
“The bivariate Sarmanov distribution for insurance claim frequencies and average severities”

Raluca Vernic, Catalina Bolancé and Ramon Alemany

The Research Institute of Applied Economics (IREA) in Barcelona was founded in 2005, as a research institute in applied economics. Three consolidated research groups make up the institute: AQR, RISK and GiM, and a large number of members are involved in the Institute. IREA focuses on four priority lines of investigation: (i) the quantitative study of regional and urban economic activity and analysis of regional and local economic policies, (ii) study of public economic activity in markets, particularly in the fields of empirical evaluation of privatization, the regulation and competition in the markets of public services using state of industrial economy, (iii) risk analysis in finance and insurance, and (iv) the development of micro and macro econometrics applied for the analysis of economic activity, particularly for quantitative evaluation of public policies.

IREA Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. For that reason, IREA Working Papers may not be reproduced or distributed without the written consent of the author. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of IREA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

Abstract

Real data studies emphasized situations where the classical independence assumption between the frequency and the severity of claims does not hold in the collective model. Therefore, there is an increasing interest in defining models that capture this dependence. In this paper, we introduce such a model based on Sarmanov's bivariate distribution, which has the ability of joining different types of marginals in flexible dependence structures. More precisely, we join the claims frequency and the average severity by means of this distribution. We also suggest a maximum likelihood estimation procedure to estimate the parameters and illustrate it both on simulated and real data.

JEL classification: G22, G52.

Keywords: Dependence, Sarmanov distribution, Frequency, Severity, Parameters estimation.

Raluca Vernic: Faculty of Mathematics and Computer Science, Ovidius University of Constanta, and Institute for Mathematical Statistics and Applied Mathematics, Calea 13 Septembrie 13, 050711 Bucharest, Romania. Email: rvernic@univ-ovidius.ro

Catalina Bolancé: Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain. Email: bolance@ub.edu

Ramon Alemany: Department of Econometrics, Riskcenter-IREA, University of Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain. Email: ralemany@ub.edu

1 Introduction

When modeling aggregate claims with the classical collective model, the usual assumption is that claim frequency and severity are independent, an assumption which facilitates the corresponding computations. In practice, however, claim frequency and severity tend to be dependent, albeit minimally. For example, in auto insurance data, some negative or positive dependence could be found; on one hand, a high frequency can be associated with an urban driving area where the costs are low or, on the other hand, the same high frequency can be associated with daily journeys on secondary roads where accident costs are usually higher. Another example is found in health insurance data, where the dependence between frequency and severity is usually positive. Furthermore, the sample estimation of the dependence between these two variables is not easy to measure; classical correlation coefficient can provide distorted results that can be affected by a few events. For all these reasons, recently, there is an increasing interest in exploring models that account the dependence between frequency and severity. In this sense, two different approaches can be distinguished: on one hand, a model is defined for the average claim size distribution using the number of claims as covariate (see Frees and Wang, 2006; Gschlößl and Czado, 2007; Frees et al., 2011; Garrido et al., 2016; Valdez et al., 2018); as a second approach, the frequency and severity (or average severity) components are related through a copula (see Erhardt and Czado, 2012; Czado et al., 2012; Krämer et al., 2013; Hua, 2015; Lee and Shi, 2019; Oh et al., 2019; Shi et al., 2015).

In this paper, as in Czado et al. (2012), we introduce dependence between the number of claims and the corresponding average claim size. Nevertheless, in contrast to these authors, who modeled this dependence by a Gaussian copula, we assume a Sarmanov dependence between the frequency and the average severity. As Czado et al. (2012) did, to estimate the parameters we propose a maximization by parts of the log-likelihood function, but given our bounded parametric space, to optimize each part we use the `optim()` function of R and validate our algorithm with a simulation study.

Due to its ability to join different marginals in flexible dependence structures, Sarmanov's multivariate distribution recently gained a lot of attention in the actuarial literature in several aspects, like: modeling continuous claim sizes (see Bahraoui et al., 2015); modeling discrete claim frequencies (see Abdallah et al., 2016; Bolancé and Vernic, 2019); in the evaluation of ruin probabilities (see, for example, Yang and Yuen, 2016; Guo et al., 2017) etc. In some of the just mentioned papers, the Sarmanov distribution has been fitted in its bivariate and trivariate forms to real insurance data and it proved to provide a better fit than other distributions, including Copula ones.

In this paper, we make particular use of the special capacity of the Sarmanov distribution to join marginals of different types, more precisely, one marginal will be of discrete type, corresponding to the claim frequency, and a second marginal will be continuous, representing the average severity. This flexibility, associated with combining various marginal distributions, allows us to propose alternative models that mix a count data distribution for the frequency with a Gamma distribution for the severity, as follows: Poisson-Gamma, Negative Binomial-Gamma, Zero Inflated Poisson-Gamma and Zero Inflated Negative Binomial-Gamma compound distributions. We use the Gamma distribution because it has flexibility and allows us to model a right skewed distribution and to deduce closed type expressions for the main results of our models.

The proposed models take into account that a cost only exists if the claim frequency is 1 or more. Therefore, they are specified in two parts: the first part corresponds to the probability of 0

frequency and severity, and the second part to the bivariate probability of frequency and severity larger than 0.

A possible limitation of our compound Sarmanov-based distributions is that the dependency is related to a bounded parameter, which in some cases does not allow fitting strong correlations. However, our experience has shown that the correlation between the number and the amount of claims is not very high - a correlation lower than 0.5 is common. For example, Czado et al. (2012), using Mixed Copula models, estimated a correlation parameter equal to 0.1366; although statistically significant, even lower correlations can be found. Specifically, we illustrate this using a real data set consisting of a random sample of auto insurance policyholders.

The rest of the paper is organized as follows: in Section 2, we describe the proposed Sarmanov distribution, its properties, particular cases and estimation procedure. In Section 3, we present the results of a simulation study to evaluate the estimated parameters using a two parts log-likelihood maximization. An application to a real data set containing auto insurance number and average cost of claims is discussed in Section 4. Finally, we conclude in Section 5. The paper ends with an appendix containing the proofs.

2 Collective model with dependent number and average size of claims

We shall introduce dependence between the number of claims N and the corresponding average claim size X of a portfolio or of a certain policy. Letting S denote the aggregate claims, clearly

$$S = NX. \quad (1)$$

We let p denote the probability mass function (pmf) of N . In respect of the random variable (r.v.) X , its distribution will have both an absolutely continuous component with probability density function (pdf) denoted by f_X and a probability mass at 0. Therefore, the distribution of S also has a probability mass at 0 and a pdf that we denote by f_S . We denote the cumulative distribution function (cdf) of a r.v. by F indexed with the name of that r.v..

2.1 Sarmanov dependence

We assume a Sarmanov dependence between N and X as follows

$$f_{X,N}(x, n) = \begin{cases} p(0), & n = x = 0 \\ p(n) f(x) (1 + \omega \psi(n) \phi(x)), & n \geq 1, x > 0 \end{cases}, \quad (2)$$

where f is a pdf, ψ and ϕ are bounded non-constant kernel functions and $\omega \in \mathbb{R}$. We call the pdf (2) mixed because it joins the continuous pdf f and the discrete pmf p . Also, in order for (2) to define a proper pdf, we impose the conditions

$$\sum_{n \geq 1} \psi(n) p(n) = \int_{\mathbb{R}} \phi(x) f(x) dx = 0 \text{ and} \quad (3)$$

$$1 + \omega \psi(n) \phi(x) \geq 0, \text{ for all } n \geq 1, x > 0. \quad (4)$$

To simplify the writing, we denote by Y a r.v. having pdf f and representing $X > 0$. Letting $m_1 = \inf_{n \geq 1} \psi(n)$, $m_2 = \inf_{x > 0} \phi(x)$, $M_1 = \sup_{n \geq 1} \psi(n)$, $M_2 = \sup_{x > 0} \phi(x)$, condition (4) restricts ω to the following interval

$$\max \left\{ -\frac{1}{m_1 m_2}, -\frac{1}{M_1 M_2} \right\} \leq \omega \leq \min \left\{ -\frac{1}{m_1 M_2}, -\frac{1}{M_1 m_2} \right\}. \quad (5)$$

The following proposition presents the distributions of X , of S and conditional distributions. All the proofs are given in the appendix.

Proposition 1. *Under the Sarmanov dependence condition (2), it holds that*

$$\begin{aligned} \text{i) } \Pr(X = 0) &= p(0), \\ f_X(x) &= (1 - p(0)) f(x), \quad x > 0. \\ \text{ii) } \Pr(X = 0 | N = n) &= \begin{cases} 1, & n = 0 \\ 0, & n \geq 1 \end{cases}, \\ f_{X|N=n}(x) &= f(x) (1 + \omega \psi(n) \phi(x)), \quad x > 0, \quad n \geq 1. \\ \text{iii) } \Pr(N = n | X = x) &= \begin{cases} 1, & n = x = 0 \\ \frac{p(n)}{1 - p(0)} (1 + \omega \psi(n) \phi(x)), & n \geq 1, \quad x > 0 \end{cases}. \\ \text{iv) } \Pr(S = 0) &= p(0), \\ f_S(s) &= \sum_{n \geq 1} \frac{p(n)}{n} f\left(\frac{s}{n}\right) \left(1 + \omega \psi(n) \phi\left(\frac{s}{n}\right)\right), \quad s > 0. \end{aligned}$$

The first two moments of S are given in the following result; note that they are expressed in terms of the r.v. Y .

Proposition 2. *Under the Sarmanov dependence condition (2), the expected value and variance of S are given respectively, by*

$$\begin{aligned} \mathbb{E}S &= \mathbb{E}N\mathbb{E}Y + \omega \mathbb{E}[N\psi(N)] \mathbb{E}[Y\phi(Y)], \\ \text{Var}S &= \mathbb{E}[Y^2] \text{Var}N + (\mathbb{E}N)^2 \text{Var}Y - \omega^2 \mathbb{E}^2[N\psi(N)] \mathbb{E}^2[Y\phi(Y)] \\ &\quad + \omega (\mathbb{E}[N^2\psi(N)] \mathbb{E}[Y^2\phi(Y)] - 2\mathbb{E}N \mathbb{E}[N\psi(N)] \mathbb{E}Y \mathbb{E}[Y\phi(Y)]). \end{aligned}$$

Proposition 3. *The correlation coefficient of the pdf (2) is given by*

$$\text{corr}(X, N) = \frac{\omega \mathbb{E}[N\psi(N)] \mathbb{E}[Y\phi(Y)] + p(0) \mathbb{E}N\mathbb{E}Y}{\sqrt{(1 - p(0)) (\text{Var}Y + p(0) \mathbb{E}^2[Y]) \text{Var}N}}. \quad (6)$$

We propose to use exponential kernels. Regarding Sarmanov's pdf in (2), we shall consider in particular the exponential kernels satisfying condition (3). More precisely, $\phi(y) = e^{-\gamma y} - \mathcal{L}_Y(\gamma)$, where \mathcal{L}_Y denotes the Laplace transform of the r.v. Y . Furthermore, we let $\psi(n) = e^{-\delta n} - k$, and to find k , we write

$$\begin{aligned} \sum_{n \geq 1} \psi(n) p(n) &= \sum_{n \geq 1} (e^{-\delta n} - k) p(n) \\ &= \sum_{n \geq 0} e^{-\delta n} p(n) - p(0) - k \left(\sum_{n \geq 0} p(n) - p(0) \right) \\ &= \mathcal{L}_N(\delta) - p(0) - k(1 - p(0)). \end{aligned}$$

Imposing the condition expressed in (3), i.e. $\sum_{n \geq 1} \psi(n) p(n) = 0$, we obtain $k = \frac{\mathcal{L}_N(\delta) - p(0)}{1 - p(0)}$. Therefore, $\psi(n) = e^{-\delta n} - \frac{\mathcal{L}_N(\delta) - p(0)}{1 - p(0)}$ because in the second formula of the pdf (2) we have $n \geq 1$ (similar to a left truncation of N in 0).

2.2 Simulation from the collective model

To simulate values from the two parts bivariate Sarmanov distribution whose pdf is defined in (2), we use the inversion method from the conditional cdf of X given $N = n$, which easily results from (ii) in Proposition 1 as

$$\begin{aligned} F_{X|N=0}(0) &= 1, \\ F_{X|N=n}(x) &= \int_0^x f(y) (1 + \omega \psi(n) \phi(y)) dy \\ &= F_Y(x) + \omega \psi(n) \int_0^x f(y) \phi(y) dy, \quad n \geq 1, x > 0. \end{aligned} \quad (7)$$

Hence, we simulate the value n from the distribution of N . If $n = 0$ then clearly $x = 0$; otherwise, we generate an uniform $U(0, 1)$ value u and solve the equation $F_{X|N=n}(x) = u$ for x . This yields the generated pair (n, x) .

2.3 Parameters estimation

Let $(n_i, x_i)_{i=1}^K$ be a random bivariate sample of the number and average amount of claims. Let θ and \mathbf{v} be, respectively, the parameters vectors of the marginal distribution of N and of the continuous marginal distribution of Y , while ω is the dependence parameter of Sarmanov's distribution. Based on (2), the log-likelihood function is

$$\begin{aligned} \ln L\left((n_i, x_i)_{i=1}^K; \theta; \mathbf{v}; \omega\right) &= \sum_{\{i: n_i = x_i = 0\}} \ln p(0; \theta) + \sum_{\{i: n_i \geq 1, x_i > 0\}} [\ln p(n_i; \theta) \\ &\quad + \ln f(x_i; \mathbf{v}) + \ln(1 + \omega \psi(n_i) \phi(x_i))] \\ &= \ln L\left((n_i)_{i=1}^K; \theta\right) + \ln L(\{x_i | x_i > 0, i = 1, \dots, K\}; \mathbf{v}) \\ &\quad + \sum_{\{i: n_i \geq 1, x_i > 0\}} \ln(1 + \omega \psi(n_i) \phi(x_i)), \end{aligned} \quad (8)$$

where $L\left((n_i)_{i=1}^K; \theta\right)$ is the likelihood function corresponding to the marginal r.v. N , while $L(\{x_i | x_i > 0, i = 1, \dots, K\}; \mathbf{v})$ is the one corresponding to Y .

Maximizing the log-likelihood expressed in (8) is very difficult, mainly for two reasons. The first reason is because, given the limits of the dependency parameter ω that were defined in (5), the parametric space is bounded. The second reason is due to the strong relationship that exists between the dependence parameter and the marginal ones.

We also define $l_{(n_i, x_i)_{i=1}^K}(\theta; \mathbf{v} | \omega)$ to be the log-likelihood function corresponding to the marginal parameters given the dependence parameter ω and, similarly, $l_{(n_i, x_i)_{i=1}^K}(\omega | \theta; \mathbf{v})$ the log-likelihood function of the dependence parameter given the marginal parameters θ, \mathbf{v} . As in Bolancé and Vernic (2019), we propose to determine the Maximum Likelihood Estimation (MLE) of the parameters in two phases. The first phase consists of maximising by parts the log-likelihood function (an example in a similar context is in Czado et al., 2012). We describe the procedure below.

Phase 1 Using MLE, find initial values for the parameters of the univariate marginal distributions. The rest of the procedure in Phase 1 is divided into two steps:

Step 1 (iteration j) Given the parameters for the marginal distributions, find $\hat{\omega}^j$ within the interval defined in (5) for this dependence parameter by maximizing the log-likelihood $l_{(n_i, x_i)_{i=1}^K}(\omega | \theta; \nu)$.

Step 2 Given $\hat{\omega}^j$, obtain new values for the parameters of the marginal distributions by maximizing the log-likelihood function $l_{(n_i, x_i)_{i=1}^K}(\theta; \nu | \omega)$.

Steps 1 and 2 are repeated until convergence. If the dependence parameter is located at an extreme of the interval, recalculate these intervals using the parameters for marginals obtained in Step 2.

Phase 2 Starting with the initial parameters estimated in Phase 1, perform full MLE.

Given our bounded parametric space, optimizations in the two phases were carried out using the `optim()` function of R with the method L-BFGS-B (Byrd et al., 1995).

2.4 Particular cases

2.4.1 Counting distributions

We consider four different distributions for the r.v. number of claims, i.e., Poisson, zero inflated Poisson, Negative Binomial and zero inflated Negative Binomial. Note that if N follows a certain discrete distribution with support \mathbb{N} and \tilde{N} follows the same distribution in the zero inflated form with parameter $\pi \in (0, 1)$ (the probability of extra zeros), then the following relations hold

$$\begin{aligned} \Pr(\tilde{N} = n) &= \begin{cases} \pi + (1 - \pi) \Pr(N = 0), & n = 0 \\ (1 - \pi) \Pr(N = n), & n \geq 1 \end{cases}, \\ \mathbb{E}\tilde{N} &= (1 - \pi) \mathbb{E}N, \quad \mathbb{E}[\tilde{N}^2] = (1 - \pi) \mathbb{E}[N^2], \quad \text{Var}\tilde{N} = (1 - \pi) (\text{Var}N + \pi \mathbb{E}^2N), \\ \mathcal{L}_{\tilde{N}}(\delta) &= \pi + (1 - \pi) \mathcal{L}_N(\delta). \end{aligned}$$

Assuming that N is Poisson distributed, $N \sim Po(\lambda)$, $\lambda > 0$, we recall that

$$\mathbb{E}N = \text{Var}N = \lambda, \quad \mathbb{E}[N^2] = \lambda + \lambda^2, \quad \mathcal{L}_N(\delta) = e^{\lambda(e^{-\delta} - 1)}.$$

Then, when N is zero inflated Poisson distributed, $N \sim ZIP(\lambda, \pi)$, $\lambda > 0$, $\pi \in (0, 1)$, we easily obtain that

$$\begin{aligned} \Pr(N = n) &= \begin{cases} \pi + (1 - \pi) e^{-\lambda}, & n = 0 \\ (1 - \pi) e^{-\lambda} \frac{\lambda^n}{n!}, & n \geq 1 \end{cases}, \\ \mathbb{E}N &= (1 - \pi) \lambda, \quad \mathbb{E}[N^2] = (1 - \pi) \lambda (\lambda + 1), \quad \text{Var}N = (1 - \pi) \lambda (\lambda \pi + 1), \\ \mathcal{L}_N(\delta) &= \pi + (1 - \pi) e^{\lambda(e^{-\delta} - 1)}. \end{aligned}$$

If N is Negative Binomial distributed, $N \sim NB(r, p)$, $r > 0$, $p \in (0, 1)$, then, with $q = 1 - p$,

$$\begin{aligned}\Pr(N = n) &= \frac{\Gamma(r+n)}{n!\Gamma(r)} p^r q^n, \quad n \in \mathbb{N}, \\ \mathbb{E}N &= \frac{rq}{p}, \quad \mathbb{E}[N^2] = \frac{rq(1+qr)}{p^2}, \quad \text{Var}N = \frac{rq}{p^2}, \quad \mathcal{L}_N(\delta) = \left(\frac{p}{1-qe^{-\delta}} \right)^r.\end{aligned}$$

Considering that N is zero inflated Negative Binomial distributed, $N \sim ZINB(r, p)$, $r > 0$, $p \in (0, 1)$, $\pi \in (0, 1)$, the above formulas yield

$$\begin{aligned}\Pr(N = n) &= \begin{cases} \pi + (1-\pi)p^r, & n = 0 \\ (1-\pi) \frac{\Gamma(r+n)}{n!\Gamma(r)} p^r q^n, & n \geq 1 \end{cases}, \\ \mathbb{E}N &= (1-\pi) \frac{rq}{p}, \quad \mathbb{E}[N^2] = (1-\pi) \frac{rq(1+rq)}{p^2}, \quad \text{Var}N = (1-\pi) \frac{rq(1+\pi rq)}{p^2}, \\ \mathcal{L}_N(\delta) &= \pi + (1-\pi) \left(\frac{p}{1-qe^{-\delta}} \right)^r.\end{aligned}$$

In the following proposition, we present formulas needed to evaluate the expected value and variance of S given in Proposition 2.

Proposition 4. Let $\psi(n) = e^{-\delta n} - \frac{\mathcal{L}_N(\delta) - p(0)}{1-p(0)}$ be the exponential kernel.

i) If $N \sim Po(\lambda)$, then

$$\begin{aligned}\mathbb{E}[N\psi(N)] &= \lambda e^{-\lambda} \left(e^{\lambda e^{-\delta} - \delta} - \frac{e^{\lambda e^{-\delta}} - 1}{1 - e^{-\lambda}} \right), \\ \mathbb{E}[N^2\psi(N)] &= \lambda e^{-\lambda} \left[e^{\lambda e^{-\delta} - \delta} (\lambda e^{-\delta} + 1) - (\lambda + 1) \frac{e^{\lambda e^{-\delta}} - 1}{1 - e^{-\lambda}} \right].\end{aligned}$$

ii) If $N \sim ZIP(\lambda, \pi)$, then

$$\begin{aligned}\mathbb{E}[N\psi(N)] &= (1-\pi) \lambda e^{-\lambda} \left(e^{\lambda e^{-\delta} - \delta} - \frac{e^{\lambda e^{-\delta}} - 1}{1 - e^{-\lambda}} \right), \\ \mathbb{E}[N^2\psi(N)] &= (1-\pi) \lambda e^{-\lambda} \left[e^{\lambda e^{-\delta} - \delta} (\lambda e^{-\delta} + 1) - (\lambda + 1) \frac{e^{\lambda e^{-\delta}} - 1}{1 - e^{-\lambda}} \right].\end{aligned}$$

iii) If $N \sim NB(r, p)$, then

$$\begin{aligned}\mathbb{E}[N\psi(N)] &= \frac{rqp^r}{(1-qe^{-\delta})^r} \left(\frac{1}{e^\delta - q} - \frac{1 - (1-qe^{-\delta})^r}{p(1-p^r)} \right), \\ \mathbb{E}[N^2\psi(N)] &= \frac{rqp^r}{(1-qe^{-\delta})^r} \left[\frac{rq + e^\delta}{(e^\delta - q)^2} - (1+qr) \frac{1 - (1-qe^{-\delta})^r}{p^2(1-p^r)} \right].\end{aligned}$$

iv) If $N \sim ZINB(r, p, \pi)$, then

$$\begin{aligned}\mathbb{E}[N\psi(N)] &= (1-\pi) \frac{rqp^r}{(1-qe^{-\delta})^r} \left(\frac{1}{e^\delta - q} - \frac{1 - (1-qe^{-\delta})^r}{p(1-p^r)} \right), \\ \mathbb{E}[N^2\psi(N)] &= (1-\pi) \frac{rqp^r}{(1-qe^{-\delta})^r} \left[\frac{rq + e^\delta}{(e^\delta - q)^2} - (1+qr) \frac{1 - (1-qe^{-\delta})^r}{p^2(1-p^r)} \right].\end{aligned}$$

2.4.2 Gamma severity distribution

Let Y be Gamma distributed, $Y \sim Ga(\alpha, \beta)$, $\alpha, \beta > 0$, where β is the rate parameter. We recall that

$$\mathbb{E}Y = \frac{\alpha}{\beta}, \mathbb{E}[Y^2] = \frac{\alpha(\alpha+1)}{\beta^2}, \text{Var}Y = \frac{\alpha}{\beta^2}, \mathcal{L}_Y(\gamma) = \left(\frac{\beta}{\beta+\gamma} \right)^\alpha.$$

The following result is needed to evaluate the expected value and variance of S .

Proposition 5. Let $Y \sim Ga(\alpha, \beta)$, $\alpha, \beta > 0$, and let $\phi(x) = e^{-\gamma x} - \mathcal{L}_Y(\gamma)$ be the exponential kernel. Then

$$\begin{aligned}\mathbb{E}[Y\phi(Y)] &= -\frac{\alpha\gamma\beta^{\alpha-1}}{(\beta+\gamma)^{\alpha+1}}, \\ \mathbb{E}[Y^2\phi(Y)] &= -\frac{\alpha(\alpha+1)\gamma\beta^{\alpha-2}(2\beta+\gamma)}{(\beta+\gamma)^{\alpha+2}}.\end{aligned}$$

2.4.3 Particular compound distributions

By combining the above discussed counting distributions with the Gamma severity distribution, we obtain four particular compound distributions: compound Poisson-Gamma, compound Zero Inflated Poisson-Gamma, compound Negative Binomial-Gamma and compound Zero Inflated Negative Binomial-Gamma. The next proposition presents their pdfs; the proof is immediate, hence we omit it.

Proposition 6. Let $Y \sim Ga(\alpha, \beta)$ and let $\psi(n) = e^{-\delta n} - \frac{\mathcal{L}_N(\delta) - p(0)}{1-p(0)}$, $\phi(x) = e^{-\gamma x} - \left(\frac{\beta}{\beta+\gamma} \right)^\alpha$ be exponential kernels. Then:

i) If $N \sim Po(\lambda)$, then the compound Poisson-Gamma pdf is given by

$$f_{X,N}(x,n) = \begin{cases} e^{-\lambda}, & n=x=0 \\ \frac{\beta^\alpha e^{-\lambda}}{\Gamma(\alpha)} \frac{\lambda^n}{n!} x^{\alpha-1} e^{-\beta x} \left(1 + \omega \left(e^{-\delta n} - e^{\lambda(e^{-\delta}-1)} \right) \phi(x) \right), & n \geq 1, x > 0 \end{cases}.$$

ii) If $N \sim ZIP(\lambda, \pi)$, then the compound zero inflated Poisson-Gamma pdf is

$$f_{X,N}(x,n) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, & n=x=0 \\ (1-\pi) \frac{\beta^\alpha e^{-\lambda}}{\Gamma(\alpha)} \frac{\lambda^n}{n!} x^{\alpha-1} e^{-\beta x} \left[1 + \omega \left(e^{-\delta n} - \pi - (1-\pi)e^{\lambda(e^{-\delta}-1)} \right) \phi(x) \right], & n \geq 1, x > 0. \end{cases}$$

iii) If $N \sim NB(r, p)$, then the compound Negative Binomial–Gamma pdf results as

$$f_{X,N}(x, n) = \begin{cases} p^r, & n = x = 0 \\ \frac{\beta^\alpha p^r \Gamma(r+n)}{\Gamma(\alpha) \Gamma(r) n!} q^n x^{\alpha-1} e^{-\beta x} \left[1 + \omega \left(e^{-\delta n} - \frac{p^r}{(1-qe^{-\delta})^r} \right) \phi(x) \right], & n \geq 1, x > 0 \end{cases} .$$

iv) If $N \sim ZINB(r, p, \pi)$, then the compound zero inflated Negative Binomial–Gamma pdf is

$$f_{X,N}(x, n) = \begin{cases} \pi + (1 - \pi) p^r, & n = x = 0 \\ (1 - \pi) \frac{\beta^\alpha p^r \Gamma(r+n)}{\Gamma(\alpha) \Gamma(r) n!} q^n x^{\alpha-1} e^{-\beta x} \left[1 + \omega \left(e^{-\delta n} - \pi - \frac{(1-\pi)p^r}{(1-qe^{-\delta})^r} \right) \phi(x) \right], & n \geq 1, x > 0. \end{cases}$$

To simulate values from such compound distributions by inversion, we use formula (7) of the conditional cdf under the assumptions that $Y \sim Ga(\alpha, \beta)$ and $\phi(x) = e^{-\gamma x} - \left(\frac{\beta}{\beta + \gamma}\right)^\alpha$. We have

$$\begin{aligned} \int_0^x f(y) \phi(y) dy &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^x y^{\alpha-1} e^{-\beta y} \left(e^{-\gamma y} - \left(\frac{\beta}{\beta + \gamma}\right)^\alpha \right) dy \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left[\int_0^x \left(y^{\alpha-1} e^{-(\beta+\gamma)y} \right) dy - \left(\frac{\beta}{\beta + \gamma}\right)^\alpha \int_0^x y^{\alpha-1} e^{-\beta y} dy \right], \end{aligned}$$

hence, letting $F_{Ga(\alpha, \beta)}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^x y^{\alpha-1} e^{-\beta y} dy$ denote the $Ga(\alpha, \beta)$ cdf, this yields for $n \geq 1, x > 0$,

$$F_{X|N=n}(x) = \left[1 - \omega \psi(n) \left(\frac{\beta}{\beta + \gamma}\right)^\alpha \right] F_{Ga(\alpha, \beta)}(x) + \omega \psi(n) \left(\frac{\beta}{\beta + \gamma}\right)^\alpha F_{Ga(\alpha, \beta + \gamma)}(x).$$

Therefore, as discussed before, to simulate a pair (n, x) , we first simulate the value n from the distribution of N , and if $n \geq 1$, we generate an uniform $U(0, 1)$ value u and solve the equation $F_{X|N=n}(x) = u$ for x .

3 Simulation Study

To evaluate our proposed estimation procedure, we summarize the results of a simulation study. We compare the Mean Square Error (MSE) of the estimated parameters associated to the different bivariate Sarmanov distributions that we have analyzed in the previous sections for modeling the dependence between claims frequency and claims average severity.

We generated 500 bivariate samples of sizes $K = 500$, $K = 5000$ and $K = 50000$ from the following compound Sarmanov models: Poisson–Gamma (CPG), Negative Binomial–Gamma (CNBG), Zero Inflated Poisson–Gamma (CZIPG) and Zero Inflated Negative Binomial–Gamma (CZINBG). We have selected different parameters for the analyzed distribution such that the expected number of claims is around 0.1 or 0.2. In all the simulated models, we assumed the same parameters for the Gamma marginal distribution: shape $\alpha = 0.3$ and rate $\beta = 0.0006$. Concerning the frequency and dependence parameters, we used those shown in Table 1, considering four distinct cases for each model, denoted M1, ..., M4; in this table, we also show the values of the correlation coefficient defined in (6). In general, the analyzed models present low although statistically significant correlation. In fact, we wanted to check if our estimation procedure worked with these low correlations.

In Tables 2 and 3, we show the results of the MSE divided by the true corresponding parameter in each case, i.e. we show a relativized MSE. The estimated parameters for each sample are obtained using the procedure described in Subsection 2.3. We observe that the relativized MSE decreases when the sample size increases, except for some cases of Gamma parameters of the CZIP distribution with mean 0.1. In these cases, the number of nonzero values is very small and some considerable errors associated with some random selected samples can be found. To conclude, we can affirm that our proposed procedure works well when we have large samples, for example in most insurance database. Also, the runtime is fast, to obtain 500 replicates with $n = 50,000$ we need around 1 hour (i7-6700 CPU, 3.40GHz).

Table 1: Parameters of the bivariate compound Sarmanov models. The Gamma parameters are the same in all the cases: $\alpha = 0.3$ and $\beta = 0.0006$. Dependence bounds between parentheses.

	λ		ω (-26.85,3.25)	
CPG-M1	0.20		-7.00	
CPG-M2	0.20		3.00	
	λ		ω (-25.99,3.15)	
CPG-M3	0.10		-7.00	
CPG-M4	0.10		3.00	
	r	p	ω (-14.13,3.47)	
CNBG-M1	0.30	0.60	-12.00	
CNBG-M2	0.30	0.60	3.00	
	r	p	ω (-15.90,3.38)	
CNBG-M3	0.15	0.60	-12.00	
CNBG-M4	0.15	0.60	3.00	
	λ	π	ω (-24.61,3.48)	
CZIPG-M1	0.40	0.50	-12.00	
CZIPG-M2	0.40	0.50	3.00	
	λ	π	ω (-24.61,3.48)	
CZIPG-M3	0.20	0.50	-12.00	
CZIPG-M4	0.20	0.50	3.00	
	r	p	π	ω (-8.95,4.05)
CZINBG-M1	0.30	0.43	0.50	-8.00
CZINBG-M2	0.30	0.43	0.50	3.00
	r	p	π	ω (-14.13,3.47)
CZINBG-M3	0.15	0.60	0.50	-8.00
CZINBG-M4	0.15	0.60	0.50	3.00

4 Numerical example

We now analyze a data set of auto insurance policyholders of an international company. This data set contains a sample of $K = 99,972$ Spanish insureds. We assume that they have a homogeneous risk profile. For each individual we have information on the number and the average cost of claims. Our aim is to fit the bivariate Sarmanov distribution and to check the effect of dependence between frequency and severity on the risk premium.

In Table 4, we display results of the initial analysis that consisted in obtaining the basic descriptives and estimated initial parameters for the marginal distributions assuming independence.

Table 2: Results of MSE divided by the true value of the parameter for compound Poisson-Gamma (CPG) and for compound Negative Binomial-Gamma (CNBG).

		Poisson		Gamma		Dependece
		λ		α	β	ω
CPG-M1	K=500	0.106		0.147	0.406	1.100
	K=5000	0.045		0.091	0.323	1.054
	K=50000	0.034		0.107	0.335	1.007
CPG-M2	K=500	0.107		0.140	0.326	2.254
	K=5000	0.034		0.052	0.130	1.171
	K=50000	0.012		0.030	0.076	1.019
CPG-M3	K=500	0.151		0.198	0.444	1.572
	K=5000	0.057		0.114	0.289	1.052
	K=50000	0.029		0.094	0.271	1.013
CPG-M4	K=500	0.150		0.192	0.403	3.905
	K=5000	0.048		0.071	0.152	1.651
	K=50000	0.017		0.031	0.063	1.063
		Negative Binomial		Gamma		Dependece
		r	p	α	β	ω
CNBG-M1	K=500	0.436	0.139	0.379	0.468	1.028
	K=5000	0.106	0.046	0.233	0.322	1.014
	K=50000	0.031	0.016	0.188	0.277	1.002
CNBG-M2	K=500	0.421	0.138	0.328	0.362	2.867
	K=5000	0.107	0.045	0.125	0.142	1.343
	K=50000	0.015	0.008	0.029	0.050	1.074
CNBG-M3	K=500	0.596	0.172	0.520	0.546	1.049
	K=5000	0.134	0.059	0.241	0.326	1.026
	K=50000	0.043	0.019	0.196	0.269	1.013
CNBG-M4	K=500	0.588	0.173	0.484	0.473	3.490
	K=5000	0.133	0.058	0.149	0.168	1.559
	K=50000	0.041	0.018	0.067	0.084	1.106

Table 3: Results of MSE divided by the true value of the parameter for compound Zero Inflated Poisson-Gamma (CZIPG) and for compound Zero Inflated Negative Binomial-Gamma (CZINBG).

		Zero Inflated Poisson		Gamma		Dependece	
		λ	π	α	β	ω	
CZIPG-M1	K=500	0.626	0.552	0.145	0.485	1.043	
	K=5000	0.507	0.509	0.087	0.361	0.996	
	K=50000	0.502	0.500	0.071	0.131	0.971	
CZIPG-M2	K=500	0.629	0.552	0.158	0.435	1.746	
	K=5000	0.505	0.509	0.133	0.486	1.143	
	K=50000	0.504	0.501	0.144	0.563	0.974	
CZIPG-M3	K=500	0.865	7.846	0.292	0.678	1.141	
	K=5000	0.524	0.525	0.075	0.220	0.997	
	K=50000	0.503	0.503	0.062	0.164	0.973	
CZIPG-M4	K=500	0.866	7.846	0.250	0.531	3.457	
	K=5000	0.525	0.524	0.133	0.449	1.397	
	K=50000	0.504	0.503	0.139	0.507	1.030	
		Zero Inflated Negative Binomial			Gamma		Dependece
		r	p	π	α	β	ω
CZINBG-M1	K=500	0.032	0.006	0.009	0.037	0.000	10.381
	K=5000	0.008	0.001	0.003	0.003	0.000	8.664
	K=50000	0.006	0.000	0.003	0.002	0.000	8.309
CZINBG-M2	K=500	0.032	0.006	0.010	0.034	0.000	14.785
	K=5000	0.008	0.001	0.003	0.005	0.000	4.917
	K=50000	0.006	0.000	0.003	0.005	0.000	3.203
CZINBG-M3	K=500	0.073	0.057	0.037	0.072	0.000	9.802
	K=5000	0.055	0.015	0.041	0.005	0.000	9.172
	K=50000	0.043	0.015	0.049	0.007	0.000	8.168
CZINBG-M4	K=500	0.076	0.051	0.036	0.072	0.000	25.638
	K=5000	0.055	0.015	0.041	0.006	0.000	6.879
	K=50000	0.051	0.010	0.046	0.006	0.000	3.694

At the top of this table, we present the analysis of the number of claims. From the values of the Chi-square statistic we can see that the best adjustments are obtained with the NB and ZINB distributions, being somewhat better for the NB. Below the double line in Table 4, we show the basic descriptive statistics for the average cost of claims, together with the estimated parameters of the Gamma distribution for this variable. We compared the goodness of fit of the Gamma and Log-Normal distributions for the average severity and obtained that the best fit is provided by Gamma.

Table 4: Results of basic descriptive analysis and initial parameters for marginal distributions.

	Po	NB	ZIPo	ZINB	
Initial Parameters	$\lambda = 0.0887$	$r = 0.3171$ $p = 0.7814$	$\lambda = 0.3647$ $\pi = 0.7567$	$r = 11.1344$ $p = 0.9705$ $\pi = 0.7374$	
Frequency	TRUE				
0	92538.00	91482.28	92524.63	92538.00	92537.99
1	6166.00	8118.58	6285.65	6160.47	6172.32
2	1122.00	360.24	950.48	1123.51	1103.16
3	125.00	10.66	170.11	136.60	142.28
4	18.00	0.24	32.81	12.46	14.81
5	3.00	0.00	1.73	0.06	0.11
Chi-Square	99972.00	6761.20	52.81	152.02	77.09
Gamma					
Initial Parameters		$\alpha = 0.1881$ $\beta = 0.0003$			
Severity	Mean	Median	STDEV	Skewness	
	685.63	441.00	1580.81	15.73	

The Pearson correlation coefficient between the frequency and severity is 0.4152.

Table 5 contains the results of the estimated parameters for the bivariate Sarmanov CNBG and CZINBG; as expected, given the results in Table 4, the results for CNPG and CZIPG were worse. From the values of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), we note that the best fit is obtained with CZINBG, although the difference from the CNBG model is minimal. In both cases, we obtain a positive and statistically significant dependence between the frequency and average severity of claims. Furthermore, the dependence parameter is within the interval defined in (5), which indicates that the estimated Sarmanov models work. The effect of this dependence on risk premium is analyzed below.

4.1 Effect on pure and risk premiums

In insurance, the pure premium is calculated as the expected cost of the reported claims, i.e. $\mathbb{E}S = \mathbb{E}[NX]$ in our case, while the risk premium commonly consists of adding the effect of the dispersion of this variable, i.e. $VarS = Var[NX]$. For example, if we use the standard deviation criterion, we obtain the risk premium formula $\rho_R = \mathbb{E}S + \delta\sqrt{VarS}$, where $\delta > 0$ is a loading constant. Therefore, for calculating this premium we need to know the distribution of S and especially its first two moments. For our numerical example, we show in Table 6 the pure and risk premiums obtained if N and X were independent (i.e., $\omega = 0$), and by assuming that N and X are Sarmanov distributed with $\omega > 0$ and with $\mathbb{E}S$ and $VarS$ given in Proposition 2. We used the models whose parameters

Table 5: Estimation results of bivariate Sarmanov distributions for CNBG and CZINBG models

	CNBG	CZINBG
r	0.2814	11.1136
p	0.7602	0.9709
pi	0.0000	0.7337
α	0.2753	0.2742
β	0.0004	0.0004
ω	1.3386*	1.3996*
$\text{Min}(\omega)$	-26.1225	-26.2315
$\text{Max}(\omega)$	3.4484	3.4614
$\text{corr}(X, N)$	0.4159	0.4208
AIC	157522.9	157479.3
BIC	157537.9	157497.3

*Statistically significant positive dependence at 99% confidence level.

are shown in Table 5 and assumed $\delta = 0$ (pure premium) and $\delta = 1$. If we compare the premiums without and with dependence, we obtain that, for the CNBG model, with $\delta = 0$ the first one is approximately 0.57% smaller than the second, with $\delta = 1$ this same percentage is approximately 0.72% and, furthermore, when $\delta \rightarrow \infty$ the percentage increases to 1.48%. For the CZINBG model, these percentages are 0.56, 0.67% and 1.38%, respectively. If we consider that these percentages represent a loss per insured, the total losses could be large and the risk of insolvency will increase.

Table 6: Premiums obtained with CNBG and CZINBG models using $\omega = 0$ and $\omega > 0$.

	$\delta = 0$		$\delta = 1$	
	CNBG	CZINBG	CNBG	CZINBG
ρ_R with $\omega = 0$	61.0930	60.8068	580.4958	574.8728
ρ_R with $\omega > 0$	61.4424	61.1454	584.6742	571.0315

5 Conclusions

In this paper, we have shown how Sarmanov distribution allows us to mix continuous and discrete marginal distributions and to model their dependence. Specifically, we have obtained four bivariate particular cases where we assumed the Gamma distribution for the continuous marginal, and Poisson, Zero Inflated Poisson, Negative Binomial and, respectively, Zero Inflated Negative Binomial distribution for the discrete marginal. Furthermore, a two part maximum likelihood estimation method was proposed and evaluated using a simulation study. We concluded that our proposed method is consistent in terms of the MSE of the estimated parameters for the four proposed particular cases.

As a direct application, we used our model to introduce dependence between the frequency and severity of claims in the collective model. We numerically illustrated this on an auto insurance data set, for which we obtained low, but significant positive dependence between frequency and severity. We concluded that with our model, this dependence between frequency and severity leads

to an increase in premiums that could improve the company's solvency, reducing hence the ruin probability.

6 Appendix

Proof of Proposition 1. We omit the proof of (i)-(iii) being very simple. Also, clearly $\Pr(S = 0) = p(0)$, while for $s > 0$, we have

$$\begin{aligned} F_S(s) &= p(0) + \sum_{n \geq 1} p(n) \Pr(nX \leq s | N = n) \\ &= p(0) + \sum_{n \geq 1} p(n) F_{X|N=n} \left(\frac{s}{n} \right), \end{aligned}$$

hence

$$f_S(s) = \sum_{n \geq 1} \frac{p(n)}{n} f_{X|N=n} \left(\frac{s}{n} \right), s > 0,$$

and inserting here the formula of $f_{X|N=n}$ immediately yields the result in (iv). \square

Proof of Proposition 2. The expected value results easily from

$$\begin{aligned} \mathbb{E}S &= \mathbb{E}[NX] = \sum_{n \geq 1} \int_0^\infty nyp(n) f(y) (1 + \omega\psi(n)\phi(y)) dy \\ &= \mathbb{E}N\mathbb{E}Y + \omega \sum_{n \geq 1} np(n) \psi(n) \int_0^\infty yf(y) \phi(y) dy. \end{aligned}$$

For the variance, we start with

$$\begin{aligned} \mathbb{E}[S^2] &= \mathbb{E}[N^2X^2] = \sum_{n \geq 1} \int_0^\infty n^2y^2p(n) f(y) (1 + \omega\psi(n)\phi(y)) dy \\ &= \mathbb{E}[N^2] \mathbb{E}[Y^2] + \omega \sum_{n \geq 1} n^2p(n) \psi(n) \int_0^\infty y^2f(y) \phi(y) dy \\ &= \mathbb{E}[N^2] \mathbb{E}[Y^2] + \omega \mathbb{E}[N^2\psi(N)] \mathbb{E}[Y^2\phi(Y)]. \end{aligned}$$

Therefore, the variance follows from

$$\begin{aligned} \text{Var}S &= \mathbb{E}[S^2] - \mathbb{E}^2[S] = \mathbb{E}[N^2] \mathbb{E}[Y^2] + \omega \mathbb{E}[N^2\psi(N)] \mathbb{E}[Y^2\phi(Y)] \\ &\quad - (\mathbb{E}^2N \mathbb{E}^2Y + 2\omega \mathbb{E}N\mathbb{E}Y\mathbb{E}[N\psi(N)] \mathbb{E}[Y\phi(Y)] + \omega^2 \mathbb{E}^2[N\psi(N)] \mathbb{E}^2[Y\phi(Y)]) \\ &= (\mathbb{E}[N^2] - \mathbb{E}^2N) \mathbb{E}[Y^2] + \mathbb{E}^2N (\mathbb{E}[Y^2] - \mathbb{E}^2Y) - \omega^2 \mathbb{E}^2[N\psi(N)] \mathbb{E}^2[Y\phi(Y)] \\ &\quad + \omega (\mathbb{E}[N^2\psi(N)] \mathbb{E}[Y^2\phi(Y)] - 2\mathbb{E}N\mathbb{E}Y\mathbb{E}[N\psi(N)] \mathbb{E}[Y\phi(Y)]). \end{aligned}$$

This completes the proof. \square

Proof of Proposition 3. Using (i) from Proposition 1, it is easy to check that

$$\begin{aligned} \mathbb{E}X &= (1 - p(0)) \mathbb{E}Y, \mathbb{E}[X^2] = (1 - p(0)) \mathbb{E}[Y^2], \\ \text{Var}X &= (1 - p(0)) (\text{Var}Y + p(0) (\mathbb{E}Y)^2). \end{aligned}$$

On the other hand, from (1) and Proposition 2 we know that

$$\mathbb{E}[XN] = \mathbb{E}S = \mathbb{E}N\mathbb{E}Y + \omega\mathbb{E}[N\psi(N)]\mathbb{E}[Y\phi(Y)],$$

hence

$$\begin{aligned} \text{cov}(X, N) &= \mathbb{E}[XN] - \mathbb{E}X\mathbb{E}N = \mathbb{E}N\mathbb{E}Y + \omega\mathbb{E}[N\psi(N)]\mathbb{E}[Y\phi(Y)] - (1 - p(0))\mathbb{E}Y\mathbb{E}N \\ &= \omega\mathbb{E}[N\psi(N)]\mathbb{E}[Y\phi(Y)] + p(0)\mathbb{E}Y\mathbb{E}N, \end{aligned}$$

which, together with the above formula of $\text{Var}X$, immediately yields the stated formula of $\text{corr}(X, N)$. This completes the proof. \square

The following lemmas will be needed to prove Proposition 4; although the first lemma is given for the continuous r.v. Y , it holds for any r.v., including a discrete r.v. N , assuming that the involved expected values exist. The proof of this lemma is immediate, hence we omit it.

Lemma 1. *Let Y be some r.v. and let $\psi(x) = e^{-\delta x} - \mathcal{L}_Y(\delta)$ be the corresponding exponential kernel. Then*

$$\mathbb{E}[Y\psi(Y)] = \mathbb{E}\left[Ye^{-\delta Y}\right] - \mathcal{L}_Y(\delta)\mathbb{E}[Y], \quad (9)$$

$$\mathbb{E}[Y^2\psi(Y)] = \mathbb{E}\left[Y^2e^{-\delta Y}\right] - \mathcal{L}_Y(\delta)\mathbb{E}[Y^2]. \quad (10)$$

Lemma 2. *If the r.v. N follows a certain discrete distribution with support \mathbb{N} and \tilde{N} follows the same distribution in the zero inflated form with parameter $\pi \in (0, 1)$, then*

$$\begin{aligned} \mathbb{E}[\tilde{N}\psi(\tilde{N})] &= (1 - \pi)\mathbb{E}[N\psi(N)], \\ \mathbb{E}[\tilde{N}^2\psi(\tilde{N})] &= (1 - \pi)\mathbb{E}[N^2\psi(N)], \end{aligned}$$

where $\psi(N) = e^{-\delta N} - \frac{\mathcal{L}_N(\delta) - p(0)}{1 - p(0)}$ and $\psi(\tilde{N}) = e^{-\delta \tilde{N}} - \frac{\mathcal{L}_{\tilde{N}}(\delta) - \tilde{p}(0)}{1 - \tilde{p}(0)}$, $\tilde{p}(0) = \Pr(\tilde{N} = 0)$.

Proof of Lemma 2. The first formula easily results by applying formula (9),

$$\begin{aligned} \mathbb{E}[\tilde{N}\psi(\tilde{N})] &= \mathbb{E}\left[\tilde{N}e^{-\delta \tilde{N}}\right] - \frac{\mathcal{L}_{\tilde{N}}(\delta) - \tilde{p}(0)}{1 - \tilde{p}(0)}\mathbb{E}\tilde{N} = (1 - \pi)\sum_{n \geq 1} ne^{-\delta n}p(n) \\ &\quad - \frac{\pi + (1 - \pi)\mathcal{L}_N(\delta) - \pi - (1 - \pi)p(0)}{1 - \pi - (1 - \pi)p(0)}(1 - \pi)\mathbb{E}N \\ &= (1 - \pi)\left(\mathbb{E}\left[Ne^{-\delta N}\right] - \frac{\mathcal{L}_N(\delta) - p(0)}{1 - p(0)}\mathbb{E}N\right) = (1 - \pi)\mathbb{E}[N\psi(N)]. \end{aligned}$$

The proof of the second formula is similar, based on formula (10). \square

Proof of Proposition 4. *i)* When $N \sim Po(\lambda)$, from the proof of Lemma 4.1 in Tamraz and Vernic (2018) we know that $\mathbb{E}\left[Ne^{-\delta N}\right] = \lambda e^{\lambda(e^{-\delta} - 1) - \delta}$, hence, applying also formula (9),

$$\mathbb{E}[N\psi(N)] = \lambda e^{\lambda(e^{-\delta} - 1) - \delta} - \lambda \frac{e^{\lambda(e^{-\delta} - 1)} - e^{-\lambda}}{1 - e^{-\lambda}} = \lambda e^{-\lambda} \left(e^{\lambda e^{-\delta} - \delta} - \frac{e^{\lambda e^{-\delta}} - 1}{1 - e^{-\lambda}} \right).$$

For the second formula, we use

$$\begin{aligned}
\mathbb{E} \left[N^2 e^{-\delta N} \right] &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{n^2 \lambda^n}{n!} e^{-\delta n} = e^{-\lambda} \sum_{n=1}^{\infty} \frac{(n-1+1) (\lambda e^{-\delta})^n}{(n-1)!} \\
&= e^{-\lambda} \left[(\lambda e^{-\delta})^2 \sum_{n=2}^{\infty} \frac{(\lambda e^{-\delta})^{n-2}}{(n-2)!} + \lambda e^{-\delta} \sum_{n=1}^{\infty} \frac{(\lambda e^{-\delta})^{n-1}}{(n-1)!} \right] \\
&= e^{-\lambda} \left((\lambda e^{-\delta})^2 e^{\lambda e^{-\delta}} + \lambda e^{-\delta} e^{\lambda e^{-\delta}} \right) = \lambda e^{\lambda e^{-\delta} - \lambda - \delta} (\lambda e^{-\delta} + 1),
\end{aligned}$$

that we insert into (10) and obtain

$$\begin{aligned}
\mathbb{E} [N^2 \psi(N)] &= \lambda e^{\lambda e^{-\delta} - \lambda - \delta} (\lambda e^{-\delta} + 1) - \lambda (\lambda + 1) \frac{e^{\lambda(e^{-\delta}-1)} - e^{-\lambda}}{1 - e^{-\lambda}} \\
&= \lambda e^{-\lambda} \left[e^{\lambda e^{-\delta} - \delta} (\lambda e^{-\delta} + 1) - (\lambda + 1) \frac{e^{\lambda e^{-\delta}} - 1}{1 - e^{-\lambda}} \right].
\end{aligned}$$

ii) The formulas for $N \sim ZIP(\lambda, \pi)$ easily result by applying Lemma 2.

iii) For $N \sim NB(r, p)$, from the proof of Lemma 4.1 from Tamraz and Vernic (2018) we have that

$$\mathbb{E} \left[N e^{-\delta N} \right] = \frac{rqp^r e^{-\delta}}{(1 - qe^{-\delta})^{r+1}}. \text{ Then, based on formula (9),}$$

$$\begin{aligned}
\mathbb{E} [N \psi(N)] &= \frac{rqp^r e^{-\delta}}{(1 - qe^{-\delta})^{r+1}} - \frac{rq \left(\frac{p}{1 - qe^{-\delta}} \right)^r - p^r}{1 - p^r} \\
&= \frac{rqp^r}{(1 - qe^{-\delta})^r} \left(\frac{e^{-\delta}}{1 - qe^{-\delta}} - \frac{1 - (1 - qe^{-\delta})^r}{p(1 - p^r)} \right),
\end{aligned}$$

yielding the first formula. To obtain the second stated formula, we first evaluate

$$\begin{aligned}
\mathbb{E} \left[N^2 e^{-\delta N} \right] &= \sum_{n=0}^{\infty} \frac{\Gamma(r+n)}{n! \Gamma(r)} n^2 p^r (qe^{-\delta})^n = \sum_{n=1}^{\infty} \frac{\Gamma(r+n)(n-1+1)}{(n-1)! \Gamma(r)} p^r (qe^{-\delta})^n \\
&= \frac{p^r}{(1 - qe^{-\delta})^r} \left[\sum_{n=2}^{\infty} \frac{\Gamma(r+n)}{(n-2)! \Gamma(r)} (1 - qe^{-\delta})^r (qe^{-\delta})^n \right. \\
&\quad \left. + \sum_{n=1}^{\infty} \frac{\Gamma(r+n)}{(n-1)! \Gamma(r)} (1 - qe^{-\delta})^r (qe^{-\delta})^n \right] \\
&= \frac{p^r}{(1 - qe^{-\delta})^r} \left[\frac{r(r+1) (qe^{-\delta})^2}{(1 - qe^{-\delta})^2} + \frac{rqp^r}{1 - qe^{-\delta}} \right] \\
&= \frac{rqp^r e^{-\delta} (rqp^r e^{-\delta} + 1)}{(1 - qe^{-\delta})^{r+2}}.
\end{aligned}$$

Therefore, based on (10), we have

$$\begin{aligned}\mathbb{E}[N^2\psi(N)] &= \frac{rqpe^{-\delta}(rqe^{-\delta}+1)}{(1-qe^{-\delta})^{r+2}} - \frac{rq(1+qr)\left(\frac{p}{1-qe^{-\delta}}\right)^r - p^r}{p^2} \\ &= \frac{rqpe^{-\delta}}{(1-qe^{-\delta})^r} \left[\frac{e^{-\delta}(rqe^{-\delta}+1)}{(1-qe^{-\delta})^2} - \frac{1+qr}{p^2} \frac{1 - (1-qe^{-\delta})^r}{1-p^r} \right],\end{aligned}$$

which easily yields the stated formula.

iv) The case $N \sim ZINB(r, p, \pi)$ follows from Lemma 2, which completes the proof. \square

Proof of Proposition 5. We start with

$$\mathbb{E}[Ye^{-\gamma Y}] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty y^{\alpha+1-1} e^{-(\beta+\gamma)y} dy = \frac{\alpha\beta^\alpha}{(\beta+\gamma)^{\alpha+1}},$$

that we insert into (9) and obtain

$$\mathbb{E}[Y\phi(Y)] = \frac{\alpha\beta^\alpha}{(\beta+\gamma)^{\alpha+1}} - \frac{\alpha}{\beta} \left(\frac{\beta}{\beta+\gamma} \right)^\alpha,$$

hence the first stated formula.

Also,

$$\mathbb{E}[Y^2e^{-\gamma Y}] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty y^{\alpha+2-1} e^{-(\beta+\gamma)y} dy = \frac{\alpha(\alpha+1)\beta^\alpha}{(\beta+\gamma)^{\alpha+2}},$$

hence, according to (10),

$$\mathbb{E}[Y^2\phi(Y)] = \frac{\alpha(\alpha+1)\beta^\alpha}{(\beta+\gamma)^{\alpha+2}} - \frac{\alpha(\alpha+1)}{\beta^2} \left(\frac{\beta}{\beta+\gamma} \right)^\alpha,$$

from where we easily obtain the second stated formula. \square

References

- Abdallah, A., Boucher, J., Cossette, H., 2016. Sarmanov family of multivariate distributions for bivariate dynamic claim counts model. *Insurance: Mathematics and Economics* 68, 120–133.
- Bahraoui, Z., Bolancé, C., Pelican, E., Vernic, R., 2015. On the bivariate distribution and copula. an application on insurance data using truncated marginal distributions. *Statistics and Operations Research Transactions, SORT* 39, 209–230.
- Bolancé, C., Vernic, R., 2019. Multivariate count data generalized linear models: Three approaches based on the sarmanovdistribution. *Insurance: Mathematics and Economics* 85, 89–103.
- Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 16, 1190–1208.

- Czado, C., Kastenmeier, R., Brechmann, E.C., Min, A., 2012. A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal* 4, 278–305.
- Erhardt, V., Czado, C., 2012. Modeling dependent yearly claim totals including zero claims in private health insurance. *Scandinavian Actuarial Journal* 2, 106–129.
- Frees, E., Gao, J., Rosenberg, M., 2011. Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal* 15, 377–392.
- Frees, E.W., Wang, P., 2006. Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics* 38, 360–373.
- Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics* 70, 205–215.
- Gschlößl, S., Czado, C., 2007. Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal* 3, 360–373.
- Guo, F., Wang, D., Yang, H., 2017. Asymptotic results for ruin probability in a two-dimensional risk model with stochastic investment returns. *Journal of Computational and Applied Mathematics* 325, 198–221.
- Hua, L., 2015. Tail negative dependence and its applications for aggregate loss modeling. *Insurance: Mathematics and Economics* 61, 135–145.
- Krämer, N., Brechmann, E., Silvestrini, D., Czado, C., 2013. Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics* 53, 829–839.
- Lee, G.Y., Shi, P., 2019. A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics* 87, 115–129.
- Oh, R., Shi, P., Ahn, J.Y., 2019. Bonus-malus premiums under the dependent frequency-severity modeling. *Scandinavian Actuarial Journal* , 1–24.
- Shi, P., Feng, X., Ivantsova, A., 2015. Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics* 64, 417–428.
- Valdez, E.A., Jeong, H., Ahn, J.Y., Park, S., 2018. Generalized linear mixed models for dependent compound risk models. *Variance* .
- Yang, Y., Yuen, K.C., 2016. Finite-time and infinite-time ruin probabilities in a two-dimensional delayed renewal risk model with sarmanov dependent claims. *Journal of Mathematical Analysis and Applications* 442, 600–625.

The logo for UBIREA, featuring the text 'UBIREA' in a bold, sans-serif font. The 'U' and 'B' are white, while 'I', 'R', 'E', and 'A' are blue. The text is set against a white rounded rectangular background.


UBIREA

Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

Universitat de Barcelona

Av. Diagonal, 690 • 08034 Barcelona

WEBSITE: www.ub.edu/irea/ • **CONTACT:** irea@ub.edu

A large, faint, semi-circular graphic composed of many thin, parallel lines, mirroring the design of the UBIREA logo, positioned in the bottom right corner of the page.