Multiperil rate making for property insurance using longitudinal data

Lu Yang

University of Amsterdam, The Netherlands

and Peng Shi

University of Wisconsin–Madison, USA

[Received October 2017. Revised September 2018]

Summary. In property insurance, a contract often provides the policyholder with protection against damages to the insured properties that arise from a variety of perils. We propose a multivariate framework for pricing property insurance contracts with multiperil coverage in a longitudinal context. Specifically, a two-part model is employed to accommodate the excess of 0s and heavy tails in the insurance loss cost, and a Gaussian copula with a structured correlation is used to capture the dependence within and between perils, as well as their interaction. Using the government property insurance data from the state of Wisconsin in the USA, we show that the multiperil claim model has important implications in both experience rating and risk margin analysis.

Keywords: Gaussian copula; Multiperil rate making; Multivariate longitudinal data; Predictive distribution; Zero inflation

1. Introduction

Property insurance, as an important component in modern economy, provides homeowners and business owners with protection against losses to their buildings and the properties within. Pricing a property insurance contract accurately helps an insurer to develop strategies to identify and retain profitable business, and thus to reduce the solvency risks.

Property insurance is typically marketed through open perils coverage, i.e. it provides coverage for losses from all causes that are not specifically excluded in the policy. Common exclusions include damages due to natural disasters such as earthquake and flood, which are often covered by separate government insurance programmes. Many insurers look into loss costs by peril when pricing such contracts. The current practice is to treat each peril as an independent risk and to model them separately. Rollins (2005) suggested replacing the classical ‘indivisible premium’ with peril-based partial premium to improve economic efficiency and actuarial equity. Along this line of study, Frees et al. (2010) proposed a dependent multiperil-rate-making model to examine the risks of different perils jointly, arguing that perils are not independent because of ambiguous classification or latent policyholder characteristics. In a recent study, Frees et al. (2012a) compared a variety of models under both single-peril and multiperil modelling frameworks using out-of-sample validation and concluded that the choice of methods could have a substantial influence on an insurer’s pricing structure. On one hand, it is intuitively appealing to
decompose risks by peril because the variables that predict well for one peril might not predict well for other perils. Hence, rating by perils allows insurers to understand the determinant of each component better. In addition, claim adjustment could potentially be prioritized by type of peril. On the other hand, insurers charge a single price in the end. Pricing by peril requires more efforts on data collection and model building, and complex models with more parameters could lead to overfitting and poor prediction.

Meanwhile, insurers usually follow the experience of a policyholder over time, resulting in longitudinal data. Insurance rate making is essentially about prediction. With repeated observations, insurers seek to identify unobserved risk characteristics of a policyholder from the claim histories and thus make better predictions. Longitudinal models have become useful tools in this regard; see applications in Boucher and Denuit (2006), Shi and Valdez (2014) and Antonio et al. (2014), among others.

Combining the longitudinal and cross-sectional natures of property insurance claim data, we examine loss costs from multiple perils that are repeatedly observed over time. To the best of our knowledge, dependence modelling of longitudinal multiperil risks has not been explored in the literature. Some challenges are presented in this setting. First, the data are complex with a hierarchical structure. If one thinks that there is a latent risk characteristic for each peril, then these latent variables could be correlated across perils and, meanwhile, changing over time. Second, estimation and prediction of the multiperil risk distribution is not straightforward because of its multivariate nature. Third, loss cost data are known to feature a significant fraction of 0s and an otherwise continuous but skewed and thick-tailed component. As a result, well-studied methods such as multivariate linear models are not ready to apply.

In this paper, we propose a multivariate framework for pricing property insurance contracts with multiperil coverage in the longitudinal context by using copulas. Specifically, a two-part marginal model is employed to accommodate the excess of 0s and heavy tails in the insurance loss cost. A Gaussian copula is used to capture the serial dependence within and cross-sectional dependence between perils, as well as their interaction, i.e. the lagged cross-correlation. In addition, since Gaussian copulas are complete under marginalization, they demonstrate flexibility for handling missing values under the assumption of data missing at random. In the application, we examine a government property insurance data set from the state of Wisconsin. As will be seen in the next section, the fund works as a self-run property insurer and provides a natural experiment to study insurance operations such as underwriting and rate making. We show that the multiperil claim model has important implications in both experience rating and risk margin analysis.

The rest of the paper is organized as follows. Section 2 describes the local government property insurance fund in the state of Wisconsin and the characteristics of the claim data. Section 3 introduces the copula regression model for multiperil rate making using longitudinal data and discusses estimation and prediction for the model. Empirical analysis is performed in Section 4, and Section 5 concludes the paper. Technical details are provided in Appendix A.

2. Data

We examine the claim data set from the Local Government Property Insurance Fund (LGPIF) in the state of Wisconsin. The LGPIF was established by Chapter 605 of the Wisconsin Statutes and is administered by the Wisconsin Office of the Commissioner of Insurance. The purpose of the LGPIF is to make property insurance available for local government units, such as counties, cities, towns, villages, school districts and library boards, and the fund does not provide coverage for state government buildings. The LGPIF is designed to moderate the bud-
get effects of uncertain insurable events for local government entities with separate budgetary responsibilities.

The fund operates, to some extent, as a stand-alone property insurer in that it charges premiums and pays claims to its policyholders, i.e. local government units. The LGPIF offers three major types of coverage for local government properties: building and contents, inland marine (construction equipment) and motor vehicles. It covers all causes of property loss with certain exclusions including those resulting from flood, earthquake, wear and tear, extremes in temperature, mould, war, nuclear reactions and embezzlement or theft by an employee. In terms of size, the fund is currently insuring over 1000 entities. On average, it writes approximately $25 million in premiums and $75 billion in coverage each year. Nonetheless, the fund differs from property insurance companies in its operations. First, the fund has only one state employee who supervises the day-to-day operations by contracting for specialized services such as claim management and policy administration. Second, the LGPIF is not allowed to deny coverage, whereas local government units can buy insurance in the open market.

In this application, we focus on the insurance coverage for building and contents. We collect data on the lost costs of the government entities under coverage by type of peril and by year; data are collected for 1240 local government entities over a period of 6 years from 2006 to 2011. Because of the role of ‘residual market’ of the LGPIF, attrition is a rare event at least during our sampling period. For the same reason, the policyholders’ experience becomes particularly important since other sources of market data may not be relevant. We use the data in years 2006–2010 to develop the model and reserve the data of 2011 for validation.

To balance credibility and homogeneity, we consider three perils: water, fire and other. Water and fire are two main causes of losses in property insurance. Losses from all other causes are lumped into one category because of the small volume. Table 1 shows the distribution of the loss cost which is defined as the insurer’s total annual amount of losses arising from the loss events of a given policyholder, by peril. It contains both indemnity payments and loss adjustment expenses, and is an aggregated amount for all accidents that occurred during the policy year. It can be seen that the loss cost contains a significant fraction of 0s, which correspond to the scenario of no loss event during the year. For example, there is approximately 84% 0s in the loss cost due to water damage. In other words, the probability of any water damage for a given entity per year is about 0.16. Another noticeable feature of the loss cost is the skewness and long tails in the distribution. Given that the loss cost is strictly positive, Table 1 reports the selected percentiles by peril. The asymmetry around the median implies that the distribution is very right skewed and platykurtic, resulting in larger likelihood of extremely adverse events than the case if the loss costs had been normally distributed. This is further confirmed in Fig. 1 which illustrates the distribution of losses from the three types of peril given occurrence. For comparison Fig. 2 includes pairwise quantile–quantile (QQ-) plots of the positive loss costs from various perils.

| Peril     | Proportion of 0s | Results for the following conditional percentiles: |
|-----------|-----------------|---------------------------------------------------|
|           |                 | 0%  | 5%  | 25% | 50% | 75% | 95% | 100% |
| Water     | 0.843           | 1    | 854 | 2087 | 5144 | 13618 | 79663 | 12922218 |
| Fire      | 0.867           | 145  | 1205 | 3076 | 7686 | 20023 | 121506 | 1570619 |
| Other     | 0.884           | 1    | 945 | 2800 | 7946 | 26334 | 136724 | 4682537  |
We also obtain a set of rating variables that the fund uses to price each contract. Table 2 summarizes the descriptive statistics of the rating variables by year. Unlike the private market, the available rating variables are limited. One variable that the fund uses to differentiate risks is the type of government entity. There are six types of entity that the buildings covered might belong to: a city, county, school, town, village or a miscellaneous entity such as fire stations. Variables County, City, Town, Village, School and Misc are the corresponding indicators in the regression analysis. Another categorical rating variable is the alarm credit that is received by the policyholder. In our case, the policyholder receives a 5% discount in premium if automatic smoke alarms are installed in some of the main rooms within a building, a 10% discount if alarms are installed in all the main rooms and a 15% discount if the alarms are monitored continuously in addition. Binary variables AC00, AC05, AC10 and AC15 are created to denote 0%, 5%, 10% and 15% credit respectively. The last rating variable is the amount of coverage which measures

Fig. 1. Distributions of claim conditional on occurrence by perils: ———, fire; ······, water; ————, other
the policyholder’s exposure. In the analysis, we take the natural logarithm of the coverage amount to make it more symmetric. We report the sample mean for binary variables, and the mean and standard deviation (in parentheses) for continuous variables. For example, 5\% of the policyholders were county entities in 2006. Among the three rating variables, the type of entity is time constant, whereas alarm credit and the amount of coverage are time varying for each policyholder. The last two rows report the sample size by year and the associated attribution rate. There is a slight attrition over time; we ignore the attrition bias, relying on the assumption of values missing at random in the following analysis.

To provide intuition, we check the dependence in the loss costs across perils and the serial dependence for each peril by using rank correlation coefficients. The pairwise Kendall’s \( \tau \) is 0.22, 0.22 and 0.18 between water and fire, water and other and fire and other respectively. Table 3 displays the serial correlation for each peril. The lower triangle reports Kendall’s \( \tau \) and the upper triangle reports Spearman’s \( \rho \). Similarly to the cross-peril relationship, there are modest temporal relationships in the loss cost for all peril types. It draws our attention that the temporal correlation does not decay significantly over the observation period, indicating an exchangeable dependence structure.
Table 2. Descriptive statistics of rating variables by year

| Variable     | Results for the following years: |
|--------------|----------------------------------|
|              | 2006 | 2007 | 2008 | 2009 | 2010 |
| County       | 0.053| 0.054| 0.055| 0.064| 0.064|
| City         | 0.140| 0.139| 0.140| 0.143| 0.140|
| Town         | 0.184| 0.175| 0.171| 0.170| 0.167|
| Village      | 0.233| 0.241| 0.242| 0.234| 0.239|
| School       | 0.282| 0.285| 0.283| 0.282| 0.280|
| Misc         | 0.108| 0.106| 0.109| 0.108| 0.110|
| AC00         | 0.565| 0.537| 0.492| 0.408| 0.312|
| AC05         | 0.024| 0.025| 0.036| 0.053| 0.074|
| AC10         | 0.044| 0.050| 0.050| 0.065| 0.081|
| AC15         | 0.367| 0.388| 0.421| 0.475| 0.533|
| Coverage     | 1.951| 2.064| 2.134| 2.224| 2.232|
| (log-million)| (2.039)| (1.998)| (1.983)| (1.985)| (1.983)|
| Number of observations | 1159 | 1143 | 1130 | 1114 | 1114 |
| Attrition rate | —   | 0.014| 0.011| 0.014| 0    |

Table 3. Serial correlation of loss cost by peril

| Year | Results for the following years: |
|------|----------------------------------|
|      | 2006 | 2007 | 2008 | 2009 | 2010 |
|      |      |      |      |      |      |
| Water|      |      |      |      |      |
| 2006 | 1    | 0.239| 0.242| 0.232| 0.192|
| 2007 | 0.258| 1    | 0.320| 0.268| 0.254|
| 2008 | 0.257| 0.345| 1    | 0.281| 0.263|
| 2009 | 0.248| 0.289| 0.299| 1    | 0.246|
| 2010 | 0.206| 0.274| 0.284| 0.265| 1    |
| Fire |      |      |      |      |      |
| 2006 | 1    | 0.224| 0.265| 0.251| 0.279|
| 2007 | 0.239| 1    | 0.280| 0.247| 0.234|
| 2008 | 0.282| 0.297| 1    | 0.181| 0.288|
| 2009 | 0.268| 0.261| 0.193| 1    | 0.235|
| 2010 | 0.300| 0.251| 0.307| 0.251| 1    |
| Other|      |      |      |      |      |
| 2006 | 1    | 0.211| 0.247| 0.227| 0.220|
| 2007 | 0.222| 1    | 0.203| 0.178| 0.113|
| 2008 | 0.258| 0.215| 1    | 0.189| 0.198|
| 2009 | 0.236| 0.187| 0.197| 1    | 0.193|
| 2010 | 0.234| 0.122| 0.212| 0.205| 1    |

3. Multivariate claim modelling

3.1. Model

Consider an insurance policy that provides property coverage with $J$ types of peril; $J = 3$ in our case. We observe a portfolio of $n$ policyholders over a fixed time period of $T$ years. Let $Y_{it,j}$ be the loss cost corresponding to peril $j (= 1, \ldots, J)$ for policyholder $i (= 1, \ldots, n)$ in period
\( t (= 1, \ldots, T) \). We further summarize the associated rating variables in a vector \( x_{it,j} \), and it is allowed to vary across types of peril and over time. To facilitate model development, we assume at the moment that all the policyholders in the portfolio are observed for \( T \) years, and there are no missing values. We shall relax this assumption of balanced observations later in the section.

It is worth stressing that \( Y_{it,j} \) is semicontinuous in that it has a probability mass at zero corresponding to no claims over the year and a continuous component for positive values representing the claim amounts conditionally on occurrence. Two strategies have been proposed in the literature to handle zero inflation in insurance claims. The first is Tweedie generalized linear models that are based on the Tweedie distribution (Tweedie, 1984). The Tweedie distribution is defined as a Poisson sum of gamma-distributed random variables, and thus the probability mass at zero is embedded in the Poisson distribution. Jørgensen and de Souza (1994) and Smyth and Jørgensen (2002) explored fitting the Tweedie compound Poisson model to insurance claims in a regression set-up. The second is two-part models which decompose the claim costs into a frequency and a severity component. The frequency part examines whether there are claim(s) over the year, and the severity part focuses on claim amounts given occurrence. Frees (2014) provided a comprehensive review and discussed some variations of the approach.

Motivated by the skewness and long tails in the claim amounts, we adopt a two-part modelling strategy in this application. Denoting the density and cumulative distribution functions of \( Y_{it,j} \) as \( f_{it,j} \) and \( F_{it,j} \) respectively, we specify the following two-component mixture model:

\[
\begin{align*}
    f_{it,j}(y) &= f_j(y; \mathbf{x}_{it,j}, \mathbf{\eta}_j) \\
                 &= \begin{cases} 
                p_j(\mathbf{x}_{it,j}, \mathbf{\eta}_j^F) & y = 0, \\
                \{1 - p_j(\mathbf{x}_{it,j}, \mathbf{\eta}_j^F)\} g_j(y; \mathbf{x}_{it,j}, \mathbf{\eta}_j^S) & y > 0; 
            \end{cases}
\end{align*}
\]

\[
F_{it,j}(y) = F_j(y; \mathbf{x}_{it,j}, \mathbf{\eta}_j) = \Pr(Y_{it,j} \leq y | \mathbf{x}_{it,j}) = p_j(\mathbf{x}_{it,j}, \mathbf{\eta}_j^F) + \{1 - p_j(\mathbf{x}_{it,j}, \mathbf{\eta}_j^F)\} G_j(y; \mathbf{x}_{it,j}, \mathbf{\eta}_j^S)
\]

where \( p_j(\mathbf{x}_{it,j}, \mathbf{\eta}_j^F) \) represents the probability of zero losses for peril \( j \), and \( g_j(\cdot) \) and \( G_j(\cdot) \) are the density and distribution functions of a skewed and heavy-tailed variable defined on \((0, \infty)\). The vector \( \mathbf{\eta}_j = (\mathbf{\eta}_j^F, \mathbf{\eta}_j^S) \) summarizes the parameters in the marginal model for the \( j \)th type of peril with the superscripts F and S denoting the different effects of covariates on the frequency and severity components. In property insurance, we can think that the probability of having claims is more related to the risk profile of the property, whereas the amount of payment usually depends on the adjuster’s discretion and the type of coverage, for instance replacement cost versus actual cash value.

Repeated observations of loss costs from multiple perils suggest a multivariate modelling strategy where one hopes to examine the loss costs from different perils jointly while controlling for the peril-specific unobserved heterogeneity. In doing so, we employ a parametric copula to accommodate the cross-sectional dependence among perils, the serial correlation among the longitudinal outcomes and their interactions. A copula is a multivariate distribution function with uniform marginals and has become a popular tool in multivariate analysis. We refer to Joe (2014) for a comprehensive review of using copulas in dependence modelling. Denote \( \mathbf{Y}_{ij} = (Y_{i1,j}, \ldots, Y_{iT,j})' \) and \( \mathbf{Y}_i = (\mathbf{Y}^{'j}_{i1}, \ldots, \mathbf{Y}^{'j}_{iT})' \). According to Sklar’s theorem, there is a copula \( C_i \) such that the joint distribution of \( \mathbf{Y}_i \), which is denoted by \( H_i \), can be expressed as

\[
H_i(\mathbf{y}_i) = \Pr(\mathbf{Y}_i \leq \mathbf{y}_i) = C_i\{F_{i1,1}(y_{i1,1}), \ldots, F_{iT,T}(y_{iT,T}), \ldots, F_{i1,j}(y_{i1,j}), \ldots, F_{iT,j}(y_{iT,j})\}.
\]
Recall that $Y_{it,j}$ has a probability mass at zero. Without loss of generality, we assume that, for policyholder $i$, $K_j$ components of $y_i$ are positive and the remaining $TJ - K_j$ components are 0. Then the joint density function of $Y_i$, which is denoted by $h_i$, can be written as

$$h_i(y_i) = c_i^{K_i} \prod_{\{t,j: y_{it,j} > 0\}} f_{it,j}(y_{it,j}),$$

where

$$c_i^{K_i}(u_{i1,1}, \ldots, u_{iT,1}, \ldots, u_{i1,J}, \ldots, u_{iT,J}) = \prod_{\{t,j: y_{it,j} > 0\}} \frac{\partial^{K_i}}{\partial u_{it,j}} C_i(u_{i1,1}, \ldots, u_{iT,1}, \ldots, u_{i1,J}, \ldots, u_{iT,J}).$$

This framework allows us to preserve the flexibility in the two-part model that separates the frequency and severity components while imposing one copula to accommodate the dependence structure. The mixture approach avoids involving two copulas for each of the frequency and severity components separately.

3.2. Specification

Section 3.1 laid out the general modelling framework for the longitudinal multiperil loss costs. In this section, we provide a detailed specification for this application. In the marginal model, the probability of zero claims is formulated by using a standard generalized linear model with a logit link function:

$$p_j(x_{it,j}, \eta_j^F) = \eta_j^{-1}(x^\prime_{it,j} \eta_j^F) = \frac{\exp(x^\prime_{it,j} \eta_j^F)}{1 + \exp(x^\prime_{it,j} \eta_j^F)}, \quad j = 1, \ldots, J,$$

where $\eta_j(\cdot)$ is the link function. Other commonly used link functions include probit and log–log-links. Our analysis indicated that the selection of link functions is not critical for the application. The claim amounts often exhibit skewness and heavy tails as in Section 2, and thus we employ the generalized beta of the second kind distribution, GB2. The GB2 distribution was introduced by McDonald (1984) and has found extensive applications in the economics literature (McDonald and Xu, 1995). More recently, Frees and Valdez (2008) and Shi and Zhang (2015) considered an alternative parameterization and demonstrated its flexibility in fitting insurance claims and healthcare expenditures. GB2 is a member of the log-location–scale family with location parameter $\mu_{it,j}$, scale parameter $\sigma_j$ and shape parameters $\phi_{1j}$ and $\phi_{2j}$. The density of the GB2 distribution can be expressed as

$$g_j(y; x_{it,j}, \eta_j^S) = \frac{\exp(\phi_{1j} \lambda_{it,j})}{y[\sigma_j B(\phi_{1j}, \phi_{2j}) \{1 + \exp(\lambda_{it,j})\}]^{\phi_{1j} + \phi_{2j}}}$$

where $\lambda_{it,j} = \{\ln(y) - \mu_{it,j}\} / \sigma_j$ and $B(\phi_{1j}, \phi_{2j})$ is the Euler beta function. With four parameters, the GB2 distribution is very flexible for modelling skewed and heavy-tailed data. For instance, $\phi_{1j} > \phi_{2j}$ indicates right skewness and $\phi_{1j} < \phi_{2j}$ implies left skewness. We further specify the location parameter as a linear combination of covariates to control for the observed heterogeneity $\mu_{it,j} = x^\prime_{it,j} \eta_j^S$, with $\eta_j^S$ as a subset of $\eta_j$ corresponding to the regression coefficients.

We use a Gaussian copula to model the dependence. Gaussian copulas are the most commonly used copulas in the elliptical family. A $D$-variate Gaussian copula is

$$C(u_1, \ldots, u_D) = \Phi_D(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_D); \Sigma),$$

(3)
where $\Phi(\cdot)$ and $\Phi_D(\cdot)$ are the cumulative distribution functions of a univariate standard normal and a $D$-variate standard normal random variable respectively, and $\Sigma$ is the dispersion matrix of the multivariate normal distribution. $D = TJ$ in our application. Gaussian copulas have many appealing properties that are particularly valuable in predictive applications.

One advantage of a Gaussian copula is that the dispersion matrix allows for flexible dependence structure and meanwhile maintains the interpretability of dependence parameters, especially for data with multilevel structures (Shi et al., 2016). To specify the dispersion matrix $\Sigma$ for the multivariate outcomes $Y_i$, we adopt the strategy that was proposed by Shi (2012) and consider the dependence structure

$$
\Sigma_T = \Sigma_{(TJ) \times (TJ)} = 
\begin{pmatrix}
\Omega_{11} & \sigma_{12} \Omega_{12} & \cdots & \sigma_{1J} \Omega_{1J} \\
\sigma_{21} \Omega_{21} & \Omega_{22} & \cdots & \sigma_{2J} \Omega_{2J} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{J1} \Omega_{J1} & \sigma_{J2} \Omega_{J2} & \cdots & \Omega_{JJ}
\end{pmatrix},
$$

where $\sigma_{jj} = \sigma_{j'j}$ and $\Omega_{jj'} = \Omega_{j'j}$. The diagonal elements of $\Sigma_T$ are 1s and the off-diagonal elements indicate the lead–lag relationships between component series. For interpretation, $\sigma_{jj'}$ describes the cross-sectional dependence between types of peril $j$ and $j'$ in the same time period, i.e. $Y_{it,j}$ and $Y_{it,j'}$, which is known as the concurrent or contemporaneous correlation coefficient in time series analysis. The matrix $\Omega_{jj}$ captures the serial correlation for the claim cost of the $j$th peril, $Y_{ij}$, and the matrix $\Omega_{jj', j \neq j'}$ implies the dependence of the claim costs across peril types $j$ and $j'$, i.e. between $Y_{i,j}$ and $Y_{i,j'}$. Note that $\Omega_{jj'}$ is in general not symmetric.

For predictive purposes, we consider three serial correlation structures that are commonly used in longitudinal data modelling: an auto-regressive process of order 1 (AR(1)), an exchangeable and a banded Toeplitz structure. The AR(1) structure implies a correlation that diminishes over time. In contrast, the exchangeable structure, which is also known as compound symmetry or uniform correlation, assumes a constant dependence. In a banded Toeplitz structure, the correlation between two observations depends only on their time difference, and the correlation becomes 0 when the time difference is sufficiently large. Consider the case $T = 5$, and a bandwidth of 2 is used in the banded Toeplitz structure; then the corresponding matrix $\Omega_{jj'}$ can be specified as

$$
\Omega_{jj'}^{AR} = 
\begin{pmatrix}
1 & \rho_{j'} & \rho_{j'}^2 & \rho_{j'}^3 & \rho_{j'}^4 \\
\rho_j & 1 & \rho_{j'} & \rho_{j'}^2 & \rho_{j'}^3 \\
\rho_j^2 & \rho_j & 1 & \rho_{j'} & \rho_{j'}^2 \\
\rho_j^3 & \rho_j^2 & \rho_j & 1 & \rho_{j'} \\
\rho_j^4 & \rho_j^3 & \rho_j^2 & \rho_j & 1
\end{pmatrix},
$$

$$
\Omega_{jj'}^{EX} = 
\begin{pmatrix}
1 & \rho_{j'} & \rho_{j'} & \rho_{j'} & \rho_{j'} \\
\rho_j & 1 & \rho_{j'} & \rho_{j'} & \rho_{j'} \\
\rho_j & \rho_j & 1 & \rho_{j'} & \rho_{j'} \\
\rho_j & \rho_j & \rho_j & 1 & \rho_{j'} \\
\rho_j & \rho_j & \rho_j & \rho_j & 1
\end{pmatrix},
$$

$$
\Omega_{jj'}^{TOEP} = 
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
\rho_{1j} & 1 & \rho_{1j} & \rho_{2j} & 0 \\
\rho_{2j} & \rho_{1j} & 1 & \rho_{1j} & \rho_{2j'} \\
0 & \rho_{2j} & \rho_{1j} & 1 & \rho_{1j} \\
0 & 0 & \rho_{2j} & \rho_{1j} & 1
\end{pmatrix}.
$$
The current application assumes that all perils follow the same temporal pattern though with different parameters. This limitation in theory could be relaxed because the copula model only requires equation (4) to be a valid correlation matrix. However, the flexibility comes at a cost of interpretability. In particular, the current specification is motivated by a factor model as described in Appendix A.3. We argue that the limited specification is not a major concern for short panels. It also seems to be supported by the descriptive statistics in Table 3. However, it might not be appropriate for long panels where we might be more interested in the evolution of cross-sectional dependence over time.

Another benefit of a Gaussian copula is that it is complete under marginalization. This property enables us to modify the above model easily to accommodate unbalanced observations given values missing at random. Define \( \iota_{id} (d = 1, \ldots, TJ) \) as dummy variables with 0 indicating a missing value and \( L_i = (\iota_{i1}, \ldots, \iota_{iTJ})' \); then \( D_i = \sum_{d=1}^{TJ} \iota_{id} \) is the number of observed outcomes for policyholder \( i \). Then the copula \( C_i \) in the joint distribution of \( Y_i \) in equation (1) takes the form

\[
C_i(u_1, \ldots, u_{Di}) = \Phi_{Di} \{ \Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_{Di}); L_i \Sigma_T L_i' \}.
\] (5)

In our application, the missing values are due only to customer dropout, i.e. we have outcomes from all perils in the observed years. In what follows, we focus on our sampling mechanism, though our model can be applied in more general scenarios. Specifically, we assume that policyholder \( i \) is observed for \( T_i \leq T \) years.

We also would like to point out some limitations of Gaussian copulas. Specifically, a Gaussian copula is not capable of capturing tail dependence which is found to be critical in modelling financial risks of asset returns. In addition, with Gaussian copulas, statistical inference (estimation and prediction) is computationally heavy when the outcomes are discrete, for instance count data. We emphasize that these limitations are not of much concern to the current application. First, we are more interested in the properties of an insurance portfolio where tail dependence is less pronounced than for other financial products. Our empirical experience suggests that tail dependence is less crucial to insurance applications in that it has much less effect on key operations such as pricing and reserving. Second, in our application, we adopt a composite likelihood approach for estimation, which will be discussed in the following section. It avoids the computation of high dimensional integration and reduces the computational burden significantly. Third, the outcome variables in our study are semicontinuous. Though it has a probability mass, we show that the predictive distribution has a closed form under the Gaussian copula.

Finally, the above Gaussian copula model can be extended to the elliptical family naturally. A \( D \)-variate elliptical copula is defined as

\[
C(u_1, \ldots, u_D; \Sigma) = P_D \{ P^{-1}(u_1), \ldots, P^{-1}(u_D); \Sigma \},
\]

where \( P(\cdot) \) and \( P_D(\cdot) \) are the cumulative distribution function of a univariate and a \( D \)-variate elliptical random variable with zero mean and unit dispersion respectively. The theory of elliptical copulas is well developed; however, their value in empirical applications is not yet clear. We leave it as a future project to investigate.

3.3. Inference

Because of the discontinuity in the marginals, evaluation of the likelihood function (2) involves multi-dimensional integration that is computationally expensive. Considering the applied nature of this work, we adopt the composite likelihood method (Lindsay, 1988) to balance statistical and computational efficiency. Specifically, we use the pairwise likelihood (Cox and Reid, 2004) which...
provides relatively high estimation efficiency as shown in the literature; see Varin (2008) and Varin et al. (2011) for reviews on the composite likelihood approach. Further, in our application we assume that $C$ is constant over observations for simplicity; see alternatives in Patton (2006).

Let $\theta$ denote the parameter vector. The pairwise composite likelihood function for policy $i$ is defined as

$$l_i(\theta; y_i) = \sum_{j=1}^{J} \sum_{i<j} l(\theta; y_{it}, y_{jt}) + \sum_{j<i} \sum_{t=1}^{T_i} l(\theta; y_{it}, y_{jt})$$

where $l(\theta; y_{it}, y_{jt}) = \log \{ I(\theta; y_{it}, y_{jt}) \}$ and

$$I(\theta; y_{it}, y_{jt}) = \begin{cases} C \{ F_{it,j}(y_{it}, y_{jt}), F_{jt,i}(y_{jt}, y_{it}) \}, & \text{if } y_{it}, y_{jt} \\ \frac{1}{2} \left[ \frac{1}{u_{it}} c_1 \{ F_{it,j}(y_{it}, y_{jt}), F_{jt,i}(y_{jt}, y_{it}) \}, & \text{if } y_{it}, y_{jt} > 0 \end{cases}$$

Here $C(\theta)$ is the bivariate Gaussian copula joining $Y_{it}$ and $Y_{jt}$ with association parameter $\theta$. For instance, in the case of the AR(1) structure,

$$\tilde{\rho}_{it,j,t'} = \sigma \frac{\mathcal{I}(t < j)}{\rho_{it,j,t'}} \mathcal{I}(t < t') \mathcal{I}(t > t'),$$

where $\mathcal{I}(\cdot)$ denotes the indicator function. More insights on the dispersion matrix are provided in Appendix A.3. Specifically, we have

$$C(u_1, u_2; \rho) = \Phi_2 \{ \Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho \},$$

$$c_1(u_1, u_2; \rho) = \frac{1}{\partial u_1} \partial C(u_1, u_2) = \Phi \left\{ \frac{\Phi^{-1}(u_2) - \rho \Phi^{-1}(u_1)}{\sqrt{1 - \rho^2}} \right\},$$

$$c_2(u_1, u_2; \rho) = \frac{1}{\partial u_2} \partial C(u_1, u_2) = \Phi \left\{ \frac{\Phi^{-1}(u_1) - \rho \Phi^{-1}(u_2)}{\sqrt{1 - \rho^2}} \right\},$$

$$c(u_1, u_2; \rho) = \frac{1}{\partial u_1 \partial u_2} \partial C(u_1, u_2)$$

$$= \frac{1}{\sqrt{1 - \rho^2}} \exp \left[ \frac{2 \rho \Phi^{-1}(u_1) \Phi^{-1}(u_2) - \rho^2 \{ \Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2 \}}{2(1 - \rho^2)} \right].$$

The total composite likelihood for the portfolio of policies is calculated as

$$l(\theta; y) = \sum_{i=1}^{n} \frac{1}{T_i J - 1} l_i(\theta; y_i),$$

where $1/(T_i J - 1)$ is the weight that is assigned to the $i$th policy.

The composite likelihood estimator is defined as $\hat{\theta}_n = \arg \max_{\theta} l(\theta; y)$. The variance of $\hat{\theta}_n$ is estimated on the basis of the Godambe information matrix $\hat{\Omega}_n(\theta) = \hat{S}_n(\theta) \hat{V}_n^{-1}(\theta) \hat{S}_n(\theta)$, where $\hat{S}_n(\theta)$ and $\hat{V}_n(\theta)$ are known as the sensitivity matrix and the variability matrix respectively and are calculated by

$$\hat{S}_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i(\theta; y_i)}{\partial \theta^2} \bigg|_{\theta = \hat{\theta}_n},$$

$$\hat{V}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial l_i(\theta; y_i)}{\partial \theta} \frac{\partial l_i(\theta; y_i)}{\partial \theta'} \bigg|_{\theta = \hat{\theta}_n}.$$
Furthermore, model comparison can be done by using the composite likelihood version of the Akaike information criterion AIC (Varin and Vidoni, 2005) and the Bayesian information criterion BIC (Gao and Song, 2010), which are respectively calculated as

\[
\begin{align*}
\text{CLAIC} &= -2l(\hat{\theta}_n; y) + 2 \text{tr}\{\hat{\Sigma}_n^{-1}(\theta)\}, \\
\text{CLBIC} &= -2l(\hat{\theta}_n; y) + \log(n)\text{tr}\{\hat{\Sigma}_n^{-1}(\theta)\}.
\end{align*}
\]

3.4. Prediction

What is of particular interest to this application is prediction; our approach enjoys the benefit of straightforward prediction. From the Gaussian copula model, we can derive the predictive distribution of loss cost in year \(T + 1\). For a \(D\)-variate variable joined by a Gaussian copula, and without loss of generality, \(y_1, \ldots, y_d > 0\) and \(y_{d+1} = \ldots = y_D = 0\), we first define the general form

\[
R(d, D, \Sigma) = \Phi_{D-d}\{(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}')^{-1/2}(\mathbf{z}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{z}_1)\}\frac{\phi_d(\mathbf{z}_1)}{\prod_{i=1}^d \phi(\Phi^{-1}\{F(y_i)\})},
\]

\[
\bar{R}(d, D, \Sigma) = \Phi_{D-d}\{(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}')^{-1/2}(\mathbf{z}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{z}_1)\}
\]

where \(\mathbf{z}_1 = (\Phi^{-1}\{F(y_1)\}), \ldots, \Phi^{-1}\{F(y_d)\}\), \(\mathbf{z}_2 = (\Phi^{-1}\{F(y_{d+1})\}), \ldots, \Phi^{-1}\{F(y_D)\}\) and

\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}.
\]

Using these formulas with permutation, for policyholder \(i\), the distribution of \(Y_{iT+1}\) given \((Y_{iT+1}^i, \ldots, Y_{iT}^i)'\) is

\[
f(y_{iT+1}|y_i, \ldots, y_{iT}) = \frac{f(y_1, \ldots, y_{iT}, y_{iT+1})}{f(y_1, \ldots, y_{iT})} = \frac{R\{K_i + k_i, (T + 1)J, \Sigma_{T+1}\}}{R(K_i, TJ, \Sigma_T)} \prod_{j:y_{iT+1}^j > 0} f_{iT+1,j}(y_{iT+1,j}),
\]

\[
F(y_{iT+1}|y_i, \ldots, y_{iT}) = \frac{\text{Pr}(Y_{iT+1} = y_{iT+1}|Y_i = y_i, \ldots, Y_{iT} = y_{iT})}{\text{Pr}(Y_i = y_i, \ldots, Y_{iT} = y_{iT})} = \frac{\bar{R}\{K_i, (T + 1)J, \Sigma_{T+1}\}}{R(K_i, TJ, \Sigma_T)}
\]

where \(k_i\) perils have positive loss cost in year \(T + 1\). The detailed proof for results (7) and (8) can be found in Appendix A. When part of \((Y_{iT+1}^i, \ldots, Y_{iT}^i)'\) is missing, we can use the marginal distribution of the observed subset for computing the predictive distribution, as in expression (5).

The predictive mean is commonly used in insurance pricing. To calculate the predictive mean for peril \(j\),

\[
E[Y_{iT+1,j}|Y_i = y_i, \ldots, Y_{iT} = y_{iT}] = \int_0^\infty 1 - \text{Pr}(Y_{iT+1,j} \leq y|Y_i = y_i, \ldots, Y_{iT} = y_{iT}) \, dy
\]

\[
= \int_0^\infty 1 - \frac{\bar{R}(K_i, TJ + 1, L_j \Sigma_{T+1} L_j')}{R(K_i, TJ, \Sigma_T)} \, dy,
\]

where \(\Sigma_{T+1,j}\) is the correlation matrix of \((Y_{iT+1}^i, \ldots, Y_{iT}^i, Y_{iT+1,j})\) and is a submatrix of \(\Sigma_{T+1}\).
4. Empirical analysis

4.1. Estimation

We use the copula regression model that was proposed in Section 3 to analyse the longitudinal multiperil claim data from the LGPIF. To recap, the marginal model for the loss cost from each peril is a mixture of a logit and a GB2 regression model, and the joint model for the multivariate outcomes is a Gaussian copula with structured dependences. The cross-sectional association between types of peril, the temporal association within each peril and their interactions are of our interest. In this study, an unstructured dependence is specified for the between-peril correlation. For prediction, structured dependence is considered for the temporal within-peril correlation. Specifically, we fit models with AR(1), exchangeable and banded Toeplitz serial correlations. The parameters are estimated by using the pairwise composite likelihood method that was described in Section 3. Further, we consider two commonly used estimation approaches for copula regression models: inference for margins and full maximum likelihood estimation. In inference for margins, parameters in the marginals are estimated in the first stage and, conditionally on these estimates, dependence parameters are estimated in the second stage (Joe, 2005), whereas, in full maximum likelihood estimation, both marginal and copula parameters are estimated simultaneously.

The goodness-of-fit statistics, CLAIC and CLBIC, are summarized in Table 4. Along with the copula models, we also report in Table 4 the results of the independence model which fits only marginals. It is not surprising to observe the superior fit of copula models over the independence model, indicating significant association between the longitudinal multiperil loss costs. Both goodness-of-fit statistics suggest that the exchangeable correlation outperforms the two alternatives, which is consistent with the non-decaying trend in Table 3. It appears that there is not much gain from the extra flexibility of the Toeplitz correlation.

Table 5 presents the estimates of the parameters for the Gaussian copula model with exchangeable temporal dependence from full maximum likelihood estimation. The entity type captures the heterogeneity in both the frequency and the severity distributions for all perils. The amount of coverage is found to be an important predictor. Recall that the frequency component models the probability of zero claim. Hence the negative sign of the coefficient implies a higher chance of loss events for an entity with larger coverage. Intuitively, the coverage is a function of the number and the area of the properties in building and contents insurance, and thus more coverage means larger exposure to risk. Similar arguments apply to the severity component. In the GB2 regression, we note that the shape parameters satisfy \( \phi_1 > \phi_2 \) for all perils, indicating the right skewness of the loss cost distributions, which is consistent with Fig. 1. Furthermore, the distributions of loss cost for all perils are heavy tailed in the sense that the second moments of the fitted GB2 distribution do not exist because \( \phi_2 < 2\sigma \).

| Model             | Log-likelihood | Results for inference for margins | Results for maximum likelihood estimation |
|-------------------|----------------|-----------------------------------|------------------------------------------|
|                   | CLAIC          | CLBIC                             | CLAIC                              | CLBIC                  |
| Independence      | -30215         | 60431                             | 60613                                | 61078                   |
| Auto-regressive   | -30202         | 60407                             | 60585                                | 61048                   |
| Exchangeable      | -30197         | 60396                             | 60574                                | 61034                   |
| Banded Toeplitz   | -30200         | 60403                             | 60582                                | 61044                   |

Table 4. Goodness-of-fit statistics of alternative dependence models
Table 5. Estimation of the Gaussian copula model with exchangeable dependence

| Variable          | Results for fire | Results for water | Results for others |
|-------------------|------------------|-------------------|-------------------|
|                   | Estimate         | Standard error    | Estimate          | Standard error    | Estimate          | Standard error    |
| (Intercept)       | 3.702            | 0.289             | 3.930             | 0.277             | 4.212             | 0.316             |
| County            | −1.293           | 0.310             | −1.839            | 0.287             | −1.368            | 0.315             |
| City              | −1.109           | 0.282             | −0.961            | 0.273             | −0.942            | 0.295             |
| Town              | −0.607           | 0.327             | −0.069            | 0.369             | −0.239            | 0.389             |
| Village           | −0.893           | 0.278             | −0.976            | 0.262             | −0.635            | 0.298             |
| School            | 0.105            | 0.281             | −0.087            | 0.263             | −0.575            | 0.289             |
| AC05              | −0.077           | 0.241             | −0.213            | 0.269             | −0.183            | 0.265             |
| AC10              | −0.270           | 0.228             | −0.085            | 0.231             | −0.194            | 0.211             |
| AC15              | −0.268           | 0.116             | −0.081            | 0.121             | −0.096            | 0.122             |
| log(coverage)     | −0.472           | 0.051             | −0.485            | 0.050             | −0.500            | 0.056             |
| (Intercept)       | 7.272            | 0.293             | 7.442             | 0.341             | 7.498             | 0.388             |
| County            | −0.643           | 0.275             | −0.727            | 0.353             | −0.790            | 0.321             |
| City              | −0.289           | 0.251             | −0.904            | 0.325             | −0.493            | 0.304             |
| Town              | −0.150           | 0.302             | −0.514            | 0.302             | 0.006             | 0.356             |
| Village           | −0.258           | 0.240             | −0.383            | 0.280             | −0.248            | 0.276             |
| School            | −0.629           | 0.275             | −0.746            | 0.339             | 0.004             | 0.311             |
| AC05              | 0.435            | 0.326             | 0.002             | 0.160             | 0.332             | 0.265             |
| AC10              | −0.119           | 0.178             | 0.424             | 0.209             | −0.532            | 0.249             |
| AC15              | 0.026            | 0.131             | 0.042             | 0.114             | −0.136            | 0.187             |
| log(coverage)     | 0.350            | 0.071             | 0.436             | 0.058             | 0.442             | 0.075             |
| σ                 | 0.585            | 0.128             | 0.722             | 0.166             | 0.927             | 0.232             |
| φ₁                | 0.958            | 0.363             | 1.475             | 0.574             | 1.387             | 0.618             |
| φ₂                | 0.558            | 0.153             | 0.795             | 0.248             | 1.064             | 0.408             |
| Dependence model  | ρ₁               | 0.273             |             | 0.219             | 0.036             |
| ρ₂               | 0.217             | 0.029             |             | 0.303             | 0.038             |
| ρ₃               | 0.234             | 0.042             |             | 0.191             | 0.040             |

The results in Table 5 also confirm the statistical significance of the associations in the loss costs both between and within types of peril. For example, the association between fire and water damages is about 22%, and the temporal correlations for the three perils are within the range 20–30%. Although the cross-sectional and serial dependences are mild, their implications in the rate-making applications are substantial, which we shall demonstrate in the next section. Another note is that, because of the multiplicative specification in the dispersion matrix (4), the moderate association implies relatively weak predictive power across types of peril. For example, the association between the fire loss cost in the current year and the water loss cost in the following year is $\rho_1 \sigma_{12} = 0.273 \times 0.219 \approx 6\%$. However, the findings for this particular data set should not discount the value of the model proposed.

Because of the parametric nature of the model, we now perform some diagnostic analysis to validate the model assumptions. For the marginal model, we refer to the Q-Q-plot based on generalized Cox–Snell residuals (Cox and Snell, 1968) $r_{it,j} = \Phi^{-1}\{F_j^{*}(Y_{it,j}; x_{it,j}, \eta_j)\}$, where $F_j^{*}(*)$ is the distribution function of

$$Y_{it,j}^* = \begin{cases} Y_{it,j} & Y_{it,j} > 0, \\ Y_{it,j} - U_{it,j} & Y_{it,j} = 0, \end{cases}$$
and $U_{it,j}$ is a uniform random variable independent of $Y_{it,j}$. Because of the zero inflation in $Y_{it,j}$, it is known that the regular probability integral transform $F_{it,j}(Y_{it,j})$ is not uniformly distributed. Hence, we transform $Y_{it,j}$ into continuous variable $Y_{it,j}^{*}$, and the resulting distribution function is

$$F_j^*(y) = \begin{cases} p_j + (1 - p_j)G_j(y) & y > 0, \\ (1 + y)p_j & y < 0. \end{cases} \quad (10)$$

The validity of the generalized probability integral transform is provided in Appendix A, and the results are displayed in Fig. 3. Figs 3(a)–3(c) show the histogram of $F_j^*(Y_{it,j}^{*}, x_{it,j}, \eta_j)$ for the three perils, and we expect them to show uniform trend. Figs 3(d)–3(f) show the QQ-plots of $r_{it,j}$ which suggests that the zero-inflated GB2 regression model is a sensible choice for the loss cost of each peril. For comparison, we present the QQ-plots for zero-inflated gamma regression models in Figs 3(g)–3(i). It is clear that the gamma distribution does not handle the skewness

Fig. 3. Histograms and QQ-plots of transformed residuals: (a) water, GB2; (b) fire, GB2; (c) others, GB2; (d) water, GB2; (e) fire, GB2; (f) others, GB2; (f) water, gamma; (g) fire, gamma; (h) others, gamma
and heavy tails in the loss cost as well as the GB2 distribution does. The results also support our preference for the two-part framework over the Tweedie compound Poisson regression for modelling the loss cost in the first place.

Provided that the marginal models are correctly specified, we examine the goodness of fit of the copula model. Since our model imposes serial and cross-sectional dependence, which impacts the prediction and aggregation of losses, we perform a uniform test based on the
predictive distribution of the aggregated loss from three perils. Specifically, from expression (7), we can compute the predictive distribution of the total loss cost in year $T+1$ for policyholder $i$: $s_{i,T+1} = \sum_{j=1}^{J} y_{i,T+1,j}$. This is done by using Monte Carlo simulation where we generate loss costs by peril sequentially from equation (8) and then aggregate them. To generate the multiperil losses, we use a conditional approach. For instance, we first simulate $y_{i,T+1,1}$ conditioning on the lost cost in the past $T$ years, then we update expression (8) by adding $y_{i,T+1,1}$ to the conditioning set and simulate $y_{i,T+1,2}$, and so on until $y_{i,T+1,J}$ is generated. We then compute $s_{i,T+1}$ through aggregation. With 5000 replications, we can have an empirical distribution of $s_{i,T+1}$, which is denoted as $\hat{F}_{i,T+1,s}$. In our analysis, we leave the data of 2011 as the hold-out sample, which is denoted as $s_{0,i,T+1}$. We construct a sequence of $p_i = \hat{F}_{i,T+1,s}(s_{0,i,T+1})$, where $s_{0,i,T+1}$ is generated from $s_{i,T+1}$ through the transformation, and $\hat{F}_{i,T+1,s}$ is constructed from $\hat{F}_{i,T+1,s}$ by using expression (10) to handle the point mass at zero. We expect that $p_i$ is a random sample from a uniform distribution on $[0, 1]$. Fig. 4 displays the supporting results. Fig. 4(a) is the histogram of $p_i$, and Fig. 4(b) is the uniform QQ-plot with confidence bands.

4.2. Prediction

To evaluate the effect of the proposed multiperil rate-making model on experience rating and risk margin analysis, we do two experiments using the hold-out sample here. The first demonstrates the predictive performance at the policyholder level. The second emphasizes the influence of dependence on predictive distribution at the portfolio level.

In experience rating, an insurer incorporates policyholders’ claim experience into pricing. Because of regulatory constraints or lack of advanced analytics, the risk classification system of the insurer is imperfect. The mechanism of experience rating is that the insurer makes a Bayesian update on the belief about a policyholder’s risk by using his or her claim history. In this application, experience rating is implemented on the basis of the same notion of predictive distribution in a Bayesian context and is derived from the copula regression model. Using equation (9), we calculate the predictive mean of the total loss cost from all perils for each policyholder for year 2011 and compare it with the actual losses in the hold-out sample. Because of the zero inflation in the loss cost, we employ the Gini index to evaluate the out-of-sample performance. The Gini index, which was proposed in Frees et al. (2012b), compares the loss distribution and the ordered premium distribution, and hypothesizes whether a challenging premium performs better than the base premium. In this study, we make two sets of comparisons. In the first, a constant premium is used as base, and we consider two challenging premiums: the predictive means from the marginal model and the copula model. The constant premium treats all risks as homogeneous and charges the average cost for each risk. In contrast, both challenging models use risk classification to take observed heterogeneity into account. The Gini statistics are 72.48 and 73.11 with standard error 4.75 and 5.34 respectively. The statistical significance indicates that both challenging premiums lead to a better separation of risks than the base, which is expected. This test further confirms the predictive power of the explanatory variables in the marginal model. In the second comparison, the base premium is the contract premium which is the actual premium that the property fund charges for the government entities. The Gini statistics in this case are $-8.26$ and 32.66 from the marginal model and the copula model with standard errors 11.61 and 9.37 respectively, which suggests that the copula model outperforms the contract premium whereas the marginal model does not. This finding is particularly interesting. First, the preference of the contract premium to the marginal model is explained by the fact that, in addition to the rating variables, the property insurance fund might use some form of experience rating to incorporate policyholders’ claim experience in the rate making. Second,
the superiority of the copula model to the contract premium seems to suggest that, aside from experience rating, the multiperil rate-making approach provides additional lift in the prediction.

To visualize the results, we present the ordered Lorenz curves that are associated with these tests in Fig. 5. The Gini index is twice the area between the Lorenz curve and the 45° line. A curve under or above the line corresponds respectively to a positive or negative statistic.

The second application concerns risk margin analysis for an insurance portfolio. The risk margin provides a cushion against the adverse outcomes and is determined for the entire portfolio...
of policies as a whole and then allocated to each policy. The common approaches to calculating the risk margin use tail risk measures such as value at risk or conditional tail expectation. To demonstrate the effect of the dependence between and within perils on risk margin, we display the predictive distribution of the loss cost of the LGPIF in Fig. 6. The distribution is derived by using simulation where the loss cost for each government entity is generated sequentially as described in Section 3 and then aggregated for the portfolio. To compare, we also report in Fig. 6 the predictive distribution under the independence assumption. To control for the inherent randomness due to the GB2 distribution, we use the same pseudonumbers for simulation from the two models. As expected, the diversification effect is dampened by the positive dependence, and thus the insurer is exposed to higher tail risk. For instance, the 95th percentiles are $4.39 \times 10^7$ and $6.22 \times 10^7$ for the independence model and the copula model respectively. In addition, the vertical line in Fig. 6 represents the actual loss cost of the portfolio in the hold-out sample, which corresponds to the 75th and 62nd percentiles in the predictive distributions from the independence model and the copula model respectively.

5. Concluding remarks

In this paper, we explored dependent multiperil rate making for property insurance by using longitudinal data. Multiperil rate making by itself is interesting in that it maintains actuarial equity and economic efficiency. Intuitively, if perils are correlated with each other, the information on one peril will help us to learn about others. In addition, dependence between risks has important implications for risk aggregation and risk margin analysis. The availability of longitudinal data in our application makes multiperil rate making even more appealing because repeated observations over time enable insurers to implement experience rating in the multiperil
pricing context. An insurer can incorporate the policyholders’ claim history not only from the peril being priced but also other correlated perils in the price. It is worth stressing that longitudinal modelling shows how to harness the dependence between perils in prediction, since the cross-sectional correlation between perils is not ready to use for predictive purposes.

This paper demonstrated using advanced statistical models to provide solutions to insurance operations. We employed a Gaussian copula to accommodate the temporal dependence within each peril and cross-sectional dependence between perils, and thus to account for the unobserved heterogeneity. The usage of Gaussian copulas in our application is novel in two perspectives: first, the copula was proposed to apply to zero-inflated outcomes to preserve the flexibility in marginals. The discontinuity also motivated the novel statistical inference. Second, to balance complexity and interpretability, a structured correlation was proposed to describe the dependence in the hierarchical data.

Note that the outcome of interest in this study represents the annual ground-up loss for each policyholder. However, incomplete observations due to censoring and truncation are common in insurance claims data; see, for example, Antonio and Plat (2014). The applicability of the approach proposed depends on the features of the insurance contract. For instance, if the censoring or truncation applies to the annual loss cost, such as an aggregate deductible per year, the model proposed is ready to apply in that one just needs to adjust the two operations in the marginal mixture regression. However, if the censoring or truncation applies to each claim, such as per occurrence deductible, one must develop a model by using more granular claim data.

Finally, we point out the potential broader applications of the proposed modelling framework. Within insurance, the approach is expected to be useful in pricing bundling products such as an umbrella policy providing both homeowner and automobile coverage. Beyond insurance, the framework could be used to study consumer purchasing behaviour in marketing research where no purchase creates the zero inflation. In healthcare, the same modelling framework is promising for studying utilization of different types of care services.

Acknowledgements

We thank the Joint Editor and two reviewers for their comments that helped to improve the presentation of the paper. Peng Shi acknowledges the support by the Centers of Actuarial Excellence research grant from the Society of Actuaries and the Fall Research Competition from the University of Wisconsin–Madison.

Appendix A

A.1. Probability integral transform

Let \( Y \) denote a zero-inflated outcome that is 0 with probability \( p \) and is generated from distribution \( G(\cdot) \) with probability \( 1 - p \). Assume that \( G(\cdot) \) is smooth and defined on \( (0, \infty) \). The distribution function of \( Y \) is

\[
F(y) = p + (1 - p)G(y).
\]

Because of the probability mass at zero, \( F(Y) \), the regular probability integral transform, is not uniformly distributed. We consider a generalized probability integral transformation. Define

\[
Y^* = \begin{cases} 
  Y - U & Y = 0, \\
  Y & Y > 0
\end{cases}
\]

where \( U \) is uniformly distributed on \((0, 1)\) and is independent of \( Y \). It is straightforward to show the distribution of \( Y^* \) as
Because $Y^*$ is continuous, the probability integral transformation is ready to apply. Therefore, $F^*(Y^*)$ is uniformly distributed.

A.2. Predictive distribution

This section derives the predictive distribution (7). We examine a $D$-variate distribution:

$$F(y_1, \ldots, y_D) = C\{F(y_1), \ldots, F(y_D)\},$$

where $C(\cdot)$ is a Gaussian copula as in equation (3). Without loss of generality, assume that $y_1, \ldots, y_d > 0$ and $y_{d+1} = \ldots = y_D = 0$. Thus we have

$$f(y_1, \ldots, y_D) = \prod_{j=1}^d f(y_j)C^d\{F(y_1), \ldots, F(y_D)\}.$$ 

Let $v_1 = (u_1, \ldots, u_d)$ and $v_2 = (u_{d+1}, \ldots, u_D)$. Denote $z_1 = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))$ and $z_2 = (\Phi^{-1}(u_{d+1}), \ldots, \Phi^{-1}(u_D))$. It can be shown that

$$C^d(u_1, \ldots, u_D) = \frac{\partial C(u_1, \ldots, u_D)}{\partial v_1} = \frac{\partial \Phi_D\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_D); \Sigma\}}{\partial v_1} = \frac{\partial \Phi_D\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_D); \Sigma\}}{\partial z_1} \frac{\partial z_1}{\partial v_1} = F(z_2|z_1) f(z_1) \frac{1}{\prod_{i=1}^d \phi(\Phi^{-1}(u_i))} = \frac{\phi_d(z_1)}{\prod_{i=1}^d \phi(\Phi^{-1}(u_i))}.$$ 

The last equality is based on $z_2|z_1 \sim N(\mu_{z_2|z_1}, \Sigma_{z_2|z_1})$, where

$$\mu_{z_2|z_1} = \Sigma_{z_2|z_1}^{-1} z_1,$$
$$\Sigma_{z_2|z_1} = \Sigma_{z_2} - \Sigma_{z_2} \Sigma_{z_1}^{-1} \Sigma_{z_1}'.$$

A.3. Dispersion matrix

To gain further insights on the dispersion matrix, here we provide justification for the AR(1) structure and the link between the dispersion matrix and a factor model. Assume that our semicontinuous observations $Y_{it,j}$ are driven by a latent continuous variable $Y^*_{it,j}$ and

$$Y_{it,j} = Y^*_{it,j} I(Y^*_{it,j} > 0).$$

Define $\eta_{it,j} = \Phi^{-1}(F^*_{it,j}(Y^*_{it,j}))$, where $F^*_{it,j}$ is the distribution function of $Y^*_{it,j}$; then $\eta_{it,j}$ follows a standard normal distribution. With our hierarchical structure, we can assume that $\eta_{it,j} = \rho_{i,t-1,j} + \epsilon_{it,j}$, where the error terms are assumed to be correlated cross-sectionally with $\text{cov}(\epsilon_{it,j}, \epsilon_{i't,j'}) = \lambda_{jj'}$, and let $\sigma_{j't'} = \sqrt{(1 - \rho_j^2)(1 - \rho_{j'}^2)}$. One can verify that the temporal correlation for peril $j$ is $\text{corr}(\eta_{it,j}, \eta_{it',j'}) = \rho_{j'}$ and, when $t' > t$ without loss of generality and $j \neq j'$, $\text{corr}(\eta_{it,j}, \eta_{i't,j'}) = \rho_{j'}$ and $\text{corr}(\eta_{it,j'}, \eta_{i't,j'}) = \sigma_{j't'}$, i.e. the correlation between $\eta_{it,j}$ and $\eta_{i't,j'}$ is $\rho_{j't}$ as given in Section 3.3. Since $\eta$ follows a multivariate normal distribution, $Y^*$ has a Gaussian copula with dependence structure $\Sigma_T$, denoted as $C^*$. 

\[ F^*(y) = \Pr(Y^* \leq y) \]
\[ = \Pr(Y = 0)\Pr(Y^* \leq y|Y = 0) + \Pr(Y > 0)\Pr(Y^* \leq y|Y > 0) \]
\[ = p \Pr(-U \leq y|Y = 0) + (1 - p)\Pr(Y \leq y|Y > 0) \]
\[ = \begin{cases} p + (1 - p)G(y) & y > 0, \\ (1 + y)p & y < 0. \end{cases} \]
For any $y_i \geq 0$,

$$
\Pr(Y_i \leq y_i) = \Pr(Y_i^* \leq y_i) \\
= C^* \{ F_{1,1}^*(y_{1,1}), \ldots, F_{T,1}^*(y_{T,1}), \ldots, F_{1,J}^*(y_{1,J}), \ldots, F_{T,J}^*(y_{T,J}) \} \\
= C^* \{ F_{1,1}(y_{1,1}), \ldots, F_{T,1}(y_{T,1}), \ldots, F_{1,J}(y_{1,J}), \ldots, F_{T,J}(y_{T,J}) \}.
$$

Comparing this equation with equation (1) yields that $C = C^*$ in the range $[p_{1,1}, 1] [p_{2,1}, 1] \ldots [p_{T,J}, 1]$, where $p_{i,j}$ is the zero probability.

References

Antonio, K. and Plat, R. (2014) Micro-level stochastic loss reserving for general insurance. Scand. Act. J., no. 7, 649–669.

Antonio, K., Shi, P. and van Berkum, F. (2014) Longitudinal data and experience rating. In Computational Actuarial Science with R (ed. A. Charpentier), pp. 100–125. Boca Raton: Chapman and Hall–CRC.

Boucher, J.-P. and Denuit, M. (2006) Fixed versus random effects in Poisson regression models for claim counts: a case study with motor insurance. Astin Bull., 36, 285–301.

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

Cox, D. R. and Snell, E. J. (1968) A general definition of residuals (with discussion). J. R. Statist. Soc. B, 30, 248–275.

Frees, E. (2014) Frequency and severity models. In Predictive Modeling Applications in Actuarial Sciences, vol. I, Predictive Modeling Techniques (eds E. Frees, G. Meyers and R. Derrig), pp. 138–166. Cambridge: Cambridge University Press.

Frees, E. W., Meyers, G. and Cummings, A. D. (2010) Dependent multi-peril ratemaking models. Astin Bull., 40, 699–726.

Frees, E. W., Meyers, G. and Cummings, A. D. (2012a) Predictive modeling of multi-peril homeowners insurance. Variance, 6, 11–31.

Frees, E. W., Meyers, G. and Cummings, A. D. (2012b) Summarizing insurance scores using a Gini index. J. Am. Statist. Ass., 106, 1085–1098.

Frees, E. and Valdez, E. (2008) Hierarchical insurance claims modeling. J. Am. Statist. Ass., 103, 1457–1469.

Gao, X. and Song, P. X.-K. (2010) Composite likelihood Bayesian information criteria for model selection in high-dimensional data. J. Am. Statist. Ass., 105, 1531–1540.

Joe, H. (2005) Asymptotic efficiency of the two-stage estimation method for copula-based models. J. Multiv. Anal., 94, 401–419.

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

Jørgensen, B. and de Souza, M. (1994) Fitting Tweedie’s compound Poisson model to insurance claims data. Scand. Act. J., no. 1, 69–93.

Lindsay, B. (1988) Composite likelihood methods. Contemp. Math., 80, 220–239.

McDonald, J. (1984) Some generalized functions for the size distribution of income. Econometrica, 52, 647–663.

McDonald, J. and Xu, Y. (1995) A generalization of the beta distribution with applications. J. Econmetr., 66, 133–152.

Patton, A. J. (2006) Modelling asymmetric exchange rate dependence. Int. Econ. Rev., 47, 527–556.

Rollins, J. (2005) A modern architecture for residential property insurance ratemaking. Proc. Caslty Act. Soc., 92, 486–578.

Shi, P. (2012) Multivariate longitudinal modeling of insurance company expenses. Insur. Math. Econ., 51, 204–215.

Shi, P., Feng, X. and Boucher, J.-P. (2016) Multilevel modeling of insurance claims using copulas. Ann. Appl. Statist., 10, 834–863.

Shi, P. and Valdez, E. A. (2014) Longitudinal modeling of insurance claim counts using jitters. Scand. Act. J., no. 2, 159–179.

Shi, P. and Zhang, W. (2015) Private information in healthcare utilization: specification of a copula-based hurdle model. J. R. Statist. Soc. A, 178, 337–361.

Smyth, G. and Jørgensen, B. (2002) Fitting Tweedie’s compound Poisson model to insurance claims data: dispersion modelling. Astin Bull., 32, 143–157.

Tweedie, M. (1984) An index which distinguishes between some important exponential families. In Statistics: Applications and New Directions (eds J. Ghosh and J. Roy), pp. 579–604. Calcutta: Indian Statistical Institute.

Varin, C. (2008) On composite marginal likelihoods. Adv. Statist. Anal., 92, 1–28.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. Statist. Sin., 21, 5–42.

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