Multivariate negative binomial models for insurance claim counts

Peng Shi a,*, Emiliano A. Valdez b

a Department of Actuarial Science, Risk Management, and Insurance, School of Business, University of Wisconsin - Madison, Madison, WI 53706, United States
b Department of Mathematics, Michigan State University, East Lansing, MI 48824, United States

ABSTRACT

It is no longer uncommon these days to find the need in actuarial practice to model claim counts from multiple types of coverage, such as the ratemaking process for bundled insurance contracts. Since different types of claims are conceivably correlated with each other, the multivariate count regression models that emphasize the dependency among claim types are more helpful for inference and prediction purposes. Motivated by the characteristics of an insurance dataset, we investigate alternative approaches to constructing multivariate count models based on the negative binomial distribution. A classical approach to induce correlation is to employ common shock variables. However, this formulation relies on the NB-1 distribution which is restrictive for dispersion modeling. To address these issues, we consider two different methods of modeling multivariate claim counts using copulas. The first one works with the discrete count data directly using a mixture of max-id copulas that allows for flexible pair-wise association as well as tail and global dependence. The second one employs elliptical copulas to join continuized data while preserving the dependence structure of the original counts. The empirical analysis examines a portfolio of auto insurance policies from a Singapore insurer where claim frequency of three types of claims (third party property damage, own damage, and third party bodily injury) are considered. The results demonstrate the superiority of the copula-based approaches over the common shock model. Finally, we implemented the various models in loss predictive applications.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Modeling insurance claim counts is a critical component in the ratemaking process for property-casualty insurers. Typically insurance companies keep a comprehensive record of the claim history of their customers and have access to an additional set of personal information. The frequency of claim counts to a great extent reveals the riskiness of the insureds. Thus by examining the relation between claim counts and policyholders’ characteristics, the insurer classifies the policyholders and determines the fair premium according to their risk level. For example, in the standard frequency-severity framework, one examines the number of claims and then the size of each claim given occurrence. In addition, the insurer could detect the presence of private information such as moral hazard and adverse selection through analyzing the claim behavior of policyholders, which is useful for the design of an insurance contract.

In practice, it is not uncommon for an insurer to observe claim counts of multiple types from a policyholder when various types of coverage are bundled into one single policy. For example, a homeowner’s insurance could compensate losses from multiple perils, an automobile insurance could offer protection against third party and own damages, and so on. Our goal is to develop multivariate count regression models that accommodate dependency among different claim types.

Following the seminal contribution of Jorgenson (1961), count data regression techniques have been greatly extended and applied in various fields of studies. In general, three classes of
approaches are discussed in the literature: the semi-parametric approach based on the pseudolikelihood method (Nelder and Wedderburn, 1972; Gourieroux et al., 1984), the parametric count regression models (Hausman et al., 1984), and the quantile regression that is a non-parametric approach (Machado and Silva, 2005). Comprehensive reviews for count regression can be found in Cameron and Trivedi (1998) and Winkelmann (2008). One straightforward approach of introducing correlation amongst multivariate count outcomes is through a common additive error. Kocherlakota and Kocherlakota (1992) and Johnson et al. (1997) provided a detailed discussion for the one-factor multivariate Poisson model. Along this line of study, Winkelmann (2000) proposed a multivariate negative binomial regression to account for overdispersion, Karlis and Meligkitisidou (2005) considered a multivariate Poisson model with a combination of common shocks to allow for a more flexible covariance structure, and Bermúdez and Karlis (2011) examined zero-inflated versions of the Poisson model. An alternative way to incorporate correlation among count data is the mixture model with a multiplicative error that captures unobserved individual-specific heterogeneity, for example, see Hausman et al. (1984) and Dey and Chung (1992). A common limitation of the above models is that the covariance structure is restricted to non-negative correlation. Such a limitation could be addressed through multi-factor models with examples that include the multivariate Poisson-log-normal model (Aitchison and Ho, 1989) and the latent Poisson-normal model (van Ophem, 1999). An emerging approach to constructing a general discrete multivariate distribution to support complex correlation structures is to use copulas. Despite of its popularity in dependence modeling, the application of copulas for count data is still in its infancy (Genest and Neslehová, 2007). A relevant strand of studies of multivariate count data is regarding the longitudinal data. Unlike the genuine multivariate outcomes, longitudinal data usually has a large cross-sectional dimension but a small time dimension. See Boucher et al. (2008) for a recent survey on models of insurance claim counts with time dependence.

For purposes of risk classification and predictive modeling, we are more interested in the entire conditional distribution just as in many other applied research. Thus our study will limit to the parametric modeling framework that is based on probabilistic count distributions. Motivated by the characteristics of a claim data set from a Singapore automobile insurer, we are particularly interested in models based on negative binomial distributions.

A count variable $N$ is known to follow a negative binomial (NB) distribution if its probability function could be expressed as

$$Pr(N = n) = \frac{1}{\Gamma(n+\phi)} \frac{\Gamma(n+\phi+\eta)}{\Gamma(n+1)} \left( \frac{1}{1 + \psi} \right)^\phi \left( \frac{\psi}{1 + \psi} \right)^n,$$

for $n = 0, 1, 2, \ldots$

and is denoted as $N \sim NB(\psi, \eta)$ for $\psi, \eta > 0$. The mean and variance of $N$ are $E(N) = \eta / \psi$ and $Var(N) = \eta / \psi^2$, respectively. Compared with the Poisson distribution, the negative binomial accommodates overdispersion via parameter $\psi$. As $\psi \to 0$, overdispersion vanishes and the negative binomial converges to the Poisson distribution. For regression purposes, it is helpful to consider a mean parameterization that could be specified in terms of covariates as in $\lambda = \eta \psi = \exp(x \beta)$, where $x$ denotes the vector of explanatory variables and $\beta$ denotes the vector of regression coefficients. Then the negative binomial regression could come in two different parameterizations. The NB-I model is obtained for $\eta = \sigma^2 \exp(x \beta)$ and takes the form

$$f_{NB-I}(n|x; \beta, \sigma^2) = \frac{\Gamma(\sigma^2 \exp(x \beta) + n)}{\Gamma(\sigma^2 \exp(x \beta)) \Gamma(n + 1)} \left( \frac{1}{1 + \sigma^2} \right)^{\sigma^2 \exp(x \beta)} \left( \frac{\sigma^2}{1 + \sigma^2} \right)^n.$$

The NB-II model is obtained for $\eta = \sigma^{-2}$ and takes the form

$$f_{NB-II}(n|x; \beta, \sigma^2) = \frac{\Gamma(\sigma^{-2} + n)}{\Gamma(\sigma^{-2}) \Gamma(n + 1)} \left( \frac{1}{1 + \sigma^2 \exp(x \beta)} \right)^{-\sigma^{-2} \exp(x \beta)} \left( \frac{\sigma^2}{1 + \sigma^2} \right)^n.$$

Though both models assume the same mean structure of the count variable, their difference could be characterized in terms of a dispersion function $\phi$ such that $Var(N|x) = \phi E(N|x)$. The NB-I model implies a constant dispersion $\phi = 1 + \sigma^2$, while the NB-II model allows for subject heterogeneity in the dispersion $\phi = 1 + \sigma^2 \exp(x \beta)$ (see Winkelmann, 2008).

We examine three methods of constructing multivariate negative binomial models that allow for flexible pair-wise association and explore the possibility of using either NB-I or NB-II formulations. The first approach is to use common shock variables. Since this method relies on the additivity of the count distribution, only NB-I is suitable for this formulation. The other two approaches are based on parametric copulas: one working with the discrete count data directly with the mixture of max-id copulas that allows for flexible pair-wise association as well as global dependence, and the other employing elliptical copulas to join continuousized data while preserving the dependency among the original counts. In the empirical analysis, we look into an insurance portfolio from a Singapore auto insurer where claim frequency of three types of claims (third party property damage, own damage, and third party bodily injury) are considered, and we show that the copula-based approaches outperform the common shock model.

The rest of the paper focuses on the theory and applications of multivariate negative binomial regression models, and it is structured as follows: Section 2 describes the motivating dataset of insurance claim counts from a Singaporean automobile insurer. Section 3 briefly discusses the multivariate model using the combination of common shock variables. Section 4 explores the possibilities of using copulas to construct multivariate models with flexible dependence structures. The estimation and inference results are summarized in Section 5. Section 6 presents predictive applications and compares the performance of alternative models. Section 7 concludes the paper.

2. Data structure and characteristics

The motivating dataset of insurance claim counts is from a major automobile insurer in Singapore. According to the General Insurance Association of Singapore (GIA), automobile insurance is one of the largest lines of business underwritten by general insurers and the gross premium income accounts for over one third of the entire insurance market.

As in most developed countries, automobile insurance protects insureds from various types of financial losses. The protection in Singapore comes in hierarchies. The minimum level of protection, which is also a mandatory coverage for all car owners, covers death or bodily injury to third parties. Although not mandated by law, third party coverage often provides protection against the costs that may arise as a result of damage to third party properties. On top of third party benefits, fire and theft, the policy also covers damage from these respective hazards. The maximum protection is offered by a comprehensive policy, which additionally compensates for losses of the insured vehicle, and in many cases, the associated medical expenses for the insured.

To study the dependency among claim types and also to construct a homogeneous portfolio of policyholders, we limit the analysis to those individuals with comprehensive coverage. Our final sample includes one year of observation of 9739 individuals. For a given accident, there are three possible types of claims,
third party bodily injury, own damage (including both property damage and bodily injury), and third party property damage, and thus seven possible combinations. Instead of investigating the number of accidents and the type of each, we take an alternative approach. Specifically, the claim experience of the insurer allows us to classify the frequency of claims into three categories: the number of claims of third-party bodily injury ($N_1$), the number of claims of own damage ($N_2$), and the number of claims of third-party property damage ($N_3$). The descriptive statistics of claim counts are displayed in Table 1. It is not surprising to see that own damage has the highest claim frequency, because not all accidents would involve a third party. For each type of claim, the variance is slightly larger than the mean, indicating the potential overdispersion typically associated with insurance data.

In addition to the claim history, the insurer has access to a rich set of information for each risk class, including the policyholder’s characteristics (age, gender, marital status, driving experience etc.), the vehicle’s characteristics (model, year, capacity etc.), as well as the experience rating scheme in the insurance market. Such information is useful for risk classification and ratemaking purposes and also provides a set of covariates for our regression type of analysis. Through preliminary investigation, we select a group of explanatory variables that could attribute to the likelihood of incurring accidents. The age of the driver is indicated by variable young, which equals one if the policyholder is under 35. lowncd equals one if NCD is less than 20. Here NCD, which stands for ‘no claims discount’, is a similar experience rating method to the bonus-malus system in the European motor insurance markets. On one hand, the NCD is introduced to encourage safe driving where policyholders will be compensated by a discount in the premium. On the other hand, it is also a good indicator of the policyholder’s claim history, at least for the recent past years. The vehicle age is captured by a continuous variable vage, measured in years. private and vlux are binary variables indicating whether or not the vehicle is a private car and a luxury car, respectively. If the vehicle capacity is small, variable smallcap is set to one. The capacity, measured in terms of engine displacement, is defined as small if less than 1000cc for private vehicles and less than 1 ton for goods vehicles.

The mean and standard deviation of the covariates are also presented in Table 1. Less than half of the individuals are young drivers and about 36% have a low NCD score. The average age of insured vehicles is about seven years. The majority of the policies are written for private vehicles and a small percentage of them are luxury cars. Small vehicles account for only 10% of the portfolio. The low percentage might be due to our restriction to comprehensive policies, because the owners of bigger vehicles and of higher values tend to purchase more coverage.

To motivate the multivariate negative binomial models, we initially perform a marginal analysis on the claim frequency of each type. Specifically, we fit four regression models for the number of claims: the Poisson, the zero-inflated Poisson (ZIP), the negative binomial of type I (NegBin-I), and the negative binomial of type II (NegBin-II). The goodness-of-fit results of the marginal models are exhibited in Table 2, where for each type of count variable, we compare the observed and the corresponding fitted claim frequencies. For example, for the type of third-party bodily injury, 7461 policyholders are observed to have no accidents over the year. The claim frequencies predicted by the four candidate models are 7453, 7460, 7461, and 7461 respectively. Because of the subject heterogeneity introduced by covariates, the fitted frequency is calculated as the summation of marginal probability of each individual in the portfolio. The smaller distance between observed and fitted frequencies suggests a better fit. The formal $\chi^2$-statistic is also provided in Table 2. The NB-I and NB-II fit all types of claim counts consistently well. As anticipated, we observe the poor performance of the Poisson model, which is explained by the equidispersion constraint. The ZIP, accounting for the overdispersion, improves the model fit when compared to the Poisson model. However, the results suggest that the excess of zeros is not an issue for this dataset or it is accommodated well by the negative binomial models if it exists at all. The characteristics of the claim counts data inspire us to build a trivariate negative binomial model that could additionally capture the dependency among the different claim types. This is the focus of the next two sections.

### 3. Multivariate models using common shocks

For policyholder $i$, let $N_{i1}$ denote the number of claims of third party bodily injury, $N_{i2}$ denote the number of claims of own damage, and $N_{i3}$ denote the number of claims of third party property damage. Motivated by the data characteristics in Section 2, we aim to build a multivariate count regression model that uses the negative binomial distribution for marginals and incorporates the dependency among the claim types. One way to introduce correlation among multivariate counts is to use common shock variables and this approach has been extensively examined for the Poisson model. A simple version is the multivariate Poisson model described by Johnson et al. (1997) and Tsionas (2001), where a common covariance term is shared by each pair of count variables. Karlis and Meligkotsidou (2005) proposed an extension by allowing for full covariance...
among Poisson counts via a combination of common shocks. The main drawback of multivariate Poisson models is the limitation of modeling overdispersion that is typically observed in insurance claim counts. To address this issue, the recent study of Bermúdez and Karlis (2011) considered zero-inflated versions of these common shock models. Although these models could capture excess of zeros and overdispersion, only positive correlation is permitted and the interpretation of covariance structure is not straightforward.

The multivariate Poisson models rely on the property that the sum of independent Poisson random variables follows the Poisson distribution. The generalization could be easily done for the negative binomial distribution. A simple trivariate model with a common covariance term could be constructed by assuming

\[
\begin{align*}
N_1 &= U_1 + U_0 \\
N_2 &= U_2 + U_0 \\
N_3 &= U_3 + U_0.
\end{align*}
\]

(1)

Apparently, if one assumes \(U_j \sim NB(\psi, \eta_j)\) for \(j \in \{0, 1, 2, 3\}\), then all marginals follow negative binomial distribution, i.e. \(N_j \sim NB(\psi, \eta_j + \eta)\) for \(j \in \{1, 2, 3\}\). The key assumption in this formulation is the common parameter \(\psi\) for all negative binomial variables. From the perspective of regression, the NB-I distribution is the only appropriate choice for the multivariate negative binomial model and mean parameterizations are often used such that \(\psi = 1\) and \(\eta_j = \lambda_j/\psi\) for \(j \in \{0, 1, 2, 3\}\). The covariates could be incorporated as the usual log-linear model \(\lambda_j = \text{exp}(\mathbf{x}_j^T \boldsymbol{\beta})\), where \(\mathbf{x}_j\) denotes the vector of explanatory variables for the type \(j\) claim count and \(\boldsymbol{\beta}\) is the corresponding vector of regression coefficients. Note that different sets of covariates could be used for each type \((j = 1, 2, 3)\) of claims. Typically, we do not use covariates for \(\lambda_0\) (i.e. \(\lambda_0 = \lambda\)) because this potentially complicates the interpretation of the regression coefficients. Under these assumptions, the count vector \(\mathbf{N} = (N_1, N_2, N_3)\) follows a multivariate negative binomial regression model and has a joint probability function as:

\[
f_j(n_1, n_2, n_3|x) = \sum_{s=0}^{\min(n_1,n_2,n_3)} f_0(s) f_j(n_1-s|x_1) \times f_2(n_2-s|x_2) f_j(n_3-s|x_3)
\]

(2)

where \(x_1 = (x_{11}, x_{12}, x_{13})\), \(f_0(\cdot) = f_{\text{NB-I}}(\cdot; 1, \ln \lambda_0, \sigma^2)\), and \(f_j(\cdot|x_i) = f_{\text{NB-I}}(\cdot|x_i^T \boldsymbol{\beta}_j, \sigma^2)\) for \(j \in \{1, 2, 3\}\). It is straightforward to show the following relations:

\(E(N_j|x) = \lambda_j + \lambda_0\)

and \(\text{Cov}(N_j, N_{j'}) = \begin{cases} (1 + \sigma^2)(\lambda_j + \lambda_0) & \text{if } j = j' \\ (1 + \sigma^2)\lambda_0 & \text{if } j \neq j' \end{cases}\)

Similar to the Poisson model in Tsionas (2001), the above model relaxes the independence assumption among negative binomial counts through the common shock variable \(U_0\). The common shock, serving similar purposes as a random effect, implies an identical positive correlation for all pairs of claim type. Although the symmetric correlation could be useful in modeling claim counts with time dependence (see Boucher et al., 2008), our analysis shows that it is not appropriate for modeling asymmetric association among claim types. Compared with the analogy of the multivariate Poisson regression, the negative binomial model allows for overdispersion. However, the modeling of dispersion is very restrictive because the same overdispersion \((1 + \sigma^2)\) is superimposed on all types of claim counts.

A more flexible covariance structure could be accommodated by extending model (1) with a combination of common shock variables:

\[
\begin{align*}
N_1 &= U_{11} + U_{12} + U_{13} \\
N_2 &= U_{21} + U_{12} + U_{23} \\
N_3 &= U_{31} + U_{13} + U_{33}.
\end{align*}
\]

(3)

To guarantee the marginal count variable \(N_{ij}\), \(j = 1, 2, 3\), to follow the negative binomial distribution, one could assume that \(U_{ij} (j \in \{1, 2, 3\})\) and \(U_{ik} (j, k \in \{1, 2, 3\} \text{ and } j < k)\) are independent NB-I variables. Denoting \(N_{ij} \sim NB(\sigma^2, \lambda_{ij} \sigma^{-2})\) and \(U_{ik} \sim NB(\sigma^2, \lambda_{ik} \sigma^{-2})\), we can easily show that \(N_{ij} \sim \sigma^2 (\lambda_{ij} + \lambda_{ij} + \lambda_{jk} \sigma^{-2})\) for \(j, k, l \in \{1, 2, 3\}\) and \(j \neq k \neq l\). As in model (1), explanatory variables could be injected in the mean parameters through \(\ln \lambda_j = \mathbf{x}_j^T \boldsymbol{\beta}\). For the purposes of easy interpretations, we do not add covariates in parameters \(\lambda_{ijk}\) (i.e. \(\lambda_{ijk} = \lambda_{jk}\)) though it is theoretically permissible.

The count variables \(N_{11}, N_{12}, N_{13}\) then follow a trivariate negative binomial distribution with the joint probability function

\[
f_j(n_1, n_2, n_3|x) = \sum_{s_1=0}^{n_1} \sum_{s_2=0}^{n_2} \sum_{s_3=0}^{n_3} f_{j1}(s_1)f_{j2}(s_2)f_{j3}(s_3) \times f_1(n_1-s_1-s_2|x_1)f_2(n_2-s_1-s_3|x_2) \\
\times f_3(n_3-s_2-s_3|x_3))
\]

(4)

where \(r_1 = \min(n_1, n_2), r_2 = \min(n_1, n_1), r_3 = \min(n_2, n_2, n_3 - s_2), x_1\) is the vector of covariates for individual \(i\) as defined above, \(f_{jk}(\cdot) = f_{\text{NB-I}}(\cdot; 1, \ln \lambda_{jk}, \sigma^2)\), and \(f_{j}(\cdot|x_i) = f_{\text{NB-I}}(\cdot|x_i^T \boldsymbol{\beta}_j, \sigma^2)\). The computation of the joint probability is more expensive in this case. However, as in model (1), it only involves the evaluation of univariate negative binomial distributions. In the trivariate negative binomial regression, each marginal follows the NB-I model with conditional mean

\(E(N_{ij}|x) = \lambda_{ij} + \lambda_{jk} + \lambda_{jk}\), \(j, k, l \in \{1, 2, 3\}\) and \(j \neq k \neq l\)

and the marginals are associated through common shocks with covariance matrix

\(\text{Cov}(N_j, N_{j'}) = (1 + \sigma^2) \begin{pmatrix}
\lambda_{11} + \lambda_{12} + \lambda_{13} & \lambda_{12} & \lambda_{13} \\
\lambda_{12} & \lambda_{22} + \lambda_{12} + \lambda_{23} & \lambda_{23} \\
\lambda_{13} & \lambda_{23} & \lambda_{23} + \lambda_{13} + \lambda_{33}
\end{pmatrix}\)

(4)

Model (4) is a natural generalization of the multivariate Poisson model by Karlis and Meligkotsidou (2005). It allows for flexible correlation structure and accommodates overdispersion. Constructed from the summation of independent negative binomial variables, model (4) is subject to the same criticisms and restrictions as model (1). For example, negative relation among claim types is prohibited and an identical dispersion is assumed for all marginals.

4. Multivariate models using copulas

Copulas provide a flexible way to construct multivariate distributions and are employed to build negative binomial models in this section. By definition, a copula is simply a multivariate joint distribution defined on a \(d\)-dimensional cube \([0, 1]^d\) such that every marginal follows uniform distribution on interval \([0, 1]\). A copula captures both linear and nonlinear relationship and has been widely employed in multivariate analysis. The advantage is that it separates the modeling of marginal and dependence structure, see Joe (1997) and Nelsen (2006) for more details on copulas and dependence modeling. The modeling framework more relevant to this study is the copula regression technique, where distributions are specified conditional on a set of regressors. Copula regressions have been extensively employed in various applied disciplines, with an emerging trend on discrete outcomes. Some recent applications in actuarial science include the modeling of insurance claims (Frees and Valdez, 2008; Frees et al., 2009), the prediction of loss reserves (Zhao and Zhou, 2010; Shi and Frees, 2011), the examination of asymmetric information (Shi and Valdez, 2011; Shi et al., 2012), and the analysis of insurance expenses (Shi and Frees, 2010; Shi, 2012), among others.
In this application, the variable of interest is the count vector \( N_i = (N_1, N_2, N_3) \) for the \( i \)th individual, with \( N_j \) following a negative binomial distribution. Using the concept of conditional copulas, the joint distribution of \( N_i \) given covariates \( \mathbf{x} \) could be expressed in terms of a copula function as:

\[
F_i(n_1, n_2, n_3 | \mathbf{x}) = C(F_1(n_1 | \mathbf{x}_1), F_2(n_2 | \mathbf{x}_2), F_3(n_3 | \mathbf{x}_3) | \mathbf{x}; \Theta)
\]

where \( F_j \) indicates the marginal count distribution, and \( C(\cdot | \mathbf{x}; \Theta) \) denotes the conditional copula with the vector of association parameters \( \Theta \) capturing the dependency among the marginals. The focus of the following sections will be on the alternative copula models that allow for flexible pair-wise associations for the multivariate claim counts. Note that although copula models permit arbitrary distributional choice of marginals, our analysis limits to the two formulations of the negative binomial distribution as motivated by the data. Compared with common shock models, the copula approach does not prohibitively require a common dispersion for all marginals.

### 4.1. Archimedean and related copulas

The copula regression model is fully parametric so that standard likelihood-based methods are readily applicable upon the selection of an appropriate copula and marginal count distributions. For discrete variables such as count data, the likelihood is the joint probability mass function. From Eq. (5), the likelihood contribution of the \( i \)th individual is

\[
f_i(n_1, n_2, n_3 | \mathbf{x}) = \sum_{j_1=0}^1 \sum_{j_2=0}^1 \sum_{j_3=0}^1 (-1)^{j_1+j_2+j_3} \times C(u_{1j_1}, u_{2j_2}, u_{3j_3} | \mathbf{x}; \Theta)
\]

where \( u_{ij} = F_j(n_j | \mathbf{x}_j) \) and \( u_{1j} = F_j(n_j - 1 | \mathbf{x}_j) \) for \( j = 1, 2, 3 \). Since dependency modeling via copulas preserves the shape of marginals, the selection of copulas and marginals could be performed separately. Our procedure is driven by the characteristics of the insurance claims data. Due to the favorable fit of negative binomial distributions, we consider both the NB-I and NB-II models for the marginals. The evaluation of the likelihood in (6) involves the evaluation of the cdf of copulas. Thus copulas with closed-form copula’s would be helpful from a computational perspective. For this reason, the family of Archimedean copulas and its extensions are prime legitimate candidates. We limit our discussion to trivariate copulas because of the three claim types motivated by our empirical data.

For a univariate cdf \( M \) with Laplace transformation \( \psi(\cdot) \), consider the power mixture

\[
\int_0^\infty H_1^j H_2^k H_3^l dM(\gamma) = \psi(-\log H_1 - \log H_2 - \log H_3)
\]

where \( H_j, j \in \{1, 2, 3\} \), are univariate cdf’s, the term \( \gamma \) captures unobserved heterogeneity, and the cdf \( M \) is known as the mixing function. A detailed discussion of the mixture of powers is provided by Joe (1993), following which, the univariate marginals of (7) are \( G_j = \psi(-\log H_j) \). Copulas are distribution functions with uniform \((0, 1)\) margins so that if \( H_j(u_j) \) is chosen to be \( \exp(-\psi^{-1}(u_j)) \), (7) leads to the trivariate Archimedean copula:

\[
C(u_1, u_2, u_3) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2) + \psi^{-1}(u_3)).
\]

One advantage of this family is that both positive and negative associations are permitted by some members, such as the Frank copula. However, with only one dependency parameter associated with \( \psi \), this class of copulas assumes an exchangeable dependence structure and thus could be very restrictive in empirical studies. For example, it is unable to segregate the pair-wise relationship among claims count pairs \((N_1, N_2)\), \((N_1, N_3)\), and \((N_2, N_3)\) in our insurance application.

One extension of Archimedean copulas to allow for non-exchangeable association is given by Joe (1993). This class is known as partially symmetric copulas and has been used by Zimmer and Trivedi (2006) to model discrete data. In general, the non-exchangeable structure in a \( d \)-variate copula is accommodated by \( d-1 \) dependence parameters. The trivariate copula could be derived from the following mixtures of powers:

\[
\int_0^\infty \int_0^\infty H_1^j H_2^k H_3^l dM(\alpha; \gamma) H_1^j dM_1(\gamma)
\]

where the mixing functions \( M_1 \) has Laplace transformation \( \psi_1 \) and \( M_2 \) has Laplace transformation \( \psi_2 \). Similarly, by setting \( H_1(u_1) = \exp(-\psi_1(u_1)) \), \( H_2(u_2) = \exp(-\psi_2(u_2)) \), and \( H_3(u_3) = \exp(-\psi_2(u_3)) \), one has the extension of (8) as:

\[
C(u_1, u_2, u_3) = \psi_1(\psi_1^{-1}(u_1) + \psi_2^{-1}(u_2) + \psi_2^{-1}(u_3)).
\]

Copula (10) is less restrictive than (8) in the sense that the dependencies among three marginals are partially symmetric. To be more specific, the dependence structure is determined by two parameters defined by \( \psi_1 \) and \( \psi_2 \) respectively. In this formulation, \( \psi_2 \) measures the association of \( (U_1, U_2) \) and \( \psi_2 \) measures the association of pairs \((U_1, U_3) \) and \((U_1, U_2) \). Additional constraints on dependence parameters are required for this copula: First, the two association parameters are non-negative, i.e. only positive relation is permitted. Second, the dependence determined by \( \psi_1 \) is stronger than \( \psi_2 \). Note that expression (10) is not the only representation of the partially symmetric trivariate copula. The copula could be symmetric with respect to \((U_1, U_2)\) (or \((U_1, U_3)\)) but not with respect to \((U_1, U_2)\). The above representation is chosen because the dependence symmetry is supported by our data although other constraints make it additionally inappropriate for our empirical analysis. A detailed discussion will be provided in Section 5.

Joe and Hu (1996) proposed the family of maximum-infinitely divisible (max-id) copulas based on mixtures of max-id distributions. This family allows for a more flexible dependence structure but has been almost overlooked in the literature (see, for example, Nikoloulopoulos and Karlis, 2008, 2009 for recent applications). The trivariate max-id copula, to illustrate, is constructed from the mixture:

\[
\int_0^\infty \prod_{1 \leq j < k \leq 3} \int_j \ H_1^{\alpha_j} dM(\gamma)
\]

where \( R_k(\cdot) \), \( 1 \leq j < k \leq 3 \), are bivariate copulas that are max-id. Recall that a multivariate cdf is said to be max-id if all its positive powers are cdf’s (see Joe and Hu, 1996 for more details). Parameter \( \alpha_j \) can be negative only if some of \( R_k \) are product copulas. The univariate marginals in (11) are \( G_j = \psi(-\omega_j + 2 \log H_j) \) for \( j = 1, 2, 3 \). Henceforth, by choosing \( H_j(u_j) = \exp(-\psi^{-1}(u_j)) \) with \( \psi_j = \omega_j + 2 \) for \( j = 1, 2, 3 \), the associated copula could be derived and shown to be:

\[
C(u_1, u_2, u_3) = \psi(-\sum_{1 \leq j < k \leq 3} \log R_k(e^{-\psi_j^{-1}(u_j)}, e^{-\psi_k^{-1}(u_k)}) - \sum_{j=1}^3 \alpha_j \psi_j^{-1}(u_j)).
\]
\[ \phi \text{ introduces a global pairwise association, and on top of that,} \]
the bivariate copula \( R_{uv} \) provides for additional dependence for pairs \((U_j, U_k)\). Note that the mixture of max-id is a general class of copulas with flexible choices of the Laplace transformation and the bivariate copula. The selection of various specifications is out of the scope of this study. Instead, we are more interested in the accommodation of dependency implied by the claims data. As shown in the empirical analysis, the choice of the copula within the family is not a matter of importance as long as the specification allows for the flexible association structure.

Once the marginal and copula distributions are specified, the total log-likelihood function could be derived by summing up the contribution of each individuals. The parameters could then be estimated via standard maximum likelihood approach:

\[
\hat{\Lambda} = \arg \max_A \mathbb{L}(\Lambda) = \arg \max_A \sum_{i=1}^{n} \log f_i(n_{1i}, n_{2i}, n_{3i}| z_i)
\]

where \( f_i(z_i) \) is defined according to (6) with \( f_j(n_j|x_j) = \sum_{k=0}^{\infty} f_{\text{NB}-1}(k|x_j; \beta_j, \sigma_j^2) \) for \( j = 1, 2, 3 \), and \( \Lambda = (\beta_1, \sigma_1^2, \beta_2, \sigma_2^2, \theta, \Theta) \) denoting the vector of parameters in the marginals and the copula function. Note that the estimation of the copula regression model is performed via inference function of marginals (IFM) in most existing studies. The IFM is a two-step estimation approach that separates the inference for the parameters in the marginals and the copula function. We believe that increased efficiency could be gained by estimating all parameters simultaneously and this argument is supported by the improvement in the likelihood in the empirical analysis. In this application, one of our goals is to explore the appropriateness of using copulas for multivariate claim counts. We have briefly discussed the pros and cons of different families of Archimedean copulas and it is easy to see that the max-id family may offer the most flexibility in the dependence structure. In the inference section of this paper, we show that such flexibility is necessary and the (partially) symmetric association is not appropriate at least for our dataset in consideration.

### 4.2. Elliptical copulas

One ideal property of a copula in multivariate analysis is the ability of accommodating a wide range of dependence, including the flexibility of allowing for pair-wise association. The family of elliptical copulas is such an example, providing unstructured dependence and allowing for both positive and negative relations. Elliptical copulas are built from multivariate elliptical distributions. The trivariate copula has a distribution function of the form:

\[
F(u_1, u_2, u_3) = \int_{-\infty}^{h_{1}^{-1}(u_1)} \int_{-\infty}^{h_{2}^{-1}(u_2)} \int_{-\infty}^{h_{3}^{-1}(u_3)} h(z_1, z_2, z_3) dz_1 dz_2 dz_3.
\]

Let \( z = (z_1, z_2, z_3) \), then \( h(z) = \kappa_3 |\Sigma|^{1/2} g_3 \left( \frac{z}{\sqrt{\Sigma}} \right) \) denotes the density of a trivariate standard elliptical distribution with location \( \mathbf{0} \) and correlation \( \Sigma \). Here \( \kappa_3 \) is a normalizing constant and \( g_3(\cdot) \) is known as the density generator function (see Fang et al., 1990 for details on multivariate elliptical distributions). \( h_j \) is the cdf of the \( j \)th marginal, \( j = 1, 2, 3 \). For copula (14), the primary association is captured by the dispersion matrix \( \Sigma \), and the parameters in \( g_3(\cdot) \) may accommodate additional dependency. Apparently, the matrix \( \Sigma \) serves as a natural device to accommodate flexible pair-wise association among multivariate claim counts.

As already foreshadowed in Section 4.1, copula (14) may not be computationally tractable for multivariate discrete outcomes because the evaluation of likelihood (6) involves repeated multi-dimensional numerical integration. On the other hand, we notice that the trivariate elliptical copula has a closed-form density function

\[
c(u_1, u_2, u_3) = h(H_1^{-1}(u_1), H_2^{-1}(u_2), H_3^{-1}(u_3))
\]

\[
\times \prod_{j=1}^3 \frac{1}{h_j(H_j^{-1}(u_j))}
\]

where \( h_j \) is the density function associated with \( H_j \). For example, the commonly used Gaussian and t copulas have respective densities as expressed below:

Gaussian copula: \( c(u_1, u_2, u_3) = \frac{1}{\sqrt{|\Sigma|}} \exp \left( -\frac{1}{2} \theta (\Sigma^{-1} - \mathbf{1}) \right) \)

\[
t \text{ copula: } c(u_1, u_2, u_3) = \frac{1}{\sqrt{|\Sigma|}} \Gamma((\tau + 3)/2) \Gamma((\tau + 1)/2) \prod_{j=1}^3 \left( 1 + \sigma_j^2/\tau \right)^{-(\tau + 3)/2} \]

where \( \theta = (\theta_1, \theta_2, \theta_3) \) with \( \theta_j = \Phi^{-1}(u_j) \) for the Gaussian copula and \( \theta_j = \tau_j^{-1}(u_j) \) for the t copula. \( \Phi(\cdot) \) denotes the cdf of standard normal distribution and \( \tau_j(\cdot) \) denotes the cdf of standardized Student’s t distribution with \( \tau \) degrees of freedom.

To circumvent the problem of using (14) directly, we adopt a “jitter” approach for the multivariate claim counts. The key idea is to discretize the variables and then apply (15) to the jittered continuous variables. The jitter approach was inspired by the pioneer work of Denuit and Lambert (2005), where it has been demonstrated that the concordance-based association measures, such as Kendall’s \( \tau \), are preserved when jittering discrete data. The preservation of association also allows us to naturally interpret the dependence parameter in the copula. Based on this property, Madsen and Fang (2011) proposed using Gaussian copulas to model discrete longitudinal data and illustrated the method with binary outcomes. Shi and Valdez (2012) employed similar techniques for longitudinal count data.

In this application, the jitter approach is customized for modeling multivariate insurance claim count. Specifically, for individual \( i \), define jittered counts \( \tilde{N}_i = N_i - U_i \) for \( j = 1, 2, 3 \). Here \( U_j \) for \( i = 1, \ldots, n \) and \( j = 1, 2, 3 \) are independent uniform (0,1) variables. Then elliptical copulas could be used to model the continuous vector \( \tilde{N}_i = (\tilde{N}_{i1}, \tilde{N}_{i2}, \tilde{N}_{i3}) \). The joint distribution of \( \tilde{N}_i \) has the following copula representation:

\[
\tilde{F}(\tilde{N}_{i1}, \tilde{N}_{i2}, \tilde{N}_{i3}| x_i; \Theta) = C(\tilde{F}_1(\tilde{N}_{i1}| x_i), \tilde{F}_2(\tilde{N}_{i2}| x_i), \tilde{F}_3(\tilde{N}_{i3}| x_i)| \Theta).
\]

where \( C(\cdot; \Theta) \) is a conditional elliptical copula of form (14).

Taking derivatives of (16) gives the joint density function:

\[
\tilde{f}(\tilde{N}_{i1}, \tilde{N}_{i2}, \tilde{N}_{i3}| x_i) = c(\tilde{F}_1(\tilde{N}_{i1}| x_i), \tilde{F}_2(\tilde{N}_{i2}| x_i), \tilde{F}_3(\tilde{N}_{i3}| x_i)| \Theta) \prod_{j=1}^3 \tilde{f}_j(\tilde{N}_{ij}| x_i).
\]

In the above, where \( \tilde{F}_j(\cdot|x_i) \) and \( \tilde{f}_j(\cdot|x_i) \) indicate the cdf and density function of the jittered continuous random variable \( \tilde{N}_{ij} \), respectively. From the relation between \( N_i \) and \( \tilde{N}_i \), one could derive the distribution of \( \tilde{N}_j \) for \( j = 1, 2, 3 \) as:

\[
\tilde{F}_j(\tilde{N}_j|x_i) = F_j((\tilde{N}_j + 1)| x_i) - F_j((\tilde{N}_j + 1)| x_i) + \tilde{N}_j(\tilde{N}_j + 1)| x_i)
\]

and

\[
\tilde{f}_j(\tilde{N}_j|x_i) = f_j(\tilde{N}_j + 1)| x_i).
\]

provided that the distributions of original count variables \( F_j(\cdot|x_i) \) or \( f_j(\cdot|x_i) \) are specified, be it either a NB-I or NB-II distribution. It
is straightforward to see that the original claim count \( N_i \) could be retrieved from \( N_i = [N_{ij} + 1] \), and the parameters in \( \tilde{f}_i(\cdot | x_i) \) and \( \tilde{F}_i(\cdot | x_i) \) are retained in \( \tilde{f}_i(\cdot | x_i) \) and \( \tilde{F}_i(\cdot | x_i) \) respectively. Hence no information is lost in the jittering process.

The likelihood-based method could be used to estimate the model. One needs to be careful that (17) is not the contribution to the likelihood function of the \( i \)th individual, because the jittered counts \( \tilde{N}_i \) are not observable. Instead, the \( i \)th individual’s contribution is the joint probability mass function of original count vector \( N_i = (N_{1i}, N_{2i}, N_{3i}) \), which could be derived by averaging over the jitters \( U_{ij} = (U_{1ij}, U_{2ij}, U_{3ij}) \):

\[
\tilde{f}_i(n_{1i}, n_{2i}, n_{3i} | x_i) = E_{U_i}(\tilde{f}_1(n_{1i} - U_{1ij} | x_{1j}), \tilde{f}_2(n_{2i} - U_{2ij} | x_{2j}), \\
\tilde{F}_1(n_{1i} - U_{1ij} | x_{1j}) \cdot \prod_{j=1}^3 \tilde{f}_j(n_{ij} - U_{ijj} | x_{ij})).
\]

The parameters in the marginals and the copula are simultaneously estimated by maximizing the total log-likelihood function in the similar manner as (13). Monte Carlo simulation is required for the evaluation of the likelihood function; we refer to Madsen and Fang (2011) and Shi and Valdez (2012) for the details. As shown in Denuit and Lambert (2005), the concordance-based association measures for \( (N_{1i}, N_{2i}, N_{3i}) \) are the same as \( (N_{1i}, N_{2i}, N_{3i}) \). Thus one can interpret the dependence implied by the copula as that of the insurance claim counts. Furthermore, in the Appendix we show that in general, the likelihood in (18) is equal to the joint probability mass function. This was proved in Madsen and Fang (2011) but only for the case of the Gaussian copula. It is worth stressing that the likelihood in (18) is that of the original observations rather than that of the “jittered” data. It is also straightforward to see that regularity conditions on the model are satisfied such that the usual asymptotic properties apply to the MLEs.

5. Estimation results

We fit the multivariate negative binomial models discussed in Sections 3 and 4 to the automobile insurance claim count empirical data described in Section 2. The main estimation results are summarized in this section. The common shock approach relies on the NB-I distribution and the corresponding estimates are displayed in Table 3. Because of the restriction in the dependence structure, it is not difficult to deduce the poor performance of model (1). Thus we only report the estimation for model (4). The copula models are estimated for both NB-I and NB-II distributions, and the estimates are exhibited in Tables 4 and 5, respectively, for the max-id and elliptical copulas.

We compare the results of the different methods along three perspectives. First, we examine the model fit indicated by the value of the log-likelihood function and the goodness-of-fit statistics including AIC and BIC reported at the bottom of each table. All these performance measures suggest the superiority of the copula approach to the common shock approach, presumably due to the limitation of its linear relationship. In Table 4, the model is estimated with the bivariate Gumbel copula and the associated Laplace transformation. The estimates in Table 5 correspond to the model with the Gaussian copula. Though not reported here, we also fit the model with different mixtures of max-id copula in Joe and Hu (1996) and the t copula in the elliptical family. Our analysis showed that the effect of choosing copulas is non-substantive and immaterial so long as the copula is capable of capturing the underlying dependency. We also notice the closeness of the NB-I and NB-II models for various copulas, which is consistent with the marginal analysis performed in Section 2.

Secondly, we turn to parameter estimates. In general, the regression coefficients are quite consistent within the copula approach. As anticipated, age and driving history are good indicators of the risk level of the policyholder. Young drivers and those with lower NCD scores tend to make more claims irrespective of the type. The vehicle age shows a nonlinear effect though it is not significant for the type of property damage. The usage of vehicle is also a significant predictor. Not very intuitively, private cars are more likely to have claims of third party bodily injury and property damage, but less likely to have claims of own damage. Small cars are associated with fewer claims. This could be explained by the deductible effect or the moral hazard effect on the lower severity associated with compact vehicles. On the other hand, we observe some slight differences between the estimates of the common shock approach and the copula approach. For example, the effect of age is not significant for the third party property damage and the effect of engine size is not significant for the third party bodily injury.

Finally, we look into the dependency among claim types. Ideally we hope to have a universal association measure that informs the relation among the three types of claim counts, so that we can make a direct comparison across different models. Unfortunately, the comparison between the common shock and copula approaches is not straightforward. The common shock model only captures the linear relationship and the correlation is formulated conditional on the covariates. In contrast, the copula model accommodates both linear and nonlinear dependencies, and the association is interpreted as the relation with the effects of covariates purged off.

Despite the above differences, the dependency estimates to some extent are consistent. All methods suggest that the associations between third party bodily injury and own damage, and between third party bodily injury and property damage are comparable, and are greater than the association between own damage and third party property damage. In the common shock model, the correlation is measured by \( \lambda_{13} \) in Table 3, where \( \lambda_{12} \) and \( \lambda_{13} \) are close to \( \exp(-3.5) \) and are greater than \( \lambda_{23} \). In the max-id copula, \( \theta \) measures the global association among all types, and due to its statistical insignificance, \( \theta_{23} \) is set to zero to reflect the weaker relation between the two types. In doing so, we specify \( \omega_2 = \omega_3 = -1 \) and \( \omega_1 = 0 \). For example, in the NB-II formulation in Table 4, the association between type II and type III claims is captured only by the global dependence parameter \( \theta = 1.42 \). On
Estimates of the multivariate NB models using the Max-ID copula.

<table>
<thead>
<tr>
<th>Copula</th>
<th>Body injury</th>
<th>Own damage</th>
<th>Property damage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate t-ratio</td>
<td>Estimate t-ratio</td>
<td>Estimate t-ratio</td>
</tr>
<tr>
<td>NB-I</td>
<td>Intercept</td>
<td>−2.293</td>
<td>−14.53</td>
</tr>
<tr>
<td></td>
<td>Young</td>
<td>0.281</td>
<td>2.691</td>
</tr>
<tr>
<td></td>
<td>Lowncd</td>
<td>0.451</td>
<td>4.681</td>
</tr>
<tr>
<td></td>
<td>Vage</td>
<td>−0.071</td>
<td>−2.14</td>
</tr>
<tr>
<td></td>
<td>Vage × vage</td>
<td>0.002</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>−0.234</td>
<td>−1.58</td>
</tr>
<tr>
<td></td>
<td>Vlux</td>
<td>0.017</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Smallcap</td>
<td>−0.442</td>
<td>−2.54</td>
</tr>
</tbody>
</table>

Goodness-of-fit:

<table>
<thead>
<tr>
<th>Loglikelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max-ID-NB1</td>
<td>4287.56</td>
<td>8635.12</td>
</tr>
</tbody>
</table>

Table 5

Estimates of the multivariate NB models using the elliptical copula.

<table>
<thead>
<tr>
<th>Copula</th>
<th>Body injury</th>
<th>Own damage</th>
<th>Property damage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate t-ratio</td>
<td>Estimate t-ratio</td>
<td>Estimate t-ratio</td>
</tr>
<tr>
<td>NB-I</td>
<td>Intercept</td>
<td>−2.404</td>
<td>−12.71</td>
</tr>
<tr>
<td></td>
<td>Young</td>
<td>0.291</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>Lowncd</td>
<td>0.478</td>
<td>4.70</td>
</tr>
<tr>
<td></td>
<td>Vage</td>
<td>−0.072</td>
<td>−2.04</td>
</tr>
<tr>
<td></td>
<td>Vage × vage</td>
<td>0.003</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>−0.190</td>
<td>−1.16</td>
</tr>
<tr>
<td></td>
<td>Vlux</td>
<td>0.017</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Smallcap</td>
<td>−0.465</td>
<td>−2.49</td>
</tr>
</tbody>
</table>

Goodness-of-fit:

<table>
<thead>
<tr>
<th>Loglikelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elliptical-NB1</td>
<td>4293.58</td>
<td>8647.16</td>
</tr>
</tbody>
</table>

Table 6

Estimates of association parameters for selected formulations of the mixture. Recall that for both Gumbel and Joe copulas, the similarity in the corresponding dependency structure. To show the robustness of the dependency estimation, we also report in Table 6 the association parameters for selected formulations of the mixture. Recall that for both Gumbel and Joe copulas, the boundary value of one corresponds to the independence case. Thus the significant association is indicated by a parameter greater than one. A similar pattern is also observed for the Gaussian copula, where the pairwise association is described by the dispersion matrix $\rho_v$ in Table 5. In the t copula model, results of which are not reported to avoid overwhelming the reader, the estimate of the degree of freedom $\tau$ is 351.78, with a standard error of 76.09, indicating the adequacy of the Gaussian copula.

We conclude the analysis with the Vuong test to examine the closeness of the alternative multivariate negative binomial models (see Vuong, 1989 for more details on the Vuong test). The test statistics and the corresponding $p$-values are presented in Table 7. A larger statistic suggests that two models are further apart. As anticipated, one observes that the copula-based models significantly outperform the common shock model. There is no substantial difference between the copula models using NB-I and NB-II distributions, which has already been foreshadowed by the goodness-of-fit statistics reported in Table 2.
displayed in Tables 4 and 5. When comparing the product copula with other copulas, it is noticeable that the independence assumption does not provide strong bias on the mean estimates. This is not at all surprising because it is well-known that the copula model preserves the marginals. However, the independence assumption leads to a lower variance estimates when compared with other copulas, which is explained by the positive dependence among claim types.

The expectation of total claims is the building block for the ratemaking process. For example, percentile premiums could be derived from the distribution of $E(L)$ (see, for example, Bermúdez and Karlis, 2011). Though not using a Bayesian approach, the parameter uncertainty could be incorporated via Monte Carlo simulation. Presumably our sample size is large enough so that the asymptotic property of MLEs applies. Let $\hat{\theta}$ denote the vector of parameter estimates of the copula model and $\hat{\Sigma}$ denote its asymptotic variance. We simulate parameter vector $\tilde{\theta}$ from the multivariate normal distribution with mean $\hat{\theta}$ and covariance $\hat{\Sigma}$, and then we calculate the expected claims $E(L) = \mu_1(\tilde{\theta}(\chi)) + \mu_2(\tilde{\theta}(\chi_2)) + \mu_3(\tilde{\theta}(\chi_3))$. By repeating the procedure a large number of times, one has the distribution of $E(L)$.

Fig. 1 displays the distribution of $E(L_i)$ for the five risk classes under different combinations of copulas and marginal distributions. Apparently, the risk level of each class is fairly reflected in the distributions of the expected claim. Except for the third one, a riskier class is generally associated with higher volatility in the expected claims. For illustration purposes, Table 11 presents the two commonly used percentile premiums based on the distribution in Fig. 1: the gross premium based on the 75th percentile and the risk-based premium based on the 95th percentile. In line with previous analysis, Fig. 1 and Table 11 show that the two copula approaches have similar implications on the ratemaking process.

7. Concluding remarks

In this article, we considered alternative approaches to construct multivariate count regression models based on the negative binomial for the marginal distributions. The work was motivated by the characteristics of a dataset of multivariate claim counts from an automobile insurer in Singapore. We showed that both NB-I and NB-II are superior among various count regression models in capturing the overdispersion and excess of zeros in the count data. Trivariate negative binomial models were proposed in order to accommodate the dependency among three types of claims: the third party bodily injury, own damage, and third-party property damage.

We have emphasized the flexibility of the copula approach in the modeling of dispersion and dependence structure. Specifically, both formulations, NB-I and NB-II, of negative binomial regression models are permitted, and both positive and negative associations could be captured by the copula models. In contrast, the class of multivariate negative binomial models based on common shock variables relies on the additivity of the NB-I distribution. Thus it requires a common dispersion for all marginals and only allows for positive linear correlation among marginals. As expected, the superiority of the copula approach was supported by the better model fit in the calibration of our empirical data.

From the perspective of dependency modeling, we studied the mixture of max-id copulas and the family of elliptical copulas. Both families in some sense have the flexibility of allowing for unstructured pair-wise associations. The max-id copulas have closed-form cdf which facilitates the likelihood evaluation. For elliptical copulas, a jitter approach was adopted to avoid the multi-dimensional integration in the cdf. In the empirical analysis, we
Table 9
Joint distribution of claim frequency for hypothetical individuals.

<table>
<thead>
<tr>
<th></th>
<th>Excellent</th>
<th>Very good</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MVNB</td>
<td>MaxID</td>
<td>Ellipt</td>
<td>Product</td>
<td>MVNB</td>
</tr>
<tr>
<td>ooo</td>
<td>0.9031</td>
<td>0.9469</td>
<td>0.9472</td>
<td>0.9247</td>
<td>0.8452</td>
</tr>
<tr>
<td>oo+</td>
<td>0.0015</td>
<td>0.0010</td>
<td>0.0018</td>
<td>0.0062</td>
<td>0.0069</td>
</tr>
<tr>
<td>o+o</td>
<td>0.0234</td>
<td>0.0232</td>
<td>0.0237</td>
<td>0.0406</td>
<td>0.0376</td>
</tr>
<tr>
<td>+oo</td>
<td>0.0121</td>
<td>0.0092</td>
<td>0.0105</td>
<td>0.0268</td>
<td>0.0478</td>
</tr>
<tr>
<td>o++</td>
<td>0.0029</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.0031</td>
</tr>
<tr>
<td>+o+</td>
<td>0.0227</td>
<td>0.0018</td>
<td>0.0019</td>
<td>0.0002</td>
<td>0.0226</td>
</tr>
<tr>
<td>++o</td>
<td>0.0327</td>
<td>0.0123</td>
<td>0.0123</td>
<td>0.0012</td>
<td>0.0343</td>
</tr>
<tr>
<td>+++</td>
<td>0.0017</td>
<td>0.0052</td>
<td>0.0025</td>
<td>0.0000</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

Table 10
Mean and variance of total claims by risk class.

<table>
<thead>
<tr>
<th></th>
<th>Excellent</th>
<th>Very good</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E(L_i)</td>
<td>Var(L_i)</td>
<td>E(L_i)</td>
<td>Var(L_i)</td>
<td>E(L_i)</td>
</tr>
<tr>
<td>MVNB</td>
<td>0.102</td>
<td>0.165</td>
<td>0.169</td>
<td>0.232</td>
<td>0.205</td>
</tr>
<tr>
<td>MaxID-I</td>
<td>0.087</td>
<td>0.178</td>
<td>0.181</td>
<td>0.394</td>
<td>0.273</td>
</tr>
<tr>
<td>MaxID-II</td>
<td>0.081</td>
<td>0.151</td>
<td>0.166</td>
<td>0.349</td>
<td>0.277</td>
</tr>
<tr>
<td>Ellip-I</td>
<td>0.078</td>
<td>0.136</td>
<td>0.154</td>
<td>0.294</td>
<td>0.263</td>
</tr>
<tr>
<td>Ellip-II</td>
<td>0.073</td>
<td>0.122</td>
<td>0.145</td>
<td>0.271</td>
<td>0.267</td>
</tr>
<tr>
<td>Indep-I</td>
<td>0.081</td>
<td>0.085</td>
<td>0.131</td>
<td>0.137</td>
<td>0.280</td>
</tr>
<tr>
<td>Indep-II</td>
<td>0.079</td>
<td>0.080</td>
<td>0.126</td>
<td>0.129</td>
<td>0.280</td>
</tr>
</tbody>
</table>

Table 11
Quantile premiums by risk class.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Excellent</th>
<th>Very good</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MaxID</td>
<td>Ellipt</td>
<td>MaxID</td>
<td>Ellipt</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NB-I</td>
<td>NB-II</td>
<td>NB-I</td>
<td>NB-II</td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td>0.102</td>
<td>0.094</td>
<td>0.094</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>95th</td>
<td>0.124</td>
<td>0.117</td>
<td>0.117</td>
<td>0.111</td>
<td></td>
</tr>
<tr>
<td>Very good</td>
<td>0.215</td>
<td>0.200</td>
<td>0.188</td>
<td>0.178</td>
<td></td>
</tr>
<tr>
<td>95th</td>
<td>0.266</td>
<td>0.251</td>
<td>0.239</td>
<td>0.228</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>0.289</td>
<td>0.292</td>
<td>0.282</td>
<td>0.285</td>
<td></td>
</tr>
<tr>
<td>95th</td>
<td>0.314</td>
<td>0.314</td>
<td>0.303</td>
<td>0.307</td>
<td></td>
</tr>
<tr>
<td>Fair</td>
<td>0.531</td>
<td>0.515</td>
<td>0.488</td>
<td>0.478</td>
<td></td>
</tr>
<tr>
<td>95th</td>
<td>0.639</td>
<td>0.640</td>
<td>0.600</td>
<td>0.596</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>0.711</td>
<td>0.726</td>
<td>0.682</td>
<td>0.694</td>
<td></td>
</tr>
<tr>
<td>95th</td>
<td>0.813</td>
<td>0.847</td>
<td>0.799</td>
<td>0.812</td>
<td></td>
</tr>
</tbody>
</table>

showed that the performance of the two families are comparable in terms of model fit and dependence measure.

Another advantage of the copula approach is the freedom in the choice of marginals. Although we have used negative binomial distributions for all types of claim count, it is worth stressing that distinct regressions are allowed for the three marginals in the copula models. For prediction purposes, such flexibility could be useful when, for example, overdispersion is important for some marginals but not for others, or excess of zeros is critical for some marginals but not for others.

Acknowledgments

The authors thank the Editor and the reviewers for their comments which helped improve the quality of the final version of this paper. Emiliano also wishes to extend his appreciation for the participants of the Non-Gaussian Multivariate Statistical Models workshop held at the Banff International Research Station in May 2013, for their additional comments and encouragement.

Appendix

Consider a d-dimensional count random vector \((N_1, N_2, \ldots, N_d)\) with a corresponding copula \(C\) so that

\[
\Pr(N_1 \leq n_1, N_2 \leq n_2, \ldots, N_d \leq n_d) = C(F_1(n_1), F_2(n_2), \ldots, F_d(n_d))
\]
where $f_j$ are the corresponding marginals of $N_j$ for $j = 1, 2, \ldots, d$. Equivalently, we know that

$$\Pr(N_1 = n_1, N_2 = n_2, \ldots, N_d = n_d)$$

$$= \sum_{j_1=0}^{1} \sum_{j_2=0}^{1} \ldots \sum_{j_d=0}^{1} (-1)^{j_1+j_2+\cdots+j_d} C(u_1, u_2, \ldots, u_d)$$

where $u_j = f_j(n_j)$ for $j = 1, 2, \ldots, d$. Now, define the jitters $N_j = N_j - U_j$, for $j = 1, 2, \ldots, d$, where $U_j$ is uniform on $(0, 1)$. Conditional on $U_j$, the random vector $(\tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_d)$, with $\tilde{Z}_j = n_j - U_j$ for $j = 1, 2, \ldots, d$, has the joint distribution function expressed by

$$\Pr(\tilde{Z}_1 \leq z_1, \tilde{Z}_2 \leq z_2, \ldots, \tilde{Z}_d \leq z_d)$$

$$= \Pr(F_1(\tilde{Z}_1) \leq F_1(z_1), F_2(\tilde{Z}_2) \leq F_2(z_2), \ldots, F_d(\tilde{Z}_d) \leq F_d(z_d))$$

$$= C(F_1(z_1), \ldots, F_d(z_d)),$$

so that its joint density is

$$f_{\tilde{Z}}(z_1, z_2, \ldots, z_d) = C(F_1(z_1), F_2(z_2), \ldots, F_d(z_d))$$

$$\times \prod_{j=1}^{d} f_j(z_j).$$

Averaging the density over the jitters $(U_1, U_2, \ldots, U_d)$, we get

$$E_U(f_{\tilde{Z}}(z_1, z_2, \ldots, z_d))$$

$$= E_U \left( C(F_1(z_1), F_2(z_2), \ldots, F_d(z_d)) \times \prod_{j=1}^{d} f_j(z_j) \right)$$

$$= E_U \left( C(\tilde{F}_1(n_1 - U_1), \tilde{F}_2(n_2 - U_2), \ldots, \tilde{F}_d(n_d - U_d)) \right)$$

$$\times \prod_{j=1}^{d} f_j(n_j - U_j)$$

which is equivalent to the joint probability mass function

$$\Pr(N_1 = n_1, N_2 = n_2, \ldots, N_d = n_d)$$

$$= E_U \left( f_{\tilde{Z}}(n_1 - U_1, n_2 - U_2, \ldots, n_d - U_d) \right)$$

$$= E_U \left( f_{\tilde{Z}}(z_1, z_2, \ldots, z_d) \right)$$

because for $j = 1, 2, \ldots, d$, $\tilde{N}_j = N_j - U_j$ and $\tilde{Z}_j = n_j - U_j$, conditional on $U_j$.  

**Fig. 1.** Predictive distribution of $E(L)$ under different models.
References