

# Predictive Modeling of Multivariate Longitudinal Insurance Claims Using Pair Copula Construction

Peng Shi

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

Zifeng Zhao

Department of Statistics  
University of Wisconsin - Madison  
Email: zifeng@stat.wisc.edu

January 13, 2018

## Abstract

The bundling feature of a nonlife insurance contract often leads to multiple longitudinal measurements of an insurance risk. Assessing the association among the evolution of the multivariate outcomes is critical to the operation of property-casualty insurers. One complication in the modeling process is the non-continuousness of insurance risks.

Motivated by insurance applications, we propose a general framework for modeling multivariate repeated measurements. The framework easily accommodates different types of data, including continuous, discrete, as well as mixed outcomes. Specifically, the longitudinal observations of each response is separately modeled using pair copula constructions with a D-vine structure. The multiple D-vines are then joined by a multivariate copula. A sequential approach is employed for inference and its performance is investigated under a simulated setting.

In the empirical analysis, we examine property risks in a government multi-peril property insurance program. The proposed method is applied to both policyholders' claim count and loss cost. The model is validated based on out-of-sample predictions.

**Keywords:** Non-continuous outcome, Predictive Distribution, Property Insurance, D-Vine, Zero/one inflation

# 1 Introduction

In the past decade, insurance industry, especially property and casualty insurance, has been advancing the use of analytics to leverage big data and improve business performance. Statistical modeling of insurance claims has become an important component in the data-driven decision making in various insurance operations (Frees (2015)). This study focuses on predictive modeling for nonlife insurance products with a “bundling” feature.

Bundling is a common design in modern short-term nonlife insurance contracts and it takes different forms in practice. For instance, in automobile insurance, a comprehensive policy provides coverage for both collision and third-party liability; in property insurance, an open peril policy covers losses from different causes subject to certain exclusions; furthermore, in personal lines of business, car insurance and homeowner insurance are often marketed to households as a package. In the context of predictive modeling, the bundling feature of nonlife insurance contracts often leads to multiple longitudinal measurements of an insurance risk. The goal of this study is to promote a modeling framework for evaluating the association among the evolution of the multivariate risk outcomes.

Assessing the association among multivariate longitudinal insurance claims is critical to the operation of property-casualty insurers. Two types of association among claims are in particular of interest to researchers and have been studied in separate strands of literature. On one hand, one is interested in the temporal relation of insurance claims. The availability of repeated measurements over time allows an insurer to adjust a policyholder’s premium based on the claim history, known as experience rating in insurance (Pinquet (2013)). Understanding the temporal dependence is essential to the derivation of the predictive distribution of future claims and thus the determination of the optimal rating scheme (see, for example, Dionne and Vanasse (1992) and Boucher and Inoussa (2014)). On the other hand, insurers are also interested in the contemporaneous association among the bundling risks. Correlated claims induce concentration risk in the insurance portfolio on the liability side of an insurer’s book in contrast to the investment portfolio on the asset side. Measuring the contemporaneous dependency is necessary to the quantification of the inherent risk in the bundling products, which offers valuable inputs to the insurer’s claim management and risk financing practice (Frees and Valdez (2008) and Frees et al. (2009)).

As noted above, the majority of existing studies tends to examine the contemporaneous and temporal dependence in insurance claims in distinct contexts. We argue that an analysis of both types of association in a unified framework will deliver unique insights to the decision making that cannot be gained from a silo approach. For instance, a joint analysis of multivariate longitudinal outcomes provides a prospective measure of the association among dependent risks as opposed to a retrospective measure, which is more relevant to the insurer’s operation such as ratemaking, risk retention, and portfolio management. Despite of the appealing benefit, studies in this line are still sparse, with one recent exception of Shi et al. (2016). There are several possible explanations: first, a separate analysis provides more focus in problem-driven studies; second, researchers might not have access to the relevant data; third, the discreteness in insurance claims data brings additional

challenge to the modeling process.

Setting apart from the existing literature, this work examines the contemporaneous and temporal association embedded in the multivariate longitudinal insurance claims in a unified framework. Specifically, we adopt a regression approach based on parametric copulas, where the longitudinal observations of each response is separately modeled using pair copula constructions with a D-vine structure, and the multiple D-vines are then joined by a multivariate copula. The proposed framework is very general and has a much broader audience in that it easily accommodates different types of data, including continuous, discrete, as well as mixed outcomes. In our context, we are interested in both the number of claims (discrete outcome) and the loss cost of claims (mixed outcome) of individual policyholders.

A number of approaches to joint modeling of multivariate longitudinal outcomes have been found in statistical literature. We refer to Verbeke et al. (2014) for a recent review. All the methods imply specific structures on the serial association of repeated measurements of a response, the contemporaneous association among multiple responses at the same time point, as well as the lead-lag association between the multiple responses at different points in time. In general, dependence among multivariate longitudinal outcomes can be accommodated via three mechanisms.

The first is to use random effects models where latent variables at either time-dimension or outcome dimension are specified to introduce association. For example, Reinsel (1984) and Shah et al. (1997) focused on linear models but limited with balanced data, and Roy and Lin(2000, 2002) relaxed this restriction to allow for unbalanced data. In theory, non-Gaussian outcomes can be accommodated using generalized linear mixed models. However, their applications to multivariate longitudinal data are far less found in the literature, arguable due to the difficulties in parameter interpretation, model diagnostics, and computational complexity.

The second approach is to directly specify the multivariate distribution for the outcome vector. The conventional case is the multivariate linear regression for Gaussian outcomes where the flexible dependence can be captured by appropriately structuring the covariance matrix (see, for instance, Galecki (1994)). Full specification of the distribution for discrete outcomes is more challenging and difficult to implement, with Molenberghs and Lesaffre (1994) being one example on ordinal categorical data. A more general class of models that is particularly useful for non-Gaussian outcomes is the copula regression. See Nelsen (2006) for an introduction to copula and Joe (2014) for recent advances. Applications of copulas for analyzing multivariate longitudinal outcomes include Lambert and Vandenhende (2002), Shi (2012), and Shi et al. (2016).

The third strategy is marginal models using generalized estimating equations(GEE) (Liang and Zeger (1986)). Compared to the above two methods, the advantage of using GEE is that the specification of full distribution can be avoid and inference for regression parameters is robust with respect to misspecification. For example, Rochon (1996) considered a bivariate model to jointly analyze a binary outcome and a continuous outcome that are both repeatedly measured over time. Gray and Brookmeyer (1998) and Gray and Brookmeyer (2000) proposed multivariate longitudinal models for continuous and discrete/time-to-event response variables respectively.

For the purpose of our application, we adopt the copula approach. On one hand, common measurements for insurance risks are neither Gaussian nor continuous, for example, the number of claims and the loss cost of claims in this study. Due to some unique features such as zero-inflation and heavy tails in these measurements, standard statistical methods including generalized linear models are not ready to apply. In this case, implementing a random effect approach is not straightforward and computationally challenging. On the other hand, prediction plays a central role in insurance business. From the inference perspective, forecasting carries no less weight than estimating regression coefficients for the current task. Hence the GEE approach is not appropriate since it treats association as nuisance and measures it using working correlation.

The proposed copula model in this paper is novel and differs from the literature in that pair copula construction offers a wider range of dependency compared to existing studies on multivariate longitudinal data that limit to the elliptical family of copulas. For example, Shi et al. (2016) employed a factor model to specify the dispersion matrix of elliptical copulas. The factor model implies restricted dependence structure although it is particularly helpful when the data exhibit a hierarchical feature. Lambert and Vandenhende (2002) used two separate copulas in the model formulation, one for the temporal association and the other for the (conditional) contemporaneous association. With an extra copula, the model adds more flexibility to the dependence structure. However, both temporal and contemporaneous association are still limited to the modeling of the correlation matrix in a Gaussian copula. This work contributes to the literature from both the methodological and the applied perspectives. From the methodological standpoint, we introduce a unified framework for modeling dependence in multivariate longitudinal outcomes based on pair copula construction. From the applied standpoint, the unique predictive application in property insurance motivates the proposed modeling framework and advocates its usage for the general two-dimensional data and in much broader disciplines.

The rest of the article is structured as follows. Section 2 describes the Wisconsin local government property insurance fund that drives the demand for advanced statistical methods. Section 3 proposes the pair copula construction approach to modeling multivariate longitudinal outcomes and discusses inference issues. Section 4 investigates the performance of the estimation procedure using simulation studies. Section 5 presents results of the multivariate analysis on the number of claims and the loss cost of claims for the policyholders in the Wisconsin government property insurance program. Section 6 concludes the article.

## 2 Motivating Dataset

This study is motivated by the operation of the local government property insurance fund from the state of Wisconsin. The fund was established to provide property insurance for local government entities that include counties, cities, towns, villages, school districts, fire departments, and other miscellaneous entities, and is administered by the Wisconsin Office of the Insurance Commissioner. It is a residual market mechanism in that the fund cannot deny coverage, although local government

units can secure insurance in the open market. See Frees et al. (2016) and Shi and Yang (2017) for more detailed description of the insurance property fund data.

We focus on the coverage for buildings and contents that is similar to a combined home insurance policy, where the building element covers for the physical structure of a property including its permanent fixtures and fittings, and the contents element covers possessions and valuables within the property that are detached and removable. It is an open-peril policy such that the policy insures against loss to covered property from all causes with certain exclusions. Examples of such exclusions are earthquake, war, wear and tear, nuclear reactions, and embezzlement.

For the sustainable operation of the property fund, the fund manager is particularly interested in two measurements for each policyholder, the claim frequency and the loss cost. The former measures the riskiness of the policyholder and provides insights for practice on loss control. The latter serves as the basis for determining the premium rate charged for a given risk. Policyholder level data are collected for 1019 local government entities over a six-year period from 2006 to 2011. Because of the role of residual market of the property fund, attrition is not a concern for the current longitudinal study. The open-peril policy type allows us further decompose the claim count and loss cost by peril types which provide a natural multivariate context to examine the longitudinal outcomes. Following Shi and Yang (2017), we consider three perils, water, fire, and other. We use the data in years 2006-2010 to develop the model and the data in year 2011 for validation.

One common feature shared by the claim count and the loss cost is the zero inflation, i.e. a significant portion of zeros associated with policyholders of no claims. Table 1 presents the descriptive statistics of the two outcomes by peril. Specifically, the table reports the percentage of zeros for each peril, and the mean and standard deviation of the outcome given there are at least one claim over the year. On average, about 85% government entities have zero claim per year for a given peril. Conditional on occurrence, we observe over-dispersion in the claim frequency except for fire peril, and skewness and heavy tails in the claim severity for all perils.

Table 1: Descriptive statistics of claim count and loss cost by peril

	% of Zeros	Claim Count		Loss Cost	
		Mean	SD	Mean	SD
Water	83.81	3.35	13.16	48,874	520,147
Fire	86.20	1.41	0.82	33,705	118,149
Other	87.95	1.60	2.86	50,824	263,241

The property fund data also contain basic risk classification variables that allow us to control for observed heterogeneity in claim count and loss cost. There are two categorical variables, entity type and alarm credit. Entity type is time constant, indicating whether the covered buildings belong to a city, county, school, town, village, or a miscellaneous entity such as fire stations. Alarm credit is time varying, reflecting the discount in premium received by a policyholder based on the features of the fire alarm system in the building. Available levels of discounts are 5%, 10%, and 15%. There is an additional continuous explanatory variable, amount of coverage, that measures

the exposure of the policyholder. Due to its skewness, we use the coverage amount in log scale in the regression analysis.

Unobserved heterogeneity in the multivariate longitudinal outcomes is accommodated by modeling the temporal association within each peril, and the contemporaneous and lead-lag association between different perils. To obtain intuitive knowledge of dependence, we visualize in Figure 1 and Figure 2 the pair-wise rank correlation for claim count and loss cost respectively. First, it is not surprising to observe modest serial correlation for each peril. This is consistent with the presence of unobserved peril-specific effect that is often attributed to the imperfection of the insurer’s risk classification. Second, there exists moderate contemporaneous dependence among the three perils, which suggests some relation among the peril-specific effects and thus a diversification effect on the insurer’s liability portfolio. In addition, the lead-lag correlation across perils implies that the temporal and contemporaneous dependence are interrelated. Similar dependence patterns are observed for the claim count and the loss cost, which is explained by the presence of excess zeros in both outcomes.

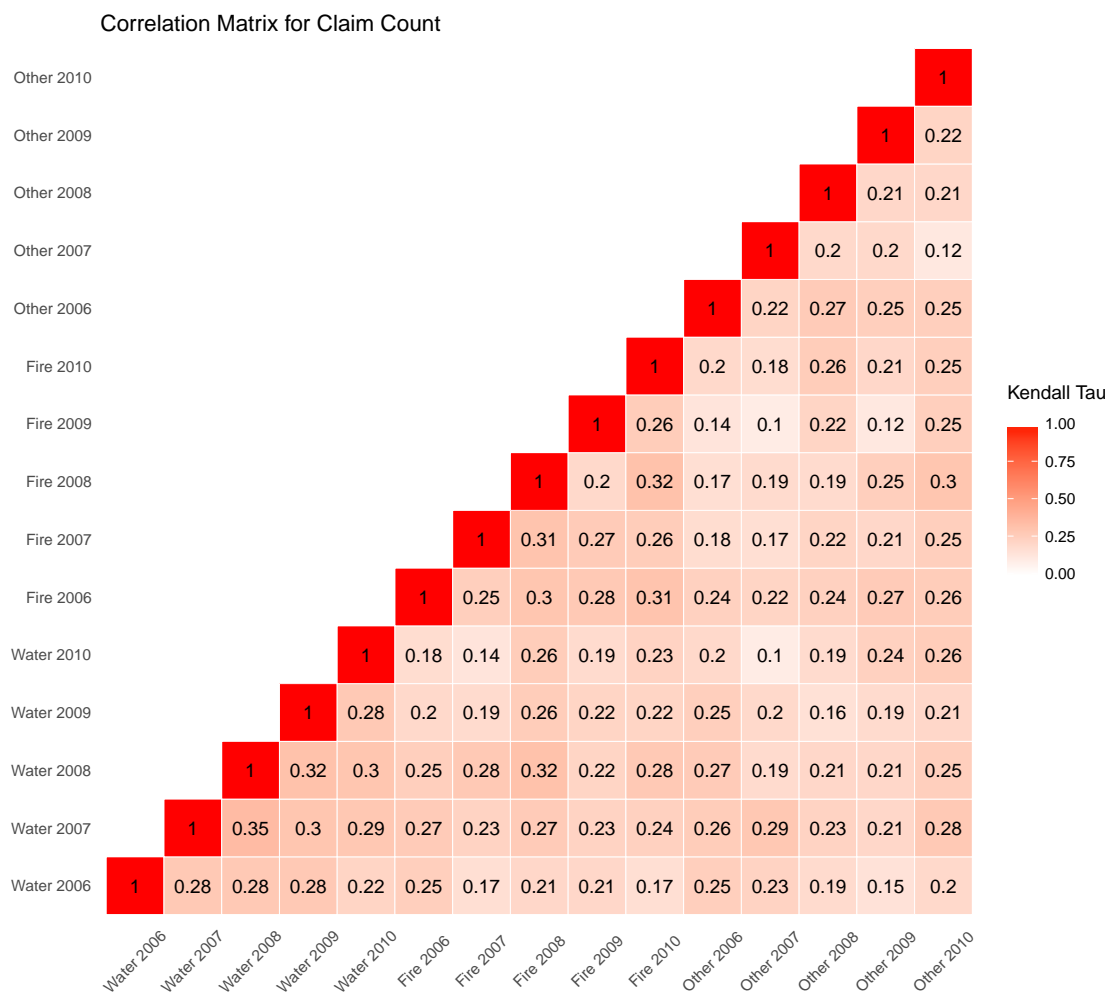


Figure 1: Rank correlation matrix for the claim count.

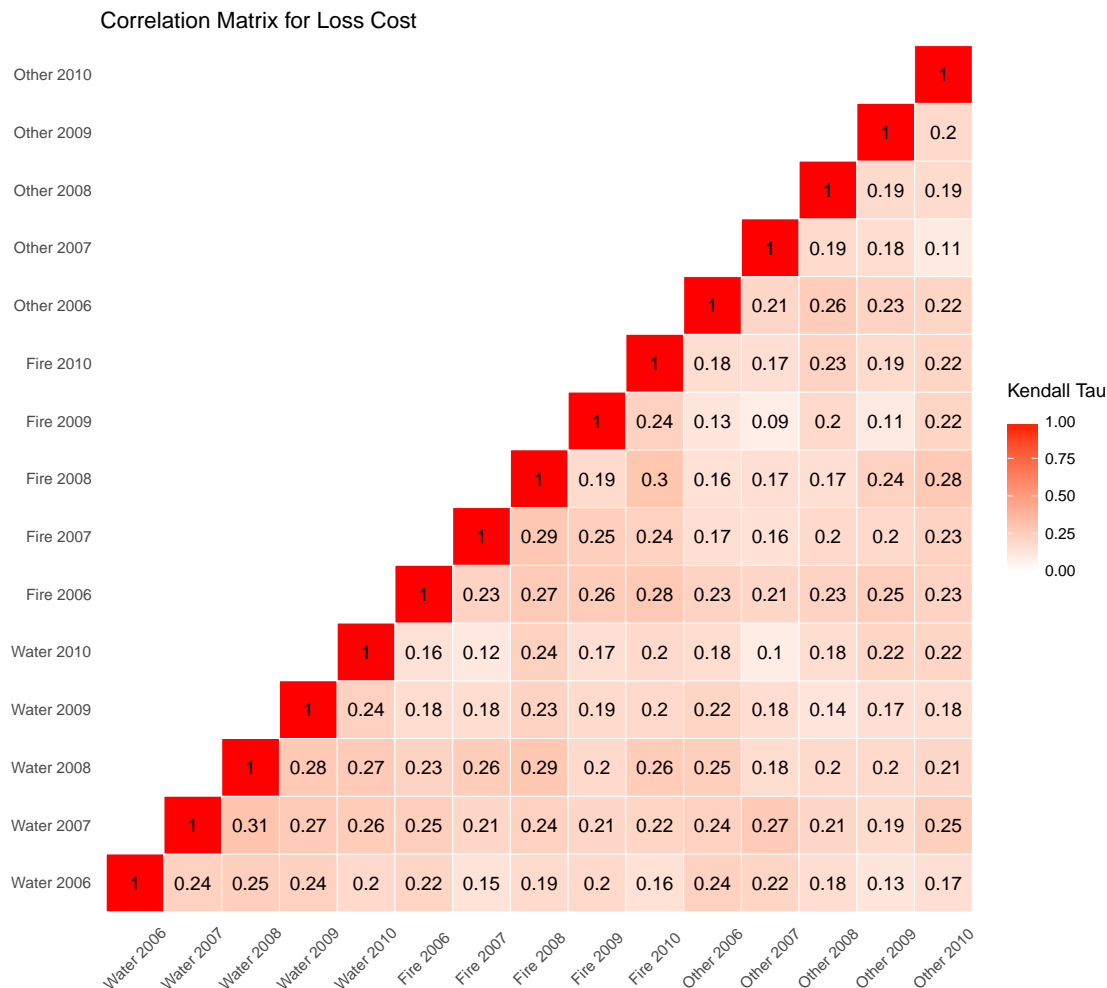


Figure 2: Rank correlation matrix for the loss cost.

### 3 Statistical Method

#### 3.1 Modeling Framework

Let  $y_{it}^{(j)}$  denote the  $j$ th ( $= 1, \dots, J$ ) response in the  $t$ th ( $= 1, \dots, T$ ) period for subject  $i$  ( $= 1, \dots, n$ ), and  $\mathbf{x}_{ijt}$  be the associated vector of explanatory variables. In this study, the outcomes of interest are the number of claims and the loss cost of claims from multiple perils for local government entities, and these outcomes are repeated observed over years. Specifically, index  $i$  corresponds to government entity,  $j$  peril type, and  $t$  policy year.

Due to the unique features exhibited in the data, standard regression models often fail for insurance data. Therefore, we look into customized models for capturing such interesting data

characteristics. For the claim count, we consider a zero-one inflated count regression:

$$f_{ijt}(y) = p_{ijt}^0 I(y = 0) + p_{ijt}^1 I(y = 1) + (1 - p_{ijt}^0 - p_{ijt}^1) g_{ijt}(y), \quad (1)$$

where  $p_{ijt}^k$  ( $k = 0, 1$ ) is specified using a multinomial logistic regression:

$$p_{ijt}^k = \frac{\exp(\mathbf{x}'_{ijt} \boldsymbol{\beta}_k)}{1 + \sum_{k=0}^1 \exp(\mathbf{x}'_{ijt} \boldsymbol{\beta}_k)}, \quad k = 0, 1,$$

and  $g_{ijt}(\cdot)$  is a standard count regression such as Poisson or negative binomial models. This specification allows to accommodate the excess of both zeros and ones in the claim count.

For the loss cost, we employ a mixture formulation:

$$f_{ijt}(y) = q_{ijt} I(y = 0) + (1 - q_{ijt}) g_{ijt}(y), \quad (2)$$

where  $q_{ijt}$  corresponds to a binary regression such as logit or probit for accommodating the mass probability at zero, and  $g_{ijt}(y)$  is a long-tailed regression to capture the skewness and heavy tails. Commonly used examples are generalized linear models or parametric survival models.

To introduce the dependence model for the multiple longitudinal outcomes, we adopt the pair copula construction approach based on the graphical model known as vine (see Bedford and Cooke (2001, 2002)). Mainly due to its flexibility especially in terms of tail dependence and asymmetric dependence, vine copula has received extensive attention in the recent literature of dependence modeling (see, for instance, Joe and Kurowicka (2011)). In particular, Kurowicka and Cooke (2006) and Aas et al. (2009) are among the first to exploit the idea of building a multivariate model through a series of bivariate copulas and the original effort has been focused on the continuous data. Building on the framework for continuous outcome, Panagiotelis et al. (2012) introduced the pair copula construction approach to the discrete case, Stöber et al. (2015) examined the vine method for multivariate responses including both continuous and discrete variables, Shi and Yang (2017) employed the pair copula construction to model the temporal dependence among hybrid variables, and Barthel et al. (2018) further extended the vine approach to the event time data with censoring. Despite of existing studies, using pair copula construction for data with multilevel structure is still sparse in the literature. Recently, Brechmann and Czado (2015) and Smith (2015) discussed possible strategies of constructing vine trees for multivariate time series. In contrast, we propose below a strategy for multiple longitudinal outcomes.

Define  $H_{it-}^{(j)}$  as the history of the  $j$ th outcome prior to time  $t$  for subject  $i$ , i.e.  $H_{it-}^{(j)} = (y_{i1}^{(j)}, \dots, y_{it-1}^{(j)})$ , and denote  $H_{it-} = (H_{it-}^{(1)}, \dots, H_{it-}^{(J)})$ . Letting  $\mathbf{y}_{it} = (y_{it}^{(1)}, \dots, y_{it}^{(J)})$ , we express the joint distribution for subject  $i$  as:

$$f(\mathbf{y}_{i1}, \dots, \mathbf{y}_{iT}) = f(\mathbf{y}_{i1}) f(\mathbf{y}_{i2} | \mathbf{y}_{i1}) \cdots f(\mathbf{y}_{iT} | \mathbf{y}_{i1}, \dots, \mathbf{y}_{iT-1}) = \prod_{t=1}^T f(\mathbf{y}_{it} | H_{it-}). \quad (3)$$



In the above  $f(\cdot|\cdot)$  in (3) denotes either joint (conditional) pdf or (conditional) pmf depending on the scale of the outcomes. Note that this conditional decomposition is generic and it does not impose any constraint on model specification.

Let  $F$  denote the cdf, be it either univariate or multivariate. The joint distribution of  $\mathbf{y}_{it}$  conditioning on  $H_{it-}$  in (3) is specified by a  $J$ -variate copula  $C^J$ :

$$F(\mathbf{y}_{it}|H_{it-}) = C^J \left( F \left( y_{it}^{(1)} | H_{it-}^{(1)} \right), \dots, F \left( y_{it}^{(J)} | H_{it-}^{(J)} \right) \right). \quad (4)$$

The conditional distribution  $F \left( y_{it}^{(j)} | H_{it-}^{(j)} \right)$ ,  $j = 1, \dots, J$  in (4) can be derived from the joint distribution of the outcomes  $\left( y_{i1}^{(j)}, \dots, y_{iT}^{(j)} \right)$  of type  $j$ . To allow for flexible temporal dependence, we further construct this distribution using the pair copula construction approach based on a D-vine structure:

$$f \left( y_{i1}^{(j)}, \dots, y_{iT}^{(j)} \right) = \prod_{t=1}^T f \left( y_{it}^{(j)} \right) \prod_{t=2}^T \prod_{s=1}^{t-1} \frac{f \left( y_{is}^{(j)}, y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right)}{f \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) f \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right)}. \quad (5)$$

We emphasize that the above proposed framework is general in that it accommodates marginals of different types, including continuous, discrete, and mixed distributions. Thus similar to (3), function  $f(\cdot|\cdot)$  in (5) could denote either pdf or pmf depending on the scale of the response variable. Below we discuss in detail the specifications for both continuous and non-continuous including discrete and semi-continuous cases.

### 3.1.1 Continuous Case

To introduce the idea, we start with the case of continuous outcomes. Using a  $J$ -variate copula  $C^J$  to accommodate the dependence among the multiple outcomes, the joint density associated with (4) becomes

$$f(\mathbf{y}_{it}|H_{it-}) = c^J \left( F \left( y_{it}^{(1)} | H_{it-}^{(1)} \right), \dots, F \left( y_{it}^{(J)} | H_{it-}^{(J)} \right) \right) \prod_{j=1}^J f \left( y_{it}^{(j)} | H_{it-}^{(j)} \right) \quad (6)$$

In (6),  $c^J$  denotes the density of copula  $C^J$ . It could be specified using an elliptical copula with an unstructured dispersion matrix to allow for flexible pair-wise association. As an alternative, one could use pair copula construction with a regular vine, especially when the dimension is high or tail and asymmetric dependence are of more interest.

Next, for each of the multiple outcome, we employ a unique D-vine structure to model the

temporal association. The conditional bivariate distribution in (5) can be expressed as

$$\begin{aligned}
& f\left(y_{is}^{(j)}, y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right) \\
&= f\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right) \times f\left(y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right) \\
&\quad \times c_{s,t|(s+1):(t-1)}^{(j)}\left(F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right), \tag{7}
\end{aligned}$$

where  $c_{s,t|(s+1):(t-1)}^{(j)}$  denotes the bivariate copula density for the  $j$ th responses in period  $s$  and  $t$  conditioning on the responses in the time periods in between. The D-vine specification offers a balance between interpretability and complexity. First, the natural ordering embedded in the longitudinal observations motivates the D-vine structure in modeling the temporal dependence for each outcome. Second, the conditional distribution  $F\left(y_{it}^{(j)} | H_{it-}^{(j)}\right)$  for  $j = 1, \dots, J$  in (6) is an intermediate product in the evaluation of (5). This link between the two components of the proposed dependence model implies no extra computational complexity in the evaluation of the likelihood function.

### 3.1.2 Non-continuous Case

Building upon the above idea, we modify the model to accommodate both discrete outcome and semi-continuous outcome that are more relevant to our application. For discrete outcomes such as the number of insurance claims, (6) will be replaced by

$$f(\mathbf{y}_{it} | H_{it-}) = \sum_{k_1=0}^1 \dots \sum_{k_J=0}^1 (-1)^{k_1 + \dots + k_J} C^J \left( F\left(y_{it}^{(1)} - k_1 | H_{it-}^{(1)}\right), \dots, F\left(y_{it}^{(J)} - k_J | H_{it-}^{(J)}\right) \right), \tag{8}$$

and in (5), (7) will be replaced by

$$\begin{aligned}
& f\left(y_{is}^{(j)}, y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right) \\
&= C_{s,t|(s+1):(t-1)}^{(j)}\left(F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right) \\
&\quad - C_{s,t|(s+1):(t-1)}^{(j)}\left(F\left(y_{is}^{(j)} - 1 | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right) \\
&\quad - C_{s,t|(s+1):(t-1)}^{(j)}\left(F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)} - 1 | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right) \\
&\quad + C_{s,t|(s+1):(t-1)}^{(j)}\left(F\left(y_{is}^{(j)} - 1 | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)} - 1 | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right), \tag{9}
\end{aligned}$$

where  $C_{s,t|(s+1):(t-1)}^{(j)}$  is the bivariate copula associated with  $c_{s,t|(s+1):(t-1)}^{(j)}$ .

For semi-continuous outcomes, one can think of a hybrid variable where a mass probability of zero is incorporated into an otherwise continuous variable, for instance, a policyholder's loss cost.

In this case, (6) will be replaced by

$$\begin{aligned}
& f(\mathbf{y}_{it}|H_{it-}) \\
= & \begin{cases} C^J \left( F \left( y_{it}^{(1)}|H_{it-}^{(1)} \right), \dots, F \left( y_{it}^{(J)}|H_{it-}^{(J)} \right) \right), & y_{it}^{(1)} = \dots = y_{it}^{(J)} = 0 \\ \partial^{(L)} \circ C^J \left( F \left( y_{it}^{(1)}|H_{it-}^{(1)} \right), \dots, F \left( y_{it}^{(J)}|H_{it-}^{(J)} \right) \right) \prod_{j=1}^L f \left( y_{it}^{(j)}|H_{it-}^{(j)} \right), & y_{it}^{(j)} > 0 \text{ for } 1 \leq j \leq L \\ C^J \left( F \left( y_{it}^{(1)}|H_{it-}^{(1)} \right), \dots, F \left( y_{it}^{(J)}|H_{it-}^{(J)} \right) \right) \prod_{j=1}^J f \left( y_{it}^{(j)}|H_{it-}^{(j)} \right), & y_{it}^{(j)} = 0 \text{ for } L+1 \leq j \leq J \\ & y_{it}^{(1)} > 0, \dots, y_{it}^{(J)} > 0 \end{cases} .
\end{aligned} \tag{10}$$

In the second scenario in (10), we define

$$\partial^{(L)} \circ C^J(u_1, \dots, u_J) := \frac{\partial^L}{\partial u_1 \dots \partial u_L} C^J(u_1, \dots, u_J)$$

where, without loss of generality, we assume that the first  $L$  ( $0 < L < J$ ) outcomes are positive and the remaining  $J - L$  are zero. Further, in (5), (7) will be replaced by

$$\begin{aligned}
& f \left( y_{is}^{(j)}, y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) \\
= & \begin{cases} C_{s,t|(s+1):(t-1)}^{(j)} \left( F \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right), F \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) \right) & y_{is}^{(j)} = 0, y_{it}^{(j)} = 0 \\ f \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) & y_{is}^{(j)} > 0, y_{it}^{(j)} = 0 \\ \times c_{1,s,t|(s+1):(t-1)}^{(j)} \left( F \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right), F \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) \right) & \\ f \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) & y_{is}^{(j)} = 0, y_{it}^{(j)} > 0, \\ \times c_{2,s,t|(s+1):(t-1)}^{(j)} \left( F \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right), F \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) \right) & \\ f \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) f \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) & y_{is}^{(j)} > 0, y_{it}^{(j)} > 0 \\ \times c_{s,t|(s+1):(t-1)}^{(j)} \left( F \left( y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right), F \left( y_{it}^{(j)} | y_{is+1}^{(j)}, \dots, y_{it-1}^{(j)} \right) \right) & \end{cases} ,
\end{aligned} \tag{11}$$

where  $c_{k,s,t|(s+1):(t-1)}^{(j)}(u_1, u_2) = \partial C_{s,t|(s+1):(t-1)}^{(j)}(u_1, u_2) / \partial u_k$  for  $k = 1, 2$ .

### 3.2 Inference

Because of the parametric nature of the model, we perform a likelihood-based estimation. The log-likelihood function for a portfolio of  $N$  policyholders can be expressed as:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{t=1}^T \log f(\mathbf{y}_{it}|H_{it-}), \tag{12}$$

where the marginal and dependence models are specified respectively by (1) and (8)(9) for claim count, and (2) and (10)(11) for loss cost. The hierarchical nature of the model and the separability assumption for dependence parameters motivate us to adopt a stage-wise estimation strategy in the

same spirit of inference function for margins (IFM) (Joe (2005)). In the first stage, one estimates the parameters in the marginal regression models assuming independence among observations both cross-sectionally and intertemporally. The second stage concerns the inference for the D-vine for each outcome. Either a simultaneous estimation or a tree-based estimation can be implemented. In the third stage, we estimate the copula for the contemporaneous association holding parameters in both marginal and temporal dependence models fixed. It is noted that estimation by stage allows us to gain substantial computational efficiency but at the cost of small loss in statistical efficiency. Our argument is that the stage-wise estimation is more feasible for predictive applications where the statistical efficiency is of secondary concern, especially in the case of big data. We investigate the performance of the estimation using simulation studies. Below is a detailed summary of the algorithms that are used to evaluate the likelihood:

**Step I:** For  $j = 1, \dots, J$ , we evaluate the following:

- (i) For  $t = 1, \dots, T$ , evaluate  $F\left(y_{it}^{(j)}\right)$  and  $f\left(y_{it}^{(j)}\right)$  using the marginal models of the  $J$  outcomes.
- (ii) For  $t = 1, \dots, T - 1$ , evaluate  $f\left(y_{it}^{(j)}, y_{it+1}^{(j)}\right)$  using (7) for continuous outcome or (9) for discrete outcome or (11) for semi-continuous outcome.
- (iii) For  $t = 2, \dots, T - 1$ , evaluate the following for  $s = 1, \dots, T - t$  recursively:
  - (a) Calculate  $f\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)$  and  $f\left(y_{is+t}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)$  using:

$$f\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = \frac{f\left(y_{is}^{(j)}, y_{is+t-1}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)}{f\left(y_{is+t-1}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)}$$

$$f\left(y_{is+t}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = \frac{f\left(y_{is+1}^{(j)}, y_{is+t}^{(j)} | y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)}{f\left(y_{is+1}^{(j)} | y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)}$$

- (b) Calculate  $F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)$  using:

- Continuous

$$F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = c_{2,s,s+t-1|s+1,\dots,s+t-2}^{(j)}\left(F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right), F\left(y_{is+t-1}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)\right)$$

- Discrete

$$F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = \left[ C_{s,s+t-1|s+1,\dots,s+t-2}^{(j)}\left(F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right), F\left(y_{is+t-1}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)\right) - C_{s,s+t-1|s+1,\dots,s+t-2}^{(j)}\left(F\left(y_{is}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right), F\left(y_{is+t-1}^{(j)} - 1 | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)\right) \right] / \left[ F\left(y_{is+t-1}^{(j)} | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right) - F\left(y_{is+t-1}^{(j)} - 1 | y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right) \right]$$

- Semi-continuous

$$F\left(y_{is}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = \begin{cases} \frac{C_{s,s+t-1|s+1,\dots,s+t-2}^{(j)}\left(F\left(y_{is}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right), F\left(y_{is+t-2}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)\right)}{F\left(y_{is+t-1}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)} & y_{is+t-1}^{(j)} = 0 \\ c_{2,s,s+t-1|s+1,\dots,s+t-2}^{(j)}\left(F\left(y_{is}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right), F\left(y_{is+t-1}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-2}^{(j)}\right)\right) & y_{is+t-1}^{(j)} > 0 \end{cases}$$

(c) Following the similar procedure in (b), calculate  $F\left(y_{is+t}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)$ .

- Continuous

$$F\left(y_{is+t}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = c_{1,s+1,s+t|s+2,\dots,s+t-1}^{(j)}\left(F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right), F\left(y_{is+t}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)\right)$$

- Discrete

$$F\left(y_{is+t}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = \left[ C_{s+1,s+t|s+2,\dots,s+t-1}^{(j)}\left(F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right), F\left(y_{is+t}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)\right) - C_{s+1,s+t|s+2,\dots,s+t-1}^{(j)}\left(F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right), F\left(y_{is+t}-1|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)\right) \right] / \left[ F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) - F\left(y_{is+1}-1|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) \right]$$

- Semi-continuous

$$F\left(y_{is+t}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right) = \begin{cases} \frac{C_{s+1,s+t|s+2,\dots,s+t-1}^{(j)}\left(F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right), F\left(y_{is+t}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)\right)}{F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)} & y_{is+1}^{(j)} = 0 \\ c_{1,s+1,s+t|s+2,\dots,s+t-1}^{(j)}\left(F\left(y_{is+1}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right), F\left(y_{is+t}^{(j)}|y_{is+2}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)\right) & y_{is+1}^{(j)} > 0 \end{cases}$$

(d) Calculate  $f\left(y_{is}^{(j)}, y_{is+t}^{(j)}|y_{is+1}^{(j)}, \dots, y_{is+t-1}^{(j)}\right)$  using (7), (9), and (11) for continuous, discrete, and semi-continuous outcomes respectively.

**Step II:** For  $j = 1, \dots, J$ , calculate the following for  $t = 2, \dots, T$ :

(i) Calculate  $f\left(y_{it}^{(j)}|H_{t-}^{(j)}\right)$  using:

$$f\left(y_{it}^{(j)}|H_{t-}^{(j)}\right) = \frac{f\left(y_{i1}^{(j)}, y_{it}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)}{f\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)}$$

(ii) Calculate  $F\left(y_{it}^{(j)}|H_{t-}^{(j)}\right)$  using:

- Continuous

$$F\left(y_{it}^{(j)}|H_{t-}^{(j)}\right) = c_{1,1,t|2,\dots,t-1}^{(j)}\left(F\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right)$$

- Discrete

$$F\left(y_{it}^{(j)}|H_{t-}^{(j)}\right) = \left[ C_{1,t|2,\dots,t-1}^{(j)}\left(F\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right) - C_{1,t|2,\dots,t-1}^{(j)}\left(F\left(y_{i1}^{(j)} - 1|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right) \right] / \left[ F\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right) - F\left(y_{i1}^{(j)} - 1|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right) \right]$$

- Semi-continuous

$$F\left(y_{it}^{(j)}|H_{t-}^{(j)}\right) = \begin{cases} \frac{C_{1,t|2,\dots,t-1}^{(j)}\left(F\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right)}{F\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)} & y_{i1}^{(j)} = 0 \\ c_{1,1,t|2,\dots,t-1}^{(j)}\left(F\left(y_{i1}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right), F\left(y_{it}^{(j)}|y_{i2}^{(j)}, \dots, y_{it-1}^{(j)}\right)\right) & y_{i1}^{(j)} > 0 \end{cases}$$

**Step III:** For  $t = 1, \dots, T$ , calculate  $f(\mathbf{y}_{it}|H_{it-})$  using (6), (8), and (10) for continuous, discrete, and semi-continuous cases respectively.

## 4 Simulation

This section examines the performance of the stage-wise estimation using numerical experiments. To mimic insurance applications in the study, we consider two scenarios, one for a count outcome, and the other for a semi-continuous outcome. Because dependence modeling using copulas separate the marginal model and the dependence model, the simulation focuses on the estimation of association parameters.

In the first experiment, the data generating process is defined by model (1), (8), and (9). For each subject, we assume that there are three outcomes ( $J = 3$ ) and each is observed for four years ( $T = 4$ ). The three components of the model are specified as below:

- (1) The marginal models for  $Y_{it}^{(j)}$  follow a Poisson generalized linear model with

$$\ln(\lambda_{ijt}) = \beta_{j0} + \beta_{j1}X_{1,ijt} + \beta_{j2}X_{2,ij}.$$

We set  $(\beta_{j0}, \beta_{j1}, \beta_{j2}) = (-1, 0.5, 0.5)$  for  $j = 1, 2, 3$ , and we assume  $X_{1,ijt} \sim$  i.i.d.  $N(0, 1)$  and  $X_{2,ij} \sim$  i.i.d.  $Bernoulli(0.4)$ .

(2) In the D-vine for the three outcomes, the bivariate copula for all (conditional) pairs are assumed to be rotated Joe copula. For a given D-vine, with  $T = 4$ , there are three trees. We assume a common association parameter in all the bivariate copulas in the same tree. Let  $\zeta_j = (\zeta_{j1}, \zeta_{j2}, \zeta_{j3})$  denote the association parameters for the  $j$ th outcome, we set  $\zeta_1 = (1.77, 1.44, 1.19)$ ,  $\zeta_2 = (3.83, 2.22, 1.44)$ , and  $\zeta_3 = (18.74, 3.83, 1.77)$ . These parameters are corresponding to the Kendall's  $\tau$  being  $(0.3, 0.2, 0.1)$ ,  $(0.6, 0.4, 0.2)$ , and  $(0.9, 0.6, 0.3)$ , respectively.

(3) The copula that joins the three outcomes is set to be a Gaussian copula with unstructured correlation. Assume  $(\rho_{12}, \rho_{13}, \rho_{23}) = (0.2, 0.5, 0.8)$  where  $\rho_{jj'}$  denotes the pair-wise correlation between the  $j$ th and  $j'$ th responses.

In the second experiment, the data generating process is defined by model (2), (10), and (11). Similar to the first one, we assume  $J = 3$  and  $T = 4$ . The marginal model for  $Y_{it}^{(j)}$  in this experiment is a mixture of a *Bernoulli*( $p_{ijt}$ ) distribution and a *Gamma*( $\alpha_j, \theta_{ijt}$ ) distribution. The binary outcome is generated from a logistic regression:

$$\ln(p_{ijt}) = \delta_{j0} + \delta_{j1}X_{1,ijt} + \delta_{j2}X_{2,ij},$$

with  $(\delta_{j0}, \delta_{j1}, \delta_{j2}) = (2, -1, -2)$  for  $j = 1, 2, 3$ . In the gamma regression, let  $\mu_{ijt} = \alpha_j\theta_{ijt}$ , and we consider a log link function:

$$\ln(\mu_{ijt}) = \gamma_{j0} + \gamma_{j1}X_{1,ijt} + \gamma_{j2}X_{2,ij},$$

with  $(\gamma_{j0}, \gamma_{j1}, \gamma_{j2}) = (10, 1, 0.5)$  and  $\alpha_j = 5000$  for  $j = 1, 2, 3$ . Furthermore,  $X_{1,ijt}$  and  $X_{2,ij}$  are assumed to be i.i.d.  $N(0, 1)$  and *Bernoulli*(0.4) respectively. The dependence models for the temporal association and contemporaneous association are assumed to follow the same specification as in the first experiment. That is, the rotated Joe copula is used in the D-vine for each outcome with identical copulas in the same tree, and a Gaussian copula with unstructured correlation is employed to join the three D-vines.

For both experiments, model parameters are estimated sequentially. Step one estimates parameters in the marginal model, step two parameters in the D-vines, and step three parameters in the Gaussian copula. Note that the first two steps can be combined into one step with a higher computational cost. In the last step, a pair-wise composite likelihood approach is further used instead of a full MLE to speed up the computation. The classical plug-in estimators for the asymptotic variance of the sequential estimators can be complicated to obtain due to the stage-wise estimation strategy. However, the variance can be readily estimated by a parametric bootstrap, see for example Zhao and Zhang (2017). The C.I.'s for the sequential estimators can then be constructed based on the estimated variance. In the following, we estimate the variance based on 100 times parametric bootstrap.

We conduct simulation experiments with sample size  $N$  (number of subjects) to be 500 and 1000. For each level of  $N$ , we repeat the experiment 500 times. The results are shown in Table 2 and Table 3 for the count outcome and the semi-continuous outcome respectively. The tables

present the bias, standard deviation, as well as coverage probabilities at different confidence levels. The standard deviation is calculated using the estimates from the 500 replications, and the coverage probability is for the confidence intervals constructed by the estimated variance from the parametric bootstrap. For brevity, we only report the estimation for the parameters in the dependence model. The simulation study suggests that the stage-wise method provides consistent estimates for the association parameters at a small price of efficiency loss. We argue that the benefit outweighs cost for applied studies with big or high dimensional data.



Table 2: Performance of stage-wise MLE for the dependence model with count marginals

$N = 500$	$\zeta_{11}$	$\zeta_{12}$	$\zeta_{13}$	$\zeta_{21}$	$\zeta_{22}$	$\zeta_{23}$	$\zeta_{31}$	$\zeta_{32}$	$\zeta_{33}$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$
Est.	1.780	1.440	1.198	3.863	2.232	1.431	17.867	3.707	1.722	0.198	0.492	0.778
Bias	0.010	0.000	0.008	0.033	0.012	-0.009	-0.873	-0.123	-0.048	-0.002	-0.008	-0.022
S.D.	0.149	0.106	0.132	0.369	0.126	0.192	2.382	0.367	0.221	0.041	0.045	0.042
CI 90%	0.88	0.87	0.84	0.87	0.87	0.84	0.79	0.88	0.80	0.90	0.89	0.93
CI 95%	0.91	0.93	0.89	0.92	0.94	0.89	0.87	0.93	0.89	0.92	0.92	0.96
CI 99%	0.98	1.00	0.99	0.99	0.97	0.96	0.95	0.96	0.95	0.99	0.97	0.98
$N = 1000$	$\zeta_{11}$	$\zeta_{12}$	$\zeta_{13}$	$\zeta_{21}$	$\zeta_{22}$	$\zeta_{23}$	$\zeta_{31}$	$\zeta_{32}$	$\zeta_{33}$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$
Est.	1.772	1.437	1.185	3.829	2.235	1.447	18.217	3.728	1.730	0.194	0.492	0.787
Bias	0.002	-0.003	-0.005	-0.001	0.015	0.007	-0.523	-0.102	-0.040	-0.006	-0.008	-0.013
S.D.	0.096	0.071	0.090	0.186	0.138	0.119	1.085	0.198	0.108	0.027	0.024	0.018
CI 90%	0.87	0.91	0.85	0.88	0.91	0.91	0.83	0.86	0.88	0.92	0.88	0.88
CI 95%	0.92	0.92	0.90	0.92	0.95	0.96	0.88	0.95	0.92	0.97	0.92	0.94
CI 99%	0.98	0.98	0.95	0.98	0.98	0.97	0.93	0.97	0.98	1.00	0.96	1.00

Table 3: Performance of stage-wise MLE for the dependence model with semi-continuous marginals

$N = 500$	$\zeta_{11}$	$\zeta_{12}$	$\zeta_{13}$	$\zeta_{21}$	$\zeta_{22}$	$\zeta_{23}$	$\zeta_{31}$	$\zeta_{32}$	$\zeta_{33}$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$
Est.	1.762	1.432	1.189	3.859	2.219	1.459	18.452	3.889	1.789	0.201	0.499	0.796
Bias	-0.008	-0.008	-0.001	0.029	-0.001	0.019	-0.288	0.059	0.019	0.001	-0.001	-0.004
S.D.	0.121	0.113	0.112	0.308	0.135	0.177	2.248	0.524	0.357	0.038	0.041	0.025
CI 90%	0.89	0.89	0.82	0.91	0.88	0.88	0.90	0.95	0.93	0.89	0.90	0.88
CI 95%	0.90	0.94	0.87	0.98	0.94	0.97	0.94	0.99	0.96	0.93	0.97	0.91
CI 99%	0.97	0.99	0.97	1.00	0.99	0.97	1.00	1.00	0.99	0.97	0.99	0.98
$N = 1000$	$\zeta_{11}$	$\zeta_{12}$	$\zeta_{13}$	$\zeta_{21}$	$\zeta_{22}$	$\zeta_{23}$	$\zeta_{31}$	$\zeta_{32}$	$\zeta_{33}$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$
Est.	1.772	1.439	1.198	3.844	2.239	1.453	18.457	3.813	1.800	0.196	0.500	0.797
Bias	0.002	-0.001	0.008	0.014	0.019	0.013	-0.283	-0.017	0.030	-0.004	-0.000	-0.003
S.D.	0.086	0.072	0.081	0.161	0.122	0.144	1.274	0.302	0.136	0.025	0.028	0.014
CI 90%	0.92	0.90	0.87	0.88	0.92	0.93	0.85	0.88	0.90	0.90	0.85	0.92
CI 95%	0.98	0.93	0.92	0.92	0.93	0.96	0.91	0.92	0.96	0.95	0.93	0.97
CI 99%	0.99	0.99	0.97	0.98	0.99	0.98	0.98	0.99	0.98	1.00	0.98	1.00

## 5 Empirical Results

We apply the proposed copula regression model to the multivariate longitudinal measurements of insurance claims described in Section 2. This section provides detailed analysis on both claim frequency and loss cost of local government entities from the building and contents coverage under the Wisconsin government property insurance program.

### 5.1 Marginal Analysis

For the number of claims of each peril, we consider the zero-one inflated count regression and its nested cases, including standard Poisson and negative binomial regression (NB2), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), and zero-one inflated Poisson (ZOIP) and zero-one inflated negative binomial (ZOINB).

Table 4 summarizes the goodness-of-fit statistics for various specifications for the claim frequency by the type of peril. For each peril, we compare the empirical claim frequency with the frequency implied from the fitted regression that is calculated as the sum of claim probabilities over all heterogeneous policyholders. The comparison is reported in the first half of the table for the claim counts for the peril of water for illustration. The second half of the table shows the percent of zeros, the percent of ones, and the chi-square statistics. The best model is selected based on the chi-square statistics and is highlighted in the table. The estimation results for the selected models are displayed in Table 5.

To summary, we choose the zero-one inflated NB, the NB2, and the one-inflated NB regression for the perils of water, fire, and other, respectively. Not surprisingly, The number of claims from all peril types exhibit strong evidence of overdispersion, a feature commonly observed in insurance claim counts. Since insurance operation is based on risk pooling, the overdispersion in claim counts is usually related to the excess of zeros corresponding to the large number of policyholders without any claims. Overdispersion and excess of zeros are usually accommodated well by a negative binomial regression or zero-inflated count models. An an interesting finding for property insurance fund data is that, in addition to zero inflation, there is a significant portion of ones at least for the water and other perils.

As suggested by Table 5, there exists significant difference in claim frequency across entity types, although the effects are heterogeneous among the three perils. The alarm credit is not predictive regardless of the peril type. We keep the variable in the analysis to emphasize its different effects on claim frequency and claim severity. It is intuitively understandable that the alarm credit is a more effective tool for loss control rather than loss prevention. The amount of coverage shows substantive predictive power in all components (zero-, one-, and NB2- regression) of the regression model, which is not unexpected since it measures the risk exposure for the policyholder.

For the loss cost of policyholders, we employ the two-component mixture model where the zero component is modeled using a logistic regression and the continuous component is modeled using a GB2 regression. Refer to Shi and Yang (2017) for details on the estimation and diagnostics for

Table 4: Goodness-of-fit statistics for claim frequency by peril

Water	Empirical	Poisson	NB2	ZIP	ZINB	ZOIP	ZOINB <sup>†</sup>
0	4270	3949.18	4304.92	4299.29	4305.45	4268.14	4278.91
1	554	681.15	434.14	405.31	441.95	567.66	543.78
2	132	219.79	145.79	159.27	146.83	118.98	103.52
3	44	94.06	68.95	79.70	68.31	53.65	49.92
4	19	48.40	38.73	46.17	37.77	28.33	28.82
5	21	28.68	24.18	29.02	23.26	16.74	18.53
6	13	18.76	16.21	18.98	15.42	10.69	12.80
7	11	13.00	11.46	12.71	10.80	7.23	9.31
8	3	9.27	8.43	8.73	7.88	5.10	7.04
9	1	6.70	6.40	6.20	5.94	3.70	5.49
10	1	4.85	4.98	4.59	4.59	2.74	4.38
11	1	3.50	3.97	3.53	3.63	2.04	3.57
12	4	2.50	3.21	2.79	2.93	1.52	2.96
13	2	1.76	2.64	2.26	2.39	1.13	2.48
14	1	1.21	2.20	1.85	1.98	0.84	2.11
$\geq 15$	18	12.21	18.61	14.61	15.76	6.53	20.81
% of zeros	0.838	0.775	0.845	0.844	0.845	0.838	0.840
% of ones	0.109	0.134	0.085	0.080	0.087	0.111	0.107
$\chi^2 - stat$		151.139	69.313	110.195	61.891	39.489	<b>24.604</b>
Fire							
% of zeros	0.862	0.845	0.861	0.861	0.861	0.861	—
% of ones	0.101	0.127	0.106	0.102	0.106	0.102	—
$\chi^2 - stat$		63.953	<b>8.178</b>	8.945	8.267	8.945	—
Other							
% of zeros	0.879	0.857	0.882	0.880	0.882	0.880	0.880
% of ones	0.095	0.113	0.085	0.084	0.085	0.094	0.093
$\chi^2 - stat$		100.393	51.157	81.211	51.160	48.520	<b>26.734</b>

<sup>†</sup> For peril “Other”, one-inflated NB regression is used instead of ZOINB.

Table 5: Estimation of count regression models for claim frequency by peril

	Water		Fire		Other	
	EST.	S.E.	EST.	S.E.	EST.	S.E.
(Intercept)	-5.985	0.365	-3.987	0.229	-5.276	0.401
TypeCity	1.270	0.318	1.072	0.216	0.739	0.340
TypeCounty	0.570	0.341	1.622	0.223	0.974	0.350
TypeSchool	-0.273	0.321	0.170	0.217	0.299	0.338
TypeTown	1.767	0.450	0.093	0.326	0.325	0.593
TypeVillage	1.381	0.332	1.047	0.221	0.690	0.364
AC05	-0.166	0.338	0.098	0.228	0.324	0.312
AC10	-0.099	0.273	0.276	0.179	0.107	0.285
AC15	0.065	0.144	0.141	0.102	0.086	0.158
log(Coverage)	1.225	0.061	0.477	0.036	0.808	0.057
$\phi$	0.279	0.044	1.142	0.177	0.370	0.062
Zero Model						
(Intercept)	-4.175	1.452				
log(Coverage)	0.495	0.204				
One Model						
(Intercept)	-3.356	0.152			-4.254	0.268
log(Coverage)	0.159	0.055			0.316	0.072

the marginal models for the three peril types.

## 5.2 Dependence Analysis

In the multivariate longitudinal context, two types of dependence that are embedded within the multilevel structure of the data and are interrelated with each other are temporal association and contemporaneous association. We employ a pair copula construction based on D-vine to accommodate the temporal dependence for each longitudinal response. Note that different functional forms are used for claim count and loss cost due to the scale of the data. In each D-vine, there are (at least) four trees for the data collected over five years. An identical bivariate copula is specified for all the pairs in the same tree. This constraint can be easily relaxed but is necessary for the purpose of prediction because stationary condition is required. The bivariate copula for each pair is selected based on AIC and copula selection is performed sequentially from lower to higher trees. After fixing the parametric forms of the bivariate copulas, we then estimate the parameters following the procedure in Section 4.

We report the selected bivariate copulas and the estimated association parameters for the number of claims in Table 6. The D-vine model for the loss cost follows Shi and Yang (2017). The implied Kendall's  $\tau$  for both claim count and loss cost are summarized in Table 7. First, the results support the statement that the pair copula construction allows for more flexible dependence structure, especially tail and asymmetric dependence. Second, the dependence decreases from lower to higher trees. The diminishing dependence pattern is consistent with the fundamental idea of the

graphical model in that pairs in higher trees are conditioning on a larger set of correlated variables. In particular, a vine becomes truncated when pairs in all higher trees are conditionally independent. For example, as shown in Table 6, the D-vines for claim count from both water and other perils are truncated, with the former at the third tree and the latter at the second tree.

Table 6: Selective bivariate copulas and estimated parameters in D-vine for claim count<sup>†</sup>

	Water		Fire		Other	
	Copula	Parameter	Copula	Parameter	Copula	Parameter
$T_1$	Rotated Gumbel	1.547 (0.054)	Rotated Gumbel	1.299 (0.045)	Clayton	0.920 (0.170)
$T_2$	Rotated Gumbel	1.321 (0.053)	Rotated Gumbel	1.312 (0.052)	Rotated Gumbel	1.252 (0.057)
$T_3$	Frank	1.529 (0.342)	Rotated Gumbel	1.216 (0.056)		
$T_4$			Rotated Clayton	0.172 (0.052)		

<sup>†</sup> Standard errors are presented in parenthesis.

Table 7: Kendall's  $\tau$  in the D-vines for claim count and loss cost

	Claim Count			Loss Cost		
	Water	Fire	Other	Water	Fire	Other
$T_1$	0.359	0.222	0.315	0.347	0.150	0.258
$T_2$	0.246	0.231	0.201	0.276	0.271	0.255
$T_3$	0.166	0.179		0.231	0.135	0.062
$T_4$		0.080		0.146	0.219	

The (conditional) contemporaneous association among three peril types are accommodated using a Gaussian copula with an unstructured correlation matrix. The estimated correlation coefficients for claim count and loss cost are shown in Table 8. The results suggest nonignorable association across different perils which is consistent with the observations in Figure 1 and Figure 2. Note that many alternative specifications other than a Gaussian copula are allowed in the proposed framework to model the contemporaneous association among perils. For instance, both pair copula construction and hierarchical Archimedean copulas are appealing choices. In this application with low dimensionality, the Gaussian copula is in favor due to its balance between interpretability and computational difficulty. In addition, data analysis shows that the Gaussian copula sufficiently captures the association among insurance claim measurements of different perils.

### 5.3 Prediction

Recall that the above model is developed using the insurance claims data from 2006 to 2010, and the data in year 2011 are reserved for out-of-sample validation. We propose a validation procedure

Table 8: Estimation of contemporaneous dependence for claim count and loss cost

		Water-Fire	Water-Other	Fire-Other
Claim Count	$\rho$	0.126	0.199	0.070
	$t$ -stat	3.558	5.083	1.887
Loss Cost	$\rho$	0.085	0.148	0.024
	$t$ -stat	2.607	4.212	0.646

based on the aggregate outcome defined as

$$S_{iT+1} = Y_{iT+1}^{(1)} + \dots + Y_{iT+1}^{(J)}.$$

In our context,  $S_{iT+1}$  represents either the total number of claims or the total loss cost for the  $i$ th policyholder in period  $T + 1$ . From the copula model, one derives the predictive distribution of  $S_{iT+1}$  given claim history  $H_{iT-} = (\mathbf{y}_{i1}, \dots, \mathbf{y}_{iT})$ , denoted by  $F(\cdot|H_{iT-})$ . Let  $s_{iT+1}^o$  denote the realized value of  $S_{iT+1}$  that is observed in the hold-out sample. Define  $u_{iT+1} = F(s_{iT+1}^o|H_{iT-})$ . Our test is based on the idea that  $\{u_{iT+1}; i = 1, \dots, N\}$  is approximately a random sample of a uniform distribution provided the model (both marginal and dependence) is correctly specified.

There are two challenges in the current application. First, the distribution  $F(\cdot|H_{iT-})$  has no analytical form. Second, the aggregate outcome  $S_{iT+1}$  might be discrete or have a mass probability at zero. To address these issues, we derive the predictive distribution using simulation. Furthermore, we consider a generalized distribution transformation to overcome the potential zero mass and discreteness in the distribution (see Rüschendorf (2009)). The uniformity is tested using the Kolmogorov-Smirnov statistics and the results are presented in Table 9. To emphasize dependence, we compare two models, the independence model versus the proposed copula model, with the same marginal specifications. The test statistics support the dependence assumption for both claim count and loss cost outcomes.

Table 9: Uniform test for claim count and loss cost<sup>†</sup>

	Independence	Copula
Claim Count	0.046 (0.025)	0.027 (0.441)
Loss Cost	0.059 (0.002)	0.023 (0.661)

<sup>†</sup>  $p$ -values are presented in parenthesis.

Assuring that the copula model is better suited for the property insurance claims data than the independence model, we compare the predictive performance between the two cases. The prediction is based on the aggregate outcome  $S_{iT+1}$  defined above. In the assessment, we consider three scoring rules as described in Czado et al. (2009), the ranked probability score (RPS), the quadratic score (QS), and the spherical score (SPHS). The essential idea is to quantify the closeness between the realized outcome  $s_{iT+1}^o$  in the hold-out sample and the predictive distribution of  $S_{iT+1}$ . Thus the lower the score, the more accurate is the prediction. The three scores are calculated for each of the 1,019 policyholders under both the independence model and the copula model. Given the

independence assumption among policyholders, we examine the probability that the copula model outperforms the independence model. Table 10 reports the empirical probability for both claim count and loss cost. Only RPS is reported for the loss cost because the other two measures are not well defined for semi-continuous outcomes. The copula model demonstrates superior prediction to the independence model about 70% and 65% of the time for claim count and loss cost, respectively. A one-sided binomial test further confirms the statistical significance.

Table 10: Empirical probability of superior prediction of copula model

	Claim Count	Loss Cost
RPS	69.09%	64.77%
QS	68.40%	
SPHS	68.99%	

## 6 Concluding Remarks

Motivated by the predictive applications for the non-life insurance products with bundling features, we proposed a dependence modeling framework using pair copula constructions to account for the temporal and contemporaneous association in the multivariate longitudinal outcomes. The proposed framework is generic in that it easily accommodates measurements of different scales. In particular, we demonstrated its flexibility for the zero-one inflated claim count and semi-continuous loss cost for a government property insurance program. Although in our analysis, the marginal distributions of the multiple measurements are of the same scale, be it continuous, discrete, or semi-continuous, the proposal framework certainly accommodates the case where the multivariate outcomes are of different types, for instance, one marginal is a continuous distribution while the other is a discrete distribution.

A separate but related strand of literature is regarding pair copula construction for multivariate time series data. For example, Brechmann and Czado (2015) and Smith (2015) discussed possible strategies of constructing vines for jointly modeling multiple time series outcomes. It is worth noting the difference between the two groups of studies. The time series literature focuses on the dynamics for marginal and/or dependence, and thus stationarity is necessary for statistical inference. In contrast, the longitudinal analysis in our applications uses data over a shorter and fixed time periods where stationarity is not required and dynamics is less of a concern.

The proposed framework can be easily adopted for a much broader range of topics. One area is the multivariate time series data that is discussed above. Another area is the spatial data with geocoding. For instance, Erhardt et al. (2015) employed a regular vine approach for spatial time series data, and Krupskii and Genton (2016) proposed a copula model for replicated multivariate spatial data. Both types of outcomes can potentially be accommodated by the copula model presented in this application.



## 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.
- Barthel, N., C. Geerdens, M. Killiches, P. Janssen, and C. Czado (2018). Vine copula based likelihood estimation of dependence patterns in multivariate event time data. *Computational Statistics & Data Analysis* 117, 109–127.
- Bedford, T. and R. M. Cooke (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence* 32(1), 245–268.
- Bedford, T. and R. M. Cooke (2002). Vines—a new graphical model for dependent random variables. *Annals of Statistics* 30(4), 1031–1068.
- Boucher, J.-P. and R. Inoussa (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin: The Journal of the International Actuarial Association* 44(3), 587–612.
- Brechmann, E. C. and C. Czado (2015). COPAR—multivariate time series modeling using the copula autoregressive model. *Applied Stochastic Models in Business and Industry* 31(4), 495–514.
- Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Biometrics* 65(4), 1254–1261.
- Dionne, G. and C. Vanasse (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics* 7(2), 149–165.
- Erhardt, T. M., C. Czado, and U. Schepsmeier (2015). R-vine models for spatial time series with an application to daily mean temperature. *Biometrics* 71(2), 323–332.
- Frees, E., P. Shi, and E. Valdez (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin: The Journal of the International Actuarial Association* 39(1), 165–197.
- Frees, E. and E. Valdez (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103(484), 1457–1469.
- Frees, E. W. (2015). Analytics of insurance markets. *Annual Review of Financial Economics* 7, 253–277.
- Frees, E. W., G. Lee, and L. Yang (2016). Multivariate frequency-severity regression models in insurance. *Risks* 4(1), 1–36.
- Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods* 23(11), 3105–3119.

- Gray, S. M. and R. Brookmeyer (1998). Estimating a treatment effect from multidimensional longitudinal data. *Biometrics*, 976–988.
- Gray, S. M. and R. Brookmeyer (2000). Multidimensional longitudinal data: estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association* 95(450), 396–406.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* 94(2), 401–419.
- Joe, H. (2014). *Dependence Modeling with Copulas*. New York: Chapman & Hall.
- Joe, H. and D. Kurowicka (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific.
- Krupskii, P. and M. G. Genton (2016). A copula-based linear model of coregionalization for non-gaussian multivariate spatial data. *Working Paper*.
- Kurowicka, D. and R. M. Cooke (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. John Wiley & Sons.
- Lambert, P. and F. Vandenhende (2002). A copula-based model for multivariate non-normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine* 21(21).
- Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(1), 13–22.
- Molenberghs, G. and E. Lesaffre (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association* 89(426), 633–644.
- Nelsen, R. (2006). *An Introduction to Copulas* (2nd ed.). New York: Springer.
- 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.
- Pinquet, J. (2013). Experience rating in nonlife insurance. In G. Dionne (Ed.), *Handbook of Insurance*, pp. 471–485. Springer.
- Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association* 79(386), 406–414.
- Rochon, J. (1996). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics*, 740–750.
- Roy, J. and X. Lin (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* 56(4), 1047–1054.

- Roy, J. and X. Lin (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates. *Journal of the American Statistical Association* 97(457), 40–52.
- Rüschendorf, L. (2009). On the distributional transform, sklar’s theorem, and the empirical copula process. *Journal of Statistical Planning and Inference* 139(11), 3921–3927.
- Shah, A., N. Laird, and D. Schoenfeld (1997). A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association* 92(438), 775–779.
- Shi, P. (2012). Multivariate longitudinal modeling of insurance company expenses. *Insurance: Mathematics and Economics* 51(1), 204–215.
- Shi, P., X. Feng, and J.-P. Boucher (2016). Multilevel modeling of insurance claims using copulas. *Annals of Applied Statistics* 10(2), 834–863.
- Shi, P. and L. Yang (2017). Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association* (DOI: 10.1080/01621459.2017.1330692).
- Smith, M. S. (2015). Copula modelling of dependence in multivariate time series. *International Journal of Forecasting* 31(3), 815–833.
- Stöber, J., H. G. Hong, C. Czado, and P. Ghosh (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics & Data Analysis* 88, 28–39.
- Verbeke, G., S. Fieuws, G. Molenberghs, and M. Davidian (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research* 23(1), 42–59.
- Zhao, Z. and Z. Zhang (2017). Semiparametric dynamic max-copula model for multivariate time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (DOI: 10.1111/rssb.12256).