Multilevel Modeling of Insurance Claims Using Copulas

Peng Shi
School of Business
University of Wisconsin - Madison
Email: pshi@bus.wisc.edu

Xiaoping Feng
Department of Statistics
University of Wisconsin - Madison
Email: fengx@stat.wisc.edu

Jean-Philippe Boucher
Département de Mathématiques
Université du Québec à Montréal
Email: boucher.jean-philippe@uqam.ca

June 11, 2015

Abstract

In property and casualty insurance, claims management is featured with modeling of semi-continuous insurance cost associated with individual risk transfer. This practice is further complicated by the multilevel structure of the insurance claims data, where a contract often contains a group of policyholders, each policyholder is insured under multiple types of coverage, and the contract is repeatedly observed over time. The data hierarchy introduces complex dependence structure among claims and leads to diversification in the insurer's liability portfolio.

To capture the unique features of policy-level insurance costs, we propose a copula regression for the multivariate longitudinal claims. In the model, the Tweedie double generalized linear model is employed to examine the semi-continuous claim cost of each coverage type, and a Gaussian copula is specified to accommodate the cross-sectional and temporal dependence among the multilevel claims. Inference is based on the composite likelihood approach and the properties of parameter estimates are investigated through simulation. When applied to a portfolio of personal automobile policies from a Canadian insurer, we show that the proposed copula model provides valuable insights to an insurer's claims management process.

Keywords: Longitudinal data, Multivariate regression, Tweedie distribution, Composite likelihood, Insurance claims

1 Introduction

General insurance (aka "non-life", aka "property-casualty"), protects individuals and organizations from financial losses due to property damage or legal liabilities. It allows policyholders to exchange the risk of a large loss for the certainty of smaller periodic payments of premiums. Insurers allocates the bulk of premium dollars into investment and claims payments. As it is for an insurer to manage its investment portfolio, it is equally important for the insurer to manage its claim portfolio. Claim management is the counterpart of asset management for the claims on the insurer's book.

Claim management is the analytics of insurance costs. It requires applying statistical techniques in the analysis and interpretation of the claims data. In the data-driven industry of general insurance, claim management provides useful insights for insurers to make better business decisions. For instance, analytics helps insurers in identifying risk characteristics for risk screening in the underwriting, managing claim costs and allocating resources for claims handling, refining classification ratemaking system, as well as understanding excess layers for reinsurance and retention.

The central piece of claim management is claims modeling. In this article, we provide a general framework to look into the process of modeling and estimating insurance cost with complex structure. It is well known that the insurance cost associated with individual risk transfer presents a unique semi-continuous feature where a significant fraction of zeros is incorporated into an otherwise positive continuous outcome. The portion of zeros corresponds to no claims and the positive component corresponds to the amount of claims. Two strategies are commonly used by practitioners to analyze claim distributions, the two-part approach (see, e.g., Frees (2010)) and the pure premium approach (see, e.g., Jørgensen and de Souza (1994)). The former decomposes claims cost into frequency and severity component while the latter uses the Tweedie distribution to accommodate the mass probability at zero. Each method has its own strengths and weaknesses. In addition to the statistical considerations, the selection between the two approaches often depends on the types of data available and the preference of the analyst.

Beyond their mixed character, risk- or policy-level general insurance losses are also distinctive in that they can be viewed as the sum of losses from multiple hazard or coverage types. For example, a personal automobile insurance policy could provide both liability and collision coverage. This bundling design complicates the process of claims modeling. Insurers, on one hand, must analyze claims separately by coverage type both because of the differing contract features specific to each coverage type and because predictive dimensions generally relate differently to the various coverage types. On the other hand, insurers want to analyze the multiple types of claims jointly because they are interrelated. The first effort in this line of study is due to Frees and Valdez (2008) where the authors extended the frequency-severity model to a three-component framework to incorporate claim type.

Complex design of modern insurance products brings new challenges in modeling insurance costs. One of them is the multilevel structure often encountered in property-casualty insurance, where a contract contains a group of policyholders, each policyholder is insured under multiple types of coverage, and the contract is repeatedly observed over time. For instance, a commercial

automobile insurance policy covers both bodily injury and property damage for a fleet of vehicles, a worker's compensation contract provides indemnity cost and medical care payment for all employees of an organization, an employment-based group health insurance compensates costs of medical care utilization for office-based visits, hospital stays, and emergency room usage. The data hierarchy introduces complex dependence structure among claims and leads to diversification in the insurer's liability portfolio. The data hierarchy introduces complex dependence structure among claims and leads to diversification in the insurer's liability portfolio. In our study, the claims data are from personal automobile insurance in Ontario, Canada. An insurance policy provides coverage for the motor vehicles in a household. The number of vehicles per household ranges from one to four and each vehicle are insured under four types of coverage, accident benefit, civil liability, collision, and all risk. The portfolio is observed over a 4-year period, from 2003 to 2006. In claims modeling, one expects to capture the cluster effects (household), the cross-sectional dependence among multiple claim types, as well as the serial correlation in the longitudinal context.

Motivated by the above observations, this article further advances the claims modeling in property-casualty insurance. To capture the unique features of policy-level insurance costs, we propose a copula regression for the multivariate longitudinal claims. Specifically, for the claims cost of each type, we consider using the Tweedie distribution to accommodate the massive zeros. In the Tweedie distribution, we perform regression on both mean and dispersion using the double generalized linear model framework (Jørgensen (1987)). In the insurance claims data, all available predictors are at the risk-level, such as primary owner and vehicle characteristics. We allow the set of covariates to vary by claim type.

The multilevel structure of claims are accommodated using dependence models. We use a Gaussian copula to join the mixed outcome of claim costs. Refer to Nelsen (2006) for an introduction and Joe (2014) for recent development on copulas. For our purpose, we specify three sources of dependence: the correlation among claims from multiple vehicles within the same household, the cross-sectional dependence among multiple types of claims, and the temporal association for the longitudinal claim cost of each type. These explicit relations and their implied association are specified in the dispersion matrix of the Gaussian copula, and the dependence parameters are readily interpretable. We show that the proposed dependence model has a direct link with the mixed linear model on transformed data. Another important feature of our data is the lack of balance. The unbalanced claim costs could be due to the difference in the number of vehicles of a household, type of coverage for a vehicle, or length of observation period. The Gaussian copula provides flexibility in this sense assuming that the "missing" observations are ignorable.

Because of the mixed nature of claim costs, estimation of the Gaussian copula model using the full maximum likelihood involves multidimensional integration. For a household with four cars with each being covered by a comprehensive policy, a four-year period of observation means a $4 \times 4 \times 4 = 64$ dimensional integration. As a solution, we resort to the composite likelihood method for model estimation and comparison (see Varin et al. (2011) for an overview). Before fitting the model to the insurance data, we investigate the finite sample properties of parameter estimates using

simulation. Using the Gaussian copula and composite likelihood, statistical efficiency is sacrificed to gain the computational advantage and interpretability of the dependence parameters.

In the application of the personal automobile insurance, we examine the claims distribution at both individual and portfolio level. To provide focus, we limit our study to a simplified risk classification system. We emphasize the importance of dependence modeling and its implications on an insurer's claim management practice. We show that central limit theorem collapses when aggregating correlated risks in the portfolio.

The article is organized as follows. Section 2 describes the automobile insurance claims dataset and its important characteristics that motivate the multilevel modeling framework. Section 3 proposes the statistical model and discusses the inference based on composite likelihood method. The specification of the dependence structure in the model is detailed in the Appendix. Section 4 investigates the finite sample properties of parameter estimates using simulated data. In Section 5 we fit the model to the real data and show its implications on the insurer's claim management. Concluding remarks are provided in Section 6.

2 Data

We examine an insurance claims dataset of personal automobile insurance obtained from a property-casualty insurer in Canada. The data represent the insurer's book of business written in the province of Ontario over period 2003-2006. Both public and private insurance programs coexist in Canada. Ontario uses a private insurance system. The industry is made up of more than 100 private companies that are overseen by the government agency - Financial Services Commission of Ontario. Contrary to the public system, private insurance values actuarial approach and refined risk classification. This emphasizes the importance of the statistical analysis in our study.

As in most develop countries, automobile insurance is required for all motorists and is enforced by Ontario law. An insurance contract could provide four types of coverage: (1) "accident benefit" provides the insured with medical care payments and income replacement benefits if injured in an automobile accident, regardless of who caused the accident. (2) "civil liability" is a combined bodily injury and property damage coverage. It pays claims if the insured is liable for the bodily injury or property damage of a third party. (3) "collision" covers the losses caused when an insured vehicle is involved in a collision with another object, including another vehicle. (4) "all risk" covers the losses caused by perils other than collision, such as fire, theft, and hail etc. The former two are compulsory and are included in the standard policy. The latter two are optional and available through the comprehensive policy. Policyholders of standard and comprehensive policies often shows distinct driving behavior due to different risk levels and incentives, known as information asymmetry in the economics literature (see, e.g. Chiappori and Salanié (2013)). To provide focus, we limit our analysis to comprehensive policy, and our final sample contains 87,670 policies after some screening in the preliminary analysis.

One interesting feature of the data is its multilevel structure. The level-one unit is the insurance

policy and the level-two unit is the insured vehicle. In personal automobile insurance, it is common that a single policy is purchased to insure all vehicles within the same household. The distribution of the number of insured vehicles per policy is summarized in Table 1. About 12% of policies in our data insures more than one vehicles, among which, the majority insures two vehicles and it is rare for a policy to insure more than three vehicles. Note that this percentage is lower than the actual number of households owning multiple cars. Consider a household with three cars, two of them are insured under a standard policy and the rest one is insured under a comprehensive policy. Only the vehicle in the comprehensive policy is retained in the sample and the two vehicles in the standard policy are removed for our study. In fact, because the insurance database only contains policy ID, we do not even know that theses three vehicles are from the same household.

Table 1: Distribution of the number of insured vehicles per policy

Number of vehicles	1	2	3	4	Total
Frequency	77,352	10,058	253	7	87,670
Percentage	88.23	11.47	0.289	0.01	100

The outcome variable of our interest is the insurance claims cost. The four types of claims indicates the multivariate nature of the data. We examine insurance claim cost by coverage type and look into the vector of claims cost. Figure 1 displays their distributions. The left panel shows the violin plots using data in 2003. One noticeable feature is the semi-continuity. The massive zeros correspond to no claims. In our data, this probability is about 91% regardless of coverage type. Another observation is the long tails in the individual claims cost (Klugman et al. (2012)). This is more pronounced in the liability coverage partly due to the large legal defense cost. Data in other years exhibit consistent properties. The longitudinal nature indicates another hierarchy in the multilevel data. The right panel shows the average insurance cost over time. The accident benefit coverage shows a higher variation but in general we observe a relatively stable pattern. In our application, one can think of the average cost as the pure premium for the insurance contract. The premium shows a wide range across coverage type, with civil liability and all risk being the most and least expensive coverage respectively. This relation is also true for different risk levels as shown in the data analysis. The different distributional features shown in Figure 1 motivate the insurer to analyze claims data by coverage type.

The insurance data also contains a set of predictors that could explain the variation in claims cost. It is a common practice for property-casualty insurers using indicators in the risk classification system. Hence all predictors available are binary. Table 2 summarizes the description of these predictors and their sample averages by year. Three broad categories of covariates are commonly believed to affect insurance cost: (1) Policyholder's characteristics. Our data contains indicators on the driver's age, marital status, and whether he/she is a homeowner. Because of the nonlinear age effect, we differentiate young drivers and senior citizens. (2) Driving history. Years of experience and conviction history are used in the analysis. (3) Vehicle's characteristics. Vehicle age is an indicator of ownership at purchase. The purpose of the car indicates whether it is a lease vehicle

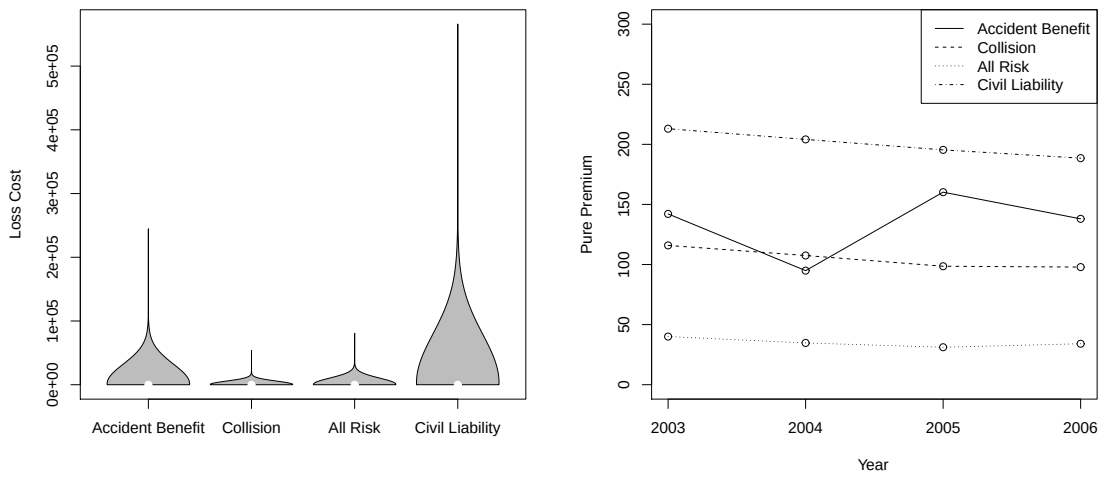

Figure 1: Distributions of claims cost by coverage type. The left panel shows the violin plot and the right panel shows the average cost over time.

and whether it is used for business. The usage of the vehicle is measured by the mileage driven and the number of drivers. For a vehicle with multiple drivers, the driver's characteristics correspond to the primary driver. As anticipated, the driver's characteristics show larger variation while the vehicle's characteristics are very consistent over time.

Table 2: Sample mean of predictors by year

Variable	Description	2003	2004	2005	2006
young	=1 if age between 16 and 25	0.028	0.022	0.018	0.016
seinor	=1 if age more than 60	0.152	0.167	0.183	0.200
marital	=1 if married	0.719	0.729	0.736	0.740
homeowner	=1 if homeowner	0.422	0.638	0.767	0.804
experience	=1 if more than ten years of experience	0.901	0.916	0.930	0.937
conviction	=1 if positive number of convictions	0.089	0.062	0.030	0.019
newcar	=1 if new car	0.895	0.896	0.896	0.897
leasecar	=1 if lease car	0.154	0.153	0.154	0.152
business	=1 if business use	0.034	0.037	0.039	0.040
highmilage	=1 if drive more than 10,000 miles	0.730	0.715	0.695	0.677
multidriver	=1 if more than two drivers	0.034	0.052	0.073	0.094

3 Modeling

3.1 Multivariate Tweedie Model

Consider an insurance portfolio consisting of N policies. For the ith $(= 1, \cdots, N)$ policy, let K_i denote the number of vehicles, J_i the number of coverage types, and T_i the number of observation

periods. Let y_{ikjt} denote the insurance cost of coverage type j in tth period for the kth vehicle in policy i. The quantity of interest is the vector of claims defined as $\boldsymbol{y}_i = (y_{ikjt})_{k=1,\cdots,K_i, j=1,\cdots,J_i, t=1,\cdots,T_i}$.

Note that y_{ikjt} follows a mixed distribution in that it consists of a discrete mass at zero and a positive continuous component. We consider the Tweedie distribution that has nonnegative support and can have a positive probability at zero (Tweedie (1984)). With appropriate parameterization, the Tweedie distribution can be shown as a member of exponential dispersion family (Jørgensen (1987)), with the density function given by

$$f(y;\mu,p,\phi) = \exp\left[\frac{1}{\phi}\left(\frac{-y}{(p-1)\mu^{p-1}} - \frac{\mu^{2-p}}{2-p}\right) + S(y;\phi)\right]$$

where

$$S(y;\phi) = \begin{cases} 0 & \text{if } y = 0 \\ \ln \sum_{n\geq 1} \left\{ \dfrac{(1/\phi)^{1/(p-1)} y^{(2-p)/(p-1)}}{(2-p)(p-1)^{(2-p)/(p-1)}} \right\}^n \dfrac{1}{n!\Gamma(n(2-p)/(p-1))y} & \text{if } y > 0 \end{cases} \tag{1}$$

With this parameterization, mean and variance of the Tweedie random variable are μ and $\phi\mu^p$, repsectively, where ϕ is the dispersion parameter and p is the power parameter that controls the variance of the distribution. This result is rather appealing because it suggests that the theories of generalized linear models are ready to apply (McCullagh and Nelder (1989)).

The Tweedie distribution becomes a Poisson distribution when $p = 1$ and a gamma distribution when $p = 2$. The more interesting range of p for our application is between 0 and 1. In this case, the Tweedie random variable can be generated from a Poisson sum of gamma random variables (Smyth (1996)). From $p = 1$ to $p = 2$, the Tweedie distribution gradually loses its mass at zero as it shifts from a Poisson distribution to a gamma distribution. The compound Poisson presentation also provides a nature interpretation for insurance claims modeling. One can think of the claims cost per year for a policyholder as sum of a series of independent gamma random variables and the number of claims in a year as a Poisson random variable.

Denote the density and cumulative distribution functions of y_{ikjt} as $f_j(y_{ikjt})$ and $F_j(y_{ikjt})$, respectively. To allow for covariates, we employ the double generalized linear model to perform regression analysis on both mean and dispersion of the Tweedie outcome. When modeling the cost of insurance claims, dispersion modeling is necessary as it increases the precision of prediction (Smyth and Jørgensen (2002)). Define $f_j(y_{ikjt}) = f(y_{ikjt};\mu_{ikjt},p_j,\phi_{ikjt})$. With log link functions, we specify

$$g_\mu(\mu_{ikjt}) = \log(\mu_{ikjt}) = \boldsymbol{x}'_{ikjt}\boldsymbol{\beta}_j,$$

$$g_\phi(\phi_{ikjt}) = \log(\phi_{ikjt}) = \boldsymbol{z}'_{ikjt}\boldsymbol{\gamma}_j.$$

Here $\boldsymbol{x}_{ikjt}$ and $\boldsymbol{z}_{ikjt}$ are vectors of covariates in the mean and dispersion regression, respectively, and $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are the associated regression coefficients. Note that we allow for different sets of covariates for the mean and dispersion. Because of the distributional differences in the coverage

types as shown in Section 2, we allow parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and p to depend on the claim type j.

Another commonly used device to incorporate mass zeros into an otherwise continuous distribution is the Tobit model (Tobin (1958)). As a censored regression, the Tobit model relies on the normality assumption. Because insurance claims are skewed and heavy-tailed, we decide not to explore this direction. However, the Tobit model can be easily extended to a multivariate context using the copula framework proposed in this section.

The multilevel structure of the insurance data is accommodated using dependence models. We use a parametric copula function to model the complex dependence embedded in the vector of claims cost. To simplify the presentation, we relabel $\boldsymbol{y}_i = (\tilde{y}_{i1}, \tilde{y}_{i2}, \cdots, \tilde{y}_{im_i})$ where $m_i = K_i \times J_i \times T_i$ denoting the total number of observations for policy i. Then the cumulative distribution function of $\boldsymbol{y}_i$ can be expressed in terms of a copula function H_i:

$$G(\boldsymbol{y_i}) = H_i(F(\tilde{y}_{i1}), \cdots, F(\tilde{y}_{im_i})) \tag{2}$$

where F is the cumulative distribution function associated with equation (1). Note that $\boldsymbol{y}_i$ is a vector of mixed random variable. Without loss of generality, assume that the first q_i components $(\tilde{y}_{i1}, \cdots, \tilde{y}_{iq_i})$ are continuous and the rest $m_i - q_i$ components $(\tilde{y}_{iq_i+1}, \cdots, \tilde{y}_{im_i})$ are discrete. The density function of $\boldsymbol{y}_i$ is shown

$$g(\boldsymbol{y_i}) = \prod_{l=1}^{q_i} f(y_l) h_i^{q_i}(F(\tilde{y}_{i1}), \cdots, F(\tilde{y}_{im_i})) \tag{3}$$

where

$$h_i^{q_i}(w_1, \cdots, w_{m_i}) = \frac{\partial^{q_i}}{\partial w_1 \cdots \partial w_{q_i}} H_i(w_1, \cdots, w_{m_i})$$

Let $m = \max\{m_1, \cdots, m_N\}$. We consider the Gaussian copula with the distributional function given by

$$H(w_1, \cdots, w_m; \boldsymbol{\Sigma}) = \Phi_m(\Phi^{-1}(w_1), \cdots, \Phi^{-1}(w_m); \boldsymbol{\Sigma})$$

where Φ_m and Φ denote the distributional function of a m-variate normal with zero mean and correlation matrix $\boldsymbol{\Sigma}$ and the standard univariate normal respectively. It can be shown that (see, e.g., Song et al. (2009))

$$\begin{aligned} & h^q(w_1, \cdots, w_m; \boldsymbol{\Sigma}) \\ = \; & (2\pi)^{-\frac{m-q}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{\Phi^{-1}(w_{q+1})} \cdots \int_{-\infty}^{\Phi^{-1}(w_m)} \exp\left\{ \frac{1}{2} (\boldsymbol{s}_1', \boldsymbol{s}_2') \boldsymbol{\Sigma}^{-1} (\boldsymbol{s}_1', \boldsymbol{s}_2')' - \frac{1}{2} \boldsymbol{s}_1' \boldsymbol{s}_1 \right\} d\boldsymbol{s}_2 \end{aligned}$$

With the Gaussian copula, the lack of balance can be easily addressed using the subclass of H and h^q. That is, for policy i, we specify $H_i(\cdot) = H(\cdot; \boldsymbol{A}_i \boldsymbol{\Sigma} \boldsymbol{A}_i')$ and $h_i^{q_i}(\cdot) = h^q(\cdot; \boldsymbol{A}_i \boldsymbol{\Sigma} \boldsymbol{A}_i')$. Here $\boldsymbol{A}_i = [\iota_1, \cdots, \iota_{m_i}, \boldsymbol{0}, \cdots, \boldsymbol{0}]_{m_i \times m}$ and ι_r is a column vector with the rth element being 1 and 0 otherwise, and $\boldsymbol{0}$ is a column vector of zeros.

The dependency among the vector of claims cost is captured by the correlation matrix $\boldsymbol{\Sigma}$ in

the Gaussian copula. In our context, one wants to accommodate three types of association, the correlation among vehicles insured under the same policy, the dependence among multiple types of claims for a given vehicle, and the temporal relationship for a particular type of coverage. To achieve these purposes, we specify $\boldsymbol{\Sigma} = \boldsymbol{B}_{K\times K} \otimes \boldsymbol{P}_{(TJ)\times(TJ)}$ with

$$\boldsymbol{B}_{K\times K} = \begin{pmatrix} 1 & \delta & \cdots & \delta \\ \delta & 1 & \cdots & \delta \\ \vdots & \vdots & \ddots & \vdots \\ \delta & \delta & \cdots & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{P}_{(TJ)\times(TJ)} = \begin{pmatrix} \sigma_{11}\boldsymbol{P}_{11} & \sigma_{12}\boldsymbol{P}_{12} & \cdots & \sigma_{1J}\boldsymbol{P}_{1J} \\ \sigma_{21}\boldsymbol{P}_{21} & \sigma_{22}\boldsymbol{P}_{22} & \cdots & \sigma_{2J}\boldsymbol{P}_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{J1}\boldsymbol{P}_{J1} & \sigma_{J2}\boldsymbol{P}_{J2} & \cdots & \sigma_{JJ}\boldsymbol{P}_{JJ} \end{pmatrix}.$$

The cluster effect is captured by an exchangeable correlation $\boldsymbol{B}_{K\times K}$ that is implied by the household-specific random effect. The dependence due to the multivariate longitudinal observations for a given vehicle is captured by $\boldsymbol{P}_{(TJ)\times(TJ)}$ where $\sigma_{jj'} = \sigma_{j'j}$ and $\boldsymbol{P}_{jj'} = \boldsymbol{P}_{j'j}$. This is a commonly used specification in models of several time series (see, e.g., Greene (2007)). $\sigma_{jj'}$ represents the cross-sectional correlation between coverage type j and j' in the same time period, known as the concurrent or contemporaneous correlation coefficient in time series analysis. $\boldsymbol{P}_{jj}$ is the serial correlation for the insurance costs of coverage j. $\boldsymbol{P}_{jj'}(j \neq j')$ is the correlation across coverage types j and j'. Note that this matrix is in general not symmetric. The diagonal elements are one and the off-diagonal elements indicate the lead-lag relationship between component series. Extending the method in Parks (1967), we specify the concurrent correlation $\sigma_{jj'}$ and serial correlation $\boldsymbol{P}_{jj}$, and let the lag corelation $\boldsymbol{P}_{jj'}$ be determined implicitly. With the AR(1) serial correlation, we have:

$$\boldsymbol{P}_{jj'} = \begin{pmatrix} 1 & \rho_{j'} & \cdots & \rho_{j'}^{T-1} \\ \rho_j & 1 & \cdots & \rho_{j'}^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{T-1} & \rho_j^{T-2} & \cdots & 1 \end{pmatrix} \quad \text{and} \quad \sigma_{jj'} = \begin{cases} 1 & \text{if } j = j' \\ \frac{\tau_{jj'}\sqrt{1-\rho_j^2}\sqrt{1-\rho_{j'}^2}}{1-\rho_j\rho_{j'}} & \text{if } j \neq j' \end{cases}.$$

We detail the specification of $\boldsymbol{\Sigma}$ and establish its connection to the linear model on transformed data in Appendix.

3.2 Inference

For inference purposes, we employ the composite likelihood method (Lindsay (1988)). Because of the mixed nature of the insurance cost, the likelihood function of model (3) involves multi-dimensional integration. In our application, an insurance contract covering four vehicles would imply a 64-dimensional integration. Thus full maximum likelihood estimation is computationally challenging and the computational difficulty increases as the number of time periods becomes larger. To minimize the computational burden, we use the pair-wise likelihood (Cox and Reid (2004)). See Varin(2008, 2011) for reviews on composite likelihood inference.

On another note, the trade-off between the computational challenge and the efficiency loss using the composite likelihood method is due to the estimation of the probability mass function

of the Gaussian copula. One alternative strategy could be to explore more flexible dependence modeling approach such as the pair-wise copula construction based on vines (see, e.g., Aas et al. (2009), Smith et al. (2010), Panagiotelis et al. (2012)). However, we find the Gaussian copula is particularly useful in our application in that it is ready to apply to the unbalanced data and the dependence parameters have intuitive interpretations. Considering the applied nature of this work, we make sacrifices to balance the interpretability, complexity, and computation of the model.

The pair-wise composite likelihood function for policy i is defined as

$$\begin{aligned} cl_i(\boldsymbol{\theta}; \boldsymbol{y}_i) &= \sum_{k=1}^{K_i} \left(\sum_{j=1}^{J_i} \sum_{t<t'} \ell(\boldsymbol{\theta}; y_{ikjt}, y_{ikjt'}) + \sum_{j<j'} \sum_{t,t'=1}^{T_i} \ell(\boldsymbol{\theta}; y_{ikjt}, y_{ikj't'}) \right) \\ &\quad + \sum_{k<k'} \sum_{j,j'=1}^{J_i} \sum_{t,t'=1}^{T_i} \ell(\boldsymbol{\theta}; y_{ikjt}, y_{ik'j't'}) \end{aligned}$$

where $\ell(\boldsymbol{\theta}; y_{ikjt}, y_{ik'j't'}) = \log(L(\boldsymbol{\theta}; y_{ikjt}, y_{ik'j't'}))$ and

$$L(\boldsymbol{\theta}; y_{ikjt}, y_{ik'j't'}) = \begin{cases} H(F_j(y_{ikjt}), F_{j'}(y_{ik'j't'}); \tilde{\rho}_{kjtk'j't'}) & \text{if } y_{ikjt} = 0 \text{ and } y_{ik'j't'} = 0 \\ f_j(y_{ikjt}) h_1(F_j(y_{ikjt}), F_{j'}(y_{ik'j't'}); \tilde{\rho}_{kjtk'j't'}) & \text{if } y_{ikjt} > 0 \text{ and } y_{ik'j't'} = 0 \\ f_{j'}(y_{ik'j't'}) h_2(F_j(y_{ikjt}), F_{j'}(y_2); \tilde{\rho}_{kjtk'j't'}) & \text{if } y_{ikjt} = 0 \text{ and } y_{ik'j't'} > 0 \\ f_j(y_{ikjt}) f_{j'}(y_{ik'j't'}) h(F_j(y_{ikjt}), F_{j'}(y_{ik'j't'}); \tilde{\rho}_{kjtk'j't'}) & \text{if } y_{ikjt} > 0 \text{ and } y_{ik'j't'} > 0 \end{cases}$$

with $\tilde{\rho}_{kjtk'j't'} = \delta^{\mathcal{I}(k \neq k')} \sigma_{jj'}^{\mathcal{I}(j \neq j')} \rho_{j'}^{\mathcal{I}(t<t')|t-t'|} \rho_j^{\mathcal{I}(t>t')|t-t'|}$. Then the total composite likelihood for the portfolio of policies can be expressed as

$$cl(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \frac{1}{m_i - 1} cl_i(\boldsymbol{\theta}; \boldsymbol{y}_i), \tag{4}$$

where $1/(m_i - 1)$ is weight assigned for the ith policy (see, e.g., Zhao and Joe (2005), Joe and Lee (2009)).

The composite likelihood estimator is defined as $\hat{\boldsymbol{\theta}}_N = \text{argmax}_{\boldsymbol{\theta}} cl(\boldsymbol{\theta}; \boldsymbol{y})$. Denote the composite score function as $S_N(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} cl_i(\boldsymbol{\theta}; \boldsymbol{y}_i)$. To estimate of the variance of $\hat{\boldsymbol{\theta}}_N$, we use the Godambe information matrix (Godambe (1960)), defined as

$$\boldsymbol{G}_N(\boldsymbol{\theta}) = \boldsymbol{R}_N(\boldsymbol{\theta}) \boldsymbol{\Omega}_N^{-1}(\boldsymbol{\theta}) \boldsymbol{R}_N(\boldsymbol{\theta}), \tag{5}$$

where $\boldsymbol{R}_N(\boldsymbol{\theta}) = -E\big(\frac{\partial}{\partial \boldsymbol{\theta}'} S_N(\boldsymbol{\theta})\big)$ and $\boldsymbol{\Omega}_N(\boldsymbol{\theta}) = \text{Var}\big(S_N(\boldsymbol{\theta})\big)$. Under regularity conditions on the bivariate log-likelihood functions, we can apply the central limit theorem to the composite likelihood score statistic, leading to the result that the composite likelihood estimator, $\hat{\boldsymbol{\theta}}_N$, is asymptotically normally distributed

$$\sqrt{N} \boldsymbol{G}_N^{1/2}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}). \tag{6}$$

The sample estimate of sensitivity matrix $\boldsymbol{R}_N(\boldsymbol{\theta})$ given by

$$\hat{\boldsymbol{R}}_N(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 cl_i(\boldsymbol{\theta};\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}},$$

and the numerical Hessian matrix is used to approximate the second order derivative. The sample estimate of variability matrix $\boldsymbol{\Omega}_N(\boldsymbol{\theta})$ is expressed by the outer product of the composite score functions as

$$\hat{\boldsymbol{\Omega}}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial cl_i(\boldsymbol{\theta};\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}}\frac{\partial cl_i(\boldsymbol{\theta};\boldsymbol{y}_i)}{\partial\boldsymbol{\theta}'}.$$

Thus the asymptotic covariance matrix can be approximated by $\hat{\boldsymbol{G}}_N^{-1}(\hat{\boldsymbol{\theta}}_N)/N$. Furthermore, model comparison is based on the composite likelihood version of AIC (Varin and Vidoni (2005)) and BIC (Gao and Song (2010)) which are respectively defined as

$$\begin{aligned} CLAIC &= -2cl(\boldsymbol{\theta};\boldsymbol{y}) + 2tr(\boldsymbol{\Omega}(\boldsymbol{\theta})\boldsymbol{R}^{-1}(\boldsymbol{\theta})).\\ CLBIC &= -2cl(\boldsymbol{\theta};\boldsymbol{y}) + \log(N)tr(\boldsymbol{\Omega}(\boldsymbol{\theta})\boldsymbol{R}^{-1}(\boldsymbol{\theta})). \end{aligned}$$

4 Simulation

The properties of the composite likelihood estimates are investigated using simulated data. In the simulation, we set $J = 2$, $K = 2$, and $T = 4$, that is, a policy covers two vehicles and provides two types of coverage for each vehicle. We consider different sample sizes (number of policies) N and report the results for $N = 200$ and 500. Data are generated from the multivariate Tweedie model in Section 3.1. In the marginal distribution, we use $Tweedie(\mu_j, p_j, \phi_j)$ with the following specification for coverage type $j = 1$ and 2:

$$\begin{aligned} \mu_j &= \exp(\beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2)\\ \phi_j &= \exp(\gamma_{j0} + \gamma_{j1}X_1 + \gamma_{j2}X_2) \end{aligned}$$

where $X_1 \sim Bernoulli(0.5)$ and $X_2 \sim Bernoulli(0.6)$ independently. In the joint distribution, we use the Gaussian copula with the correlation matrix specified as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix} \otimes \begin{pmatrix} \begin{pmatrix} 1 & \rho_1 & \rho_1^2 & \rho_1^3 \\ \rho_1 & 1 & \rho_1 & \rho_1^2 \\ \rho_1^2 & \rho_1 & 1 & \rho_1 \\ \rho_1^3 & \rho_1^2 & \rho_1 & 1 \end{pmatrix} & \sigma_{12}\begin{pmatrix} 1 & \rho_2 & \rho_2^2 & \rho_2^3 \\ \rho_1 & 1 & \rho_2 & \rho_2^2 \\ \rho_1^2 & \rho_1 & 1 & \rho_2 \\ \rho_1^3 & \rho_1^2 & \rho_1 & 1 \end{pmatrix} \\ \sigma_{12}\begin{pmatrix} 1 & \rho_1 & \rho_1^2 & \rho_1^3 \\ \rho_2 & 1 & \rho_1 & \rho_1^2 \\ \rho_2^2 & \rho_2 & 1 & \rho_1 \\ \rho_2^3 & \rho_2^2 & \rho_2 & 1 \end{pmatrix} & \begin{pmatrix} 1 & \rho_2 & \rho_2^2 & \rho_2^3 \\ \rho_2 & 1 & \rho_2 & \rho_2^2 \\ \rho_2^2 & \rho_2 & 1 & \rho_2 \\ \rho_2^3 & \rho_2^2 & \rho_2 & 1 \end{pmatrix} \end{pmatrix}$$

Here, $\sigma_{12} = \tau_1 2\sqrt{1-\rho_1^2}\sqrt{1-\rho_2^2}/(1-\rho_1\rho_2)$. The true parameters and simulation results are displayed in Table 3.

The results are based on 500 replications. For each simulated dataset, we estimate parameters by maximizing the composite likelihood function. The mean and standard deviation (SD) of these point estimates for each parameter are reported. The average estimates are very close to the corresponding true parameters for both $N = 200$ and 500. We further confirm this relation by calculating the relative bias of the estimates. As expected, increasing sample size reduces the estimation bias, and when $N = 500$ the biases for most parameters are almost zero. Next we examine the standard error of the estimator. In each replication, the standard error is estimated using the Godambe information matrix described in Section 3.2. Its average (denoted by SE in the table) is comparable with the nominal standard deviation (SD) of point estimates, indicating the accuracy of this estimator. Notice that the SD and SE are decreasing when sample size increases from $N = 200$ to $N = 500$. When we further increase the sample size to $N = 1000$, the change in both SD and SE are ignorable. This observation provides insights on the finite-sample performance of the composite likelihood estimates and suggests that their asymptotic distributions are approached with about 500 policies. Finally, we report the mean squared error (MSE) of the parameter estimates. Consistent results are observed that a larger sample size leads to more accurate estimates and that estimates with less uncertainty can be obtained with a larger number of policies.

Table 3: Simulation results for sample size (number of policies) $N = 200$ and 500

Parameter	Estimate (mean)		Relative Bias		SD		SE		MSE	
	$N = 200$	500	200	500	200	500	200	500	200	500
$\beta_{10} = 1$	0.898	0.970	-0.102	-0.030	0.361	0.244	0.369	0.234	0.141	0.060
$\beta_{11} = 1.5$	1.515	1.510	0.010	0.007	0.201	0.122	0.208	0.136	0.041	0.015
$\beta_{12} = 0.5$	0.565	0.510	0.130	0.020	0.290	0.176	0.295	0.187	0.088	0.031
$\beta_{20} = 1$	0.989	0.982	-0.011	-0.018	0.282	0.190	0.278	0.183	0.080	0.036
$\beta_{21} = 0.5$	0.494	0.502	-0.012	0.003	0.269	0.163	0.218	0.142	0.072	0.026
$\beta_{22} = 2$	2.006	1.995	0.003	-0.002	0.200	0.121	0.194	0.129	0.040	0.015
$p_1 = 1.2$	1.187	1.196	-0.011	-0.003	0.025	0.016	0.022	0.014	0.001	0.000
$p_2 = 1.4$	1.397	1.397	-0.002	-0.002	0.028	0.018	0.027	0.017	0.001	0.000
$\gamma_{10} = 5$	4.972	4.974	-0.006	-0.005	0.127	0.081	0.119	0.080	0.017	0.007
$\gamma_{11} = 1$	1.015	1.017	0.015	0.017	0.093	0.057	0.093	0.060	0.009	0.004
$\gamma_{12} = -1$	-0.972	-0.977	-0.028	-0.023	0.120	0.077	0.114	0.077	0.015	0.006
$\gamma_{20} = 4$	3.993	3.999	-0.002	0.000	0.127	0.084	0.119	0.079	0.016	0.007
$\gamma_{21} = 0$	0.002	0.001	NA	NA	0.110	0.073	0.108	0.071	0.012	0.005
$\gamma_{22} = 1$	0.998	1.010	-0.002	0.010	0.138	0.083	0.121	0.079	0.019	0.007
$\rho_1 = 0.8$	0.791	0.795	-0.011	-0.006	0.045	0.030	0.042	0.028	0.002	0.001
$\rho_2 = 0.8$	0.791	0.793	-0.011	-0.009	0.044	0.026	0.038	0.026	0.002	0.001
$\tau = 0.3$	0.288	0.289	-0.041	-0.037	0.129	0.085	0.113	0.080	0.017	0.007
$\delta = 0.6$	0.591	0.590	-0.015	-0.017	0.083	0.053	0.073	0.049	0.007	0.003

5 Application

5.1 Estimation Results

The proposed approach is applied to the portfolio of automobile insurance policies introduced in Section 2. The composite likelihood estimates are summarized in Table 4. In the Tweedie marginals, we fit log-linear models to both the mean and the dispersion for each type of claims. Exploratory analysis reveals that the effects of covariates on claim frequency and severity differ either in direction or size. Not surprisingly, we observe their significant effects on the mean as well as the dispersion. For instance, the length of driving experience shows a negative effect on the mean but positive effect on the dispersion of claims cost regardless of the coverage type. We allow the set of covariates to vary by coverage types. Exploratory analysis indicates the statistical significance of covariates and we report a more parsimonious model by retaining the important predictors. There are common factors affecting all types of claims (such as senior and conviction in the mean, and homeowner in the dispersion). Their effects across coverage types are noticeably consistent in the direction but could differ substantially in the size. In the dependence structure, we observe mild serial correlation in claims cost, which is explained by the short sampling period and the sparsity in the mixed outcome. Strong cross-sectional association are found among various types of claims. The cluster effect is statistically significant though relatively weak.

Table 5 compares the proposed model with alternative model specifications. Because dependence modeling is of the primary interest for our application, we consider nested dependence structure to emphasize the effect of various types of association among claims cost. These nested cases are: $M0$ assumes total independence ignoring all types of dependence among claims; $M1$ assumes no serial correlation in either of type of the claims; $M2$ examines the longitudinal claims cost of each type separately by assuming independence among claim types; $M3$ ignores the cluster effect, assuming all vehicles under the same policy are independent. To examine the effect of dispersion, we also look into the Tweedie GLM without dispersion modeling ($M4$). The dependence structure in the mean regression $M4$ is the same as in the proposed copula model $M5$. We report in the table the CLAIC and CLBIC statistics described in Section 3.2. Smaller statistics of the proposed model indicate a better fit. The goodness-of-fit statistics of $M0$ and $M5$ are close, suggesting modeling dispersion and dependence are equally important in terms of the reported statistics. Consistent with the size of the dependence parameters reported in Table 4, ignoring the cross-sectional correlation among claim types results in the largest penalty in the model fit.

Table 4: Composite likelihood estimates of the multilevel Tweedie model

	Accident Benefit		Collision		All Risk		Civil Liability		Dependence Model		
Mean	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Parameter	Est.	S.E.
intercept	5.732	0.127	4.467	0.066	3.485	0.092	5.337	0.063	ρ_1	0.163	0.022
young	-0.911	0.221							ρ_2	0.051	0.013
senior	-0.466	0.105	-0.142	0.038	-0.467	0.054	-0.183	0.039	ρ_3	0.099	0.015
marital			-0.102	0.030			-0.155	0.030	ρ_4	0.101	0.011
homeowner	-0.225	0.074			-0.163	0.038	0.065	0.028	τ_{12}	0.436	0.010
experience	-0.324	0.115	-0.383	0.045	-0.221	0.068	-0.227	0.042	τ_{13}	0.049	0.018
conviction	0.807	0.140	0.209	0.057	0.368	0.078	0.136	0.054	τ_{14}	0.641	0.009
newcar			0.399	0.050	0.327	0.067	0.136	0.047	τ_{23}	0.013	0.012
leasecar			0.428	0.034	0.544	0.048	0.170	0.034	τ_{24}	0.351	0.007
business					0.261	0.091	0.342	0.064	τ_{34}	0.021	0.010
highmilage	-0.578	0.075	0.194	0.031			0.082	0.029	δ	0.082	0.020
multidriver			0.402	0.048			0.428	0.050			
p	1.703	0.005	1.440	0.003	1.631	0.003	1.577	0.003			
Dispersion	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.			
intercept	7.107	0.050	6.751	0.025	6.505	0.053	6.330	0.031			
young					-0.288	0.068					
senior	0.151	0.048	-0.049	0.019	0.098	0.026	0.187	0.019			
marital	0.174	0.036	0.068	0.016	-0.100	0.021					
homeowner	-0.099	0.035	0.051	0.015	0.072	0.019	0.043	0.014			
experience	0.489	0.047	0.097	0.023	-0.169	0.039	0.216	0.021			
conviction							-0.095	0.027			
newcar					-0.093	0.032	-0.149	0.023			
leasecar			-0.116	0.017			-0.094	0.017			
business											
highmilage					-0.096	0.020					
multidriver	-0.237	0.060	-0.184	0.025							

Table 5: Goodness-of-fit statistics for alternative dependence specification

Model	Description	CLAIC	CLBIC
$M0$	independence	958,513	959,142
$M1$	no temporal correlation	957,747	958,375
$M2$	no cross-sectional dependence	958,501	959,130
$M3$	no cluster effect	957,734	958,363
$M4$	no dispersion	958,491	959,120
$M5$	the proposed model	957,730	958,358

5.2 Prediction

To demonstrate the prediction, we consider a simplified risk classification system. Ranking from low to high there are five risk class, Superior, Excellent, Good, Fair, and Poor as defined in Table 6. The quantities of interest are the mean and dispersion of claims cost, as they provide insights on the frequency and severity of claims. We report these quantities in Figure 2. Each panel corresponds to one coverage type, and within each panel, the mean and dispersion of insurance cost are displayed by risk class. To incorporate uncertainty in parameter estimates, we show their distributions instead of point estimates. The distributions are derived based on the Monte Carlo simulation from the asymptotic distributions of composite likelihood estimators. As anticipated, the expected claims costs of the four types of coverage increase from low risk to high risk class, and the different among risk classes are statistically significant. The dispersion varies by risk classes as well though not in a linear manner.

Table 6: Risk profile of hypothetical ratemaking classes

	Ratemaking Classes				
	Superior	Excellent	Good	Fair	Poor
young	0	0	0	0	0
seinor	1	1	1	1	0
marital	1	1	1	1	0
homeowner	1	1	1	1	0
experience	1	1	0	1	0
conviction	0	0	1	1	1
newcar	0	1	1	1	1
leasecar	0	1	0	1	1
business	0	1	1	1	1
highmilage	1	0	1	0	0
multidriver	0	0	0	1	1

As a second application, we are interested in the distribution of insurance costs for a block of business. Consider a hypothetical portfolio consisting of 5,000 policies that are evenly distributed in the five risk classes defined in Table 6. The claim distribution of the portfolio is provided in Figure 3. We examine the effect of dispersion modeling and dependence modeling on the portfolio

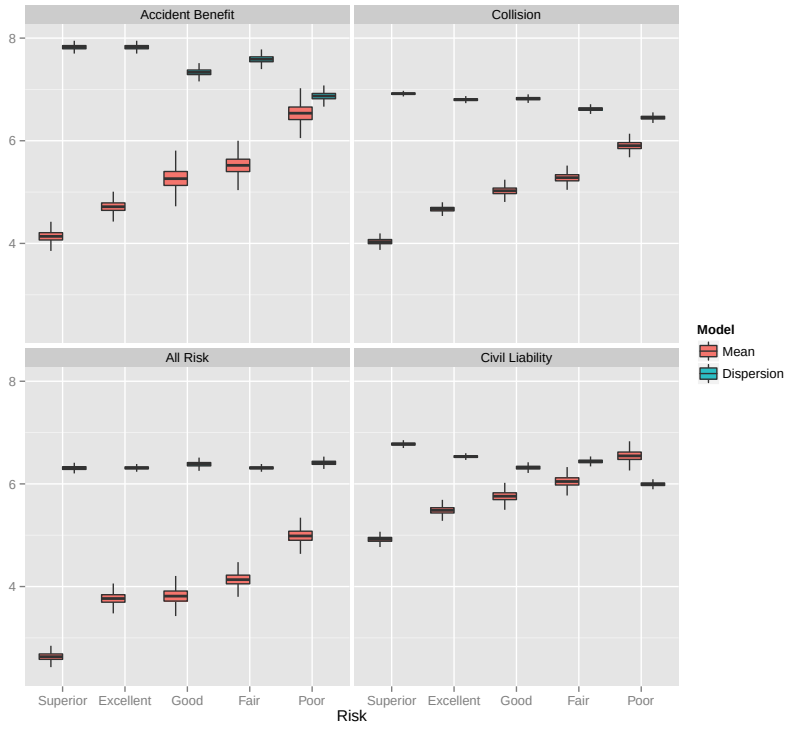

Figure 2: Distribution of mean and dispersion of different risk classes by coverage type.

risk. The left panel compares the claim distribution from the Tweedie GLM with and without dispersion modeling. To focus on this effect, we assume independence among claims. As central limit theorem predicts, the effect of dispersion modeling is less pronounced on the portfolio risk than the individual risk. The right panel compares the claim distribution from the independent and the copula-based Tweedie double GLM. In contrast, central limit theorem collapse in this case and the dependence modeling plays a critical role in risk aggregation. This results help the insurer make better business decisions such as to allocate risk capital among business lines and to determine the excess of loss reinsurance.

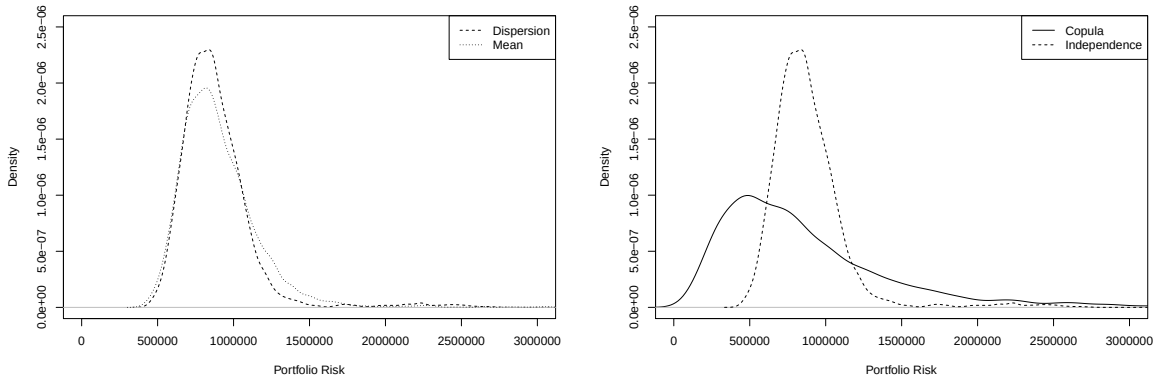

Figure 3: Comparison of portfolio risk. The left panel compares the predictive distribution of portfolio claims from the mean and dispersion models. The right panel compares the predictive distribution of portfolio claims from the copula and independence models.

6 Conclusion

In this work, we advanced modeling of insurance claims with complex data structure that often exhibits in property casualty insurance. The data is complex in that it is both multivariate and multilevel. The multivariate nature is because a vehicle is insured by multiple types of coverage. The multilevel structure is because a policy covers more than one vehicle and they are observed over time. The proposed multivariate regression model is sufficiently flexible to handle our complex insurance data.

The main contribution of this article is the introduction of the regression framework for the multivariate semi-continuous claims in the multilevel context. We used the Tweedie distribution to accommodate the semi-continuous nature of claims cost while at the same time, allowing for covariates in both mean and dispersion. One innovation in our approach is the employment of dependence modeling to accommodate the complex relationship among insurance claims. We used the Gaussian copula because of its flexibility in handling unbalanced data and the interpretability of the dependence parameters. It is worth pointing out that other copulas in the elliptical family possess similar flexibility in dependence modeling as the Gaussian copula, and thus are sensible candidates for our data. Applications of elliptical copulas other than Gaussian and t are rarely found in the literature. Some investigation is worthwhile in the future research.

The modeling approach developed in this article was motivated by the claims data in personal automobile insurance. However, it finds applications in much broader context. As pointed out already, the multilevel structure exhibited by our claims data is very common in property casualty insurance, including major personal lines (personal auto and homeowner) and commercial lines (worker's compensation, commercial multi-peril, commercial auto). The property-casualty insurance represents an important sector in the developed economy. The size of the market provides sufficient motivation for our work. Beyond insurance market, the proposed model has potential application in the modeling of health care utilization, where a household in private health plan or

an employer in group health plan forms the cluster, and the consumption of various types of care services is the outcome of interest. This provides additional motivation for the proposed method.

There are other possible approaches to modeling this type of data. One strategy is to use techniques from multivariate longitudinal data (see e.g. Fahrmeir and Tutz (2001)). Because the Tweedie density is not is not analytically tractable, the likelihood-based method for the Tweedie linear mixed model is difficult (see, e.g., Dunn and Smyth (2005; 2008) and Zhang (2013)). The dispersion model in this context is another challenge. Another possibility is the two-part model for the semi-continuous longitudinal data (see, e.g. Olsen and Schafer (2001)). However, the two-part framework is not readily to apply due to the multivariate and multilevel nature of our data. Since both strategies involves inference on the prediction of random quantities, we feel that the proposed approach is more flexible and easier to implement, especially when focus of the application is the predictive distribution of the outcome variables.

Appendix

This section provides foundation for the dependence structure used in the Gaussian copula model. Consider the linear model for the transformed data $\varepsilon_{ikjt} = \Phi^{-1}(F(y_{ikjt}; \mu_{ikjt}, p_j, \phi_{ikjt}))$:

$$\varepsilon_{ikjt} = \rho_j \varepsilon_{ikjt-1} + u_{ikjt} + v_{ijt}$$

with $\text{Var}(\varepsilon_{ikjt}) = 1$. Denote $\boldsymbol{u}_{ikt} = (u_{ik1t}, \cdots, u_{ikJt})'$ and $\boldsymbol{v}_{it} = (v_{i1t}, \cdots, v_{iJt})'$, and

$$\text{Var}\left(\boldsymbol{u}_{ikt}\right) = \begin{pmatrix} \sigma_u^{(11)} & \cdots & \sigma_u^{(1J)} \\ \vdots & \ddots & \vdots \\ \sigma_u^{(J1)} & \cdots & \sigma_u^{(JJ)} \end{pmatrix} \text{ and } \text{Var}\left(\boldsymbol{v}_{ikt}\right) = \begin{pmatrix} \sigma_v^{(11)} & \cdots & \sigma_v^{(1J)} \\ \vdots & \ddots & \vdots \\ \sigma_v^{(J1)} & \cdots & \sigma_v^{(JJ)} \end{pmatrix}$$

We show that the above model implies the dependence structure $\boldsymbol{\Sigma} = \boldsymbol{B}_{K\times K} \otimes \boldsymbol{P}_{(TJ)\times(TJ)}$ specified in Section 3.1 iff $\text{Var}\left(\boldsymbol{v}_{it}\right) = \lambda \text{Var}\left(\boldsymbol{u}_{ikt}\right)$.

Denote $\boldsymbol{\varepsilon}_{ikj} = (\varepsilon_{ikj1}, \cdots, \varepsilon_{ikjT})'$. We consider four scenarios in dependence analysis. The first is regarding the serial correlation among insurance costs for each coverage type. It is straightforward to show:

$$\text{Var}\left(\boldsymbol{\varepsilon}_{ikj}\right) = \begin{pmatrix} 1 & \rho_j & \cdots & \rho_j^{T-1} \\ \rho_j & 1 & \cdots & \rho_j^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{T-1} & \rho_j^{T-2} & \cdots & 1 \end{pmatrix} = \boldsymbol{P}_{jj}$$

The second is regarding dependence among different types of claims cost for a given vehicle.

For the same time period $t = t'$:

$$\mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ikj't}) = \frac{\sigma_u^{(jj')} + \sigma_v^{(jj')}}{1 - \rho_j \rho_j'} = \frac{1+\lambda}{\lambda} \frac{\sigma_v^{(jj')}}{1 - \rho_j \rho_j'} := \sigma_{jj'}$$

For the different time periods $t \neq t'$:

$$\mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ikj't'}) = \begin{cases} \rho_j^{t-t'} \mathrm{Cov}(\varepsilon_{ikjt'}, \varepsilon_{ikj't'}) = \rho_j^{t-t'} \sigma_{jj'} & \text{if } t > t' \\ \rho_{j'}^{t'-t} \mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ikj't}) = \rho_{j'}^{t'-t} \sigma_{jj'} & \text{if } t < t' \end{cases}$$

Thus we have

$$\mathrm{Cov}\left(\boldsymbol{\varepsilon}_{ikj}, \boldsymbol{\varepsilon}_{ikj'}\right) = \begin{pmatrix} 1 & \rho_{j'} & \cdots & \rho_{j'}^{T-1} \\ \rho_j & 1 & \cdots & \rho_{j'}^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{T-1} & \rho_j^{T-2} & \cdots & 1 \end{pmatrix} = \sigma_{jj'} \boldsymbol{P}_{jj'}$$

The third is the dependence between losses of a particular type of coverage but from different vehicles under the same policy. For the same period $t = t'$:

$$\mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ik'jt}) = \frac{\sigma_v^{(jj)}}{1 - \rho_j^2} = \frac{\lambda}{1+\lambda} := \delta$$

For the different time periods $t \neq t'$:

$$\mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ikj't'}) = \begin{cases} \rho_j^{t-t'} \mathrm{Cov}(\varepsilon_{ikjt'}, \varepsilon_{ik'jt'}) = \delta \rho_j^{t-t'} & \text{if } t > t' \\ \rho_{j'}^{t'-t} \mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ik'jt}) = \delta \rho_{j'}^{t'-t} & \text{if } t < t' \end{cases}$$

Hence one obtains

$$\mathrm{Cov}\left(\boldsymbol{\varepsilon}_{ikj}, \boldsymbol{\varepsilon}_{ik'j}\right) = \delta \begin{pmatrix} 1 & \rho_j & \cdots & \rho_j^{T-1} \\ \rho_j & 1 & \cdots & \rho_j^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{T-1} & \rho_j^{T-2} & \cdots & 1 \end{pmatrix} = \delta \boldsymbol{P}_{jj}$$

The fourth is the dependence between losses of different coverage types and from different vehicles insured by the same contract. For the same period $t = t'$:

$$\mathrm{Cov}(\varepsilon_{ikjt}, \varepsilon_{ik'j't}) = \frac{\sigma_v^{(jj')}}{1 - \rho_j \rho_{j'}} = \frac{\lambda}{1+\lambda} \sigma_{jj'} = \delta \sigma_{jj'}$$

For the different time periods $t \neq t'$:

$$\text{Cov}(\varepsilon_{ikjt}, \varepsilon_{ik'j't'}) = \begin{cases} \rho_j^{t-t'} \text{Cov}(\varepsilon_{ikjt'}, \varepsilon_{ik'j't'}) = \rho_j^{t-t'} \delta\sigma_{jj'} & \text{if } t > t' \\ \rho_{j'}^{t'-t} \text{Cov}(\varepsilon_{ikjt}, \varepsilon_{ik'j't}) = \rho_{j'}^{t'-t} \delta\sigma_{jj'} & \text{if } t < t' \end{cases}$$

One also notes that

$$\text{Cov}(\varepsilon_{ikjt}, \varepsilon_{ik'j't}) = \frac{\sigma_v^{(jj')}}{1-\rho_j\rho_{j'}} = \frac{\tau_{jj'}\sqrt{\sigma_v^{(jj)}}\sqrt{\sigma_v^{(j'j')}}}{1-\rho_j\rho_{j'}} = \delta\tau_{jj'}\frac{\sqrt{1-\rho_j^2}\sqrt{1-\rho_{j'}^2}}{1-\rho_j\rho_{j'}}$$

This justifies the reparameterization $\sigma_{jj'} = \tau_{jj'}\sqrt{1-\rho_j^2}\sqrt{1-\rho_{j'}^2}/(1-\rho_j\rho_{j'})$ and provides a natural interpretation of parameter $\tau_{jj'}$. Therefore one has:

$$\text{Cov}\left(\boldsymbol{\varepsilon}_{ikj}, \boldsymbol{\varepsilon}_{ik'j'}\right) = \delta\sigma_{jj'} \begin{pmatrix} 1 & \rho_{j'} & \cdots & \rho_{j'}^{T-1} \\ \rho_j & 1 & \cdots & \rho_{j'}^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{T-1} & \rho_j^{T-2} & \cdots & 1 \end{pmatrix} = \delta\sigma_{jj'}\boldsymbol{P}_{jj'}$$

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics 44*(2), 182–198.

Chiappori, P.-A. and B. Salanié (2013). Asymmetric information in insurance markets: predictions and tests. In G. Dionne (Ed.), *Handbook of Insurance*, pp. 397–422. Springer.

Cox, D. and N. Reid (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika 91*(3), 729–737.

Dunn, P. K. and G. K. Smyth (2005). Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing 15*(4), 267–280.

Dunn, P. K. and G. K. Smyth (2008). Series evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing 18*(1), 73–86.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd ed.). New York: Springer.

Frees, E. (2010). *Regression Modeling with Actuarial and Financial Applications*. New York: Cambridge University Press.

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

Gao, X. and P. X.-K. Song (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association 105*(492), 1531–1540.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 1208–1211.

Greene, W. (2007). *Econometric Analysis*. New Jersey: Prentice Hall.

Joe, H. (2014). *Dependence Modeling with Copulas*. New York: Chapman & Hall.

Joe, H. and Y. Lee (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis 100*(4), 670–685.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological) 49*(2), 127–162.

Jørgensen, B. and M. de Souza (1994). Fitting Tweedies compound poisson model to insurance claims data. *Scandinavian Actuarial Journal 1*(1), 69–93.

Klugman, S., H. Panjer, and G. Willmot (2012). *Loss Models: from Data to Decisions* (4th ed.). New York: Wiley.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics 80*(1), 220–239.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2nd ed.). New York: Chapman & Hall/CRC.

Nelsen, R. (2006). *An Introduction to Copulas* (2nd ed.). New York: Springer.

Olsen, M. K. and J. L. Schafer (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association 96*(454), 730–745.

Panagiotelis, A., C. Czado, and H. Joe (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association 107*(499), 1063–1072.

Parks, R. (1967). Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated. *Journal of the American Statistical Association 62*(318), 500–509.

Smith, M., A. Min, C. Almeida, and C. Czado (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association 105*(492), 1467–1479.

Smyth, G. and B. Jørgensen (2002). Fitting Tweedie's compound poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin 32*(1), 143–157.

Smyth, G. K. (1996). Regression analysis of quantity data with exact zeros. In R. Wilson, S. Osaki, and D. Murthy (Eds.), *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management*, pp. 17–19. Gold Coast, Australia.

Song, P. X.-K., M. Li, and Y. Yuan (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics 65*(1), 60–68.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Tweedie, M. (1984). An index which distinguishes between some important exponential families. In J. Ghosh and J. Roy (Eds.), *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pp. 579–604. Indian Statistical Institute, Calcutta.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis 92*(1), 1–28.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika 92*(3), 519–528.

Zhang, Y. (2013). Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models. *Statistics and Computing 23*(6), 743–757.

Zhao, Y. and H. Joe (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics 33*(3), 335–356.