



Copula-based bivariate finite mixture regression models with an application for insurance claim count data

Lluís Bermúdez¹  · Dimitris Karlis²

Received: 23 July 2021 / Accepted: 8 April 2022 / Published online: 4 May 2022

© The Author(s) 2022

Abstract

Modeling bivariate (or multivariate) count data has received increased interest in recent years. The aim is to model the number of different but correlated counts taking into account covariate information. Bivariate Poisson regression models based on the shock model approach are widely used because of their simple form and interpretation. However, these models do not allow for overdispersion or negative correlation, and thus, other models have been proposed in the literature to avoid these limitations. The present paper proposes copula-based bivariate finite mixture of regression models. These models offer some advantages since they have all the benefits of a finite mixture, allowing for unobserved heterogeneity and clustering effects, while the copula-based derivation can produce more flexible structures, including negative correlations and regressors. In this paper, the new approach is defined, estimation through an EM algorithm is presented, and then different models are applied to a Spanish insurance claim count database.

Keywords Zero-inflation · Overdispersion · EM algorithm · Automobile insurance · Frank copula

Mathematics Subject Classification 62H05 · 62H12 · 62J05 · 62P05

✉ Lluís Bermúdez
lbermudez@ub.edu

Dimitris Karlis
karlis@aueb.gr

¹ Departament de Matemàtica Econòmica, Financera i Actuarial, RISKcenter-IREA, Universitat de Barcelona (UB), Av. Diagonal, 690, 08034 Barcelona, Spain

² Department of Statistics, Athens University of Economics and Business, 76 Patission str., Athens, Greece

1 Introduction

Bivariate count data regression models are appropriate in situations where paired counts are correlated and joint estimation is required. In this paper, in the application section, an insurance claim count dataset is used to model the number of claims for two different types of claim.

The starting point in the literature for modeling bivariate count data is the bivariate Poisson regression model. We denote this bivariate Poisson distribution as $BP(\lambda_1, \lambda_2, \lambda_3)$, where $\lambda_1 + \lambda_3$ and $\lambda_2 + \lambda_3$ are the marginal means and λ_3 is the covariance parameter. Based on this distribution, there are many applications in economics and related fields. However, this distribution has two limitations: It does not allow for overdispersion (marginal distributions follow a Poisson distribution and thus the conditional mean of each count variable equals the conditional variance) or a negative correlation (the model only accommodates for nonnegative correlation between the counts).

As in the univariate case, when using mixed Poisson models, bivariate count models can be generalized in different ways to allow for overdispersion. Mixtures of bivariate Poisson distribution can be considered in at least two different ways. In the first, we start with a $BP(a\lambda_1, a\lambda_2, a\lambda_3)$ distribution where a follows some distribution. Munkin and Trivedi (1999) described multivariate mixed Poisson regression models based on this type of mixing and a gamma mixing distribution. Gurmu and Elder (2000) used an extended gamma density as a mixing distribution. Ghitany et al. (2012) proposed a general EM algorithm to facilitate estimation for the multivariate negative binomial, Poisson inverse Gaussian and Poisson lognormal regression models. The second way starts with a $BP(a_1\lambda_1, a_2\lambda_2, a_3\lambda_3)$ distribution, but now assuming that the vector $(a_1, a_2, a_3)^T$ follows a trivariate (or bivariate if we assume that $\lambda_3 = 0$) distribution. The case when $\lambda_3 = 0$ has received attention primarily because of its simplicity. Steyn (1976) proposed the use of a bivariate normal distribution as the mixing distribution. Later on, Aitchison and Ho (1989) proposed using the bivariate lognormal distribution instead of the simple bivariate normal distribution.

In both of the above models, the specification of the random-effects distribution is continuous although it may be a discrete or a finite distribution. Karlis and Meligkotsidou (2007) considered the latter case assuming that the joint distribution for the random effects is a finite distribution, i.e., the case in which only a finite number of points have positive probabilities. Such an assumption gives rise to a finite mixture of multivariate Poisson distributions. In Bermúdez and Karlis (2012), a finite mixture of bivariate Poisson regression models was proposed to allow for overdispersion. The novelty of this approach lies in the fact that it is assumed that the data consist of subpopulations with different regression structures. A potential use of such a model is for examining the clustering of observations. The model can also be used to model negative correlation and embraces zero-inflated bivariate Poisson regression models.

In parallel, there are other models that are designed to allow for negative correlations. A bivariate Poisson distribution that allows for negative, zero or positive correlations was introduced by Lakshminarayana et al. (1999) with a distribution defined as the product of Poisson marginals with a multiplicative factor parameter. Extension to other marginal are also possible (Famoye 2010; Zamani et al. 2016).

Berkhout and Plug (2004) derived an alternative count data model using conditional probabilities. More recently, Genest et al. (2018) proposed an alternative model extending the trivariate reduction to achieve negative correlation.

A different approach involves modeling dependence through copula functions to allow for greater flexibility for the dependence structure. As an example, Cameron et al. (2004) defined bivariate count distributions from copula functions. Shi and Valdez (2014) proposed different multivariate negative binomial regression models using copulas to model insurance claim counts.

Finally, it is worth mentioning that count data often display more zero outcomes than would be expected with the Poisson regression model. Indeed, the zero-inflation phenomenon is a very specific type of overdispersion, and zero-inflated Poisson regression models have been used to handle zero-inflated count data (Li et al. 1999; Wang et al. 2003).

The present paper proposes copula-based bivariate finite mixture regression models, extending the knowledge gained from the papers above, particularly from Bermúdez and Karlis (2012) and Shi and Valdez (2014). First, using a finite mixture, the proposed model allows for overdispersion (embracing zero-inflation) and, simultaneously, for an elegant interpretation based on the typical clustering application of finite mixture models. Second, by using copula functions, it allows for a more flexible dependence structure, including negative correlations.

Finite mixtures with copulas were introduced in Kosmidis and Karlis (2016), and the current paper follows their main ideas. Here, for each cluster, a bivariate distribution defined through copulas is used, extending the previous reference by allowing for regressors as well. In particular, a bivariate distribution is used for each cluster, with marginals defined as having any discrete distribution where their means are related to covariates while the two marginals are joined via some copula function.

This new approach, using copulas, offers a series of improvements such as allowing a bivariate regression to be used for each cluster that can also have a negative correlation, allowing for a wide range of marginal distributions that means not necessarily restricting ourselves to simple bivariate Poisson distributions, and allowing for different dependence structures via the selection of copulas. In parallel, using a finite mixture, this approach allows for cluster analysis, and hence, for a better interpretation of count data analysis.

We describe our approach in Sect. 2 and provide an estimation with an EM algorithm in Sect. 3. In order to show the aforementioned improvements, a real data application is given in Sect. 4. Finally, Sect. 5 presents our concluding remarks.

2 The proposed model

We base our model on copulas. Copulas are quite common for continuous data but not widely used for discrete data (see, e.g., Nikoloulopoulos 2013, for a discussion of the challenges in applying copulas to multivariate counts). In the present paper, we follow the approach in Kosmidis and Karlis (2016) where the multivariate distribution in the finite mixture is constructed via copulas. As a novelty, we assume a finite mixture of such distributions but also allowing for covariates in order to consider heterogeneity.

Our approach generalizes the work in Bermúdez and Karlis (2012) where mixtures of bivariate Poisson regressions were assumed by considering a less flexible dependence structure. In order to expand the applicability of this model, our new approach can allow for different marginal distributions for each type of claim and different dependence structures via the selection of copulas.

2.1 Copulas

A copula is a function $C(\cdot)$ that can “couple” the marginals to model the dependence structure. A p -dimensional copula $C(u_1, \dots, u_p)$ is a distribution function with uniform marginals. The importance of copulas in statistical modeling stems from Sklar’s theorem (see Nelsen 1997, §2.3), which shows that every multivariate distribution can be represented via the choice of an appropriate copula and, more importantly, that a copula provides a general mechanism for constructing new multivariate models in a straightforward manner. Note that for continuous distributions the copula is unique, but for the discrete case the copula is not unique, which is generally not a problem in applications (see the discussion in Genest and Nešlehová (2007)).

For instance, consider the probability mass function $P(x; \mu_k)$, where μ_k denotes the parameter(s) of each distribution, for $k = 1, \dots, p$. The cumulative distribution function (cdf) is given by

$$F(x; \mu_k) = \sum_{m=0}^x P(m; \mu_k), \quad k = 1, \dots, p.$$

We can then couple the marginals to add a dependence structure. For example, we can have a bivariate distribution with marginals $F(x_1; \mu_1)$ and $F(x_2; \mu_2)$, given by

$$F(x_1, x_2; \mu_1, \mu_2, \alpha) \equiv C(F(x_1; \mu_1), F(x_2; \mu_2); \alpha), \tag{1}$$

where α now denotes the copula parameter(s) that explain(s) the association between the two random variables. This is a well-defined distribution function with a dependence structure depending on the chosen copula. However, to describe the joint probability mass function, we need to calculate finite differences from the cdf shown above. For the general case, when working with copulas for discrete data, the probability mass function is given in expression (1.2) in Panagiotelis et al. (2012), and results from finite differences in the distribution function. In its general form, it is written as

$$P(\mathbf{x}) = \sum_{\mathbf{m}} \text{sgn}(\mathbf{m}) C(F_1(m_1), \dots, F_p(m_p)), \tag{2}$$

with $\mathbf{m} = (m_1, \dots, m_p)$ vertices, where each m_t is equal to either x_t or $x_t - 1$ ($t = 1, \dots, p$), and

$$\text{sgn}(\mathbf{m}) = \begin{cases} 1, & \text{if } m_t = x_t - 1 \text{ for an even number of } t\text{'s} \\ -1, & \text{if } m_t = x_t - 1 \text{ for an odd number of } t\text{'s} \end{cases}.$$

For example, for the bivariate case we have

$$\begin{aligned}
 P(X_1 = x_1, X_2 = x_2; \mu_1, \mu_2, \alpha) &= F(x_1, x_2; \mu_1, \mu_2, \alpha) \\
 &\quad - F(x_1 - 1, x_2; \mu_1, \mu_2, \alpha) \\
 &\quad - F(x_1, x_2 - 1; \mu_1, \mu_2, \alpha) \\
 &\quad + F(x_1 - 1, x_2 - 1; \mu_1, \mu_2, \alpha) \quad (3)
 \end{aligned}$$

where μ_1, μ_2 refer to marginal parameters and α to the copula parameter(s). In practice, since copulas are distribution functions, this is a bivariate integral calculated in some rectangular bivariate space. Depending on the choice of copula, this can be simplified to avoid the need for integration.

2.2 The finite mixture model

Consider the general case with p variables. We want to model the vector $\mathbf{X} = (X_1, \dots, X_p)^T$. The probability mass function of a finite mixture model is defined as

$$P(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{j=1}^G \pi_j P_j(\mathbf{x}; \boldsymbol{\theta}_j), \quad (4)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_G^T)^T \in \Theta_1 \times \dots \times \Theta_G$, and $\pi_j \in (0, 1)$ with $\sum_{j=1}^G \pi_j = 1$. Appropriate choices of $P_j(\mathbf{x}; \boldsymbol{\theta}_j)$ can result in flexible and low complexity models. The book of McNicholas (2016) provide a detailed treatment of the framework of mixture modeling.

We emphasize that $\boldsymbol{\theta}_j$ is a vector containing all the parameters associated with the j -th component. Hence, it contains both the parameters from the marginals and the association parameters of the copula.

Our novel approach assumes that $P_j(\mathbf{x}; \boldsymbol{\theta}_j)$ is a bivariate distribution defined through a copula. The probability mass function is calculated using (3) with marginal distributions that are some discrete distribution with a mean λ_{ijk} for the i -th observation, j -th component and k -th variable, not necessarily from the same family. We assume that $i = 1, \dots, n$ (i.e. we have n observations), $j = 1, \dots, G$ (i.e., G components) and $k = 1, 2$ for the two variables at hand. In addition, we add covariates in the mean of the k -th marginal distribution and hence

$$\log \lambda_{ijk} = \beta_{jk}^T \boldsymbol{\omega}_{ik},$$

where β_{jk} is a vector of regression coefficients. Also note that the covariates in $\boldsymbol{\omega}_{ik}$ refer to the k -th variable and they maybe different for different λ 's (variables); for simplicity here we assume that they are the same and drop the second subscript. Note that the marginal distributions may depend on some other parameter; for example, in the case of a negative binomial marginal, we can have an overdispersion parameter s_{jk} for the j -th component of the k -th variable.

Note that one may also consider covariates on the mixing proportion vector $\mathbf{B} = (\mathbf{B}_1, \dots, \mathbf{B}_{G-1})$ based on some multinomial regression models. We do not pursue this on the present paper.

Hence, now we can further define the vector of component specific parameters as $\theta_j = (\beta_{j1}, \beta_{j2}, \alpha_j, \phi_{j1}, \phi_{j2})$ where $\beta_{jk}, k = 1, 2$, are the regression coefficient vectors for the j -th component, α_j is the copula parameter for the j -th component, and ϕ_j 's relate to other parameters associated with the j -th component and each and depend on the choice of the marginal distributions. They may differ if the marginals assumed are not the same.

Finally, we need to define the copula function to couple the marginal distributions. In the present paper, we use three different copulas but we emphasize that the literature contains several other choices that reflect different types of association between the marginals. First, we use a Frank copula with a cdf given by

$$C_\alpha(u, v) = -\frac{1}{\alpha} \left(1 + \frac{(\exp(-u\alpha) - 1)(\exp(-v\alpha) - 1)}{(\exp(-\alpha) - 1)} \right), \tag{5}$$

where $0 \leq u, v \leq 1, \alpha \in (-\infty, \infty) - \{0\}$. In fact, α is the copula parameter, and being negative implies negative correlation. Second, we also use a Gaussian copula defined as

$$C_\alpha(u, v) = \Phi_\alpha \left(\Phi^{-1}(u), \Phi^{-1}(v) \right),$$

where Φ is the $N(0,1)$ cdf, Φ^{-1} is the functional inverse of Φ and Φ_α is the bivariate standard normal cdf with correlation α . Finally, we use a Gumbel copula defined as

$$C_\alpha(u, v) = \left[(\tilde{u}^\alpha + \tilde{v}^\alpha)^{1/\alpha} \right], \tag{6}$$

with $\tilde{u} = -\log(u)$ and $\tilde{v} = -\log(v)$.

So, for example, for the bivariate Poisson based on the Frank copula we need to evaluate (3) using (5) and the cdf of Poisson distributions for F .

We emphasize that the choice of marginal distributions is a separate task from the dependence modeling and hence one may choose different marginals that offer considerable flexibility. Also the copulas need not be the same for each component of the mixture. For a thorough discussion about such a model, see the discussion in Kosmidis and Karlis (2016).

3 ML estimation via an EM algorithm

In this section, we describe ML estimation of the model. Being a finite mixture of regressions type model, estimation is based on an EM algorithm. The additional complexity here is the use of a copula function. As in Kosmidis and Karlis (2016) we can still apply the EM algorithm with limited additional complexity. The algorithm is as follows.

As usual we assume as unobserved data the vectors $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iG})$ with values $Z_{ij} = 1$ if the i -th observation belongs to the j -th component and 0 otherwise. The EM proceeds as usual, by estimating the unobserved component memberships at the E-step and maximizing the complete likelihood at the M-step. So the E-step is the typical step that estimate the component memberships. It turns out (see Kosmidis and Karlis 2016) that the M-step consists of G weighted likelihood steps. Formally, the algorithm is as follows.

Suppose that a sample of n p -vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ is available, which are assumed to be realizations of independent random variables X_1, \dots, X_n each distributed with a probability mass function as defined by (4) and (3). The maximization of the likelihood function based on that sample can be performed using the expectation–maximization algorithm. At the ℓ -th iteration of the algorithm ($\ell = 2, \dots$),

– *E-step*: Calculate

$$w_{ij}^{(\ell+1)} = \frac{\pi_j^{(\ell)} P_j(\mathbf{x}_i; \boldsymbol{\theta}_j^{(\ell)})}{\sum_{j=1}^G \pi_j^{(\ell)} P_j(\mathbf{x}_i; \boldsymbol{\theta}_j^{(\ell)})} \quad i = 1, \dots, n; j = 1, \dots, G.$$

– *M-step 1*: Set

$$\pi_j^{(\ell+1)} = \frac{\sum_{i=1}^n w_{ij}^{(\ell+1)}}{n} \quad j = 1, \dots, G.$$

– *M-step 2*: Maximize

$$l(\boldsymbol{\theta}; \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{j=1}^G \sum_{i=1}^n w_{ij}^{(\ell+1)} \log \{P_j(\mathbf{x}_i; \boldsymbol{\theta}_j)\},$$

with respect to $\boldsymbol{\theta}$ to get an updated value $\boldsymbol{\theta}^{(\ell+1)}$ for the copula and marginal parameters. Note that this step implies that we have to maximize G weighted likelihoods, one for each component, using w 's as weights. Hence the complexity and the difficulty of this step is equivalent to fitting a bivariate copula-based regression model for the chosen marginal distributions. Also since we maximize in parallel G weighted likelihoods, efficient evaluations by parallel computing are feasible.

In the above $P_j(\cdot)$ can be any bivariate discrete distribution suitable to model bivariate count data. In our case, we use bivariate Poisson and negative binomial distributions defined through copulas. Note that to allow for covariates we fit negative binomial distributions with mean λ and variance $\lambda + \lambda^2/s$, which implies that s is the overdispersion parameter. For $s \rightarrow \infty$, we get the Poisson model as special case, so in some sense the Poisson model is included. An interesting use of copulas is that the marginals may be of different type. For example, one may assume a Poisson distributions for the first marginal and a negative binomial for the second one.

The algorithm iterates between the E-step and the M-step until some convergence (or lack of progress) criterion is satisfied. For the example in the current paper the

terminating criterion that is used is that the relative increase of the log-likelihood in two successive iterations is less than a small positive real $\epsilon = 10^{-8}$.

Initial values can be set as described in Kosmidis and Karlis (2016) but also in more heuristic ways, such as by fitting models sequentially. See also Papastamoulis et al. (2016) for a relevant model.

4 Application

4.1 The data

We used the same automobile insurance claims dataset as used in Bermúdez and Karlis (2012). The data refer to a sample from an automobile portfolio of a major insurance company operating in Spain in which only cars categorized as being for private use were considered. For our purpose here, we selected policyholders with full coverage, i.e., policies including third-party liability (claimed and counted as N_1 type), a set of basic guarantees such as emergency roadside assistance or legal and medical assistance (claimed and counted as N_2 type) and, finally, comprehensive coverage (damage to one’s vehicle caused by any unknown party, for example, damage resulting from theft, flood or fire) and collision coverage (damage resulting from a collision with another vehicle or object when the policyholder is at fault), also claimed and counted as N_2 type. The cross-tabulation for the number of claims for third-party liability (N_1) and number of claims for the rest of the guarantees (N_2) is shown in Table 1.

The data contain information from 28,590 policyholders. For each policy, the initial information and the total number of claims (for the two types of claim) from policyholders at fault were reported on an annual basis. Nine exogenous variables were considered for ratemaking purposes. The exogenous variables are described in Table 2. For the vectors (N_{i1}, N_{i2}) , we assume the finite mixture model defined through copulas that include covariates.

Table 1 Cross-tabulation of data

N_1	N_2						
	0	1	2	3	4	5	6
0	24408	1916	296	69	12	6	0
1	1068	317	61	21	6	2	2
2	203	71	18	6	2	1	1
3	49	14	8	3	3	1	0
4	11	6	2	0	1	0	0
5	2	0	0	0	0	0	1
6	1	0	0	1	0	0	0
8	0	0	1	0	0	0	0

N_1 : number of claims for third-party liability. N_2 : number of claims for the rest of the guarantees

Table 2 Explanatory variables used in the models

Variable	Definition
GEN	Equals 1 for women and 0 for men
URB	Equals 1 when driving in urban area, 0 otherwise
ZON1	Equals 1 when zone is medium risk (Madrid and Catalonia)
ZON2	Equals 1 when zone is high risk (Northern Spain)
LIC1	Equals 1 if the driving license is between 4 and 14 years old
LIC2	Equals 1 if the driving license is 15 or more years old
LOY	Equals 1 if the client has been with the company for more than 5 years
AGE	Equals 1 if the insured is 30 years old or younger
POW	Equals 1 if horsepower is greater than or equal to 5500cc

4.2 Results

We fitted several different models ranging from simple Poisson regression without dependence to different copula-based finite mixture regression models. Table 3 presents a summary of the models of increasing complexity together with their AIC and BIC values. We started from two independent Poisson regressions (Double Poisson) and two independent negative binomial regressions (Double Neg. Bin.), and then we added some covariance between the two types of claim (bivariate Poisson and bivariate negative binomial). The improvement in the loglikelihood was large when we added zero inflation for the (0, 0) cell (zero-inflated bivariate Poisson). To account for overdispersion, caused by unobserved heterogeneity as discussed in Bermúdez and Karlis (2012) and here, later on, when discussing clustering effects, the next model was the 2-finite mixture bivariate Poisson model which also improved our fit. Finally, following our copula-based approach and a Frank copula, we obtained the copula-based finite mixture of a bivariate Poisson and a negative binomial. The improvement from the Poisson marginals to the negative binomial marginals was limited but indicated that the additional effort and complexity of a negative binomial may be needed. Two additional copula-based finite mixtures of negative binomials, coupled with Gumbel and Gaussian copulas, were fitted without obtaining improvement with respect to the Frank copula model.

In sum, copula-based models provide the best AIC/BIC, indicating that they can capture the features of the data more accurately. Finite mixture models with more than two components ($j > 2$) were also fitted, but no improvement in terms of AIC or BIC was achieved, demonstrating that for the data at hand two components are sufficient and that the portfolio is comprised of two groups of policyholders. Note also that for all models all the available covariates were used, and no variable selection was carried out, i.e., the presented model is the one with all available information. We would like to point out that we have used all variable in all components and marginal distributions. Hence the significance need not to be judged from the standard errors along, as we shall see in Table 4.

Table 3 Various models fitted to the data

Model	Log-lik	Parameters	AIC	BIC
Double Poisson	-20328.6	20	40697.2	40862.4
Double Neg. Bin.	-19199.1	22	38442.2	38623.9
Bivariate Poisson (BP)	-19983.0	21	40008.0	40181.5
Bivariate Neg. Bin.	-18985.6	21	38013.2	38186.6
Zero-inflated BP	-19104.4	22	38252.8	38434.5
2-Finite Mixture BP	-18934.5	43	37955.0	38310.2
Copula Frank 2-FM BP	-18877.4	43	37840.8	38196.0
Copula Frank 2-FM NBIN	-18853.9	47	37801.8	38190.1
Copula Gumbel 2-FM NBIN	-18859.3	47	37812.6	38200.9
Copula Gaussian 2-FM NBIN	-18855.2	47	37804.4	38192.7

From Table 3, we can see that the best fit was obtained with the copula-based 2-finite mixture model with negative binomial marginals coupled with a Frank copula. For negative binomial components, we assumed a parametrization that leads to a dispersion parameter s . However the improvement over the same model with Poisson marginals was rather small. Table 4 shows the estimated regression coefficients for the negative binomial model. If we look at the parameter s , we note that for the second component and the first marginal, the estimated value is too large, implying that the negative binomial tends to a Poisson distribution. Note that the construction of our model through copulas allows the use of different marginal distributions.

The copula-based models also allow a more flexible structure (i.e., negative correlation for the Poisson distribution). For the present application, we used different copulas. The Frank copula allows for both positive and negative dependence and is computationally easier since it does not involve integrals. We also ran Gaussian copulas which also allow both positive and negative dependence at the cost of computational burden as they evaluate the bivariate normal integral many times, as well as a Gumbel copula to allow and check for tail dependence. The Frank copula provided the best results with the lowest computational cost.

Table 4 shows the results obtained by fitting the 2-finite mixture of bivariate negative binomial marginals coupled with the Frank copula. Coefficient regressors and their standard errors are provided for each component and type of claim. The standard errors reported were obtained from the Hessian matrix. The dispersion parameters (s) of the negative binomial marginals, mixing proportion (π) and copula parameter (α) of each component are also shown. The last column refers to the p-value for a likelihood ratio test for testing whether the entire variable can be removed from the model. Such a test checks whether the variables has any merit for the model or not.

In parallel, Fig. 1 shows the estimated means for this model for all the clients for the two components. Clients were assigned to groups based on their maximum a posteriori probability, i.e., using the w_{ij} 's and assigning to the group with larger probability for each client i . One can see that the two components are clearly separated. The group separation is characterized by a low mean with low variance (hereinafter

Table 4 Estimated coefficients and standard errors (* indicates a Z-score between -2 and 2) from the negative binomial model with the Frank copula. The last column refers to the p-value for a Likelihood Ratio Test for testing whether the entire variable can be removed from the model

	1st component						2nd component						LRT p-value
	N ₁			N ₂			N ₁			N ₂			
	Coef	St. err.		Coef	St. err.		Coef	St. err.		Coef	St. err.		
Intercept	-0.297	0.264*		-3.214	0.307		-0.675	0.428*		-3.105	0.232		-
GEN	-0.051	0.091*		0.065	0.128*		0.129	0.151*		0.020	0.076*		0.965
URB	0.023	0.070*		-0.098	0.103*		-0.138	0.122*		0.207	0.060		0.062
ZON1	-0.021	0.082*		0.221	0.120*		-0.067	0.171*		0.279	0.064		<0.001
ZON2	0.098	0.083*		0.370	0.122		0.241	0.131*		-0.269	0.077		
LIC1	0.023	0.228*		-0.155	0.247*		0.015	0.362*		0.496	0.194		<0.001
LIC2	-0.054	0.239*		-0.485	0.270*		-0.139	0.381*		0.148	0.202*		
LOY	0.093	0.086*		0.231	0.118*		0.208	0.145*		0.283	0.065		<0.001
AGE	0.070	0.124*		0.213	0.156*		0.015	0.217*		0.024	0.094*		0.969
POW	0.008	0.103*		-0.101	0.145*		-0.029	0.188*		0.310	0.105		0.012
s	6.170	1.215		0.830	0.112		1235.96	1598.45		0.425	0.092		
π	0.0745	0.0198					0.9256	0.0198					
α	0.3574	0.1523					3.3298	0.8301					

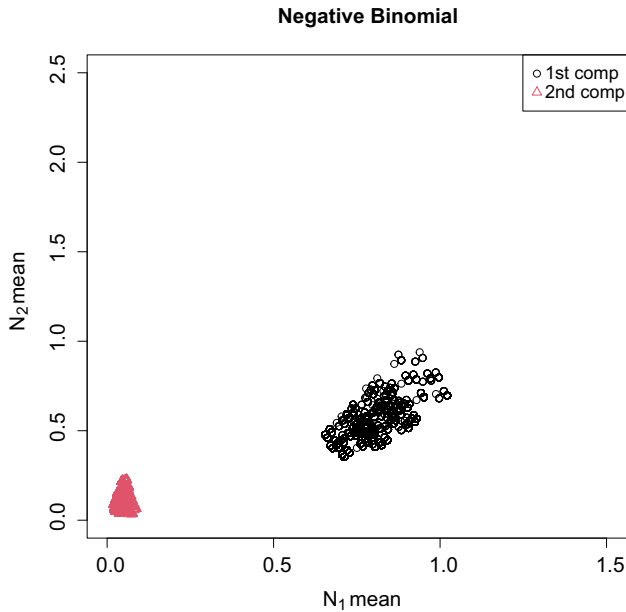


Fig. 1 Scatterplot of the fitted components

policyholders considered as “good” drivers) and a high mean with high variance (policyholders considered as “bad” drivers).

Focusing on the mixing proportion, one can see that most policyholders belong to the second component (that is, most of them are “good” drivers). Without going back over aspects related to dispersion parameters that have already been discussed, we would like to point out that the negative binomial model allows for larger variance within each component. This is appropriate for the first component (“bad drivers”) which has large variance for the two types of claim. However, for the second component (“good drivers”), especially for the N_1 claims, the overdispersion parameter of the negative binomial was too large, implying a Poisson marginal. Finally, from the copula parameter, we can see that the first component (“bad drivers”) shows a lower correlation between type of claims than the second component (“good drivers”).

It is giving careful consideration to this separation between “good” and “bad” drivers. In the context of car insurance, the Poisson regression model is usually rejected because of the presence of overdispersion and an excess of zeros. According to literature, these features of the data may be caused by differences in driving habits and behavior among policyholders that cannot be observed or measured *a priori* by the insurer. This phenomena is known as the problem of unobserved heterogeneity. For example, at the moment of signing a policy, the insurer ignores the driving ability, driving aggressiveness or the degree of obeying traffic regulations of a new client. Through the use of finite mixtures of regression models, we intend to address the problem of unobserved heterogeneity assuming that group separation can be explained by these hidden driving characteristics. In this case, we may consider those policyholders with

a safer driving behavior to belong to the second component and, hence, may be considered as a “good” drivers, while policyholders with a more dangerous driving behavior can be considered as a “bad” drivers and belong to the first component.

A common assumption in all the models used in the literature to address the problem of unobserved heterogeneity (compound frequency models and their zero-inflated or hurdle versions) is that all policyholders behave in the same way with regard to the observed (or *a priori*) variables, and thus they all have the same regression structure. The models proposed in this paper account for unobserved heterogeneity by choosing a finite number of subpopulations and allowing each component in the discrete mixture to have its own score, i.e., each group of policyholders can have a different regression structure with regard to the observed variables. To sum up, by applying finite mixtures of regression models, we aim to examine whether unobserved risk variables, such as a driver’s reflexes, aggressiveness, or compliance with Highway Code, establish the existence of subpopulations of policyholders with different *a priori* behavior.

As discussed above, here we may assume the existence of two groups of policyholders, “good” and “bad” drivers, defined by their respective driving behavior. From Table 4, one can also see that each group of policyholders exhibits a different regression structure. This is particularly noticeable for the variable *ZON2* with regard to the number of claims for the rest of guarantees (N_2). For policyholders considered to be “bad” drivers, driving in a higher risk zone (basically due to climate conditions) increases the probability of making a claim; whereas it decreases for “good” drivers. This is reasonable if we focus on their driving habits: “bad” drivers are more likely to make a claim in a zone with bad climate conditions because they drive aggressively and ignoring traffic conditions; while “good” drivers are even more careful with such a climate conditions and so they are less likely to have a claim. It is worth emphasizing that using simpler models, i.e., with one component, this covariate may not be significant as they only estimate one coefficient and the effect is canceled out, estimating an average effect that is close to zero.

Figure 2 shows the average posterior probabilities conditional on observed values of the two claims. Namely, each cell refers to a pair of the type (N_1, N_2) , $N_1, N_2 = 0, 1, 2, 3, 4$ and shows the posterior probability that such an observation belongs to the group 1 or 2 averaged over the different observations with such values. One can see that, as expected, the second component (“good drivers”) is more likely for those cells with less claims for both types, while the 1st component (“bad drivers”) corresponds to those cells with a growing number of claims. Indeed, we may observe that “good drivers” are more likely to have a claim for the rest of guarantees than having a third liability claim. As a classification rule observation, policyholders with very few counts (specially for third liability claims) are more probable to come from second component, while for large counts we are almost certain that they belong to the first component.

4.3 Prediction

In order to see the predictive ability of the model, but also to discuss about the added value offered by the proposed models, we have run a prediction experiment with

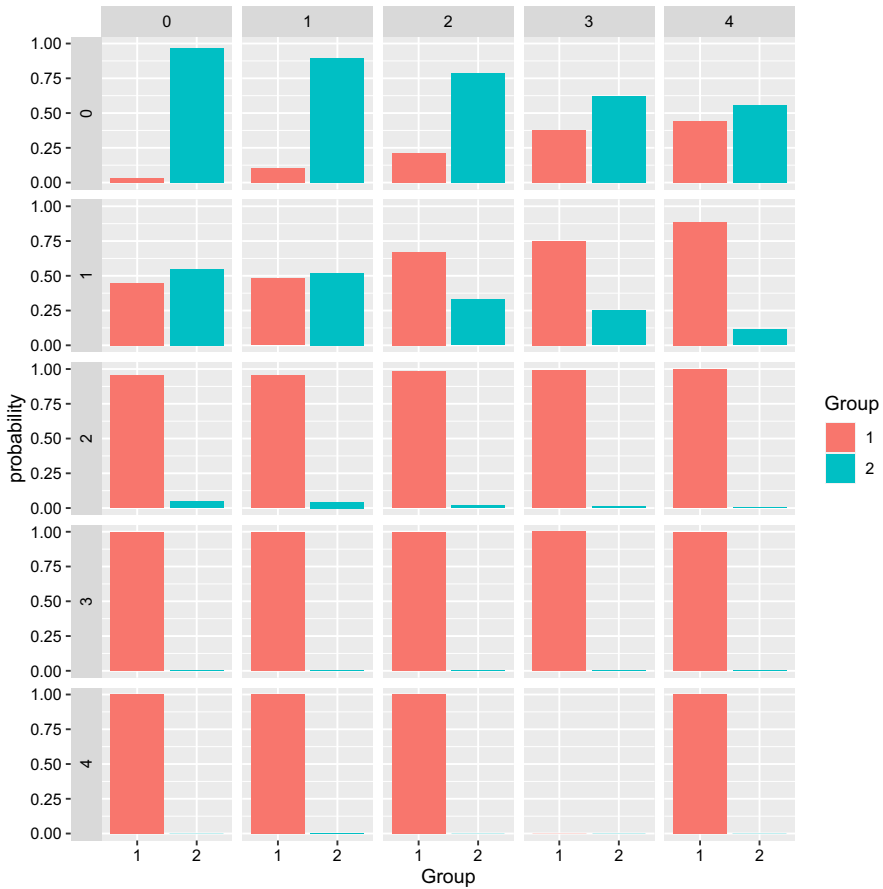


Fig. 2 Posterior probabilities conditional on observed values of the two claims. Rows refer to N_1 and columns to N_2 . Each cell refers to a pair of the type (N_1, N_2) , $N_1, N_2 = 0, 1, 2, 3, 4$ and shows the posterior probability that such an observation belongs to the component (group) 1 or 2, “bad” and “good” drivers respectively

the data. Namely, we kept randomly 20% of the data as a test set, then we fitted models to the rest (80%) and we predicted the ones in the test set based on the models derived. This firstly can show the predictive ability of the model out-of-sample but also reveals the different features of the data. In Table 5, we present results for some of the models used in Table 3. Particularly, we have fitted a double Poisson model and a double negative binomial model as baseline models; the second one captures the overdispersion we have seen in the data. Then, we fitted a 2-finite mixture of bivariate Poisson distributions, this can recover part of the correlation but also the overdispersion. Finally, from those models based on copulas we selected one with a Frank copula and negative binomial marginals.

Table 5 shows the observed frequencies and the expected frequencies for the models. It is obvious that the double Poisson fails a lot, the double negative binomial

Table 5 Observed and predicted frequencies for out-of-sample predictions for the 20% of the observations ($n = 5, 718$) for a series of models

		<i>Observed data</i>									
N_1	$N_2: 0$	1	2	3	4						
0	4918	378	41	19	2						
1	201	65	8	4	3						
2	45	11	2	1	0						
3	11	2	2	0	0						
4	2	1	0	0	0						
		<i>Double Poisson</i>					<i>Double Neg. Bin.</i>				
N_1	$N_2: 0$	1	2	3	4	0	1	2	3	4	
0	4629	578	38	2	0	4803	430	80	18	4	
1	397	50	3	0	0	269	24	5	1	0	
2	17	2	0	0	0	54	5	1	0	0	
3	1	0	0	0	0	14	1	0	0	0	
4	0	0	0	0	0	4	0	0	0	0	
		<i>2FM of Biv. Poisson</i>					<i>Frank - 2FM Neg. Bin.</i>				
N_1	$N_2: 0$	1	2	3	4	0	1	2	3	4	
0	4881	388	46	9	2	4882	380	60	12	3	
1	213	72	23	6	1	216	62	14	4	1	
2	31	21	7	2	0	39	14	4	2	1	
3	6	4	2	0	0	10	4	2	1	0	
4	1	1	0	0	0	3	1	0	0	0	

corrects this in a certain extent but it cannot capture all the features of the data. Note that the marginals are closer to the observed ones but the model fails a lot for the bivariate distribution. The 2-finite mixture of bivariate enhances a lot and accounts for the overdispersion but it is still far from the observed ones. Finally, the copula model with negative binomial marginals is closer to the observations. Overall the copula-based model predicts quite well the observed data. A further important aspect is that while the observed correlation for the test set is 0.15, the double Poisson and the double negative binomial, as expected, predict a value almost zero, while the 2-finite mixture of bivariate Poisson overestimates the correlation giving a number close to 0.22. The copula model is closer from all other models with a correlation equal to 0.16. Summarizing, the copula-based model predicts quite well the cases because it can provide characteristics of the observed data closer to the real ones due to the inherent flexibility that it offers.

5 Concluding remarks

We have presented a new approach for modeling bivariate claim data based on a finite mixture of bivariate discrete distributions defined through a copula, also allowing for covariates. In our opinion, this approach offers several advantages. On the one hand, allowing for a more flexible dependence structure we can obtain a better fit and a better predictive capacity. On the other hand, through the use of finite mixture, the model further allows for a clustering analysis that tries to address the problem of unobserved heterogeneity.

The extension of this approach to higher dimensions is more demanding. The problem is that not all the copulas can be extended in a rigorous way to higher dimensions, at least not in such a way as to allow for a fully flexible dependence structure. For example, in the case of the Frank copula, while it is possible to expand it to multidimensional cases, it may involve only one common dependence parameter, which is more restrictive. The Gaussian copula is more suitable for such generalization at the additional computational cost required to evaluate many multidimensional integrals. A possible solution would be to create an approximate pairwise model and to estimate the parameters with some other method like composite likelihood. Such an approach is being considered in ongoing research.

In terms of the actuarial implications, the above model may allow premiums to be extracted in the same way as in Bermúdez and Karlis (2012) which would help to account correctly for the dependencies between the claims. Moreover, there might be a substantive interest in classifying new clients into “good” and “bad” risks and, hence, estimating premiums according to their specific regression structure derived from the finite mixture. However, current approaches are limited by the fact that it is rather impossible to classify new clients. This may be the goal for future research. Finally, note that such an approach, while applied to count claims in this paper, can also be extended to the severities of different types of claim, by considering bivariate models with copulas and continuous marginals.

A final remark relates to identifiability issues about the model. Issues related to identifiability of finite mixtures defined with copulas have been discussed in Kosmidis and Karlis (2016). In general, it is very hard to derive theoretical results due to the complicated nature of the copulas. A natural assumption is to enforce some lexicographical ordering in the components so as to ensure that we avoid the case when interchanging the order of the components we end up with the same mixture. An additional identifiability issue in our case relates to the identifiability of mixtures of GLM regression models. Conditions for identifiability for such finite mixtures of regression and GLM models are given in Hennig (2000) and Grün and Leisch (2008). Since theoretical results are hard to establish in general, from a practical perspective, we have worked with several initial values to examine whether we became trapped with different solutions. A lexicographical ordering of the components has been used. We have not found solutions that differ substantially on the regression coefficients giving similar likelihood, adding to our belief that our model is identifiable.

Acknowledgements Research for this paper was initiated while the second author was visiting the “Risk in Finance and Insurance” Research Group at the University of Barcelona. The second author would like to thank the Laboratory of Bayesian and Computational Statistics of AUEB for giving access to its resources.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Funding We would like to thank the Fundación BBVA projects in the area of Big Data, the ICREA Academia programme, AGAUR and the Spanish Ministry of Science and Innovation grant PID2019-105986GB-C21.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison J, Ho C (1989) The multivariate Poisson-log normal distribution. *Biometrika* 76(4):643–653
- Berkhout P, Plug E (2004) A bivariate Poisson count data model using conditional probabilities. *Stat Neerl* 58(3):349–364
- Bermúdez L, Karlis D (2012) A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. *Comput Stat Data Anal* 56(12):3988–3999
- Cameron AC, Li T, Trivedi PK, Zimmer DM (2004) Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *Economet J* 7(2):566–584
- Famoye F (2010) On the bivariate negative binomial regression model. *J Appl Stat* 37(6):969–981
- Genest C, Nešlehová J (2007) A primer on copulas for count data. *ASTIN Bull* 37(2):475–515
- Genest C, Mesfioui M, Schulz J (2018) A new bivariate Poisson common shock model covering all possible degrees of dependence. *Stat Probab Lett* 140:202–209
- Ghitany M, Karlis D, Al-Mutairi D, Al-Awadhi F (2012) An EM algorithm for multivariate mixed Poisson regression models and its application. *Appl Math Sci* 6(137):6843–6856
- Grün B, Leisch F (2008) Finite mixtures of generalized linear regression models. *Recent advances in linear models and related areas*. Springer, Berlin, pp 205–230
- Gurmu S, Elder J (2000) Generalized bivariate count data regression models. *Econ Lett* 68(1):31–36
- Hennig C (2000) Identifiability of models for clusterwise linear regression. *J Classif* 17(2)
- Karlis D, Meligkotsidou L (2007) Finite multivariate Poisson mixtures with applications. *J Stat Plan Inference* 137:1942–1960
- Kosmidis I, Karlis D (2016) Model-based clustering using copulas with applications. *Stat Comput* 26(5):1079–1099
- Lakshminarayana J, Pandit S, Srinivasa Rao K (1999) On a bivariate Poisson distribution. *Commun Stat Theory Methods* 28(2):267–276
- Li C, Lu J, Park J, Kim K, Brinkley P, Peterson J (1999) A multivariate zero-inflated Poisson distribution and its inference. *Technometrics* 41(1):29–38
- McNicholas PD (2016) *Mixture model-based classification*. CRC Press, Boca Raton
- Munkin MK, Trivedi PK (1999) Simulated maximum likelihood estimation of multivariate mixed Poisson regression models, with application. *Economet J* 2(1):29–48
- Nelsen R (1997) *An introduction to copulas*. Chapman and Hall, London
- Nikoloulopoulos AK (2013) Copula-based models for multivariate discrete response data. *Copulae in mathematical and quantitative finance*. Springer, Berlin, pp 231–249
- Panagiotelis A, Czado C, Joe H (2012) Pair copula constructions for multivariate discrete data. *J Am Stat Assoc* 107(499):1063–1072

- Papastamoulis P, Martin-Magniette ML, Maugis-Rabusseau C (2016) On the estimation of mixtures of Poisson regression models with large number of components. *Comput Stat Data Anal* 93:97–106
- Shi P, Valdez EA (2014) Multivariate negative binomial models for insurance claim counts. *Insur Math Econ* 55:18–29
- Steyn H (1976) On the multivariate Poisson normal distribution. *J Am Stat Assoc* 71(353):233–236
- Wang K, Lee AH, Yau KK, Carrivick PJ (2003) A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accid Anal Prev* 35(4):625–629
- Zamani H, Faroughi P, Ismail N (2016) Bivariate generalized Poisson regression model: applications on health care data. *Empir Econ* 51(4):1607–1621

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.