the proposed approach if the context requires it.

Comparing the estimates from Table 9 with the statistics from Table 3, we can see the importance of properly considering the censoring nature of the data. All information in the database, even the unsettled claims (censored loss amounts), were used to calculate the association measures of Table 3 without considering the payments might be incomplete. However, the model structure take into account that the claim might continue to develop when estimating the association parameters of Table 9. When censoring is included in evaluating the association between the model's components, we see it is globally stronger, highlighting the risk of neglecting this aspect.

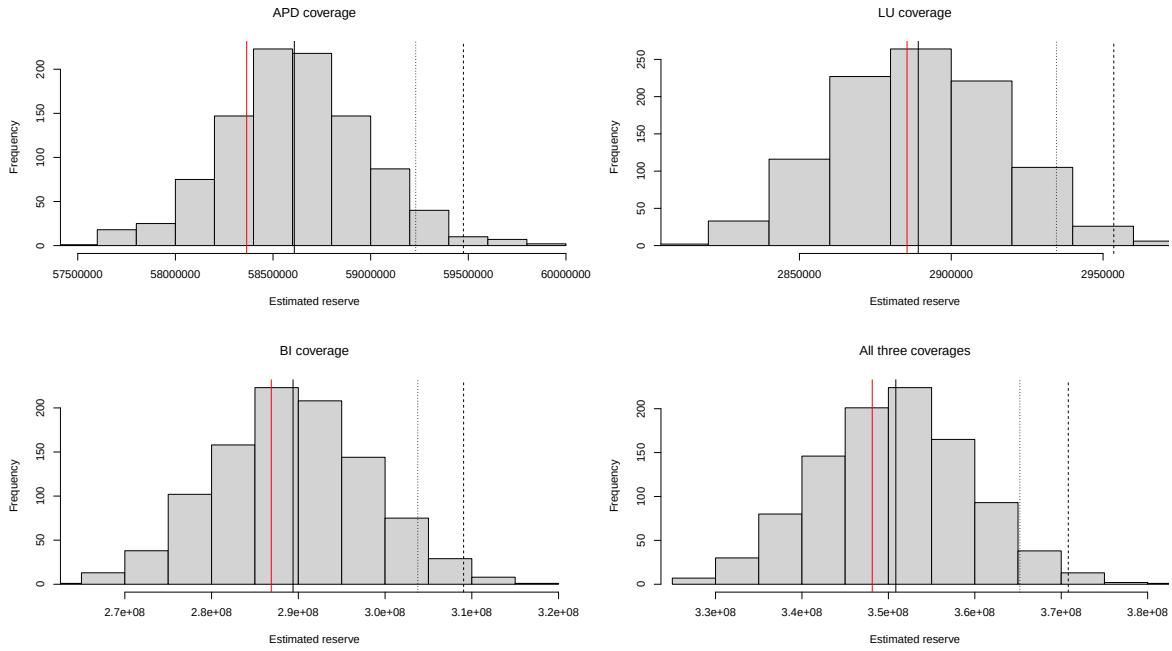

Figure 5: Predictive distribution of outstanding payments. The first three represents the payments by coverage type, and the last represents the total payments across all coverage types. The solid red line indicates the actual payments, and the solid black line indicates the median of the predictive distribution.

6 Concluding Remarks

In this paper, we have presented a copula-based granular reserving model tailored for multi-coverage insurance policies, addressing a significant gap in the literature regarding individual loss reserving methods for such policies. Our model utilizes a multivariate copula to jointly model ultimate losses from different coverage types within a claim and the settlement time of the claim. This approach not only quantifies dependence among losses of multiple coverage types but also captures the association between claim size and settlement time, a crucial relationship identified in recent literature.

By leveraging information across coverage types and settlement delays, our model enables dynamic predictions for an insurer's outstanding liability. We introduced a stage-wise estimation approach to handle challenges arising from imbalanced and censored observations in the data, and we demonstrated the efficacy of this approach through simulation studies. Moreover, we showcased the practical applicability of our model by applying it to a large portfolio of automobile insurance

claims from a Canadian insurance company, illustrating its ability to provide accurate and dynamic predictions of outstanding claim payments.

Our research highlights the importance of considering dependence among various types of losses in individual loss reserving, emphasizing its implications for prediction uncertainty and the dynamic updating of predictions. While aggregate loss reserving methods have extensively addressed dependence, our work contributes to bridging the gap in individual loss reserving methodologies, particularly for multi-coverage insurance policies.

It is important to highlight several key assumptions that the proposed method relies on. These assumptions limit the application of our approach in practice, suggesting directions for future research:

• The proposed approach focuses on reported but not settled (RBNS) claims, making it more suitable for business lines with negligible reporting delays. A separate model is needed to predict pure incurred but not reported (IBNR) claims when reporting delays are significant.

• Our model requires that all types of coverage are triggered and known to the insurer once a claim is reported. If this is not the case, one should treat each type of coverage as a separate claim, and predict the reporting of multiple coverage types.

• Our approach assumes that the cumulative paid losses of any claim is increasing over time, implying no recovery during the settlement of the claim. A stochastic process is warranted if there is significant recovery for the insurance claims.

In conclusion, our proposed copula-based granular reserving model offers a valuable tool for insurance companies to accurately estimate loss reserves for multi-coverage policies, ultimately enhancing their claims management, ratemaking, and financial reporting processes. As insurance data continue to evolve in complexity and volume, further research in this direction is warranted to advance the field of individual loss reserving and better serve the needs of the insurance industry.

References

Antonio, K. and R. Plat (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal 2014* (7), 649–669.

Arjas, E. (1989). The claims reserving problem in non-life insurance: Some structural ideas. *ASTIN Bulletin: The Journal of the IAA 19*(2), 139–152.

Avanzi, B., B. Wong, and X. Yang (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics 71*, 1–14.

Badescu, A. L., T. Chen, X. S. Lin, and D. Tang (2019). A marked cox model for the number of IBNR claims: estimation and application. *ASTIN Bulletin: The Journal of the IAA 49*(3), 709–739.

Badescu, A. L., X. S. Lin, and D. Tang (2016). A marked cox model for the number of IBNR claims: Theory. *Insurance: Mathematics and Economics 69*, 29–37.

Crevecoeur, J., K. Antonio, S. Desmedt, and A. Masquelein (2023). Bridging the gap between pricing and reserving with an occurrence and development model for non-life insurance claims. *ASTIN Bulletin: The Journal of the IAA 53*(2), 185–212.

De Jong, P. (2012). Modeling dependence between loss triangles. *North American Actuarial Journal 16*(1), 74–86.

Delong, Ł. and M. V. Wüthrich (2020). Neural networks for the joint development of individual payments and claim incurred. *Risks 8*(2), 33.

Duval, F. and M. Pigeon (2019). Individual loss reserving using a gradient boosting-based approach. *Risks 7*(3), 79.

Frees, E. W. (2015). Analytics of insurance markets. *Annual Review of Financial Economics 7*(1), 253–277.

Frees, E. W., G. Lee, and L. Yang (2016). Multivariate frequency-severity regression models in insurance. *Risks 4*(1), 4.

Frees, E. W., P. Shi, and E. A. Valdez (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin: The Journal of the IAA 39*(1), 165–197.

Frees, E. W. and E. A. Valdez (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association 103*(484), 1457–1469.

Frees, E. W. J., G. Meyers, and A. D. Cummings (2010). Dependent multi-peril ratemaking models. *ASTIN Bulletin: The Journal of the IAA 40*(2), 699–726.

Godecharle, E. and K. Antonio (2015). Reserving by conditioning on markers of individual claims: a case study using historical simulation. *North American Actuarial Journal 19*(4), 273–288.

Huang, S., J. Zhang, and W. Zhu (2023). Storm cat bond: Modeling and valuation. *North American Actuarial Journal*, DOI: 10.1080/10920277.2023.2226734.

Lopez, O., X. Milhaud, and P.-E. Thérond (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin: The Journal of the IAA 49*(3), 741–762.

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

Merz, M. and M. V. Wüthrich (2009). Prediction error of the multivariate additive loss reserving method for dependent lines of business. *Variance 3*(1), 131–151.

Michaelides, M., M. Pigeon, and H. Cossette (2023). Individual claims reserving using activation patterns. *European Actuarial Journal*, 1–33.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin: The Journal of the IAA 23*(1), 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities ii. model variations and extensions. *ASTIN Bulletin: The Journal of the IAA 29*(1), 5–25.

Okine, A. N.-A. (2023a). Individual-level loss reserving and environmental changes. *Variance 16*(1).

Okine, A. N.-A. (2023b). Ratemaking in a changing environment. *ASTIN Bulletin: The Journal of the IAA 53*(3), 596–618.

Okine, A. N.-A., E. W. Frees, and P. Shi (2022). Joint model prediction and application to individual-level loss reserving. *ASTIN Bulletin: The Journal of the IAA 52*(1), 91–116.

Shi, P., S. Basu, and G. G. Meyers (2012). A bayesian log-normal model for multivariate loss reserving. *North American Actuarial Journal 16*(1), 29–51.

Shi, P., X. Feng, J.-P. Boucher, et al. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics 10*(2), 834–863.

Shi, P. and E. W. Frees (2011). Dependent loss reserving using copulas. *ASTIN Bulletin: The Journal of the IAA 41*(2), 449–486.

Shi, P. and L. Yang (2018). Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association 113*(521), 122–133.

Taylor, G. (2019). Loss reserving models: Granular and machine learning forms. *Risks 7*(3), 82.

Taylor, G., G. McGuire, and J. Sullivan (2008). Individual claim loss reserving conditioned by case estimates. *Annals of Actuarial Science 3*(1-2), 215–256.

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal 2018*(6), 465–480.

Yang, L. and P. Shi (2019). Multiperil rate making for property insurance using longitudinal data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 182*(2), 647–668.

Yang, L., P. Shi, and S. Huang (2024). A copula model for marked point process with a terminal event: an application in dynamic prediction of insurance claims. *Working Paper*.

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

Zhao, Z., P. Shi, and X. Feng (2021). Knowledge learning of insurance risks using dependence models. *INFORMS Journal on Computing 33*(3), 1177–1196.