Journal of the American Statistical Association

March 2018 • Volume 113 • Number 521

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Pair Copula Constructions for Insurance Experience Rating

Peng Shi & Lu Yang

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 521, 122–133, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1330692

Taylor & Francis
Taylor & Francis Group

Pair Copula Constructions for Insurance Experience Rating

Peng Shi[a] and Lu Yang[b]

[a]Wisconsin School of Business, University of Wisconsin-Madison, Madison, WI; [b]Amsterdam School of Economics, University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT

In nonlife insurance, insurers use experience rating to adjust premiums to reflect policyholders' previous claim experience. Performing prospective experience rating can be challenging when the claim distribution is complex. For instance, insurance claims are semicontinuous in that a fraction of zeros is often associated with an otherwise positive continuous outcome from a right-skewed and long-tailed distribution. Practitioners use credibility premium that is a special form of the shrinkage estimator in the longitudinal data framework. However, the linear predictor is not informative especially when the outcome follows a mixed distribution. In this article, we introduce a mixed vine pair copula construction framework for modeling semicontinuous longitudinal claims. In the proposed framework, a two-component mixture regression is employed to accommodate the zero inflation and thick tails in the claim distribution. The temporal dependence among repeated observations is modeled using a sequence of bivariate conditional copulas based on a mixed D-vine. We emphasize that the resulting predictive distribution allows insurers to incorporate past experience into future premiums in a nonlinear fashion and the classic linear predictor can be viewed as a nested case. In the application, we examine a unique claims dataset of government property insurance from the state of Wisconsin. Due to the discrepancies between the claim and premium distributions, we employ an ordered Lorenz curve to evaluate the predictive performance. We show that the proposed approach offers substantial opportunities for separating risks and identifying profitable business when compared with alternative experience rating methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2016
Revised March 2017

KEYWORDS
Government insurance; Mixed D-vine; Mixture regression; Predictive distribution; Zero inflation

1. Introduction

In nonlife (property, liability, and health) insurance, insurers use experience rating to adjust premiums to reflect policyholders' past loss experience. Premiums decrease (increase) if the experience of a policyholder is better (worse) than that assumed in the manual rate—a premium rate developed from the experience of a large number of homogeneous policies defined by the insurer's risk classification system. Experience rating can be prospective or retrospective. We restrict our consideration to the former that points to a predictive modeling application.

Experience rating improves insurance market efficiency as a dynamic contract mechanism under information asymmetry, therefore providing a competitive advantage for insurers deft at its employment over the rival firms. First, an insurer's risk classification system might not be perfect. Unobserved heterogeneity remains after all underwriting criteria are accounted for. Experience rating allows insurers to further separate good risks from bad risks, and thus helps mitigate adverse selection. Second, adjusting premium based on past experience gives policyholders incentives for loss prevention, which is known as moral hazard in the economics literature.

The statistical component of experience rating is to model longitudinal insurance claims and to infer the predictive distribution of future claims given previous loss experience. This can be a difficult task when the risk distribution is complex. For instance, the distribution of claims is well known to be a mixture of zeros and a right-skewed and long-tailed distribution. The degenerate distribution at zero corresponds to no claims and the positive thick-tailed distribution describes the amount of claims given occurrence.

Insurers use credibility ratemaking to perform prospective experience rating (adjust future premium based on past experience) on a risk or a group of risks. The intuitive concept of credibility premium is to express the expected future claim of a given risk class as a weighted sum of the average claim from the risk class and the average claim over all other risk classes, which begs the question that how much of the experience of a given policyholder is due to the random variation in the underlying risk and how much is due to the policyholder being better or worse than average. The classic work of Bühlmann (1967) provided a systematic solution using what is known as random-effects framework, thereby modern theory of credibility has developed and flourished. Frees, Young, and Luo (1999) established the link between the credibility theory in actuarial science and the longitudinal data models in statistics, and noted that the credibility predictor is a special form of the shrinkage estimator in the longitudinal data framework.

Due to the zero inflation and long tails exhibited in the insurance claims, standard longitudinal data models are not ready to apply to insurers' experience rating. One attractive approach to characterizing the complex structure of semicontinuous longitudinal data is the two-part mixed models with correlated ran-

CONTACT Peng Shi pshi@bus.wisc.edu Wisconsin School of Business, University of Wisconsin-Madison, Madison, WI 53706.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

dom effects (see, e.g., Olsen and Schafer (2001)). An alternative strategy is the mixed-effects Tweedie compound Poisson model (Dunn and Smyth (2005, 2008) and Zhang (2013)). However, both approaches are subject to several difficulties in the current application of predictive modeling. First, likelihood-based estimation is computationally expensive, especially with big data such as in insurance. Second, prediction of random effects in nonlinear models is never an easy task, which further hinders the derivation of the predictive distribution. Third, the structural assumption of the subject-specific heterogeneity implies a symmetric relation among repeated observations, and thus limits the way past experience is incorporated into the prediction.

In this article, we introduce a mixed vine pair copula construction framework for modeling semicontinuous longitudinal claims. In the proposed framework, a two-component mixture regression is employed to accommodate the zero inflation and thick tails in the claim distribution. The temporal dependence among repeated observations is modeled using a sequence of bivariate conditional copulas based on a mixed D-vine. Copula has become a useful analytical tool in multivariate analysis (see Joe (2014) for the recent advancement on copulas). Early efforts of using copula regression to model longitudinal data include Frees and Wang (2005), Sun, Frees, and Rosenberg (2008), Smith et al. (2010), and Shi (2012) among others. However, predictive applications of copula model for semicontinuous longitudinal insurance claims are rarely found in the literature. Shi, Feng, and Boucher (2016) is one recent example.

The proposed approach enjoys several advantages compared with the methods available in the existing literature. First, the mixture regression combines the merits of both the two-part and the Tweedie models. Unlike the Tweedie, it allows the analyst to use different sets of predictors for the frequency and severity of claims. In the meanwhile, it does not require separate copulas for the frequency and severity components, and thus avoids the unbalanced data issue in the conditional severity model. Second, compared with the elliptical copulas in the current literature of experience rating, the mixed vine pair copula construction allows for more flexible dependence structure by using asymmetric bivariate copulas as building blocks. In addition, the computational burden is much lower than the case of elliptical copulas when there are discrete components in the response variable. Third, for the purposes of experience rating, we are interested in one particular type of statistical inference—prediction. Under the pair copula framework, it is straightforward to derive the predictive distribution of future claim given past experience without referring to the Bayesian approach. We also point out that many existing credibility predictors can be viewed in the proposed approach.

The main contribution of this article to the literature is the introduction of the vine pair copula constructions for mixed data and the novel application in insurance experience rating. Vine copulas have been studied for both continuous and discrete data. Following the seminal work of Bedford and Cooke (2001, 2002) on this new class of graphical models, 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 for continuous data. Smith et al. (2010) employed a Bayesian approach and investigated copula selection in the D-vine for longitudinal data. More recently, Panagiotelis, Czado, and Joe (2012) introduced the discrete analogue to the vine pair copula construction. Stöber (2013) and Stöber et al. (2015) studied the theory and applications of pair copula constructions for mixed data. Our work fills the blank in the literature on vine copulas for a special type of mixed outcome—hybrid data. Specifically, in Stöber's work, "mixed data" corresponds to the case where a copula is used to join a continuous distribution and a discrete distribution. In contrast, we use "mixed data" to refer to the case of a hybrid distribution, that is a random variable with both discrete and continuous components, and a copula is used to join two mixed or hybrid distributions. Note that we motivate the mixed D-vine structure using the predictive nature of the application in insurance experience rating. However, the notion of mixed vine pair copula construction easily extends to regular vines.

2. Wisconsin Local Government Property Insurance Fund

In experience rating, we examine the insurance coverage for building and contents of local government entities in Wisconsin provided by the Local Government Property Insurance Fund (LGPIF). The LGPIF is administered by the Wisconsin Office of the Commissioner of Insurance, and the purpose of it is to make property insurance available for local government units, such as counties, cities, towns, villages, school districts, and library boards. Data are collected for 1019 local government entities over six years from 2006 to 2011. Due to 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 for pricing insurance contracts because other sources of market data may not be relevant. We use data in years 2006–2010 to develop the model and reserve the data of 2011 for validation.

The quantity of interest is the entity-level cost of claims that serves as the basis for determining the pure premium. Similar to private commercial insurers, the government insurance fund keeps track of claims for its pool of policyholders, from which we derive the total claim cost of each entity for each year. In addition, the fund further breaks down the total cost of claims by the cause of losses, known as peril in property insurance. In this application, we examine the total cost as well as the cost by peril. Statistically speaking, one might prefer to analyze claims by peril presuming that more information is revealed at granular level observations. On the other hand, one might argue for the simplicity of aggregating data across perils in the sense of sufficient statistics. In practice, the choice often depends on the preference of the analyst and the type of data collected by the insurer. We view this as an empirical question and compare the predictive performance of both common practices.

Table 1 summarizes the distribution of claim cost by year. The first panel corresponds to the total claims and the other three correspond to losses caused by water, fire, and other perils, respectively. Water and fire (including smoke) damages are among those of highest frequency of occurrence. Examples of other perils include lightning strikes, windstorms and hail, explosion. All four outcomes are semicontinuous in that a significant portion of zeros is associated with an otherwise positive continuous outcome. The zeros imply no claims and

Table 1. Distribution of the claim cost by year and by peril.

Total				Water			
Year	p_0	Mean	SD	Year	p_0	Mean	SD
2006	0.723	71,338	499,690	2006	0.847	67,082	536,722
2007	0.677	51,225	178,513	2007	0.821	13,791	27,256
2008	0.716	40,439	146,224	2008	0.850	15,260	41,462
2009	0.722	36,932	143,783	2009	0.851	20,995	52,598
2010	0.626	94,784	728,353	2010	0.821	119,892	986,053
Fire				Other			
Year	p_0	Mean	SD	Year	p_0	Mean	SD
2006	0.855	19,893	58,387	2006	0.920	81,808	515,961
2007	0.863	43,441	139,471	2007	0.863	59,010	201,994
2008	0.873	40,087	167,739	2008	0.887	36,357	118,778
2009	0.879	32,975	87,335	2009	0.912	35,606	167,402
2010	0.840	33,384	111,541	2010	0.816	47,330	243,250

the positive component indicates the size of claims. In the table, we report the probability of zero claim denoted by p_0. For instance, regardless of the cause of loss, about 72.3% entities did not report any claim during year 2006. As expected, the percentage of zeros is larger when decomposing claims by peril, and water and fire damages are more common than other perils.

Conditioning on at least one claims, we also present in Table 1 the mean and standard deviation of the amount of claims. The large standard deviation is as anticipated and is indicative of the skewness and thick tails in the claim size distribution. Another noticeable feature in severity is the substantial variation across years, especially for water damages. This is in contrast with claim frequency where temporal variation is less pronounced. We attribute the temporal variation in the claim size to the heavy tails of the underlying distribution, and we accommodate such data feature by using a flexible parametric regression. To visualize the size distribution, Figure 1 displays the violin plots of the amount of claims by year and by peril. One can think of a violin plot as a marriage of a boxplot and a density trace (see Hintze and Nelson (1998) for more details). The plots suggest that the occurrence of extremely large losses is not unusual and the claims related to water damages are more volatile than fire and other perils. Overall, zero inflation and heavy tails in the claim cost distribution, as shown in Table 1 and Figure 1, motivate the two-component mixture regression in Section 3.1.

Table 2. Description and summary statistics of covariates.[†]

		Year =				
Variable	Description	2006	2007	2008	2009	2010
TypeCity	= 1 if entity type is city	0.146				
TypeCounty	= 1 if entity type is county	0.061				
TypeSchool	= 1 if entity type is school	0.291				
TypeTown	= 1 if entity type is town	0.164				
TypeVillage	= 1 if entity type is village	0.231				
AC05	= 1 indicate 5% alarm credit	0.025	0.026	0.034	0.054	0.074
AC10	= 1 indicate 10% alarm credit	0.045	0.051	0.050	0.067	0.084
AC15	= 1 indicate 15% alarm credit	0.381	0.399	0.434	0.486	0.544
Coverage	Amount of coverage in log scale	2.065	2.153	2.227	2.285	2.292
		(2.021)	(2.000)	(1.981)	(1.990)	(1.987)

[†] Standard deviation for continuous covariates is reported in parenthesis.

In a risk classification system, an insurer uses observed policyholder and contract characteristics to explain the variability in the insurance claims and then reflects such heterogeneity in the ratemaking. For example, the large claim amount could, to certain extent, relate to the size of the coverage. Table 2 presents the rating variables, their descriptions, and the associated descriptive statistics. Unlike personal lines of business (such as automobile and homeowner insurance), we have a very limited number of predictors used in the rating system, which is not unusual in commercial insurance ratemaking. One rating variable is the entity type that indicates whether the covered buildings belong to a city, county, school, town, village, or a miscellaneous entity such as fire stations. Apparently the entity type does not change over the years. For example, about 15% policyholders are city entities and 30% are school districts. We set miscellaneous entity (TypeMisc) as the reference level in the analysis. As an incentive to prevent and mitigate loss, the fund offers credits for the different types of fire alarms. 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 of the main rooms, and a 15% discount if the alarms are 24/7 monitored. No alarm

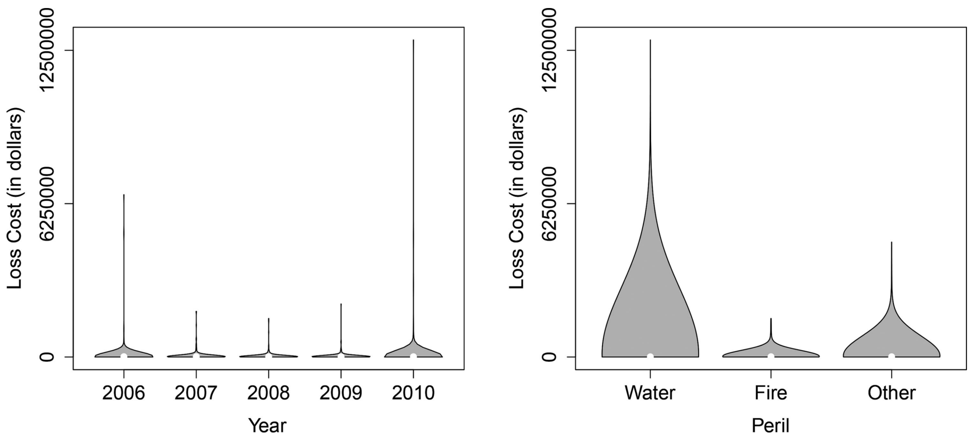

Figure 1. Violin plots of the amount of claims by year and by peril.

credit (AC00) is omitted as the reference level in the regression analysis. The alarm credit is often subject to the underwriter's discretion. The increasing temporal pattern in the alarm credit might be because policyholders are responsive to the incentives and the advanced alarm system becomes accessible at a lower cost. Because of the skewness in the amount of coverage (in million dollars), we report its mean and standard deviation (in parenthesis) of the coverage amount in the log scale. The statistics indicate a relatively small variation in coverage over time.

3. Modeling Semicontinuous Longitudinal Data

3.1. Marginal Model

Let Y_{it} denote the cost (total or by peril) of claims for policyholder i $(= 1, \dots, n)$ in year t $(= 1, \dots, T)$. We consider a two-component mixture model to accommodate the mass probability at zero, the skewness, and the long tails of the distribution. Specifically, Y_{it} is assumed as being generated from a degenerate distribution at zero with probability p_{it} and being generated from a skewed and heavy tailed distribution $G_{it}(\cdot)$ defined on $(0, +\infty)$ with probability $1 - p_{it}$. Assuming independence between the degenerate distribution and the skewed heavy-tailed distribution, the resulting variable follows a mixed distribution. Let $F_{it}(\cdot)$ and $f_{it}(\cdot)$ denote its distribution function and density function, respectively. It is shown:

$$\begin{aligned} F_{it}(y) &= p_{it} + (1 - p_{it})G_{it}(y), \\ f_{it}(y) &= p_{it} I(y = 0) + (1 - p_{it})g_{it}(y). \end{aligned} \tag{1}$$

Here $I(\cdot)$ is the indicator function and g_{it} is the density function associated with G_{it}.

In the above formulation, the zero component models the probability of incurring claims, and the continuous component models the amount of claims given occurrence. Separating the frequency and severity allows for different sets of predictors as well as different effects of the same predictor on each component. This is a common practice in pricing nonlife insurance contracts. Using property insurance as an example, one can think that the probability of having claims is more related to the risk profile of the property, while the amount of payment is, to a great extent, determined at the adjuster's discretion.

For the claim frequency, we consider a logit specification due to the straightforward interpretability of model parameters:

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) = \boldsymbol{x}'_{1it}\boldsymbol{\beta}_1,$$

where $\boldsymbol{x}_{1it}$ represents the vector of explanatory variables and $\boldsymbol{\beta}_1$ denotes the corresponding regression coefficients to be estimated. For the claim severity, we employ the generalized beta of the second kind (GB2) distribution. (See Shi (2014) for discussions of alternative strategies for handling skewness and heavy tails in insurance claims.) 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. Following this line of studies, we consider the formulation:

$$g_{it}(y) = \frac{\exp(\kappa_1 \omega_{it})}{y|\sigma|B(\kappa_1, \kappa_2)[1 + \exp(\omega_{it})]^{\kappa_1 + \kappa_2}}, \tag{2}$$

where $\omega_{it} = (\ln y - \mu_{it})/\sigma$ and $B(\kappa_1, \kappa_2)$ is the Euler beta function. The GB2 is a member of the log location-scale family with location parameter μ_{it}, scale parameter σ, and shape parameters κ_1 and κ_2. With four parameters, the GB2 distribution is very flexible to model skewed and heavy-tailed data. For instance, $\kappa_1 > \kappa_2$ indicates right skewness and $\kappa_1 < \kappa_2$ left skewness. The rth moment is $\mathrm{E}(Y^r) = \exp(\mu_{it} r)B(\kappa_1 + r\sigma, \kappa_2 - r\sigma)/B(\kappa_1, \kappa_2)$, where $-\kappa_2 < r\sigma < \kappa_2$. The location parameter is further modeled as a linear combination of covariates to control for the observed heterogeneity $\mu_{it} = \boldsymbol{x}'_{2it}\boldsymbol{\beta}_2$, with $\boldsymbol{x}_{2it}$ and $\boldsymbol{\beta}_2$ being the vector of predictors and regression coefficients, respectively.

3.2. Dependence Model

3.2.1. General Framework

Consider a vector of random variables $\boldsymbol{Z} = (Z_1, \dots, Z_m)'$ with each component following a mixed distribution. In this application, we focus on the zero inflated data that mimic the claim cost in nonlife insurance. The idea is easily extended to the general mixed case. Let $\boldsymbol{z} = (z_1, \dots, z_m)'$ denote a realization of $\boldsymbol{Z}$. Below we lay out a general framework to construct a high-dimensional mixed distribution $f(z_1, \dots, z_m)$ by using bivariate pair copulas as building blocks.

Let $\mathcal{V}_m$ denote a vine on m elements. A regular vine consists of $m - 1$ trees $\mathcal{T}_l, l = 1, \dots, m - 1$, and $\mathcal{T}_l$ is connected by nodes $\mathcal{N}_l$ and edges $\mathcal{E}_l$. Edges in a tree become nodes in the next tree, that is $\mathcal{N}_l = \mathcal{E}_{l-1}$ $(l = 2, \dots, m - 1)$. If two nodes in tree $\mathcal{T}_l$ are joined by an edge, the corresponding edges in tree $\mathcal{T}_{l-1}$ share a node. Define edge set of $\mathcal{V}_m$ as $\mathcal{E}(\mathcal{V}_m) = \mathcal{E}_1 \cup \ldots \cup \mathcal{E}_{m-1}$. To develop the mixed vine, we adopt similar notations used in Panagiotelis, Czado, and Joe (2012). Let Z be a scale element of $\boldsymbol{Z}$ and $\boldsymbol{V}$ be a subset of $\boldsymbol{Z}$ satisfying $Z \notin \boldsymbol{V}$. Let V_h be any scalar element of $\boldsymbol{V}$ and $\boldsymbol{V}_{\setminus h}$ its complement. Specify the conditional bivariate mixed distributions using copula:

$$f_{Z,V_h|\boldsymbol{V}_{\setminus h}}(z, v_h|\boldsymbol{v}_{\setminus h}) = \begin{cases} C_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right) & z = 0, v_h = 0 \\ f_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}) c_{1,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right) & z > 0, v_h = 0 \\ f_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h}) c_{2,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & z = 0, v_h > 0 \\ f_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}) f_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h}) c_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & z > 0, v_h > 0, \end{cases} \tag{3}$$

where $C_{Z,V_h;\boldsymbol{V}_{\setminus h}}(u_1, u_2)$ and $c_{Z,V_h;\boldsymbol{V}_{\setminus h}}(u_1, u_2)$ are the bivariate copula and density function associated with conditional distributions $F_{Z|\boldsymbol{V}_{\setminus h}}$ and $F_{V_h|\boldsymbol{V}_{\setminus h}}$, respectively. And $c_{k,Z,V_h;\boldsymbol{V}_{\setminus h}}(u_1, u_2) = \partial C_{Z,V_h;\boldsymbol{V}_{\setminus h}}(u_1, u_2)/\partial u_k$, for $k = 1, 2$. For inference, we require the simplifying assumption that the copula does not directly rely on the conditioning set (see, e.g., Haff, Aas, and Frigessi (2010) and Stoeber, Joe, and Czado (2013)).

To evaluate (3), we further derive the following generic conditional quantities:

$$f_{Z|\boldsymbol{V}}(z|\boldsymbol{v}) = f_{Z|V_h,\boldsymbol{V}_{\setminus h}}(z|v_h,\boldsymbol{v}_{\setminus h}) = \begin{cases} \dfrac{C_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right)}{F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})} & z=0,\, v_h=0 \\ \dfrac{f_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h})c_{1,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right)}{F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})} & z>0,\, v_h=0 \\ c_{2,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & z=0,\, v_h>0 \\ f_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h})c_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & z>0,\, v_h>0 \end{cases} \tag{4}$$

and

$$\begin{aligned} &F_{Z|\boldsymbol{V}}(z|\boldsymbol{v}) \\ &\quad = F_{Z|V_h,\boldsymbol{V}_{\setminus h}}(z|v_h,\boldsymbol{v}_{\setminus h}) \\ &\quad = \begin{cases} \dfrac{C_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right)}{F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})} & z=0,\, v_h=0 \\ \dfrac{C_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right)}{F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})} & z>0,\, v_h=0 \\ c_{2,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & z=0,\, v_h>0 \\ c_{2,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & z>0,\, v_h>0 \end{cases} \\ &\quad = \begin{cases} \dfrac{C_{Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})\right)}{F_{V_h|\boldsymbol{V}_{\setminus h}}(0|\boldsymbol{v}_{\setminus h})} & v_h=0 \\ c_{2,Z,V_h;\boldsymbol{V}_{\setminus h}}\left(F_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\right) & v_h>0 \end{cases}. \end{aligned} \tag{5}$$

Defining

$$\tilde{f}_{Z,V_h|\boldsymbol{V}_{\setminus h}}(z,v_h|\boldsymbol{v}_{\setminus h}) := \frac{f_{Z,V_h|\boldsymbol{V}_{\setminus h}}(z,v_h|\boldsymbol{v}_{\setminus h})}{f_{Z|\boldsymbol{V}_{\setminus h}}(z|\boldsymbol{v}_{\setminus h})f_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})}, \tag{6}$$

one can express the joint distribution of $\boldsymbol{Z}$ using the bivariate building blocks as:

$$f_{\boldsymbol{Z}}(z_1,\ldots,z_m) = \prod_{j=1}^{m} f_{Z_j}(z_j) \prod_{[Z,V_h|\boldsymbol{V}_{\setminus h}]\in\mathcal{E}(\mathcal{V}_m)} \tilde{f}_{Z,V_h|\boldsymbol{V}_{\setminus h}}(z,v_h|\boldsymbol{v}_{\setminus h}). \tag{7}$$

Definition (6) is the ratio of the bivariate distribution to the product of marginals given the conditioning set. Thus, one can interpret (6) as the (conditional) "dependence ratio" with a ratio of one indicating conditional independence. Each ratio corresponds to one building block in the pair copula construction.

Equation (7) shows that the joint distribution can be expressed as the product of marginals and the bivariate building blocks. Detailed discussion is provided in Appendix A.3. Formulation (7) provides a general framework for the pair copula construction in that both continuous and discrete vines can be viewed in the same framework as well. We articulate this point in more detail using the example of D-vine in sec. 3.2.2.

3.2.2. Mixed D-Vine

For this application, we focus on a specific vine—D-vine. We use the term "mixed D-Vine" to refer to a D-Vine with a distribution that is a combination of a discrete and continuous components. Due to its simplicity, the D-vine is one of the most popular vine structures used in applied studies. An example of a D-vine on five elements is exhibited in Figure 2. The key feature of the D-vine is that the nodes of each tree only connect adjacent nodes. For instance, the nodes in the first tree represent ordered marginals, and the edges in each tree become the nodes in the next tree. Each edge corresponds to a (conditional) bivariate distribution that we construct using a parametric copula. The edges of the entire vine indicate the bivariate building blocks that contribute to the pair copula constructions.

In longitudinal data, cross-sectional subjects are repeatedly observed over time. The temporal order makes D-vine a natural choice. Consider a mixed variables for T periods. The joint distribution of $(Z_1,\ldots,Z_T)$ can be expressed based on a D-vine as:

$$\begin{aligned} &f_{\boldsymbol{Z}}(z_1,\ldots,z_T) \\ &\quad = f(z_T|z_{T-1},\ldots,z_1)\times\cdots\times f(z_2|z_1)f(z_1) \\ &\quad = \prod_{t=1}^{T} f_t(z_t) \prod_{t=2}^{T}\prod_{s=1}^{t-1} \tilde{f}_{s,t|(s+1):(t-1)}(z_s,z_t|z_{s+1},\ldots,z_{t-1}), \end{aligned} \tag{8}$$

where using (6), we show:

$$\begin{aligned} &\tilde{f}_{s,t|(s+1):(t-1)}(z_s,z_t|z_{s+1},\ldots,z_{t-1}) \\ &= \begin{cases} \dfrac{C_{s,t;(s+1):(t-1)}\left(F_{s|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1}), F_{t|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1})\right)}{F_{s|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1})F_{t|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1})} & z_s=0,\, z_t=0 \\ \dfrac{c_{1,s,t;(s+1):(t-1)}\left(F_{s|(s+1):(t-1)}(z_s|z_{s+1},\ldots,z_{t-1}), F_{t|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1})\right)}{F_{t|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1})} & z_s>0,\, z_t=0 \\ \dfrac{c_{2,s,t;(s+1):(t-1)}\left(F_{s|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1}), F_{t|(s+1):(t-1)}(z_t|z_{s+1},\ldots,z_{t-1})\right)}{F_{s|(s+1):(t-1)}(0|z_{s+1},\ldots,z_{t-1})} & z_s=0,\, z_t>0 \\ c_{s,t;(s+1):(t-1)}\left(F_{s|(s+1):(t-1)}(z_s|z_{s+1},\ldots,z_{t-1}), F_{t|(s+1):(t-1)}(z_t|z_{s+1},\ldots,z_{t-1})\right) & z_s>0,\, z_t>0 \end{cases}. \end{aligned} \tag{9}$$

There are two points worth stressing. First, the decomposition in (8) is not unique. The order of these random variables determines pair copula building blocks and each decomposition

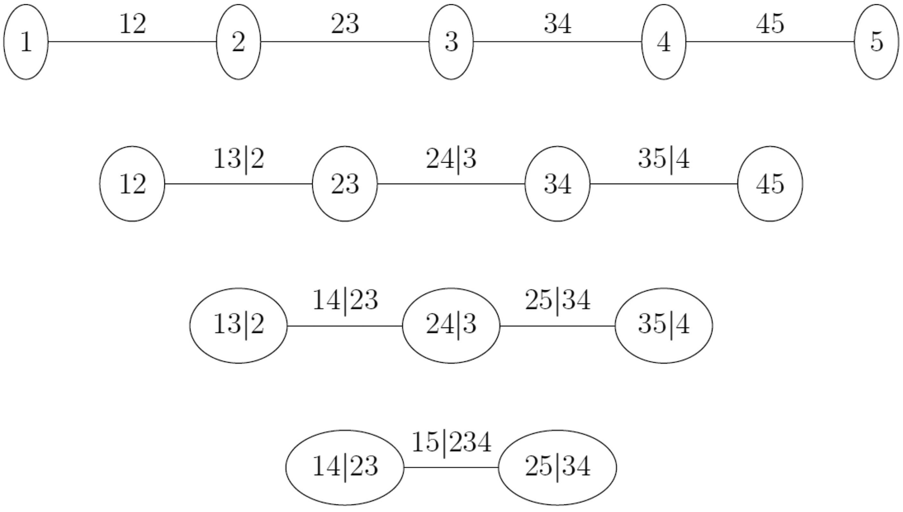

Figure 2. A five-dimension D-vine.

corresponds to a graphical model with a specific vine structure. For a T dimensional vector, there are $\frac{T!}{2} \times 2^{\binom{T-2}{2}}$ possible vine trees, which points to a vine selection problem (see, e.g. Dißmann et al. (2013), Gruber et al. (2015), Panagiotelis et al. (2015)). We choose the D-vine due to the longitudinal nature of the application. Hence, vine selection is not a concern for this study. However, the other aspect of model selection—copula selection—is of more importance and we will discuss this issue in sec. 3.3. Second, both continuous and discrete pair copula constructions can be viewed in this general framework. To be more specific, we recognize the following two cases that can be derived using (6):

(1) Continuous vine (Aas et al. (2009))

$$\begin{aligned}&\tilde{f}_{s,t|(s+1):(t-1)}(z_s, z_t|z_{s+1}, \ldots, z_{t-1})\\&\quad= c_{s,t;(s+1):(t-1)}(F_{s|(s+1):(t-1)}(z_s|z_{s+1}, \ldots, z_{t-1}),\\&\qquad\qquad F_{t|(s+1):(t-1)}(z_t|z_{s+1}, \ldots, z_{t-1}))\end{aligned}$$

(2) Discrete vine (Panagiotelis, Czado, and Joe (2012))

$$\begin{aligned}&\tilde{f}_{s,t|(s+1):(t-1)}(z_s, z_t|z_{s+1}, \ldots, z_{t-1})\\&\quad= \frac{\sum_{i_1=0,1}\sum_{i_2=0,1}(-1)^{i_1+i_2}C_{s,t;(s+1):(t-1)}\left(F_{s|(s+1):(t-1)}(z_s - i_1|z_{s+1}, \ldots, z_{t-1}), F_{t|(s+1):(t-1)}(z_t - i_2|z_{s+1}, \ldots, z_{t-1})\right)}{f_{s|(s+1):(t-1)}(z_s|z_{s+1}, \ldots, z_{t-1})\, f_{t|(s+1):(t-1)}(z_t|z_{s+1}, \ldots, z_{t-1})}.\end{aligned}$$

3.3. *Inference*

Due to the parametric nature of the proposed model, we employ likelihood-based method for estimation. Consider a portfolio of n policyholders, the total log-likelihood function is

$$\begin{aligned}ll(\boldsymbol{\theta}, \boldsymbol{\zeta}) &= \sum_{i=1}^{n}\sum_{t=1}^{T}\log f_{it}(y_{it}) + \sum_{i=1}^{n}\sum_{t=2}^{T}\sum_{s=1}^{t-1}\log \tilde{f}_{i,s,t|(s+1):(t-1)}\\&\quad\times (y_{is}, y_{it}|y_{i,s+1}, \ldots, y_{i,t-1}), \qquad (10)\end{aligned}$$

where $f_{it}(\cdot)$ and $F_{it}(\cdot)$ are specified by (1), $\tilde{f}_{i,s,t|(s+1):(t-1)}(\cdot|\cdot)$ is specified by (9), and $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ summarize the parameters in marginals and the mixed D-vine, respectively. Note that the model allows for unbalanced data provided that there are no intermittent missing values. The model can be estimated by two methods: joint maximum likelihood estimation (MLE) and inference function for margins (IFM) (see Joe (2005)). The joint MLE is a full likelihood approach and estimates all model parameters simultaneously. In a two-stage IFM, one estimates the marginal parameters ($\boldsymbol{\theta}$) from a separate univariate likelihood and then estimates the dependence parameters ($\boldsymbol{\zeta}$) from the multivariate likelihood with the marginal parameters given from the first stage. Compared with the joint MLE, the IFM is more computationally efficient by sacrificing the statistical efficiency. Therefore, the IFM is more practical for predictive applications where the statistical efficiency is of secondary concern. We examine both methods in Section A.4.

To implement the likelihood (10), one needs to evaluate the marginal densities and the bivariate building blocks (9) corresponding to each edge in the D-vine. We first calculate the marginal densities according to (1). Then we calculate (9) on a tree-by-tree basis from lower to higher orders. In the calculation of (9) for each tree, we use the copula of the current tree and the conditional cdf derived from the previous tree. An algorithm for evaluating the likelihood function for the mixed D-vine is provided in Appendix A.1.

In addition, we explore a sequential method that estimates and selects the bivariate copulas on a tree-by-tree basis. We start with the first tree, estimating the parameters and selecting the appropriate copulas from a given set of candidates. Fixing the parameters in the first tree, we then estimate the dependence parameters in the second tree for the candidate copulas and select the optimal. We continue estimating parameters and selecting copulas for the next tree of a higher order while holding the parameters fixed in all previous trees. If an independence copula is selected for a certain tree, we then truncate the vine, that is assume conditional independence in all higher order trees (see, e.g., Brechmann, Czado, and Aas (2012)). We use a heuristic procedure based on a commonly used model selection method Akaike information criterion (AIC) to select the copula. The sequential approach reduces the number of models to compare extensively and thus helps to fast select an appropriate model for applied studies. The benefit could be substantial in the case of big data or high dimensional dependence. In this application with a short panel of five-year observations, under the stationary assumption with nine candidate copulas, the sequential approach compares 9 × 4 different models in contrast to 9^4 models in an exhaustive search. The performance of the sequential method is investigated using simulation studies.

4. Application in Experience Rating

4.1. Model Fitting

Through repeated contracting, an insurer expects to gain private information regarding the risk level of its policyholders, and thus competitive advantages over its rivals. In particular, insurers hope to leverage the policyholders' past claim experience into the prediction of future claims. To this end, we explore the serial association of the claim cost over time. To motivate the specification of the mixed D-vine in Section 3.2, we report in Table 3 the partial rank correlations for the total claim cost as well as the claim cost for each peril. Specifically, the partial correlations are calculated recursively using relation:

$$\rho_{jk;l\cup V} = \frac{\rho_{jk;V} - \rho_{jl;V}\rho_{kl;V}}{\sqrt{(1-\rho^2_{jl;V})(1-\rho^2_{kl;V})}},$$

where j, k, l are distinct, V is a subset of $\{1, \dots, m\}\backslash\{j, k, l\}$, and $\rho_{jk;V}$ denotes the partial correlation between the jth and kth variables controlling for variables with indexes in V. The starting values in the recursive calculation are sample pairwise correlations.

In each correlation matrix, the upper triangle exhibits the Kendall's *tau* and the lower triangle the Spearman's *rho*. Using the upper triangle of the total claim cost as an example, the Kendall's *tau* between the claim cost in 2006 and 2007 is 0.284, between 2006 and 2008 conditioning on 2007 claims is 0.202, between 2006 and 2009 conditioning on 2007 and 2008 claims is 0.188, and so on. Two general patterns are noted from the table: First, the correlation decreases as one moves from the primary diagonal of the matrix toward its opposite corner in either upper or lower triangles, indicating that the conditioning set is more informative as two observations become further apart in time; second, the correlations along the same diagonal are of comparable size with some exceptions for the claims of other perils. These data characteristics support the D-vine specification with the stationarity assumption employed in this application.

We apply the proposed approach to the LGPIF claim data for the property coverage of building and contents. Separate models are fit for the total claim cost as well as the claim cost by peril. To summarize, the two-component mixture regression is used to accommodate the semicontinuous claim cost. The mass probability at zero is modeled using a logit regression and the amount of claims is modeled using a GB2 regression. Due to the limited number of predictors, we did not perform variable selection but instead included all available covariates in the two components. In the preliminary analysis, we explored the potential nonlinear effect of the coverage amount using scatterplot smoothing techniques (see, e.g., Ruppert, Wand, and Carroll (2003)) and we found the linear term of coverage in log scale quite satisfactory. The estimation results are summarized in Table 4. Parameters are estimated by the IFM and standard errors are calculated using the Godambe information matrix.

The results suggest that the entity type explains some heterogeneity in both claim frequency and severity. The effect of the alarm credit is a little counterintuitive which is to some extent explained by the estimation uncertainty. This counterintuitive effect could also imply some moral hazard issue. As anticipated, the odds of claim occurrence is higher for larger contract (due to higher exposure to risk), and so is the expected aggregate amount of claims. In the severity model, $\phi_1 > \phi_2$ in all the fitted GB2 distribution implying the positive skewness in the amount of claims. As indicated by the relation between ϕ_1, ϕ_2, and σ, their (theoretical) second moments do not even exist, which is consistent with the long tails in the distributions shown in Figure 1.

To demonstrate the goodness of fit of the GB2 distribution, we present in Figure 3 the *qq*-plots of the Cox-Snell residuals that is defined as $r_{it} = \Phi^{-1}(G_{it}(y_{it}))$. The match between the theoretical and empirical quantiles suggests the favorable fit of the GB2 distribution. The plots also indicate a slight lack of fit in the left tails except for the claims related to fire damages. However, for the purposes of ratemaking, we are more interested in the large claims that correspond to the right tails of the

Table 3. Serial partial correlation for the total claim cost and the claim cost by peril.

	Total						Water				
	2006	2007	2008	2009	2010		2006	2007	2008	2009	2010
2006	1	0.284	0.202	0.188	0.133	2006	1	0.243	0.188	0.157	0.086
2007	0.327	1	0.345	0.204	0.126	2007	0.261	1	0.313	0.194	0.164
2008	0.216	0.395	1	0.298	0.285	2008	0.195	0.339	1	0.284	0.211
2009	0.202	0.222	0.338	1	0.301	2009	0.163	0.205	0.303	1	0.245
2010	0.145	0.133	0.323	0.350	1	2010	0.087	0.172	0.226	0.264	1
	Fire						Other				
	2006	2007	2008	2009	2010		2006	2007	2008	2009	2010
2006	1	0.231	0.221	0.188	0.169	2006	1	0.206	0.227	0.172	0.152
2007	0.247	1	0.288	0.208	0.132	2007	0.216	1	0.186	0.152	0.049
2008	0.234	0.306	1	0.188	0.266	2008	0.236	0.197	1	0.192	0.160
2009	0.198	0.219	0.200	1	0.239	2009	0.177	0.158	0.201	1	0.200
2010	0.178	0.137	0.283	0.255	1	2010	0.160	0.051	0.170	0.222	1

Table 4. Estimates of the two-component mixture regression.

Total claim			Water			Fire			Other		
Logit	Est.	Std.	Logit	Est.	Std.	Logit	Est.	Std.	Logit	Est.	Std.
(Intercept)	2.776	0.158	(Intercept)	3.688	0.216	(Intercept)	3.985	0.247	(Intercept)	4.272	0.272
TypeCity	– 1.139	0.164	TypeCity	– 1.105	0.209	TypeCity	– 1.025	0.237	TypeCity	– 0.988	0.257
TypeCounty	– 1.813	0.203	TypeCounty	– 1.244	0.230	TypeCounty	– 1.890	0.253	TypeCounty	– 1.345	0.274
TypeSchool	– 0.162	0.160	TypeSchool	0.095	0.210	TypeSchool	– 0.137	0.237	TypeSchool	– 0.589	0.253
TypeTown	– 0.194	0.204	TypeTown	– 0.680	0.264	TypeTown	0.054	0.354	TypeTown	– 0.306	0.365
TypeVillage	– 0.887	0.156	TypeVillage	– 0.849	0.209	TypeVillage	– 1.054	0.237	TypeVillage	– 0.677	0.267
AC05	– 0.327	0.170	AC05	– 0.093	0.236	AC05	– 0.262	0.236	AC05	– 0.108	0.247
AC10	– 0.266	0.150	AC10	– 0.204	0.195	AC10	– 0.182	0.206	AC10	– 0.146	0.209
AC15	– 0.273	0.087	AC15	– 0.291	0.108	AC15	– 0.098	0.116	AC15	– 0.012	0.123
log(Coverage)	– 0.453	0.034	log(Coverage)	– 0.468	0.041	log(Coverage)	– 0.481	0.044	log(Coverage)	– 0.528	0.046
GB2	Est.	Std.	GB2	Est.	Std.	GB2	Est.	Std.	GB2		
(Intercept)	7.569	0.210	(Intercept)	7.234	0.257	(Intercept)	7.339	0.377	(Intercept)	7.306	0.424
TypeCity	– 0.483	0.184	TypeCity	– 0.263	0.237	TypeCity	– 0.858	0.286	TypeCity	– 0.538	0.337
TypeCounty	– 0.412	0.199	TypeCounty	– 0.607	0.257	TypeCounty	– 0.685	0.297	TypeCounty	– 0.724	0.358
TypeSchool	– 0.438	0.186	TypeSchool	– 0.613	0.248	TypeSchool	– 0.720	0.292	TypeSchool	0.015	0.335
TypeTown	– 0.034	0.239	TypeTown	– 0.107	0.299	TypeTown	– 0.342	0.398	TypeTown	– 0.241	0.454
TypeVillage	– 0.285	0.177	TypeVillage	– 0.246	0.235	TypeVillage	– 0.360	0.278	TypeVillage	– 0.224	0.338
AC05	0.121	0.184	AC05	0.478	0.294	AC05	0.077	0.264	AC05	0.218	0.318
AC10	– 0.239	0.158	AC10	– 0.144	0.210	AC10	0.467	0.223	AC10	– 0.555	0.267
AC15	– 0.127	0.091	AC15	– 0.009	0.123	AC15	0.078	0.125	AC15	– 0.138	0.162
log(Coverage)	0.546	0.035	log(Coverage)	0.377	0.049	log(Coverage)	0.433	0.050	log(Coverage)	0.428	0.058
σ	0.868	0.127	σ	0.593	0.116	σ	0.791	0.190	σ	1.100	0.259
κ_1	1.352	0.310	κ_1	0.941	0.268	κ_1	1.670	0.691	κ_1	2.023	0.801
κ_2	1.039	0.224	κ_2	0.569	0.145	κ_2	0.893	0.297	κ_2	1.304	0.484

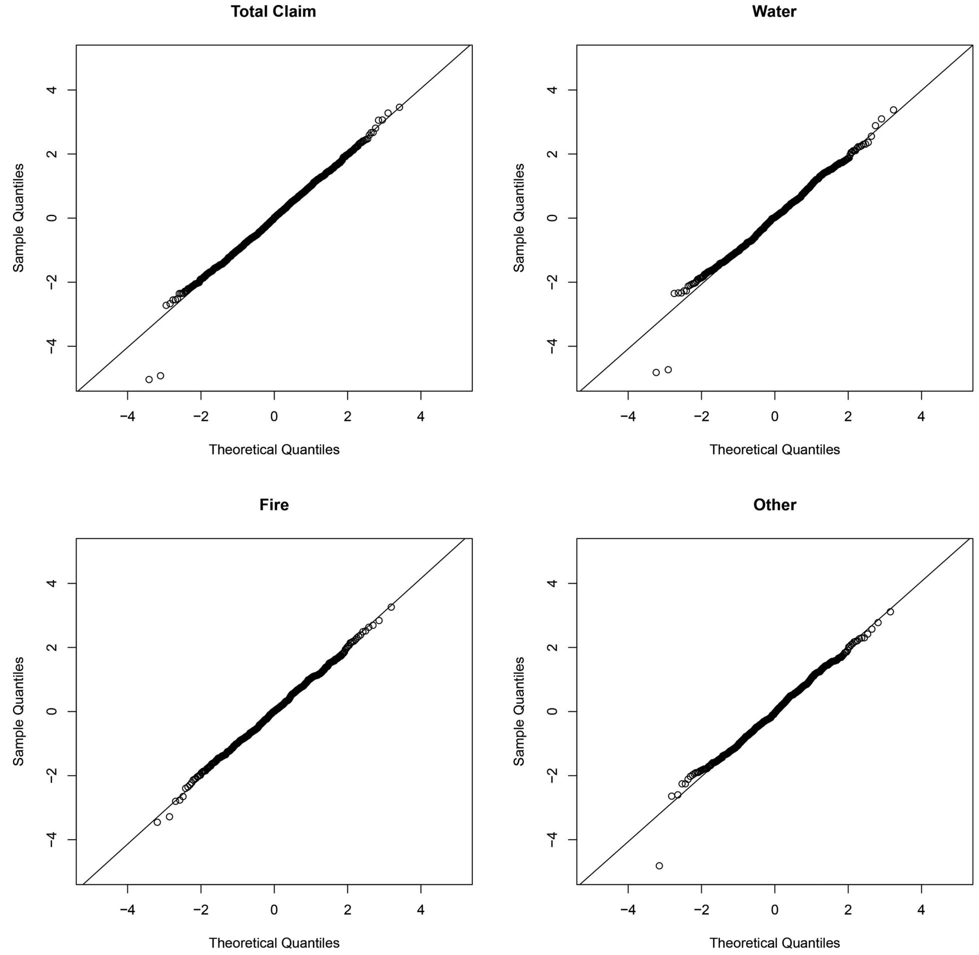

Figure 3. QQ plots of the GB2 distribution for total claims and claims by peril.

Table 5. Selected copulas for the mixed D-vine with estimated dependence.

Total Claim	Copula	Est.	Std.	Kendall's *tau*
$\mathcal{T}_1$	Rotated Joe	1.440	0.062	0.199
$\mathcal{T}_2$	Rotated Joe	1.382	0.063	0.178
$\mathcal{T}_3$	Rotated Joe	1.274	0.066	0.135
$\mathcal{T}_4$	Clayton	0.214	0.097	0.097
Water	Copula	Est.	Std.	Kendall's *tau*
$\mathcal{T}_1$	Rotated Joe	1.962	0.143	0.347
$\mathcal{T}_2$	Rotated Joe	1.685	0.126	0.276
$\mathcal{T}_3$	Rotated Joe	1.535	0.135	0.231
$\mathcal{T}_4$	Rotated Joe	1.302	0.166	0.146
Fire	Copula	Est.	Std.	Kendall's *tau*
$\mathcal{T}_1$	Frank	1.376	0.268	0.150
$\mathcal{T}_2$	Rotated Joe	1.668	0.132	0.271
$\mathcal{T}_3$	Frank	1.229	0.324	0.135
$\mathcal{T}_4$	Rotated Joe	1.500	0.194	0.219
Other	Copula	Est.	Std.	Kendall's *tau*
$\mathcal{T}_1$	Rotated Joe	1.622	0.156	0.258
$\mathcal{T}_2$	Rotated Joe	1.614	0.159	0.255
$\mathcal{T}_3$	Gaussian	0.098	0.054	0.062

distribution. The left tails represent small claims and are less of a concern in this application.

Pair copula constructions based on a mixed D-vine are used to accommodate the serial dependence among the longitudinal semicontinuous claim costs. We consider a candidate set of nine bivariate copulas as building blocks, including the Gaussian, Student's t, Gumbel, Clayton, Frank, Joe, survival Gumbel, survival Clayton, and survival Joe copulas. For the purpose of prediction, we further impose a stationarity assumption that all conditional pairs in a given tree share the same dependence. One should not view this assumption as a limitation of the proposed approach. In traditional longitudinal models, one would need a structured serial correlation such as autoregressive or exchangeable so as to borrow strength from past experience for future prediction. In the same spirit, the stationarity assumption is only required for prediction but not necessary for other types of statistical inference for the proposed longitudinal model.

Table 5 summarizes the selected bivariate copulas for the mixed D-vine, the estimated association parameters, and the corresponding Kendall's *tau*s for the total claim cost as well as the claim cost by peril. We followed the procedure in Section 3.3 to select the copula and to decide the optimal truncation. With five years of data, we have at most four trees in each of the mixed D-vines. For example, the mixed D-vine for the losses due to other perils is truncated at the third tree. The model is calibrated using the IFM. In general, the Kendall's *tau* decreases when moving from lower order trees to higher order trees. The decreasing pattern suggests that the conditioning set in higher order trees explains more of the association between the two nodes. This is consistent with the first principal of building vine trees that the (conditional) pairs with stronger association should receive higher priority. The reported Kendall's *tau* represents a partial relation in the same sense of the partial correlation. However, because of the discrete component in the marginal distributions, the inferred associations from the estimated copulas do not necessarily match the partial correlations from the data as reported in Table 3 (see Genest and Nešlehová 2007), although they present a similar decreasing pattern.

Table 6 compares the goodness-of-fit statistics of the selected D-vine with alternative modeling strategies. The first alternative that we compare with is the Gaussian copula model as in Shi, Feng, and Boucher (2016). Specifically, we report the five-dimensional Gaussian copula with AR(1) temporal dependence and the zero-inflated GB2 marginals. Further, we examine two special cases of the D-vine, a fully truncated model and a fully simplified model (Brechmann, Czado, and Aas (2012)). The former uses independence copula for all (conditional) pairs, and the latter uses Gaussian copula for all (conditional) pairs. It is not surprising that the mixed D-vine is superior to the truncated D-vine, confirming the significant temporal association in the zero-inflated longitudinal data. When compared with the simplified D-vine, the favorable fit of the mixed D-vine (smaller AIC and BIC statistics) emphasizes the value added by the flexible dependence structure (such as asymmetric and nonlinear association) embraced by the pair copula constructions. Such flexibility plays a crucial role in the dependence modeling for nonnormal outcomes such as heavy-tailed and discrete data. The Gaussian copula model is an analogy of a D-vine truncated at the first tree, which explains its undesirable fit when compared with the simplified D-vine.

4.2. Prediction

Experience rating is to incorporate policyholders' past claim experience into the future premiums. The mixed D-vine provides a natural structure to derive the predictive distribution, not just a point prediction, of the future claim cost. We stress that this is another advantage of using pair copula constructions for predictive applications. With elliptical copulas, it is not straightforward to derive the predictive distribution when there are discrete components in the marginals. For policyholder i, denoting $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$, the conditional distribution of $Y_{i,T+1}$ given $\boldsymbol{Y}_i$ is shown as

$$f_{Y_{i,T+1}|\boldsymbol{Y}_i}(y) = f_{i,T+1}(y) \prod_{t=2}^{T} \tilde{f}_{i,t,T+1|(t+1):T}(y_{it}, y|y_{i,t+1}, \ldots, y_{i,T}).$$

Here, $f_{i,T+1}(\cdot)$ and $\tilde{f}_{i,t,T+1|(t+1):T}(\cdot|\cdot)$ are defined by (1) and (9), respectively. The derivation of the predictive distribution relies on the conditional independence assumption between Y_{i1} and $Y_{i,T+1}$ given $Y_{i2}, \ldots, Y_{i,T}$. This is sensible given the

Table 6. Goodness-of-fit statistics of the mixed D-vine and alternative models.

	Total		Water		Fire		Other	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Gaussian copula	39,662	39,819	21,127	21,284	18,558	18,715	16,686	16,842
Truncated D-vine	39,708	39,859	21,191	21,341	18,584	18,735	16,701	16,851
Simplied D-vine	39,624	39,801	21,098	21,275	18,523	18,700	16,667	16,844
Mixed D-vine	39,561	39,737	21,036	21,213	18,496	18,673	16,658	16,834

pattern of the dependence in the mixed D-vine reported in Table 5. Detailed derivation of the predictive density is found in Appendix A.3. Insurers set pure premium as expected cost of the contract, thus the experience adjusted pure premium is $\mathrm{E}(Y_{i,T+1}|\boldsymbol{Y}_i = \boldsymbol{y}_i)$. The predictive mean can be estimated using the Monte Carlo simulation or the numerical integration.

The predictive performance is investigated using the hold-out sample of year 2011. It is well known that the usual loss functions are ill-suited for capturing the differences between the predicted values and the corresponding outcomes in the hold-out sample, due to the high proportion of zeros and the skewness and heavy tails in the distribution of the positive losses. Therefore, we turn to alternative statistical measures—the ordered Lorenz curve and the associated Gini index—that have been developed in the recent literature (see Frees, Meyers, and Cummings (2012) and Frees, Meyers, and Cummings (2014)). The essential idea of the ordered Lorenz curve is to measure the discrepancy between the premium and loss distributions. Let $B(\boldsymbol{x})$ be the base premium and $P(\boldsymbol{x})$ be the competing premium, both depending on a set of exogenous variables $\boldsymbol{x}$. The ordered premium and loss distributions are defined based on the relativity $R(\boldsymbol{x}) = P(\boldsymbol{x})/B(\boldsymbol{x})$ as

$$\hat{H}_P(s) = \frac{\sum_{i=1}^{n} B(\boldsymbol{x}_i) I(R(\boldsymbol{x}_i) \le s)}{\sum_{i=1}^{n} B(\boldsymbol{x}_i)} \quad \text{and}$$

$$\hat{L}_P(s) = \frac{\sum_{i=1}^{n} y_i I(R(\boldsymbol{x}_i) \le s)}{\sum_{i=1}^{n} y_i}.$$

The ordered Lorenz curve is the plot of $(\hat{H}_P(s), \hat{L}_P(s))$. The 45-degree line, known as the line of equality, indicates the percentage of losses equals the percentage of premiums. The associated Gini index is defined as twice the area between the ordered Lorenz curve and the line of equality, and it may range over $(-1, 1)$. A curve below the line of equality suggests that the insurer could look to the competing premium to identify more profitable contracts.

We make two sets of validations. The first is to compare the proposed experience rated premium with some alternative bases. Table 7 reports the Gini indices associated with the ordered Lorenz curves under three scenarios. The upper panel uses a constant premium base, the middle panel uses the contract premium in year 2011 as the base, and the lower panel uses the nonexperience adjusted premium base. Within each panel, we consider the prediction for the total claim cost as well as the claim cost by peril. Two methods are used to derive the prediction for the total claim cost. One directly predicts from the model for the total claim cost, the other predicts the claim cost for each peril and then aggregates them. The constant premium base means that the insurer does not differentiate good risks and bad risks, and charges all policyholders the average cost. Hence, it is not surprising to observe the large and significant Gini indices for all predictions in the upper panel. With both informative predictors and claim experience, insurers will achieve better risk segmentation. In practice, insurers use a finer-grained rating algorithm to classify and price the risk. Fortunately, the LGPIF data contain the actual contract premium for building and contents coverage as well as the premium for each peril. When compared with the contract premiums, the Gini indices become smaller as shown in the middle panel. However, the statistical significance suggests that the insurer can still identify profitable business when looking to the proposed experience adjusted rates. The lower panel demonstrates the importance of experience rating. The nonexperience adjusted premium is calculated based on the independence assumption among the repeated observations over time. Indeed, the results are in line with our expectations. The significant positive Gini indices implies that the claim experience provides the insurer opportunities to cream skim (or cherry-pick the low-risk policyholders).

Table 7. Gini indices (percentage) with the constant, contract, and independence premium bases.

	Total claim		Claim by peril		
Base	Total	By peril	Water	Fire	Other
Constant	76.16 (4.65)	76.57 (4.72)	69.63 (6.88)	76.58 (5.31)	78.84 (6.93)
Contract	31.34 (8.74)	33.69 (10.04)	29.29 (8.99)	16.89 (19.02)	38.92 (9.30)
Independence	36.93 (11.2)	34.19 (12.14)	32.14 (12.63)	26.52 (11.06)	29.80 (11.78)

Figure 4 displays the ordered Lorenz curves corresponding to Gini indices of the total claim cost prediction in Table 7. The left panel shows the case of the direct prediction and the right panel shows the case of the prediction by peril. For instance, the areas between the curves and the 45-degree line in the left panel are 0.38, 0.15, and 0.18 when using the constant premium, contract premium, and independence premium as bases, respectively.

The second set of validation is to compare the proposed rating algorithm with some off-the-shelf strategies for experience rating. The standard approach to incorporating past claims into future prediction is to use a random effect framework. We examine both the linear and generalized linear models. The former leads to the classic Bülhmann credibility premium. In the latter, we consider the industry benchmark—the Tweedie's compound Poisson model. We evaluate the performance of alternative approaches based on the prediction for the total claim cost for building and contents coverage. The total claim cost is either predicted directly or indirectly by aggregating the claims of different perils. Table 8 briefly summarizes the alternative predicting methods.

For model comparison, we calculate the Gini index matrix as in Table 9. The matrix summarizes the pair-wise Gini indices of the predictions from all candidate models when each of them is successively used as the base premium and the remaining as competing premiums. For example, the first row corresponds to the Gini indices using the experience-adjusted premium from the model LM.t as the base. The submatrices in the upper left and lower right corners are of our primary interest. One strategy for model selection is to use the proposed premium from the mixed D-vine model to challenge alternative premiums from

Table 8. Description of alternative experience rating models.

Model	Description
LM.t	Linear mixed model directly predicts the total claim cost
GLM.t	Tweedie mixed effects model directly predicts the total claim cost
COPULA.t	Mixed D-vine approach directly predicts the total claim cost
LM.p	Linear mixed model predicts the claim cost by peril
GLM.p	Tweedie mixed effects model predicts the claim cost by peril
COPULA.p	Mixed D-vine approach predicts the claim cost by peril

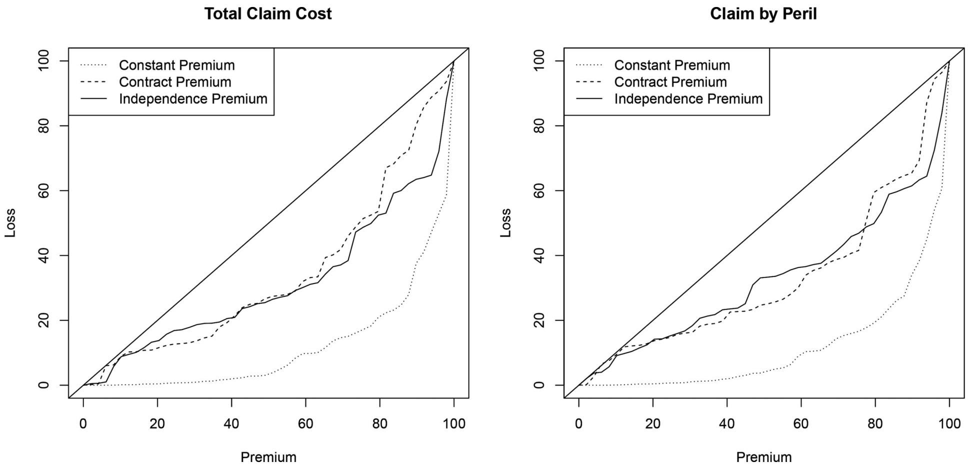

Figure 4. Ordered Lorenz curves using constant, contract, and independence premium bases. Left panel corresponds to the prediction of total claim cost and the right panel corresponds to the prediction of claim by peril.

Table 9. Gini index matrix for six alternative predictions.

	LM.t	GLM.t	COPULA.t	LM.p	GLM.p	COPULA.p
LM.t	—	7.50	32.41	1.03	6.80	32.09
		(9.61)	(11.84)	(11.44)	(9.61)	(11.77)
GLM.t	29.22	—	50.52	33.34	−39.34	51.47
	(8.92)		(8.45)	(10.85)	(9.48)	(8.78)
COPULA.t	−4.45	14.89	—	−6.20	15.87	18.22
	(11.68)	(9.05)		(13.76)	(8.98)	(11.61)
LM.p	27.33	23.77	30.53	—	23.65	29.90
	(10.68)	(9.18)	(12.86)		(9.10)	(12.48)
GLM.p	35.50	47.02	55.32	40.19	—	56.34
	(9.30)	(9.24)	(8.44)	(10.69)		(8.66)
COPULA.p	−3.99	14.72	−8.05	−5.00	15.48	—
	(12.25)	(8.95)	(12.20)	(13.34)	(8.64)	

Table 10. Rank correlations among claims of different perils.

	Water	Fire	Other
Water	1	0.233	0.228
Fire	0.250	1	0.193
Other	0.244	0.205	1

the off-the-shelf rating algorithms. The upper left matrix compares models that directly predict the total claim cost. The Gini indices are 32.41 and 50.52 for the Bülhmann premium base and the Tweedie premium base, respectively. The lower right matrix compares models that predict the claim cost by peril and then aggregate them. The Gini indices are 29.90 and 56.34 for the Bülhmann premium base and the Tweedie premium base, respectively. The statistical significance confirms that the proposed mixed D-vine model leads to a greater separation among the observations.

Alternatively, to pick the "best" model, one could use a "minimax" strategy to select the base premium model that is the least vulnerable to the competing premium models. That is, we select the model that provides the smallest of the maximal Gini indices among the challenging premiums. For the direct prediction, the maximal Gini indices are 32.41, 50.52, and 14.89 when the base premium corresponds to the linear model, Tweedie model, and copula model, respectively. The copula approach has the smallest maximal Gini index, and hence is the least vulnerable to the alternative predictions. In a similar manner, when predicting by peril, the copula approach also has the smallest maximal Gini index of 15.48. When selecting from all six models, the "minmax" approach picks the predictions by peril from the mixed D-vine model. As a summary, we attribute the superior performance of the proposed method in experience rating to: (1) The two-component mixture model provides substantial flexibility in capturing the unique features in the insurance claim data, such as zero inflation, skewness, and thick tails; (2) The pair copula constructions based on the mixed D-vine allow us to accommodate a wide range of dependence including nonlinear and asymmetric relationship, while the traditional random effects framework limits the way past claims are incorporated into the future prediction.

5. Discussion

Motivated by the experience rating practice in nonlife insurance, this article introduced pair copula constructions based on the mixed D-vine for modeling the zero-inflated longitudinal insurance claim cost. The proposed approach is shown to achieve better risk segmentation and thus improve market efficiency. The data analysis emphasized the benefits of the mixed vine approach in both fitting the observations in the training sample and predicting the observations in the hold-out sample. The size of the insurance market itself justifies the contribution of the new method. Furthermore, pair copula constructions for mixed outcomes are expected to find applications in many other disciplines. To name a few, in marketing research, retailers are interested in consumers' purchasing behavior; in healthcare, care providers are interested in the patients' consumption of medical services; in climate studies, scientists are interested in the amount of precipitation. The variables of interest from all these examples are mixed measurements.

A natural extension is to generalize the current mixed D-vine framework for the semicontinuous longitudinal data to the multivariate context. In this application, the total claim cost for building and contents coverage is decomposed into claims by water, fire, and other perils. The losses caused by these perils were treated as three independent longitudinal outcomes separately. However, if the losses from different perils are correlated, it is arguable that one can further improve prediction and experience rating scheme by borrowing strength among perils. To provide intuition, we display in Table 10 the contemporaneous cross-sectional correlations among the claim costs from various perils. The upper and lower triangles report the Kendall's *tau* and Spearman's *rho*, respectively.

The strong relation suggests some joint modeling strategy. For the prediction purposes, the association that matters most is the lead-lag correlation across perils rather than contemporaneous correlation between perils. In experience rating, one hopes that the past claims in other perils could provide prediction lift for related perils. Although not reported, the strong correlations in Table 10 are indeed associated with significant lead-lag correlations across perils. Recently, Brechmann and Czado (2015) and Smith (2015) discussed possible strategies of constructing vine trees for multivariate time series, which shed some lights on the modeling of multivariate longitudinal data. This topic is being investigated in a separate work.

Supplementary Materials

The Appendix of the article can be found in the online supplementary materials.

Acknowledgment

The authors thank the editor, the associated editor, and two anonymous reviewers for helpful comments . Peng Shi gratefully acknowledges the financial support for this study by the Centers of Actuarial Excellence (CAE) Research Grant from the Society of Actuaries and the Wisconsin Naming Partners Fellowship from the University of Wisconsin-Madison.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009), "Pair-Copula Constructions of Multiple Dependence," *Insurance: Mathematics and Economics*, 44, 182–198. [123,127]

Bedford, T., and Cooke, R. M. (2001), "Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines," *Annals of Mathematics and Artificial Intelligence*, 32, 245–268. [123]

Bedford, T., and Cooke, R. M. (2002), "Vines-A New Graphical Model for Dependent Random Variables," *Annals of Statistics*, 30, 1031–1068. [123]

Brechmann, E. C., and Czado, C. (2015), "COPAR–Multivariate Time Series Modeling Using the Copula Autoregressive Model," *Applied Stochastic Models in Business and Industry*, 31, 495–514. [133]

Brechmann, E. C., Czado, C., and Aas, K. (2012), "Truncated Regular Vines in High Dimensions with Application to Financial Data," *Canadian Journal of Statistics*, 40, 68–85. [127,130]

Bühlmann, H. (1967), "Experience Rating and Credibility," *ASTIN Bulletin: The Journal of the International Actuarial Association*, 4, 199–207. [122]

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013), "Selecting and Estimating Regular Vine Copulae and Application to Financial Returns," *Computational Statistics & Data Analysis*, 59, 52–69. [127]

Dunn, P. K., and Smyth, G. K. (2005), "Series Evaluation of Tweedie Exponential Dispersion Model Densities," *Statistics and Computing*, 15, 267–280. [123]

Dunn, P. K., and Smyth, G. K. (2008), "Series Evaluation of Tweedie Exponential Dispersion Model Densities by Fourier Inversion," *Statistics and Computing*, 18, 73–86. [123]

Frees, E., and Valdez, E. (2008), "Hierarchical Insurance Claims Modeling," *Journal of the American Statistical Association*, 103, 1457–1469. [125]

Frees, E. W., Meyers, G., and Cummings, A. D. (2012), "Summarizing Insurance Scores Using a Gini Index," *Journal of the American Statistical Association*, 106, 1085–1098. [131]

Frees, E. W., and Wang, P. (2005), "Credibility Using Copulas," *North American Actuarial Journal*, 9, 31–48. [123]

Frees, E. W., Young, V. R., and Luo, Y. (1999), "A Longitudinal Data Analysis Interpretation of Credibility Models," *Insurance: Mathematics and Economics*, 24, 229–247. [122]

Frees, E. W. J., Meyers, G., and Cummings, A. D. (2014), "Insurance Ratemaking and a Gini Index," *Journal of Risk and Insurance*, 81, 335–366. [131]

Genest, C., and Nešlehová, J. (2007), "A Primer on Copulas for Count Data," *ASTIN Bulletin: The Journal of the International Actuarial Association*, 37, 475–515. [130]

Gruber, L., and Czado, C. (2015), "Sequential Bayesian Model Selection of Regular Vine Copulas," *Bayesian Analysis*, 10, 937–963. [127]

Haff, I. H., Aas, K., and Frigessi, A. (2010), "On the Simplified Pair-Copula Construction Simply useful or too Simplistic?" *Journal of Multivariate Analysis*, 101, 1296–1310. [126]

Hintze, J. L., and Nelson, R. D. (1998), "Violin Plots: A Box Plot-Density Trace Synergism," *The American Statistician*, 52, 181–184. [124]

Joe, H. (2005), "Asymptotic Efficiency of the Two-Stage Estimation method for Copula-Based Models," *Journal of Multivariate Analysis*, 94, 401–419. [127]

——— (2014), *Dependence Modeling with Copulas*, New York: Chapman & Hall. [123]

Kurowicka, D., and Cooke, R. M. (2006), *Uncertainty Analysis with High Dimensional Dependence Modelling*, New York: Wiley. [123]

McDonald, J. (1984), "Some Generalized Functions for the Size Distribution of Income," *Econometrica*, 52, 647–63. [125]

McDonald, J., and Xu, Y. (1995), "A Generalization of the Beta Distribution with Applications," *Journal of Econometrics*, 66, 133–152. [125]

Olsen, M. K., and Schafer, J. L. (2001), "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association*, 96, 730–745. [123]

Panagiotelis, A., Czado, C., and Joe, H. (2012), "Pair Copula Constructions for Multivariate Discrete Data," *Journal of the American Statistical Association*, 107, 1063–1072. [123,125,127]

Panagiotelis, A., Czado, C., Joe, H., and Stöber, J. (2017), "Model Selection for Discrete Regular Vine Copulas," *Computational Statistics and Data Analysis*, 106, 138–152. [127]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge UK: Cambridge University Press. [128]

Shi, P. (2012), "Multivariate Longitudinal Modeling of Insurance Company Expenses," *Insurance: Mathematics and Economics*, 51, 204–215. [123]

——— (2014), "Fat-Tailed Regression Models," in *Predictive Modeling Applications in Actuarial Sciences: Volume I, Predictive Modeling Techniques*, eds. E., Frees, G. Meyers, and R. Derrig, Cambridge UK: Cambridge University Press, pp. 138–166. [125]

Shi, P., Feng, X., and Boucher, J.-P. (2016), "Multilevel Modeling of Insurance Claims Using Copulas," *Annals of Applied Statistics*, 10, 834–863. [123,130]

Shi, P., and Zhang, W. (2015), "Private Information in Healthcare Utilization: Specification of a Copula-Based Hurdle Model," *Journal of the Royal Statistical Society*, Series A, 178, 337–361. [125]

Smith, M., Min, A., Almeida, C., and Czado, C. (2010), "Modeling Longitudinal Data Using a Pair-Copula Decomposition of Serial Dependence," *Journal of the American Statistical Association*, 105, 1467–1479. [123]

Smith, M. S. (2015), "Copula Modelling of Dependence in Multivariate Time Series," *International Journal of Forecasting*, 31, 815–833. [133]

Stöber, J. (2013), "Regular Vine Copulas with the Simplifying Assumption, Time-Variation, and Mixed Discrete and Continuous Margins," Ph.D. dissertation, Technische Universität München. [123]

Stöber, J., Hong, H. G., Czado, C., and Ghosh, P. (2015), "Comorbidity of Chronic Diseases in the Elderly: Patterns Identified by a Copula Design for Mixed Responses," *Computational Statistics & Data Analysis*, 88, 28–39. [123]

Stoeber, J., Joe, H., and Czado, C. (2013), "Simplified Pair Copula Constructions Limitations and Extensions," *Journal of Multivariate Analysis*, 119, 101–118. [126]

Sun, J., Frees, E. W., and Rosenberg, M. A. (2008), "Heavy-Tailed Longitudinal Data Modeling Using Copulas," *Insurance: Mathematics and Economics*, 42, 817–830. [123]

Zhang, Y. (2013), "Likelihood-Based and Bayesian Methods for Tweedie Compound Poisson Linear Mixed Models," *Statistics and Computing*, 23, 743–757. [123]