The Annals of Applied Statistics
2020, Vol. 14, No. 1, 357–380
https://doi.org/10.1214/19-AOAS1299

REGRESSION FOR COPULA-LINKED COMPOUND DISTRIBUTIONS WITH APPLICATIONS IN MODELING AGGREGATE INSURANCE CLAIMS

BY PENG SHI[1] AND ZIFENG ZHAO[2]

[1] *Department of Risk and Insurance, Wisconsin School of Business, University of Wisconsin—Madison, pshi@bus.wisc.edu*

[2] *Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, zzhao2@nd.edu*

In actuarial research a task of particular interest and importance is to predict the loss cost for individual risks so that informative decisions are made in various insurance operations such as underwriting, ratemaking and capital management. The loss cost is typically viewed to follow a compound distribution where the summation of the severity variables is stopped by the frequency variable. A challenging issue in modeling such outcomes is to accommodate the potential dependence between the number of claims and the size of each individual claim. In this article we introduce a novel regression framework for compound distributions that uses a copula to accommodate the association between the frequency and the severity variables and, thus, allows for arbitrary dependence between the two components. We further show that the new model is very flexible and is easily modified to account for incomplete data due to censoring or truncation. The flexibility of the proposed model is illustrated using both simulated and real data sets. In the analysis of granular claims data from property insurance, we find substantive negative relationship between the number and the size of insurance claims. In addition, we demonstrate that ignoring the frequency-severity association could lead to biased decision-making in insurance operations.

1. Introduction.

In actuarial research on nonlife insurance, a task of particular interest and importance is to predict the loss cost for individual risks in an insurer's book of business. Interpretation and prediction of loss cost of individual policyholders deepens the insurer's understanding of the risk profile of the entire portfolio which further leads to better-informed decisions in various insurance operations, such as underwriting, ratemaking and capital management.

The loss cost of a policyholder is jointly determined by the number of claims and the amount of each claim during the contract period. As a result, researchers and practitioners typically view the loss cost outcome to follow a compound or generalized distribution (see Karlis and Xekalaki (2005) and Johnson, Kemp and Kotz (2005)). Specifically, the loss cost per policy year, denoted by S, can be represented as

$$S = Y_1 + \cdots + Y_N, \tag{1.1}$$

where N is a counting random variable and represents the number of claims, and Y_j ($j = 1, \ldots, N$) is a nonnegative continuous random variable and represents the size of the jth claim. The sequence of $Y_1, Y_2, \ldots$ is further assumed to be independently and identically distributed. Compound distributions have been extensively used in the actuarial science literature for modeling aggregate losses in an insurance system (see, e.g., Klugman, Panjer and Willmot (2012), Lin (2014), and Albrecher, Beirlant and Teugels (2017)). In insurance applications, N and $\{Y_j\}$ are referred to as the "frequency" and "severity" components, respectively.

Received August 2018; revised August 2019.

Key words and phrases. Insurance claim data, compound distribution, copula regression, incomplete data, two-part model.

In this article we focus on the regression method for compound distributions when both N and $(Y_1, \dots, Y_N)$ are observed. A challenging issue in modeling such outcomes in the regression setting is to accommodate the potential dependence between the number of claims and the size of each individual claim. The goal of this work is to introduce a simple yet flexible regression framework to allow for arbitrary dependence between the frequency and severity distributions.

The current regression approach to studying the aggregate loss S relies on the independence assumption between N and each Y_j. Under such independence assumption one develops regression models for the number and size of claims separately, which is known as the frequency-severity or two-part model. See Frees (2014) for discussions on various types of two-part models. As a special case, when the frequency is a Poisson variable and the severity is a gamma variable, the loss cost is known to follow a Tweedie distribution (Tweedie (1984)). Jørgensen and Paes de Souza (1994) and Smyth and Jørgensen (2002) have explored fitting the Tweedie's compound Poisson model to the loss cost data in property insurance.

In addition to actuarial science and insurance, regression models based on compound distributions have been used in many other disciplines as well. In health economics the two-part model was used to study an individual's total number of doctor visits resulting from multiple spells of illness in a given period (see, for instance, Silva and Windmeijer (2001)). In marketing, Tellis (1988) employed a special case of the frequency-severity model to study the effect of repetitive advertising on consumer purchasing choices; Aribarg, Pieters and Wedel (2010) studied consumer advertisement recognition where an individual's attention is formulated as a compound model determined by eye fixation frequency and gaze duration. In operational risk the compound distribution for aggregate losses is the foundation for the determination of the operational risk capital required by the Basel capital framework for banks (Panjer (2006) and Shevchenko (2010)). In psychology, Smithson and Shou (2014) pointed out the applications of this type of model in different areas of psychology, such as perception and decision making, where a psychological process is thought to be serially summed from observable component process outputs.

The two-part models in different scientific fields described above employ some common key assumptions, including:

(1) The distribution of N does not depend on the values of Y_j for $j = 1, \dots, N$;
(2) Conditional on $N = n > 0$, $Y_1, \dots, Y_n$ are independently distributed random variables;
(3) Conditional on $N = n > 0$, the common distribution of $Y_1, \dots, Y_n$ does not depend on n.

The (conditional) independence assumption between N and Y_j certainly leads to tractable statistical inference because it allows one to build regression models separately for the frequency and severity components. However, if N and Y_j are correlated, ignoring the association between them will lead to serious biases in the inference. First, the regression coefficients in the severity regression model will be inconsistent estimates of the marginal effect of explanatory variables. Second, there is a persistent error in the prediction for the severity given the frequency. Third, the misspecification will introduce bias in the inference for the compound distribution.

Motivated by the above observations, we introduce a novel copula-linked compound distribution and the associated two-part regression framework that allow for arbitrary dependence between the frequency and severity components. Specifically, we employ a parametric copula to construct the joint distribution of frequency and severity variables, thus relax the independence assumption in standard methods. We show that the resulting copula regression framework is able to nest several commonly used approaches as special cases, including the hurdle model, the selection model and the frequency-severity model, among others. Furthermore, we extend the basic model to accommodate the case of incomplete data due to censoring or truncation. Because of the parametric nature, likelihood-based approaches are proposed for estimation, inference and diagnostics.

The flexibility of the proposed model is illustrated using both simulated and real data sets. In the numerical experiments we showcase the impact of ignoring the frequency and severity dependence on the resulting compound distribution. In the empirical study we apply the proposed method to granular claims data in property insurance. Our analysis detects substantive negative dependency between the number and the size of insurance claims. In addition, we demonstrate the importance of such dependency in some key insurance functions, including underwriting and ratemaking, loss reserving and capital management. The results suggest that ignorance of frequency-severity dependence could lead to biased decision making in insurance operations.

To the best of our knowledge, this work is among the first efforts to explicitly incorporate the dependence between the frequency and severity variables of a compound distribution in a regression setting. Recent literature has made some development in this direction; for example, see Czado et al. (2012), Krämer et al. (2013) and Garrido, Genest and Schulz (2016) among others. The fundamental difference between our work and existing studies is that the aforementioned studies examined the relation between the frequency N and the average severity $\overline{Y} = \sum_{j=1}^{N} Y_j/N$, while the proposed method directly looks into the relation between the frequency N and the individual severity Y_j. Alternative mechanisms for introducing dependence between the frequency and individual severity variables include the correlated random-effect framework, as in Olsen and Schafer (2001), and the conditional approach, as in Frees, Gao and Rosenberg (2011). The difficulty with both methods as compared to the proposed copula approach is that it is not straightforward to handle incomplete data which is not unusual in insurance applications because of various coverage modifications.

Given that our work fits in the broader literature on multivariate modeling in insurance, it is worth discussing their differences and connections. The current literature on dependence modeling of insurance claims focuses on the joint modeling of multiple outcomes of loss cost that could arise from multiple lines of business (see Frees, Lee and Yang (2016)), multiple coverage in a single business line (see Shi, Feng and Boucher (2016)) or multiple peril types covered by a policy (see Shi and Yang (2018)). In this line of studies, each loss cost outcome is formulated using either a Tweedie model or a two-part model. Both can be viewed in the framework of the compound distribution (1.1) where the N and each Y_j are assumed to be independent with each other. Apart from the association among multiple loss cost outcomes, this work examines a single loss cost outcome, and the focus is on the dependence between the frequency and severity components in the compound model.

The rest of the paper is structured as follows. Section 2 introduces the dependent frequency-severity regression model for the compound distribution and discusses its extension for incomplete data due to censoring and truncation. The likelihood-based methods for estimation, inference and diagnostics are further discussed. Section 3 provides numerical experiments to show the impact of ignoring the frequency-severity dependence under various settings. Section 4 applies the proposed approach to the loss cost data in property insurance and shows the importance of the frequency-severity dependence in insurance operations. Section 5 concludes the article. The Supplementary Material (Shi and Zhao (2020)) contains additional technical examples, numerical studies and detailed data analysis.

2. Copula-linked compound regression.

2.1. *Basic model.* In the basic setup, we assume that complete information on $(N, Y_1, \dots, Y_N)$ is observed for each subject, where N is a count variable and $\{Y_j\}$ are continuous variables. For simplicity, we suppress the subject index in the following presentation. The joint distribution of $(N, Y_1, \dots, Y_N)$ is built upon the assumption that $(Y_1, \dots, Y_n)$ are conditionally i.i.d. given $N = n$, as opposed to the unconditional i.i.d. assumption in the standard compound distribution. There are several implications of this assumption. First, conditional independence of $(Y_1, \dots, Y_n)$ given $N = n$ introduces correlation among Y_j, which departs from the i.i.d. assumption in the standard model. Second, identical distribution of $(Y_1, \dots, Y_n)$ given $N = n$ implies identical marginal distribution of Y_j which is consistent with the i.i.d. assumption in the standard model. Third, the bivariate distribution of (N, Y_j) are identical given $N = n$ which nests the independent case in the standard model.

To facilitate presentation, we denote Y as the variable associated with the common distribution of the sequence $\{Y_j\}$. Note that Y is only defined in the sense of a distribution, not in the sense of a random variable. Under the conditional independence assumption, the associated pmf/pdf function is

$$\begin{aligned} f_{N,\boldsymbol{Y}}(n, y_1, \dots, y_n) &= \begin{cases} \Pr(N = 0) & n = 0, \\ \dfrac{\partial}{\partial y_1 \cdots \partial y_n} \Pr(N = n, Y_1 \le y_1, \dots, Y_n \le y_n) & n > 0 \end{cases} \\ &= \big[f_N(0)\big]^{I(n=0)} \big[f_N(n) \times f_{\boldsymbol{Y}|N}(y_1, \dots, y_n|n)\big]^{I(n>0)} \\ &= f_N(n) \times \left[\prod_{j=1}^{n} f_{Y|N}(y_j|n)\right]^{I(n>0)}, \end{aligned} \tag{2.1}$$

where $I(\cdot)$ is an indictor function.

The central component to define (2.1) is the joint distribution of N and Y. To allow for flexible dependence between N and Y, we take a parametric approach and employ a bivariate parametric copula to construct their joint distribution. Refer to Nelsen (2006) and Joe (2015) for an introduction to dependence modeling with copulas. According to Sklar's theorem, the joint distribution of N and Y can be expressed in terms of a bivariate copula C:

$$F_{N,Y}(n, y) = \Pr(N \le n, Y \le y) = C\big(F_N(n), F_Y(y)\big). \tag{2.2}$$

Denote $h(u, v) = \frac{\partial}{\partial v} C(u, v)$, it follows that

$$\begin{aligned} f_{N,Y}(n, y) &= \frac{\partial}{\partial y} \Pr(N = n, Y \le y) \\ &= f_Y(y)\big[h\big(F_N(n), F_Y(y)\big) - h\big(F_N(n-1), F_Y(y)\big)\big]. \end{aligned} \tag{2.3}$$

From above, one finds the conditional distribution of Y given N as

$$\begin{aligned} F_{Y|N}(y|n) &= \Pr(Y \le y|N = n) \\ &= \frac{1}{f_N(n)}\big[C\big(F_N(n), F_Y(y)\big) - C\big(F_N(n-1), F_Y(y)\big)\big], \end{aligned} \tag{2.4}$$

$$\begin{aligned} f_{Y|N}(y|n) &= \frac{\partial}{\partial y} \Pr(Y \le y|N = n) \\ &= \frac{f_Y(y)}{f_N(n)}\big[h\big(F_N(n), F_Y(y)\big) - h\big(F_N(n-1), F_Y(y)\big)\big]. \end{aligned} \tag{2.5}$$

In a regression context one wants to incorporate exogenous explanatory variables to account for observed heterogeneity in both N and Y_j. Thus, the marginal models for both N and Y_j are defined conditional on covariates. For example, in generalized linear models one could specify $g^f(\mathrm{E}(N_i|\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}^f$ and $g^s(\mathrm{E}(Y_{ij}|\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}^s$, where i is the subject index,

$\boldsymbol{x}_i$ is the vector of covariates, $\boldsymbol{\beta}$ is the regression coefficients and g denotes the link function. Superscripts f and s indicate the frequency and severity components, respectively.

As a special case, when the copula in (2.2) is an independence copula, that is, N and each Y_j are independent, model (2.1) reduces to

$$f_{N,\boldsymbol{Y}}(n, y_1, \dots, y_n) = f_N(n) \times \left[\prod_{j=1}^{n} f_Y(y_j)\right]^{I(n>0)}, \tag{2.6}$$

where the marginal models of N and Y are totally separable. Since (2.1) includes (2.6) as a special case, the usual goodness-of-fit statistics, such as the likelihood ratio test, could be used to test whether the independence assumption between N and Y_j is supported by the data.

It is worth stressing several observations in model (2.1). First, the independence assumption of Y_j given N implies a specific dependence among the sequence $\{Y_j\}$. As pointed out by Liu and Wang (2017), other types of dependence might exists between N and $\{Y_j\}$. Indeed, more flexible relation among $\{Y_j\}$ could be accommodated by further specifying a joint distribution of $\{Y_j\}$ given N. Since the focus of this work is the association between N and each Y_j rather than the association within $\{Y_j\}$, we leave this potential generalization of the current model for future investigation. Second, the proposed model is flexible such that several commonly used two-part models can be viewed in the copula framework. Specific examples include the hurdle model (Mullahy (1986)), the selection model (Smith (2003)) and the frequency-severity model (Frees (2014)). Detailed discussions can be found in Section S.1 of the Supplementary Material. Third, the current representation assumes Y to be a nonnegative continuous outcome. However, the framework is ready to accommodate discrete outcomes with suitable modifications for (2.3). For instance, Y could be a count variable in the study of health care utilization under multiple spells of illness.

2.2. *Incomplete data.* Insurance contracts typically contain some cost sharing features, such as deductible and policy limit, to reduce the cost of insurers. Due to such coverage modifications, N and/or Y are often not completely observed. Motivated by such observations, we extend the basic copula model to accommodate incomplete data.

Presumably the contract has a per-occurrence deductible d and a policy limit l. The deductible refers to the maximal amount of loss assumed by the policyholder, and the policy limit represents the maximal possible indemnification from the insurer. Note that both quantities vary by policyholders. Given that deductible and policyholder will affect the frequency and severity observed by the insurer, we denote $\widetilde{N}$ and $\widetilde{Y}$ as the corresponding modified variables. Hence, the modified aggregate loss to the insurer is

$$\widetilde{S} = \widetilde{Y}_1 + \cdots + \widetilde{Y}_{\widetilde{N}}.$$

We consider two cases of incomplete data. The first one corresponds to the per-loss scenario as defined in Klugman, Panjer and Willmot (2012). This scenario assumes that all accidents are reported to the insurer regardless of whether the loss amount exceeds the deductible. In this case the frequency component is not affected by coverage modifications; thus, $\widetilde{N} = N$. However, the severity component will be adjusted by

$$\widetilde{Y} = \begin{cases} 0 & Y \le d, \\ Y - d & d < Y \le l, \\ l - d & Y > l. \end{cases}$$

Thus, the joint distribution of $(\widetilde{N}, \widetilde{Y}_1, \dots, \widetilde{Y}_{\widetilde{N}})$ can be shown as

$$\begin{aligned} f_{\widetilde{N},\widetilde{\mathbf{Y}}}(n, y_1, \dots, y_n) &= [f_{\widetilde{N}}(0)]^{I(n=0)}[f_{\widetilde{N}}(n) \times f_{\widetilde{\mathbf{Y}}|\widetilde{N}}(y_1, \dots, y_n|n)]^{I(n>0)} \\ &= [f_{\widetilde{N}}(0)]^{I(n=0)}\left[f_{\widetilde{N}}(n) \times \prod_{j=1}^{n} f_{\widetilde{Y}|\widetilde{N}}(y_j|n)\right]^{I(n>0)}, \end{aligned} \tag{2.7}$$

where $f_{\widetilde{N}}(n) = f_N(n)$ and

$$\begin{aligned} f_{\widetilde{Y}|\widetilde{N}}(y|n) &= \begin{cases} \Pr(\widetilde{Y} = 0|\widetilde{N} = n) & y = 0, \\ \dfrac{\partial}{\partial y}\Pr(\widetilde{Y} \le y|\widetilde{N} = n) & 0 < y < l - d, \\ \Pr(\widetilde{Y} = l - d|\widetilde{N} = n) & y = l - d \end{cases} \\ &= \begin{cases} \Pr(Y \le d|N = n) & y = 0, \\ \dfrac{\partial}{\partial y}\Pr(Y \le y + d|N = n) & 0 < y < l - d, \\ \Pr(Y \ge l|N = n) & y = l - d \end{cases} \\ &= \begin{cases} F_{Y|N}(d|n) & y = 0, \\ f_{Y|N}(y + d|n) & 0 < y < l - d, \\ 1 - F_{Y|N}(l|n) & y = l - d. \end{cases} \end{aligned}$$

As pointed out by one reviewer, the copula between N and $\widetilde{Y}$ stays unchanged since censoring is a monotone increasing function of Y.

The second one corresponds to the per-payment scenario as defined in Klugman, Panjer and Willmot (2012). Differing from the former scenario, the accident with a loss amount below the deductible is unobservable to the insurer. Hence both frequency and severity are modified by coverage modifications. The relation between the original and modified variables are

$$\widetilde{N} = I(Y_1 > d) + \cdots + I(Y_N > d) \quad \text{and} \quad \widetilde{Y} = \begin{cases} - & Y \le d, \\ Y - d & d < Y \le l, \\ l - d & Y > l. \end{cases}$$

To derive the distribution of $(\widetilde{N}, \widetilde{Y}_1, \dots, \widetilde{Y}_{\widetilde{N}})$, we assume, without loss of generality, the first k $(\le \widetilde{N} = n)$ claims are below maximum indemnification, and the rest $n - k$ claims receive maximum payments, that is, $0 < y_1, \dots, y_k < l - d$ and $y_{k+1}, \dots, y_n = l - d$. Then, we have

$$\begin{aligned} &f_{\widetilde{N},\widetilde{\mathbf{Y}}}(n, y_1, \dots, y_n) \\ &\quad= \frac{\partial^k}{\partial y_1 \cdots \partial y_k}\Pr(\widetilde{N} = n, \widetilde{Y}_1 \le y_1, \dots, \widetilde{Y}_k \le y_k, \widetilde{Y}_{k+1} = \cdots = \widetilde{Y}_n = l - d) \\ &\quad= \mathrm{E}\left[\frac{\partial^k}{\partial y_1 \cdots \partial y_k}\Pr(\widetilde{N} = n, \widetilde{Y}_1 \le y_1, \dots, \widetilde{Y}_k \le y_k, \widetilde{Y}_{k+1} = \cdots = \widetilde{Y}_n = l - d|N)\right] \\ &\quad= \mathrm{E}\left[\binom{N}{n}\frac{\partial^k}{\partial y_1 \cdots \partial y_k}\Pr(d < Y_j \le y_j + d, j = 1, \dots, k,\right. \\ &\qquad\left. Y_{k+1}, \dots, Y_n > l, Y_{n+1}, \dots, Y_N \le d|N)\right] \end{aligned}$$

$$= \mathrm{E}\left[\binom{N}{n}\prod_{j=1}^{k}\Pr(Y_j = y_j + d)\prod_{j=k+1}^{n}\Pr(Y_j > l)\prod_{j=n+1}^{N}\Pr(Y_j \le d)\right]$$

$$= \mathrm{E}\left[\binom{N}{n}\prod_{j=1}^{k} f_{Y|N}(y_j + d|N)\left[1 - F_{Y|N}(l|N)\right]^{n-k}\left[F_{Y|N}(d|N)\right]^{N-n}\right].$$

Though motivated by insurance applications, the above cases are representative of two common mechanisms for incomplete observations, censoring and truncation. Our method relies on the assumption that censoring or truncation is exogenous, that is, the underlying distribution of N and Y are not affected by such mechanisms.

2.3. *Inference.* Because of the parametric nature of the proposed copula model, parameters can be estimated using likelihood-based approach. Denote model parameters by $\boldsymbol{\theta} = (\theta^f, \theta^s, \theta^c)$, where θ^f is the vector of parameters in the frequency model, θ^s is the vector of parameters in the severity model and θ^c represents association parameters in the bivariate copula. For complete data and censored data, one could employ either two-stage MLE or full MLE. However, for truncated data only full MLE is available. In the following, we give detailed estimation procedures for the case of complete data. The procedures for the censored and truncated data are similar and thus omitted.

Using the basic model (2.1), the log-likelihood function for subject i is shown as

$$l_i(\boldsymbol{\theta}) = \log f_N(n_i) + I(n_i > 0) \times \sum_{j=1}^{n_i} \log f_{Y|N}(y_{ij}|n_i).$$

Given a random sample $\{N_i, \boldsymbol{Y}_i\}_{i=1}^m = \{n_i, y_{i1}, \dots, y_{in_i}\}_{i=1}^m$, the full log likelihood for the case of complete data can be written as

$$\begin{aligned}
L(\boldsymbol{\theta}) &= \sum_{i=1}^{m} \log f_N(n_i) + \sum_{\{i:n_i>0\}} \sum_{j=1}^{n_i} \log f_{Y|N}(y_{ij}|n_i) \\
&= \sum_{i=1}^{m} \log f_N(n_i) - \sum_{\{i:n_i>0\}} n_i \log f_N(n_i) \\
&\quad + \sum_{\{i:n_i>0\}} \sum_{j=1}^{n_i} \{\log f_Y(y_{ij}) + \log[h(F_N(n_i), F_Y(y_{ij})) - h(F_N(n_i - 1), F_Y(y_{ij}))]\}.
\end{aligned}$$

One estimation strategy is the full information likelihood method. The full MLE $\hat{\boldsymbol{\theta}}$ can be obtained as the maximizer of the full log likelihood function $L(\boldsymbol{\theta})$. Under regularity conditions, for example, Theorem 3.3 in Newey and McFadden (1994), $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal. The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ can be consistently estimated using the inverse of observed information at the full MLE $\hat{\boldsymbol{\theta}}$, that is, $-[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}L(\hat{\boldsymbol{\theta}})]^{-1}$.

The above likelihood function also suggests a two-stage estimation strategy. Denote the two-stage MLE by $\hat{\boldsymbol{\theta}}^{2s} = (\hat{\theta}^f_{2s}, \hat{\theta}^s_{2s}, \hat{\theta}^c_{2s})$, and further denote

$$L_1(\theta^f) = \sum_{i=1}^{m} \log f_N(n_i), \qquad L_2(\boldsymbol{\theta}) = \sum_{i=1}^{m} I(n_i > 0) \times \left[\sum_{j=1}^{n_i} \log f_{Y|N}(y_{ij}|n_i)\right],$$

we have $L(\boldsymbol{\theta}) = L_1(\theta^f) + L_2(\boldsymbol{\theta})$. In the first stage, one estimates the count regression model $f_N(n_i)$ to obtain $\hat{\theta}^f_{2s}$ by solving $\frac{\partial}{\partial\theta^f}L_1(\theta^f) = 0$. Fixing the parameters in the first part, $\theta^f =$

$\hat{\theta}_{2s}^{f}$, the second stage estimates the conditional model $f_{Y|N}(y_{ij}|n_i)$ to obtain $\hat{\theta}_{2s}^{s}$ and $\hat{\theta}_{2s}^{c}$ by solving $\frac{\partial^2}{\partial(\theta^s,\theta^c)}L_2(\hat{\theta}_{2s}^{f},\theta^s,\theta^c)=0$. Under the regularity conditions of Theorem 6.1 in Newey and McFadden (1994), $\hat{\boldsymbol{\theta}}^{2s}$ is consistent and asymptotically normal. However, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^{2s}$ can be tedious to calculate. The advantage of the two-stage MLE is its computational efficiency. Thus, to speed up the computation, we first obtain $\hat{\boldsymbol{\theta}}^{2s}$ and then use it as the initial point for the maximization of the full likelihood.

The proposed two-stage approach differs from the inference functions for marginals (IFM) method that is widely used in copula regression (Joe (2005)). The IFM first estimates parameters in the univariate marginal models and then estimates the association parameters in the copula. In our case the parameters in the severity component and the copula shall be estimated simultaneously. Applying IFM estimation to the proposed copula model will lead to inconsistent estimation because the marginal likelihood for Y is not observed when $N=0$.

For model comparison one could refer to information-based criteria, such as AIC or BIC. To assess the goodness-of-fit of the copula model, we suggest the following steps. The adequacy of fit for the count regression can be examined using the standard Pearson's chi-squared test. The usual diagnostic analysis for neither the marginal distribution of Y nor the bivariate copula is applicable in our case, for the same reason that the two pieces must be estimated jointly. Therefore, we employ a procedure based on the conditional distribution $f_{Y|N}$. Specifically, we calculate the fitted distribution $\widehat{F}_{Y|N}(y_{ij}|n_i)$ for $i=1,\dots,m$ and $j=1,\dots,n_i$. One expects the sequence $\{\widehat{F}_{Y|N}(y_{ij}|n_i)\}$ to be a sample of uniform $(0,1)$, provided that the copula model is correctly specified. In addition, one could visualize the adequacy of fit with a normal QQ plot by graphing the empirical quantiles from $\{\Phi^{-1}(\widehat{F}_{Y|N}(y_{ij}|n_i))\}$ against the theoretical quantiles from a standard normal distribution. We demonstrate in detail the usage of the proposed diagnostic tools in Section S.3 of the Supplementary Material.

3. Numerical experiments.

3.1. *Impact of dependence between N and Y.* This section presents two numerical experiments to emphasize the importance of the dependency between N and Y. Consider a compound distribution $S=Y_1+\cdots+Y_N$, where $N\sim\text{Poisson}(\lambda=1)$, $Y\sim\text{Gamma}(\alpha=2,\gamma=500)$ and joint distribution of N and Y is specified by a parametric copula. This setting is of particular interest because of the special case where S is known as Tweedie compound Poisson distribution when N and Y are independent. As noted by Jørgensen (1987), under parameterizations $\lambda=\mu^{2-p}/[\phi(2-p)]$, $\alpha=(2-p)/(p-1)$ and $\gamma=\phi(p-1)/\mu^{p-1}$, this distribution can be expressed in the form of the exponential dispersion model with a power variance function $V(\mu)=\mu^p$ for $p\in(1,2)$.

The first experiment demonstrates the effect of frequency-severity dependency on the distribution of aggregate loss. The distribution of S is calculated using Monte Carlo simulation and is displayed in Figure 1. The first panel uses the Gaussian copula with different levels of dependence measured by Kendall's *tau*. When $tau=0$, the copula model reduces to the independence case which is equivalent to a Tweedie distribution ($\mu=1000$, $p=4/3$, $\phi=150$). The positive (negative) dependence leads to a longer (shorter) tail in the aggregate loss distribution. The second panel compares three copulas (Gaussian, Clayton and Gumbel) with the same Kendall's *tau*. One observes the effect of tail dependence (upper for Gumbel and lower for Clayton), although it is not substantial.

The second experiment examines the effect of frequency-severity dependence on the conditional severity distribution. Figure 2 reports the distribution of Y given N at different levels of dependence. In each panel we show densities $f_Y(y)$, $f_{Y|N>0}(y|N>0)$ and $f_{Y|N}(y|n)$. The

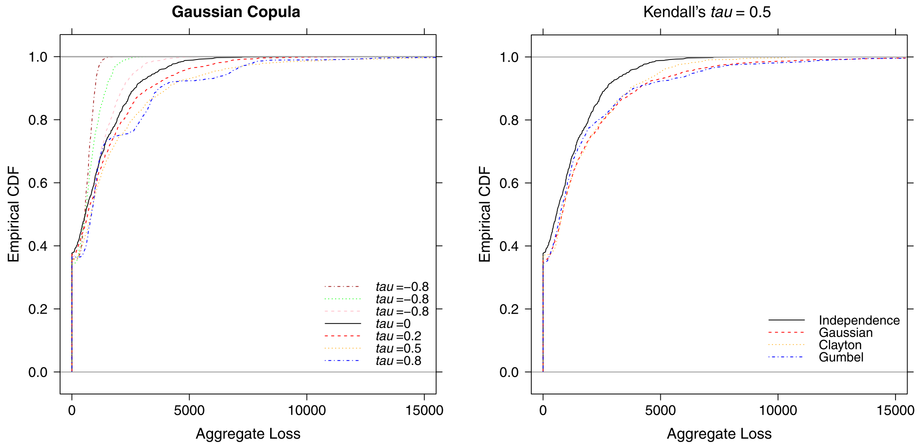

FIG. 1. *Empirical CDF of aggregate loss. The left panel simulates data from the Gaussian copula with different Kendall's tau, and the right panel simulates data from different copulas with the same Kendall's tau.*

former two cases correspond to the common practice where the claim amount is not affected by the number of claims given occurrence. The result is indicative of severe misspecification bias when the dependence between frequency and severity is ignored.

3.2. *Estimation based on the joint distribution of N and $\boldsymbol{Y}$.* This simulation study examines the finite-sample performance of the estimations based on the joint distribution of N and $\boldsymbol{Y}$ and further demonstrates the inference bias incurred by ignoring the frequency-severity dependence. We consider the Gaussian copula compound model in a regression context. The primary distribution is Poisson and the secondary distribution is gamma with

$$\begin{aligned}\text{Poisson}: &\quad \log(\mathrm{E}(N_i)) = \log(\lambda_i) = \beta_0^f + \beta_1^f X_{1i} + \beta_2^f X_{2i},\\ \text{Gamma}: &\quad \log(\mathrm{E}(Y_{ij})) = \log(\alpha\gamma_i) = \beta_0^s + \beta_1^s X_{1i} + \beta_2^s X_{2i},\end{aligned}$$

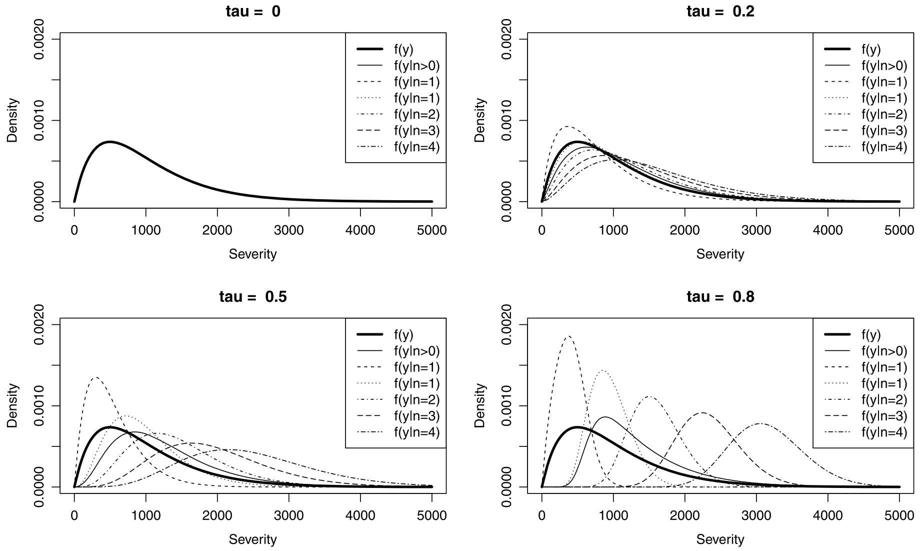

FIG. 2. *The conditional distribution of loss amount given number of claims. The four panels correspond to different levels of dependence between claim frequency and severity.*

TABLE 1

Estimation results for complete data using the two-stage approach and the joint MLE

	Independence			Two stage			Joint MLE		
Parameter	Mean	Relative bias	RMSE	Mean	Relative bias	RMSE	Mean	Relative bias	RMSE
Low Dependence									
$\beta_0^f = -1.5$	−1.515	0.010	0.107	−1.515	0.010	0.107	−1.518	0.012	0.114
$\beta_1^f = 2.5$	2.524	0.009	0.125	2.524	0.009	0.125	2.516	0.006	0.132
$\beta_2^f = 1$	0.995	−0.005	0.073	0.995	−0.005	0.073	1.002	0.002	0.075
$\beta_0^s = 5$	5.092	0.018	0.124	4.988	−0.002	0.093	4.991	−0.002	0.091
$\beta_1^s = -2.5$	−2.552	0.021	0.110	−2.493	−0.003	0.101	−2.495	−0.002	0.105
$\beta_2^s = 5$	4.977	−0.005	0.056	5.004	0.001	0.054	5.001	0.000	0.051
$\alpha = 2$	2.061	0.030	0.109	1.998	−0.001	0.097	2.005	0.003	0.093
$\rho = 0.1$				0.104	0.039	0.042	0.102	0.023	0.039
Medium Dependence									
$\beta_0^f = -1.5$	−1.487	−0.009	0.106	−1.487	−0.009	0.106	−1.500	0.000	0.098
$\beta_1^f = 2.5$	2.478	−0.009	0.118	2.478	−0.009	0.118	2.501	0.000	0.116
$\beta_2^f = 1$	1.005	0.005	0.078	1.005	0.005	0.078	0.998	−0.002	0.069
$\beta_0^s = 5$	5.419	0.084	0.429	5.002	0.000	0.082	5.002	0.000	0.079
$\beta_1^s = -2.5$	−2.733	0.093	0.262	−2.503	0.001	0.094	−2.506	0.002	0.103
$\beta_2^s = 5$	4.913	−0.017	0.104	5.001	0.000	0.057	5.005	0.001	0.053
$\alpha = 2$	2.420	0.210	0.432	2.005	0.003	0.104	2.009	0.004	0.106
$\rho = 0.5$				0.501	0.003	0.028	0.500	−0.001	0.026
High Dependence									
$\beta_0^f = -1.5$	−1.509	0.006	0.110	−1.509	0.006	0.110	−1.503	0.002	0.078
$\beta_1^f = 2.5$	2.507	0.003	0.134	2.507	0.003	0.134	2.497	−0.001	0.091
$\beta_2^f = 1$	1.003	0.003	0.080	1.003	0.003	0.080	1.002	0.002	0.058
$\beta_0^s = 5$	5.690	0.138	0.698	4.999	0.000	0.090	5.000	0.000	0.058
$\beta_1^s = -2.5$	−2.870	0.148	0.395	−2.500	0.000	0.129	−2.507	0.003	0.082
$\beta_2^s = 5$	4.855	−0.029	0.166	5.001	0.000	0.069	5.001	0.000	0.050
$\alpha = 2$	3.083	0.541	1.109	2.004	0.002	0.081	2.000	0.000	0.077
$\rho = 0.9$				0.900	0.000	0.006	0.901	0.001	0.006

where X_{1i} and X_{2i} are i.i.d. and $X_1 \sim \text{Uniform}(0, 1)$ and $X_2 \sim \text{Bernoulli}(0.5)$. In the Gaussian copula we consider different degrees of dependence. The copula model is estimated using both the two-stage method and the joint MLE, and the results are summarized in Table 1. We report the relative bias and the root mean squared error. The calculations are based on a sample size of 500 with 250 replications. There is no substantial difference in the estimates from the two approaches. For comparison, we also report in the table the results of the standard two-part model where N and Y are assumed to be independent. As anticipated, the estimates for the frequency model is consistent with the copula approach. However, the estimation assuming conditional independence introduces a long-term bias in the severity model, and this bias positively correlates with the association between N and Y.

Additional simulation studies are provided in Section S.2 of the Supplementary Material to illustrate the estimation for incomplete data. We emphasize that, in contrast to the cases of complete data and censored data, independence estimation will introduce persistent bias in both frequency and severity components of the model when data are truncated.

4. Modeling aggregate insurance claims.

In nonlife insurance (including property, casualty and health) the compound distribution (1.1) is a common approach to modeling aggregate losses in an insurance system. Examples of an insurance system include a single policyholder, a line of business or a portfolio of contracts. The compound distribution is known as collective risk model in the actuarial literature, and the frequency and severity components are the two building blocks of the model (Klugman, Panjer and Willmot (2012)).

In this application we examine the Wisconsin local government property fund which provides property insurance for local government entities in the State of Wisconsin, such as court houses, school districts, fire stations, etc. We consider the building and contents coverage 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. Similar to most nonlife insurance product, the contract provided by the property fund has a one-year term.

The insurance system in this context corresponds to a policyholder, that is, a local government entity. The outcome of interest is the aggregate loss for an entity during the policy year, determined by both the number and the size of claims. As discussed in Section 1, the collective risk model implies a frequency-severity approach for modeling the aggregate loss for each policyholder, and the current practice relies on the independence assumption between the two building blocks N and Y in the collective risk model.

The purpose of the analysis below is twofold. First, we provide empirical evidence of significant negative association between the frequency and severity of insurance claims; second, we show that ignoring the frequency-severity dependence could lead to biased decision-making in insurance operations. In the following sections we use the term "independence model" to refer to the standard frequency-severity model that assumes independence between the frequency and severity components and "copula model" to refer to the proposed copula approach in Section 2.1 that allows for flexible dependence between the frequency and severity components.

Granular insurance claim data are collected for a portfolio of local government entities for years 2009–2011. For each policyholder one observes the number of claims and the ground-up loss of each claim during each year. We use data of 2009 and 2010 to develop the model, and data of 2011 for model validation. There are 2080 and 1017 policy-year observations in the training data and validation data, respectively.

4.1. *Exploring frequency-severity association.*

To explore the relationship between claim frequency and severity, we display in Figure 3 the violin plot of claim size by the number of claims for the portfolio of government entities. To account for exposure, the claim size is normalized by the amount of coverage. First, one observes that, given occurrence, the distribution of claim severity correlates with claim frequency. Second, the violin plot suggests a negative relation between claim severity and frequency, that is, the amount of claims tends to be smaller for policyholders who have more claims.

To further motivate usage of the proposed copula model, we perform some preliminary analyses to examine the role of frequency-severity dependence in model fitting. Our starting point is the Tweedie model, given it is the industry standard in property-casualty insurance for modeling semicontinuous loss cost. Recall that the Tweedie distribution is a Poisson sum of gamma variables where the Poisson and gamma variables are assumed to be independent. To examine the role of dependence, we further allow the Possion and gamma variables in the Tweedie distribution to be correlated. Specifically, we fit a copula model for the aggregate loss where the frequency is a Poisson variable, the severity is a gamma variable and their joint distribution is specified by a bivariate Gaussian copula. The association parameter in Gaussian copula is estimated to be -0.278 with a standard error of 0.022. This result is consistent with the pattern suggested by Figure 3.

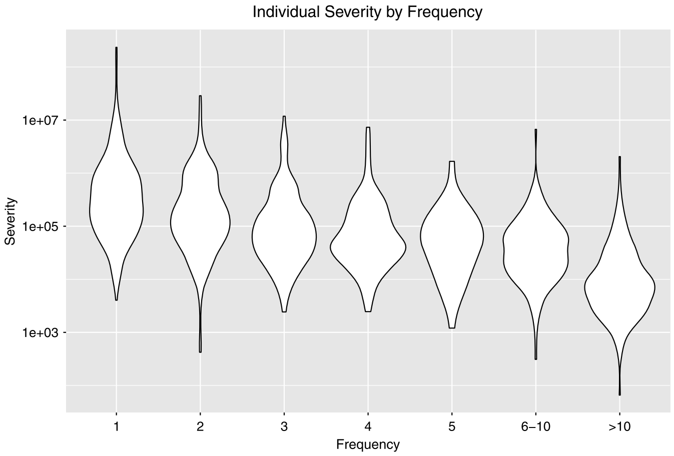

FIG. 3. *Violin plot of claim amount per* \$1000 *coverage by the number of claims.*

To compare the Tweedie and copula models, we present in Figure 4 two goodness-of-fit plots. Denote $F_S(s)$ as the cumulative distribution function (CDF) of aggregate loss. The left figure shows the fitted CDF of the aggregate loss from the two parametric models along with the empirical estimate. Since the plot of CDF emphasizes the center of the distribution, it is not ideal to visualize the effects of extremal large values. To further investigate the tail fit, the right figure plots $-\log(1 - F_S(s))$ between the empirical distribution and the two parametric (Tweedie and copula) models. On one hand, both plots indicate that the copula model exhibits superior fit to the Tweedie model, emphasizing the importance of frequency-severity dependence. On the other hand, there is still room for improvement of goodness-of-fit in both the center and the tail of the distribution. This suggests considering more flexible distributions for marginal behavior. To illustrate, we fit another copula model using zero-one inflated negative binomial distribution for claim frequency, the generalized beta of the second kind (GB2) distribution for claim severity and a Gaussian copula between the two components. The estimated association parameter is -0.207 with a standard error of 0.032. The corresponding goodness-of-fit plots are also shown in Figure 4. As anticipated, refined marginal models improve the fit, especially in the heavy right tail. Overall, the preliminary analyses suggests that there is significant negative dependence between claim frequency and severity and, accounting for such association, enhances the goodness-of-fit for the aggregate loss distribution.

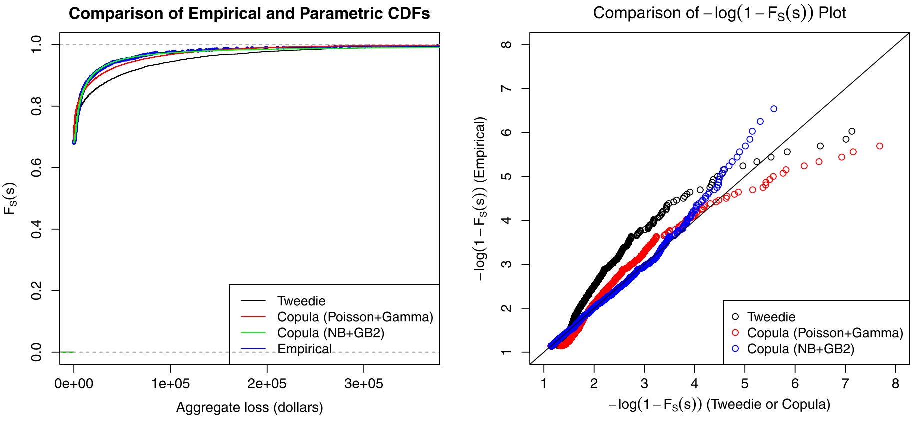

FIG. 4. *Comparison between empirical and parametric Cumulative Distribution Functions* (*CDF, denoted by* $F_S(s)$) *of aggregate loss.*

4.2. *Empirical analysis*. The observation in Section 4.1 motivates us to jointly examine the frequency and severity components in the collective risk model. Differing from the earlier preliminary analysis, first, we explore using more flexible marginal distributions for modeling the number and the size of insurance claims. Second, we incorporate covariates to account for observed heterogeneity, and thus the relation between frequency and severity is interpreted as residual dependence. Third, we consider various copula that offer different types of dependence in modeling the frequency-severity relationship.

To facilitate model specification, we examine the distributions of both claim frequency and severity, as well as their relationship with available explanatory variables. The insurance database contains policyholder-specific and claim-specific information that one could use to account for the variation in claim frequency and severity. Details of such covariate information are provided in Section S.3 of the Supplementary Material. For claim frequency we consider the policy-level characteristics, including entity type (whether a policyholder is a city, county, township, village or others), alarm credit (whether a policyholder receives a credit for alarm system), the level of deductible and the amount of coverage. For illustration we exhibit in Table 2 the empirical distribution of the number of claims per policyholder in the training data. As usually observed in insurance claims data, the majority of policyholders (about 70%) has zero claims over the year. However, this percentage is much smaller than private lines of business such as personal automobile insurance. Another striking feature of claim counts is that there is an excess of ones in addition to the zero inflation. We further break down the frequency distribution by entity type, as shown in Table 2 and visualized in Figure 5. The substantial variation suggests that entity type is an important predictor for the claim count.

Table 3 summarizes the empirical quantiles of claim amounts. There are in total 1381 claims in the sampling period. The descriptive statistics indicates that claim amount is skewed and heavy-tailed distributed. For claim severity, besides policy-level information, we look into the effects of claim-level information such as peril type, occurrence time and reporting delay. As an example, Table 3 shows the empirical distribution of claim amount by peril type and by occurrence time. The claim amount, due to fire and water damages, tends to be larger compared to other perils, and the loss events occurred in the summer is more likely to result in higher claims. The pattern is also displayed in the violin plot of the claim severity in log scale in Figure 6. The plot reinforces the skewness in the severity distribution and stresses the heterogeneity across occurrence and peril type.

TABLE 2
Distribution of claim frequency: Overall and by entity type (in percentage)

		Entity type					
Frequency	Overall	City	County	School	Town	Village	Others
0	68.08	45.67	19.67	67.11	91.95	70.33	85.45
1	19.38	24.00	31.15	23.36	6.90	20.75	12.27
2	6.54	13.33	20.49	5.26	0.86	7.05	0.91
3	2.12	4.67	6.56	2.63	0.00	1.04	0.45
4	1.49	4.00	9.84	0.66	0.00	0.62	0.00
5	0.67	2.33	4.10	0.16	0.29	0.00	0.00
≥ 6	1.73	6.00	8.20	0.82	0.00	0.21	0.91
Obs	2080	300	122	608	348	482	220

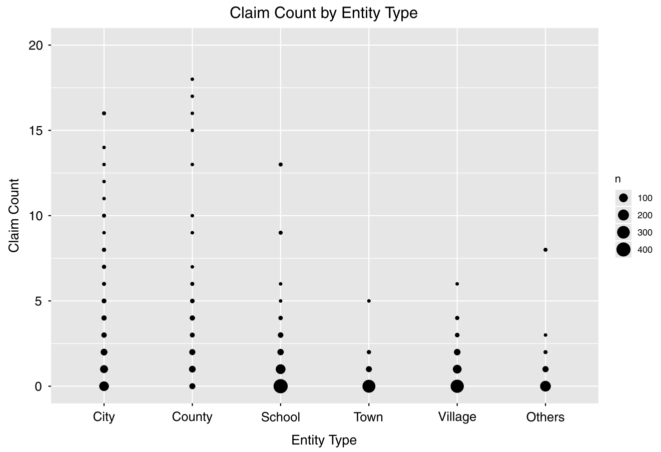

FIG. 5. *Distribution of claim count by entity type.*

In the final model, we consider a zero-one inflated negative binomial regression for claim frequency,

$$f_N(n_i) = p_i^0 I(n_i = 0) + p_i^1 I(n_i = 1) + \big(1 - p_i^0 - p_i^1\big) g_N(n_i), \tag{4.1}$$

where p_i^k $(k = 0, 1)$ is specified using a multinomial logistic regression,

$$p_i^k = \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_k^f)}{1 + \sum_{k=0}^{1} \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_k^f)}, \quad k = 0, 1$$

and $g_N(\cdot)$ is a standard negative binomial model,

$$g_N(n_i) = \frac{\Gamma(\eta + n_i)}{\Gamma(\eta)\Gamma(n_i + 1)} \left[\frac{\eta}{\eta + \exp(\boldsymbol{x}_i \boldsymbol{\beta}^f)}\right]^{\eta} \left[\frac{\exp(\boldsymbol{x}_i \boldsymbol{\beta}^f)}{\eta + \exp(\boldsymbol{x}_i \boldsymbol{\beta}^f)}\right]^{n_i},$$

with $\eta > 0$ being the dispersion parameter. This specification allows to accommodate the excess of both zeros and ones in the claim count. To accommodate the skewness and heavy-tails, a parametric regression based on GB2 distribution is employed for claim severity (for instance, see Shi (2014)) for details on GB2 regression),

$$f_Y(y_{ij}) = \frac{[\exp(w_{ij})]^{\phi_1}}{y_{ij} |\sigma| B(\phi_1, \phi_2) [1 + \exp(w_{ij})]^{\phi_2}}, \tag{4.2}$$

TABLE 3
Distribution of claim amount: Overall, by peril and by occurrence (in dollars)

		Peril			Occurrence			
Quantiles	Overall	Fire	Water	Others	Spring	Summer	Fall	Winter
10	946	1072	1009	790	991	950	945	912
25	1645	2168	1641	1418	1600	1655	1746	1666
50	3542	4989	4200	2945	3021	3859	3802	3619
75	9062	13,069	11,305	5724	7219	11,838	8852	7155
90	29,288	29,849	35,640	22,203	27,872	34,181	26,890	26,758
Obs	1381	400	389	592	290	539	289	263

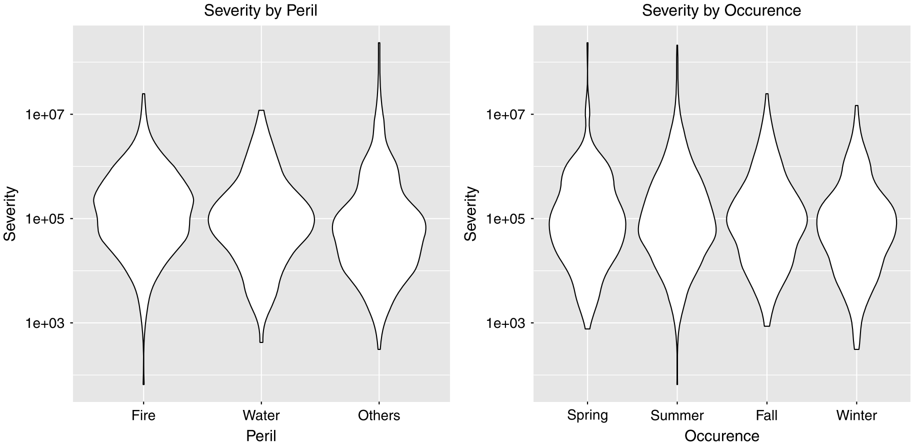

FIG. 6. *Violin plots of claim severity. The left and right panels show severity distributions by peril and occurrence, respectively.*

where ϕ_1 and ϕ_2 are shape parameters, σ is the scale parameter and $w_{ij} = (\log y_{ij} - \boldsymbol{x}_i' \boldsymbol{\beta}^s)/\sigma$. A parametric bivariate copula is employed to construct the joint distribution of N and Y. We consider commonly used bivariate copulas from the elliptical and Archimedean families, including Gaussian, t, Clayton, Frank, Gumbel and Joe. For the Archimedean copulas that only allow for positive association, we consider the associated 90 and 270 degree rotated copulas.

The copulas models are estimated using likelihood-based estimation described in Section 2.3. The corresponding goodness-of-fit statistics are reported in Table 4. The independence model is presented as a benchmark. Model selection criteria AIC and BIC recommend the Gaussian copula model. It appears that the tail dependence is not a concern in this context. The implied Kendall's *tau*, reported in the table, reinforces the negative frequency-severity dependence obtained in the earlier analysis, indicating that the claim frequency and severity are correlated after controlling for the covariates. Because the independence model is nested by the copula model, we perform a likelihood ratio test to formally evaluate the goodness-of-fit of the copula models against the independence model. The large χ^2 statistics confirm the statistical significance of the negative frequency-severity dependence.

The specification for the dependent frequency-severity model, including both the marginals and the copula, is a result of a series of model comparisons, diagnostic analysis, and robust checks. The detailed analysis is provided in Section S.3 of the Supplementary Material.

TABLE 4
Goodness-of-fit statistics for various copula models

	Kendall's *tau*	LogLik	AIC	BIC	Pearson's χ^2
Independence		−15,756	31,587	31,801	
Gaussian	−0.19	−15,720	31,518	31,738	70.77
t	−0.19	−15,719	31,519	31,744	72.38
Clayton90	−0.08	−15,723	31,523	31,743	66.06
Clayton270	−0.33	−15,739	31,555	31,775	34.09
Gumbel90	−0.29	−15,722	31,521	31,741	68.04
Gumbel270	−0.09	−15,731	31,540	31,759	49.53
Frank90/270	−0.22	−15,733	31,544	31,764	45.49
Joe90	−0.34	−15,739	31,557	31,777	32.38
Joe270	−0.05	−15,735	31,548	31,768	41.41

Table 5 reports the estimated parameters for the selected Gaussian copula model. The association parameter in the Gaussian copula is -0.29 and -0.30 using two-stage and full MLE, respectively. Given that the rating variables in insurance are highly regulated, one should regard the observed frequency-severity dependence as a result of unobserved heterogeneity, and thus the sign of the dependence could be either positive and negative. Our focus is to provide a data-driven method to capture such relationship and to show the detrimental effects of ignorant supposition of independence on statistical inference and hence insurance operations. For comparison, we also report in Table 5 the estimation results for the independence model. For the frequency component one anticipates no essential difference in estimates of regression coefficients between the independence and copula models. We observed that the two-stage MLE is identical to the independence model, and we attribute the difference from the full MLE to the finite sample property. In contrast, the difference in the estimates for the severity component is substantial between the independence and copula models (both two-stage and full MLE) which is in line with the significant negative dependence between N and Y. The analysis indicates that ignoring the frequency-severity dependence could introduce significant bias in parameter estimation.

4.3. *Implications on insurance operations.* The previous section shows the statistical significance of the dependence between frequency and severity in the collective risk model. This section focuses on the substantive significance of the frequency-severity dependence and demonstrates its impacts on the decision making in some key insurance operations (Frees (2015)).

The first operation that we consider is underwriting and ratemaking. They are two basic functions in insurance companies and are closely related to each other. The former deals with the selection of risks, and the latter deals with the determination of the price for the risks accepted. To achieve the underwriting profit target, the central task in underwriting and ratemaking is to quantify the risks of potential customers which provides the insurer a risk score of policyholders to facilitate portfolio selection. To compare performance of the independence and the copula models, we look to the policyholders in the validation data of 2011 and examine which method leads to a more profitable portfolio construction.

For the purpose of underwriting, we use the coefficient of variation to measure the risk of policyholders. For each of the 1017 policyholders in year 2011, we calculate the coefficient of variation of the loss cost, denoted $R_i = \sqrt{\mathrm{Var}[S_i]}/\mathrm{E}[S_i]$ for the ith policyholder. Given that the aggregate loss cost is specified using a collective risk model (1.1), the mean and variance of S is calculated by

$$\mathrm{E}[S] = \mathrm{E}\big[N\mathrm{E}[Y|N]\big] \overset{\text{independence}}{=} \mathrm{E}[N]\mathrm{E}[Y],$$

$$\mathrm{Var}[S] = \mathrm{E}\big[N\,\mathrm{Var}[Y|N]\big] + \mathrm{Var}\big[N\mathrm{E}[Y|N]\big] \overset{\text{independence}}{=} \mathrm{E}[N]\,\mathrm{Var}[Y] + \mathrm{Var}[N]\big(\mathrm{E}[Y]\big)^2.$$

The above calculation emphasizes the role of the dependency between the two building blocks, frequency and severity. We calculate the distribution of aggregate loss for each policyholder based on 10,000 Monte Carlo simulations. The upper panel of Figure 7 compares the risk ranking between the independence and the copula models. The first plot is the scatter plot of the ranking for each policyholder by the two methods. The second plot shows the realized aggregate losses (in log scale) with the same ranking from the two models. The risk scores from the two models are highly correlated, yet there are considerable difference in their rankings.

To evaluate whether the risk ranking points to a profitable portfolio selection strategy, we display in the lower panel of Figure 7 the cumulative loss distribution ($F_L(R_i)$) vs. the cumulative premium distribution ($F_P(R_i)$), both ordered by the riskiness of the policyholders

TABLE 5
Parameter estimation for the independence model and the copula model

	Independence				Copula-two stage MLE				Copula-full MLE			
	Frequency		Severity		Frequency		Severity		Frequency		Severity	
	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Intercept	−1.184	0.375	7.212	0.289	−1.184	0.377	7.031	0.317	−0.728	0.397	6.886	0.324
City	0.299	0.257	−0.333	0.198	0.299	0.244	−0.548	0.216	0.485	0.232	−0.616	0.216
County	0.169	0.285	−0.352	0.209	0.169	0.269	−0.637	0.231	0.391	0.260	−0.717	0.232
School	−0.872	0.262	0.141	0.205	−0.872	0.250	0.027	0.221	−0.636	0.238	−0.055	0.222
Town	0.017	0.330	−0.510	0.263	0.017	0.321	−0.610	0.274	0.121	0.312	−0.643	0.274
Village	0.247	0.253	−0.180	0.200	0.247	0.243	−0.387	0.215	0.383	0.235	−0.434	0.215
AlarmCredit05	0.328	0.216	0.060	0.201	0.328	0.215	0.024	0.201	0.316	0.212	0.026	0.200
AlarmCredit10	0.316	0.205	−0.121	0.177	0.316	0.203	−0.201	0.181	0.356	0.201	−0.217	0.180
AlarmCredit15	0.227	0.136	−0.115	0.121	0.227	0.135	−0.123	0.124	0.290	0.134	−0.147	0.124
Deductible	−0.221	0.058	0.095	0.034	−0.221	0.056	0.205	0.042	−0.322	0.064	0.235	0.044
Coverage	0.782	0.054	0.048	0.037	0.782	0.053	−0.010	0.041	0.766	0.052	−0.001	0.041
Spring			−0.110	0.106			−0.064	0.104			−0.065	0.104
Summer			−0.040	0.099			−0.023	0.097			−0.022	0.097
Fall			0.020	0.107			0.049	0.104			0.053	0.104
Fire			0.533	0.085			0.468	0.085			0.466	0.085
Water			0.316	0.084			0.290	0.082			0.288	0.082
ReportDelay			−0.001	0.001			−0.001	0.001			−0.001	0.001
Zero-inflated Regression												
Intercept	−7.834	1.406			−7.834	1.476			−8.583	2.126		
Deductible	1.097	0.185			1.097	0.195			1.126	0.266		
Coverage	−0.538	0.177			−0.538	0.173			−0.583	0.229		
One-inflated Regression												
Intercept	−7.411	1.507			−7.411	1.557			−7.084	1.829		
Deductible	0.664	0.217			0.664	0.224			0.577	0.266		
Coverage	0.020	0.182			0.020	0.184			0.016	0.201		
ρ					−0.290	0.034			−0.303	0.033		

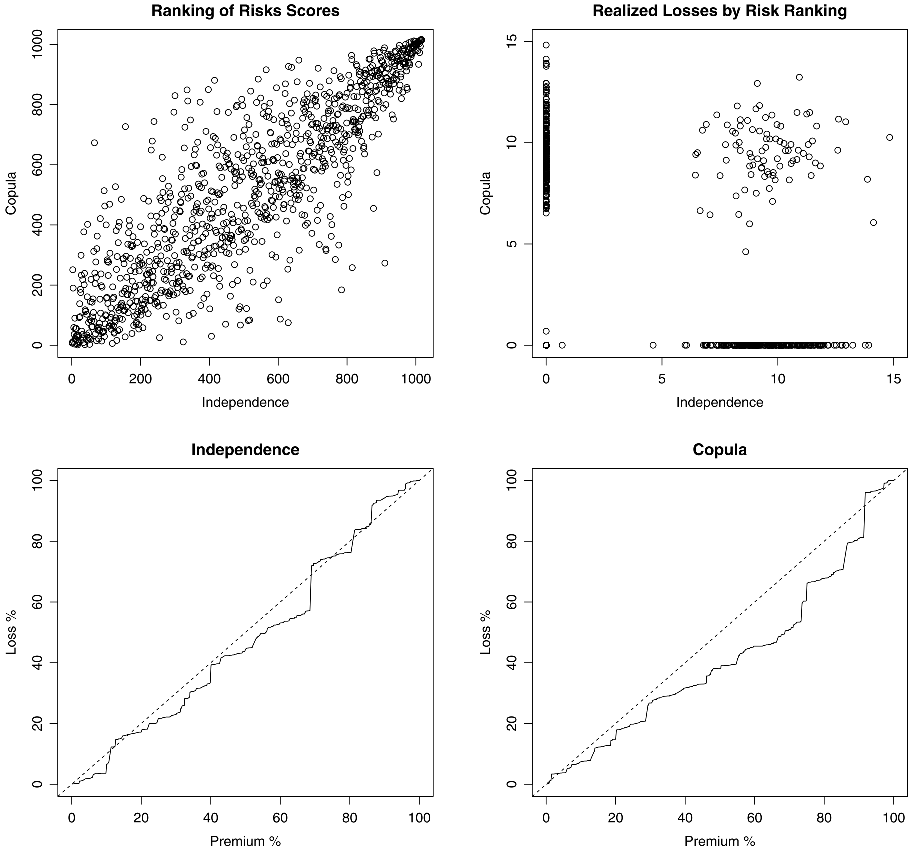

FIG. 7. *Risk ranking and portfolio selection using the independent and the copula models. The top two figures compare risk score ranking and the corresponding realized losses between the independence and copula models, respectively. The bottom two figures compare ordered Lorenz curves between the independence and copula models where the dashed line indicates perfect equality.*

R_i. This curve is known as the ordered Lorenz curve in Frees, Meyers and Cummings (2011). In Figure 7 the loss and premium distributions are calibrated using the realized losses of the policyholders and the actual premiums charged by the insurer in year 2011, respectively. The area between the curve and the 45 degree line is interpreted as an average profit or loss for the portfolio, with a convex curve for profit and a concave curve for loss. If one thinks of each underwriting strategy as retaining policies with riskiness less than or equal to R_i, the area represents an average profit in the sense that we are taking an expectation over all decision-making strategies. Furthermore, twice the area is known as the Gini index which thus has a natural economic interpretation. The Lorenz curve for the independence model is close to the 45 degree line. In contrast, the Lorenz curve for the copula model suggests a much higher average profit. Specifically, the Gini indices are 10.55% and 33.24% for the independence and the copula models, respectively. Therefore, a better underwriting strategy could be formed using the copula model, given that each policyholder is charged the contract premium.

We next compare the rates suggested by the independence and the copula models. A fair rate commensurate with the policyholder's risk mitigates adverse selection against the insurer. We perform a out-of-sample validation based on the Gini correlation in Frees, Meyers and Cummings (2011). Two base premiums are considered, the constant premium and the contract premium. The former charges average cost to each policyholder, and the latter is the premium that the property fund charges based on the basic rating variables. Table 6

TABLE 6

Gini indices for independence and copula models[†]

	Independence	Copula
Constant Premium	57.61 (6.57)	63.24 (6.82)
Contract Premium	15.93 (8.81)	26.27 (11.15)

[†] Standard errors are reported in parentheses.

presents the Gini correlation coefficients for the independence and the copula models. For both premium bases the copula model shows a higher index, implying a more refined risk classification than the independence model.

The proposed copula model can also provide insights for the practice of claims reserving. In property casualty insurance it is typical that a loss event won't be reported to the insurer immediately upon occurrence. For instance, a hail damage to the roof might be discovered by the policyholder several month later. After being reported, it further takes time for the insurer to decide coverage and finally settle the claim. Because of the long reporting and settlement delays, an insurer could be responsible for future payments associated with the loss events occurred in the policy period even post the expiration of the contract. Claims reserving or loss reserving is the process of estimating outstanding payments or the ultimate payments for which an insurer is responsible. Reserves are determined at both claim level and portfolio level (see, e.g., Antonio and Beirlant (2008) and Pigeon, Antonio and Denuit (2014)). At claim level an insurer estimates the amount for which a particular claim will ultimately be settled or adjudicated, also known as case reserve. At portfolio level an insurer also estimates its future liabilities for the entire book of business. To emphasize its importance, loss reserves typically represent the largest liability item on the balance sheet of nonlife insurers.

For reserving purposes, one is interested in the claims amount given occurrence of the loss events. As pointed out by Wüthrich and Merz (2008), because of the introduction of new supervisory guidelines (Solvency II) and financial reporting standards (IFRS 4 Phase II), the measurement of future cash flows and their uncertainty becomes more important. In this application we examine the predictive distribution of Y given N. For illustration, we display in Figure 8 the 95% prediction intervals of the claims amount for four representative risks—"poor, good, average" and "superior." The bar is determined by the 2.5th and 97.5th percentiles of the predictive distribution, and the solid dot indicates the predictive mean. The four risks are selected from the validation data based on the expected number of claims $\mathrm{E}(N)$. Specifically, they expect to have 2.37, 0.76, 0.37, 0.15 claims per year which correspond to the 95th, 75th, 50th, 25th percentiles of the frequency distribution, respectively. For comparison, we impose the corresponding prediction interval from the independence model in the figure as indicated by the dashed line. First, as expected, the predictive distribution of claim amounts, given frequency, is skewed and long-tailed. This observation emphasizes that a range estimate of reserves is more informative than a point estimate for managers to set appropriate reserves, because an insurer doesn't want to overestimate or underestimate its outstanding liabilities. Over-reserving could inflate the price and make the product less competitive, while under-reserving increases the solvency risk. Second, because of the significant negative relation between claim frequency and severity, the claims amount becomes smaller as the number of claims increases. A dynamic viewpoint is that an insurer updates its knowledge on the severity distribution based on frequency information. Third, it is apparent that ignoring the frequency-severity dependence will introduce significant bias in the reserving estimates. Under the independence assumption, not only the claim severity is invariant with respect to claim frequency but also the magnitude of the prediction could lead to poor decision making. For example, the results suggest that managers relying on the independence

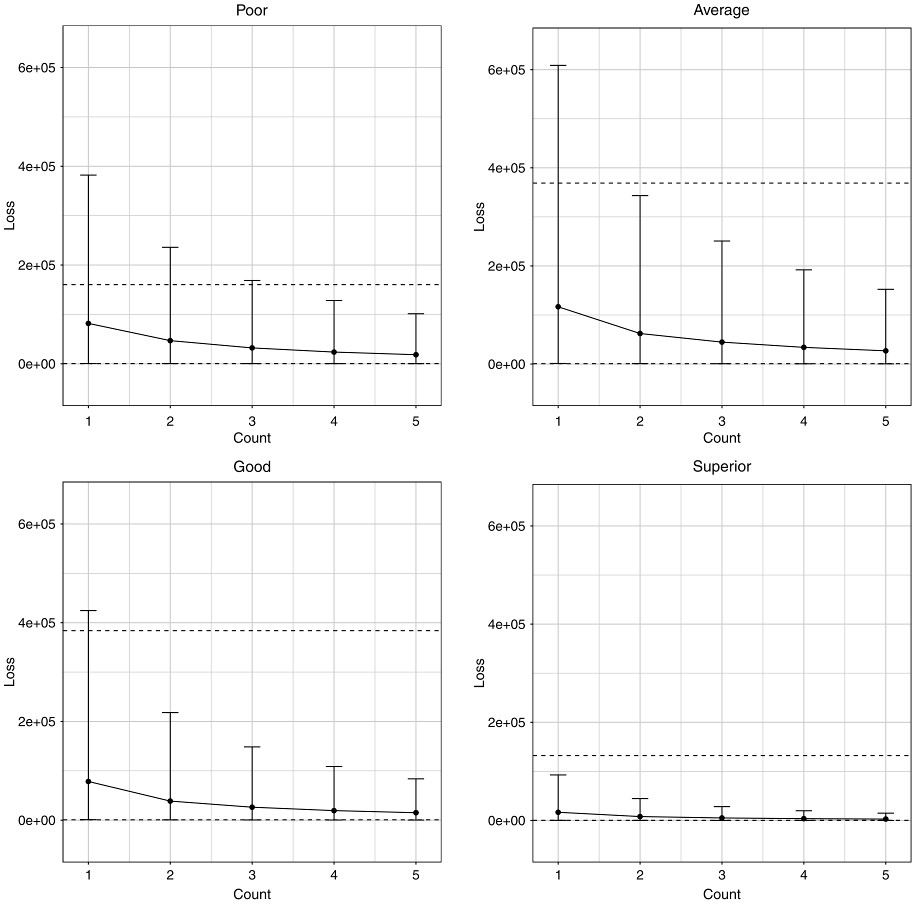

FIG. 8. *Prediction interval of conditional claims amount for four representative risks.*

model tend to over reserve for better risks. In particular, the over-reserving risk is substantial for superior risks. As described earlier, there will be negative effects on both pricing and reserving. Over prediction of unpaid losses leads to increase in price which could cause the insurer to lose profitable business.

We further test the prediction of ultimate losses given occurrence for all the policyholders in the hold-out sample. To compare the prediction from the independence model to the copula model, we employ the continuous ranked probability score (CRPS) in Gneiting and Raftery (2007) and Czado, Gneiting and Held (2009). The CRPS is a proper scoring rule that assesses the quality of probabilistic forecasts. For reserving purposes, we focus on policyholders with at least one claim, and we evaluate the prediction of the aggregate loss distribution $f_{S|S>0}(s)$. The predictive distribution is derived for each policyholder based on 10,000 Monte Carlo simulations where the aggregated loss is generated conditional on occurrence of claims. Then, the CRPS assigns a numerical score that measures the distance between the cumulative predictive distribution and the realized losses in the hold-out sample. For 73.34% of the policies in the hold-out sample, the copula model outperforms the independence model. A binomial test suggests the superior prediction of the copula model to the independence model is statistically significant.

In the third application we briefly demonstrate implications of the frequency-severity dependence on capital management. Insurance is a highly regulated industry. To mitigate solvency risk and protect public interest, insurers are required to hold minimum amount of risk capital as a buffer in case of some unexpected catastrophic events. We have already seen the

TABLE 7
Value-at-risk for the insurance portfolio ($1000)

	0.90	0.95	0.99
Independence	39,556 (38,961, 40,162)	69,124 (67,834, 70,521)	314,854 (300,348, 328,009)
Copula	41,665 (41,106, 42,210)	75,114 (73,921, 76,284)	374,234 (349,748, 397,509)
Difference	5.33%	8.67%	18.86%

consequences when the dependence between frequency and severity is unaccounted for at the individual policy level. This example emphasizes its relevance at the portfolio level since the risk capital is determined for the entire book of business.

To calculate the risk capital, we consider the value-at-risk (VaR), a risk measure widely used in the insurance and banking industry. The VaR focuses on the tail of the distribution, and, specifically, VaR(α) is defined as the 100αth percentile. Our interest is the aggregate losses for the insurance portfolio, defined as $L = \sum_{i=1}^{m} S_i$, where S_i, the loss cost for policyholder i, is specified using the collective risk model (1.1). The distribution of L is estimated using 10,000 Monte Carlo simulations. Table 7 reports the risk measure at 90%, 95% and 99% levels for both the independence and copula models. To quantify the simulation uncertainty, we replicate the simulation 100 times to obtain the 95% confidence interval. The results imply that ignoring the frequency-severity dependence in the collective risk model leads to significant underestimate of the tail risk for the portfolio.

5. Conclusion. The two-part regression model based on compound distributions is commonly used in various disciplines, including insurance, economics, marketing and psychology, among others. The current practice is to perform a marginal regression on the primary (frequency) outcome and a separate regression on the positive portion of the secondary (severity) outcome. This practice relies on the (conditional) independence assumption and causes significant biases in inference in the presence of frequency-severity dependence.

Motivated by the wide application of this type of model, this article represents the first attempt at accommodating the association between the frequency and severity components in the compound distribution and the associated regression models. We proposed the novel idea of using a parametric copula to construct the joint distribution of N and Y in the compound distribution. The copula regression is simple yet enjoys several advantages. First, the copula model allows for an arbitrary dependence between frequency and severity and, thus, includes the (conditional) independence model as a special case. Second, separating the marginal from the joint distribution, the copula model can easily accommodate nonstandard marginal regressions for complicated data structure, for instance, regressions for zero/one-inflated data or the incomplete data due to censoring and truncation. Third, the parametric nature of the model implies straightforward likelihood-based inference and, thus, facilitates data-driven model specification and diagnostics which is critical to the applications with complex and big data.

This work was motivated by the applications in insurance, where the complex and unique features of claims data provide a general setting to investigate the frequency-severity dependence in the context of the two-part model. For example, the standard count regression is not sufficient to capture the features in claim frequency, and the modifications on insurance coverage often cause observations to be incomplete. Although our empirical analysis emphasized the consequences of ignoring the frequency-severity dependence on the operations in insurance companies, the proposed model is general enough and ready to apply to other disciplines. It will be interesting to see the implications of the frequency-severity dependence on decision making in other fields as well.

Finally, we conclude the paper with some discussions on the dependence between the frequency and severity in the proposed copula model. First, the proposed copula model relies on a simplifying assumption for the dependence, that is, the association parameter in the copula is constant and does not vary across covariates. A potential extension is to use a conditional copula approach to allow the association in the copula to be dependent on covariates. See, for example, Patton (2006), Acar, Craiu and Yao (2011), Veraverbeke, Omelka and Gijbels (2011), Fermanian and Wegkamp (2012), and Castro-Camilo, de Carvalho and Wadsworth (2018) for some recent development. We note that some domain knowledge is usually needed to support the conditional copula approach, for instance, the dependence among stock markets could be time varying. We leave it as a future research topic to investigate the conditional dependence in insurance data. Second, we attribute the observed dependence in frequency and severity to unobserved heterogeneity. Regarding whether such relation is positive or negative, we think of this more as an empirical question to investigate. Often there are competing theories to support both positive and negative relationships. For the property insurance in our paper, one example of unobserved heterogeneity that induces dependence is weather related hazard. One can think of a geographical region that has frequent but modest storms vs. another region that has infrequent but very severe storms. Another example of unobserved heterogeneity is the social-economic factors. One can think of some areas with frequent but minor crimes vs. other areas with infrequent but severe crimes. Thus, it is important for the model to offer the flexibility to accommodate both positive and negative relationship and, thus, to allow for an empirical test of alternative theories.

Acknowledgments. We thank the Area Editor, the Associate Editor and two referees for their comments and suggestions that have led to a substantial improvement in the quality of this paper.

Peng Shi gratefully acknowledges the financial support from the Centers of Actuarial Excellence (CAE) Research Grant from the Society of Actuaries.

SUPPLEMENTARY MATERIAL

Supplement A (DOI: 10.1214/19-AOAS1299SUPP; .pdf). We provide additional technical examples, numerical studies and data analysis to support the paper.

REFERENCES

ACAR, E. F., CRAIU, R. V. and YAO, F. (2011). Dependence calibration in conditional copulas: A nonparametric approach. *Biometrics* **67** 445–453. MR2829013 https://doi.org/10.1111/j.1541-0420.2010.01472.x

ALBRECHER, H., BEIRLANT, J. and TEUGELS, J. L. (2017). *Reinsurance*: *Actuarial and Statistical Aspects*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3791478

ANTONIO, K. and BEIRLANT, J. (2008). Issues in claims reserving and credibility: A semiparametric approach with mixed models. *J. Risk Insur.* **75** 643–676.

ARIBARG, A., PIETERS, R. and WEDEL, M. (2010). Raising the BAR: Bias adjustment of recognition tests in advertising. *J. Mark. Res.* **47** 387–400.

CASTRO-CAMILO, D., DE CARVALHO, M. and WADSWORTH, J. (2018). Time-varying extreme value dependence with application to leading European stock markets. *Ann. Appl. Stat.* **12** 283–309. MR3773394 https://doi.org/10.1214/17-AOAS1089

CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513 https://doi.org/10.1111/j.1541-0420.2009.01191.x

CZADO, C., KASTENMEIER, R., BRECHMANN, E. C. and MIN, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scand. Actuar. J.* **4** 278–305. MR3010604 https://doi.org/10.1080/03461238.2010.546147

FERMANIAN, J.-D. and WEGKAMP, M. H. (2012). Time-dependent copulas. *J. Multivariate Anal.* **110** 19–29. MR2927507 https://doi.org/10.1016/j.jmva.2012.02.018

FREES, E. W. (2014). Frequency and severity models. In *Predictive Modeling Applications in Actuarial Science, Volume I: Predictive Modeling Techniques* (E. W. Frees, G. Meyers and R. A. Derrig, eds.) 138–166. Cambridge Univ. Press, Cambridge.

FREES, E. W. (2015). Analytics of insurance markets. *Annu. Rev. Financ. Econ.* **7** 253–277.

FREES, E. W., GAO, J. and ROSENBERG, M. A. (2011). Predicting the frequency and amount of health care expenditures. *N. Am. Actuar. J.* **15** 377–392. MR2869681 https://doi.org/10.1080/10920277.2011.10597626

FREES, E. W., LEE, G. and YANG, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks* **4** 4.

FREES, E. W., MEYERS, G. and CUMMINGS, A. D. (2011). Summarizing insurance scores using a Gini index. *J. Amer. Statist. Assoc.* **106** 1085–1098. MR2894766 https://doi.org/10.1198/jasa.2011.tm10506

GARRIDO, J., GENEST, C. and SCHULZ, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance Math. Econom.* **70** 205–215. MR3543046 https://doi.org/10.1016/j.insmatheco.2016.06.006

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548 https://doi.org/10.1198/016214506000001437

JOE, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *J. Multivariate Anal.* **94** 401–419. MR2167922 https://doi.org/10.1016/j.jmva.2004.06.003

JOE, H. (2015). *Dependence Modeling with Copulas. Monographs on Statistics and Applied Probability* **134**. CRC Press, Boca Raton, FL. MR3328438

JOHNSON, N. L., KEMP, A. W. and KOTZ, S. (2005). *Univariate Discrete Distributions*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR2163227 https://doi.org/10.1002/0471715816

JØRGENSEN, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* **49** 127–162. MR0905186

JØRGENSEN, B. and PAES DE SOUZA, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.* **1** 69–93. MR1286486 https://doi.org/10.1080/03461238.1994.10413930

KARLIS, D. and XEKALAKI, E. (2005). Mixed Poisson distributions. *Int. Stat. Rev.* **73** 35–58.

KLUGMAN, S. A., PANJER, H. H. and WILLMOT, G. E. (2012). *Loss Models: From Data to Decisions*, 4th ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3222004

KRÄMER, N., BRECHMANN, E. C., SILVESTRINI, D. and CZADO, C. (2013). Total loss estimation using copula-based regression models. *Insurance Math. Econom.* **53** 829–839. MR3130478 https://doi.org/10.1016/j.insmatheco.2013.09.003

LIN, X. S. (2014). Compound distributions. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat04411.

LIU, H. and WANG, R. (2017). Collective risk models with dependence uncertainty. *Astin Bull.* **47** 361–389. MR3654415 https://doi.org/10.1017/asb.2017.4

MULLAHY, J. (1986). Specification and testing of some modified count data models. *J. Econometrics* **33** 341–365. MR0867980 https://doi.org/10.1016/0304-4076(86)90002-3

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2197664 https://doi.org/10.1007/s11229-005-3715-x

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics, Vol. IV. Handbooks in Econom.* **2** 2111–2245. North-Holland, Amsterdam. MR1315971

OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438 https://doi.org/10.1198/016214501753168389

PANJER, H. H. (2006). *Operational Risk: Modeling Analytics. Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR2244881 https://doi.org/10.1002/0470051310

PATTON, A. J. (2006). Modelling asymmetric exchange rate dependence. *Internat. Econom. Rev.* **47** 527–556. MR2216591 https://doi.org/10.1111/j.1468-2354.2006.00387.x

PIGEON, M., ANTONIO, K. and DENUIT, M. (2014). Individual loss reserving using paid-incurred data. *Insurance Math. Econom.* **58** 121–131. MR3257343 https://doi.org/10.1016/j.insmatheco.2014.06.012

SHEVCHENKO, P. V. (2010). Calculation of aggregate loss distributions. *J. Oper. Risk* **5** 3.

SHI, P. (2014). Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science* (E. W. Frees, G. Meyers and R. A. Derrig, eds.) Cambridge Univ. Press, Cambridge.

SHI, P., FENG, X. and BOUCHER, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *Ann. Appl. Stat.* **10** 834–863. MR3528362 https://doi.org/10.1214/16-AOAS914

SHI, P. and YANG, L. (2018). Pair copula constructions for insurance experience rating. *J. Amer. Statist. Assoc.* **113** 122–133. MR3803444 https://doi.org/10.1080/01621459.2017.1330692

SHI, P. and ZHAO, Z. (2020). Supplement to "Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims." https://doi.org/10.1214/19-AOAS1299SUPP.

SILVA, J. M. C. and WINDMEIJER, F. (2001). Two-part multiple spell models for health care demand. *J. Econometrics* **104** 67–89. MR1864227 https://doi.org/10.1016/S0304-4076(01)00059-8

SMITH, M. D. (2003). Modelling sample selection using Archimedean copulas. *Econom. J.* **6** 99–123. MR1992394 https://doi.org/10.1111/1368-423X.00101

SMITHSON, M. and SHOU, Y. (2014). Randomly stopped sums: Models and psychological applications. *Front. Psychol.* **5** 1–11.

SMYTH, G. K. and JØRGENSEN, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *Astin Bull.* **32** 143–157. MR1930491 https://doi.org/10.2143/AST.32.1.1020

TELLIS, G. J. (1988). Advertising exposure, loyalty, and brand purchase: A two-stage model of choice. *J. Mark. Res.* **25** 134–144.

TWEEDIE, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics*: *Applications and New Directions* (*Calcutta*, 1981) 579–604. Indian Statist. Inst., Calcutta. MR0786162

VERAVERBEKE, N., OMELKA, M. and GIJBELS, I. (2011). Estimation of a conditional copula and association measures. *Scand. J. Stat.* **38** 766–780. MR2859749 https://doi.org/10.1111/j.1467-9469.2011.00744.x

WÜTHRICH, M. V. and MERZ, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Wiley, New York.