MULTIVARIATE MODELLING OF HOUSEHOLD CLAIM FREQUENCIES IN MOTOR THIRD-PARTY LIABILITY INSURANCE

BY

FLORIAN PECHON, JULIEN TRUFIN AND MICHEL DENUIT

ABSTRACT

Actuarial risk classification studies are typically confined to univariate, policy-based analyses: Individual claim frequencies are modelled for a single product, without accounting for the interactions between the different coverages bought by the members of the same household. Now that large amounts of data are available and that the customer's value is at the heart of insurers' strategies, it becomes essential to develop multivariate risk models combining all the products subscribed by the members of the household in order to capture the correlation effects. This paper aims to supplement the standard actuarial policy-based approach with a household-based approach. This makes the actuarial model more complex but also increases the volume of available information which eases and refines forecasting. Possible cross-selling opportunities can also be identified.

KEYWORDS

Multivariate Poisson mixture model, Poisson-LogNormal, Poisson–Gamma, negative binomial, a posteriori risk revaluation.

1. INTRODUCTION AND MOTIVATION

Actuaries now routinely analyse insurance data with the help of Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs), including their mixed model extensions with random effects capturing unexplained heterogeneity: risk selection, a priori classification, experience rating, lapse prediction, etc. can be achieved with these tools. See, e.g., Denuit *et al.* (2007) for a comprehensive account of these regression techniques in non-life insurance.

However, actuarial risk classification studies are typically confined to univariate, policy-based analyses: Individual claim frequencies are modelled for a single product, without accounting for the interactions between the different coverages bought by the members of the same household. We aim here to move from such marginal, policy-based actuarial analyses to joint,

Astin Bulletin 48(3), 969-993. doi: 10.1017/asb.2018.21

household-based risk assessment. By proper inclusion of mixed effects in Poisson model for claim frequencies, the unexplained heterogeneity as well as the dynamic nature of insurance panel data collected for all products issued to the members of the household is accounted for, allowing for periodic revaluations based on previous claim experience.

These latent factors can then be combined using multiline credibility models so that the correlation existing between the different products owned by the members of the same household can be exploited for a posteriori corrections. The proposed multivariate credibility models allow the actuary to gain access to the entire predictive loss distribution. The conditional distribution of the latent factors, given past claims history of every member of the household produces risk predictions, so that each member's predicted claim frequency depends on the numbers of claims filed by the other members of the household. Of course, we mean here technical revaluations, not necessarily commercial ones (such mechanisms may even be prohibited by law in some jurisdictions).

This is not the first proposal for a multivariate credibility model. We refer the reader to the book by Bühlmann and Gisler (2005) for a comprehensive presentation of this topic and its development up to the early 2000s. Since then, this topic has been further studied by several authors, who provided convincing applications of the multivariate credibility models. Englund *et al.* (2008) included claim history for more than one line of business in insurance pricing. They successfully applied their approach to data from two lines of business in a portfolio of a Danish insurance company. See also Englund *et al.* (2009), Frees *et al.* (2010) and Antonio *et al.* (2011). Thuring *et al.* (2012) proposed to use a multiline credibility model to identify prospects for cross-selling insurance products. This global approach allows the insurer to target customers who are expected to report fewer claims with respect to a not yet owned insurance coverage and cross-sell them that specific coverage. Besides correlated latent factors in multiline credibility models, there are other approaches to account for the correlation existing between the different products owned by the household. See, for instance, Shi (2016), Shi and Valdez (2014) and Shi *et al.* (2016) for alternative approaches based on copula modelling. Let us also mention the similarities with credibility models developed for fleets of vehicles. See, e.g., Fardilha *et al.* (2016) and the references therein.

So far, the literature about multivariate credibility models has mostly concentrated on different business lines. In this paper, we consider the same insurance products issued to several members of a household and we investigate the correlation structure of the respective numbers of claims. Barseghyan *et al.* (2016) also considered households, but assessed the dependence structure between the claim experience in motor and home insurance. In their study, the numbers of claims in motor insurance were aggregated over households, and paired with the number of claims in home insurance. Also, Shi *et al.* (2016) considered the Tweedie model for the claim costs (to accommodate the massive zeros) related to different coverages comprised in motor insurance and employed the Gaussian copula to jointly analyse the semi-continuous claim costs in a

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

multilevel context. In particular, they accounted for the correlation among claims from multiple vehicles within the same household (all covered by the same policy in the data studied by these authors).

In the present paper, we restrict our analysis to the compulsory third-party liability motor insurance and study the dependencies between the numbers of claims filed by each member of the household, parents and children. The model proposed for the respective numbers of claims filed by each household member is not new in itself. Following the literature devoted to the multivariate credibility models, we use multivariate Poisson mixtures, with correlated Gamma or Log-Normal random effects (see, e.g., Chapter 2 in Denuit *et al.*, 2007). The main contribution of this paper is more on the application side. The present study is conducted on the motor insurance portfolio of a major insurance company operating in the EU, with more than 1 million insured drivers. This extensive data set allows us to accurately analyse the correlation structure existing between the numbers of claims inside the same household. Contrary to Shi *et al.* (2016), let us notice that each policy of the current dataset is linked to a single car. This means that in case a household owns multiple cars (all insured by the company under consideration), then the dataset records for this household a separate policy for each vehicle. The same Gaussian copula is used, but here to jointly model Poisson mixing (latent) factors, whereas Shi *et al.* (2016) applied it to the observed claim costs.

Our approach can be decomposed as follows. First, we perform a marginal analysis to account for individual risk profiles. Based on a Poisson GAM regression, we predict the expected number of claims for each member of the household using information about the policyholder, his or her vehicle and the characteristics of the contract. In a second stage, we include information about the number of claims reported by the other members of the same household. To this end, we use a multivariate Poisson mixture model with the correlated latent factors inducing the correlation between individual claim histories. It turns out that the association of these latent factors is quite strong on the database used to illustrate this paper. As a consequence, the predictive distributions appear to be sensitive to the claim histories of the other members of the same household: The knowledge of the claim experience at the household level thus refines the prediction of future losses for each member.

The remainder of this paper is organized as follows. In Section 2, we introduce the dataset and define four subpopulations that represent policyholders that are typically encountered in the households: parents and young drivers. In Section 3, we start with a justification for the multivariate modelling before introducing the bivariate model which allows a joint modelling of the parents' number of claims. In Section 4, we generalize the bivariate model to a multivariate model that can capture the main households effects. In Section 5, the parameters of the models are estimated and then used in the applications such as premium corrections, detection of cross-selling opportunities and determination of underwriting rules for young drivers. The final Section 6 briefly concludes the paper. Some technical details are gathered in the appendix.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

2. Composition of the households and description of the dataset

Let us briefly describe the dataset that will be used to support our analysis. Data relate to a European motor third-party liability insurance portfolio observed during calendar years 2011 to 2013. For each policyholder, age, gender and place of residence are available. We also know the power and use of the car (recall that each policy covers a single vehicle and is associated with a main driver). Finally, we also have at our disposal information about the contract: whether premium payment has been splitted (premiums paid annually, semi-annually, quarterly or monthly) and whether material damages are covered in addition to motor TPL. The database also contains the number of claims observed during the three different years and a litigation variable, indicating whether the policyholder has had a failure to pay the premium in due time. Finally, a household code allows us to identify the policyholders belonging to the same household.

In a first step, we account for this information by running a Poisson GAM regression using all the available covariates, with the logarithm of the coverage period as offset. The effect of the continuous covariates age and power, as well as the geographic effect, are captured by splines (with an interaction between gender and age). For more information about this kind of modelling approach, we refer the reader, e.g., to Denuit and Lang (2004).

Even if we know that several policyholders belong to the same household, we do not know the relationship between them (except that they live under the same roof). Notice that only individuals covered by the insurance company are included in the database. No information is available about individuals in the household who have not subscribed an insurance contract with this company. Therefore, the number of kids at home and their respective ages are not necessarily known, nor the presence of a partner, husband or wife. As a consequence, we are not sure about the actual household composition. The only information available is whether different policyholders belong to the same household.

Some assumptions were made to establish the position in the household each policyholder holds. Henceforth, a "kid" is defined as a policyholder aged at most 23 and living with at least one policyholder who is at least 40 years older. The threshold age 23 has been selected to ensure that kids have relative risk of at least 120% compared to the reference level 100% corresponding to age 25, as it can be seen from Figure 1. Restricting to young drivers below age 23 ensures that they report significantly more claims compared to older ages. It is also in line with market practice considering the first 5 years of the driving history as the most dangerous period. The cut-off point for ages corresponding to "kids" has thus been chosen graphically based on Figure 1, as from age 24, the claim frequency tends to become more stable.

A "parent" is defined as a policyholder aged between 40 and 56, referring to the typical age range with children at home possessing a driving license. As it can be seen in Figure 2, the majority of adult policyholders living with a "kid" (as defined previously) are aged between 40 and 56 years. This explains why we have selected this age range for defining the parents. Ages between 40 and 56

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

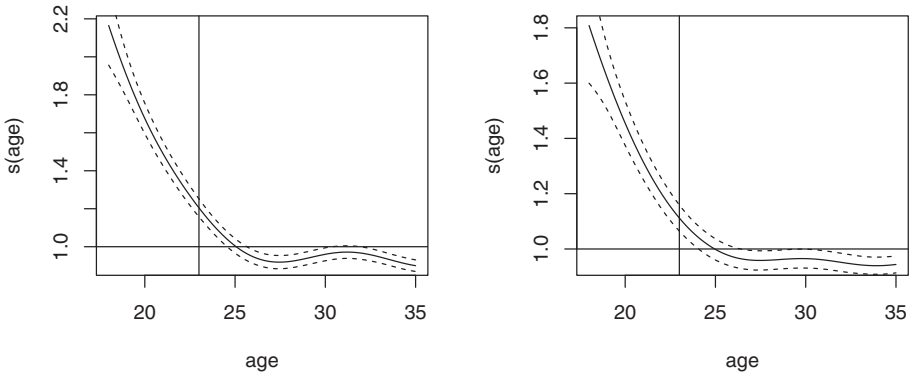

FIGURE 1: Relative impact of age on claim frequencies in a Poisson GAM regression including all covariates. Vertical line at age 23. Left: Males. Right: Females.

also typically correspond to contracts comprising an exclusive driver clause, as it can be seen from Figure 3. By virtue of this clause, the only authorized drivers of the insured vehicle are the policyholder and his or her spouse. This prevents children to drive their parents' car. Parents selecting this clause are rewarded by a significant premium discount.

Figure 3 shows a marked hump around age 45. We suspect this increase in claim frequency to be due to young people driving their parents' car. Integrating the effects of all covariates except age in the offset, we can estimate the impact of age on the claim frequency separately for the policyholders with exclusive driver clause, and for those without this clause. The resulting estimates are shown in Figure 3. Note that in order to avoid side effects at the limiting ages 40 and 56, we included policyholders aged from 38 to 60 in the analysis. Figure 3 shows that inserting the exclusive driver clause in the policy conditions decreases the claim frequency at ages at which the accident hump was visible.

In Figure 3, we also see that female policyholders with an exclusive driver clause aged between 38 and 60 appear to have lower claim frequencies compared to those without the clause. This suggests that the increase in estimated claim frequencies when the exclusive driver clause is absent comes from other drivers that are not the spouse using the policyholder's car, as for instance, young drivers who do not own a car yet and borrow their parents' one.

Since the aim of this paper is to present a household modelling, we define four subpopulations that correspond to the most typical members of households. More specifically, throughout this paper,

- P1 corresponds to "fathers", i.e., men aged between 40 and 56 years comprised in the portfolio;
- P2 corresponds to "mothers", i.e., women aged between 40 and 56 years in the portfolio;
- P3 corresponds to "sons", i.e., young male drivers, aged at most 23, living with a policyholder from P1 and/or P2;

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

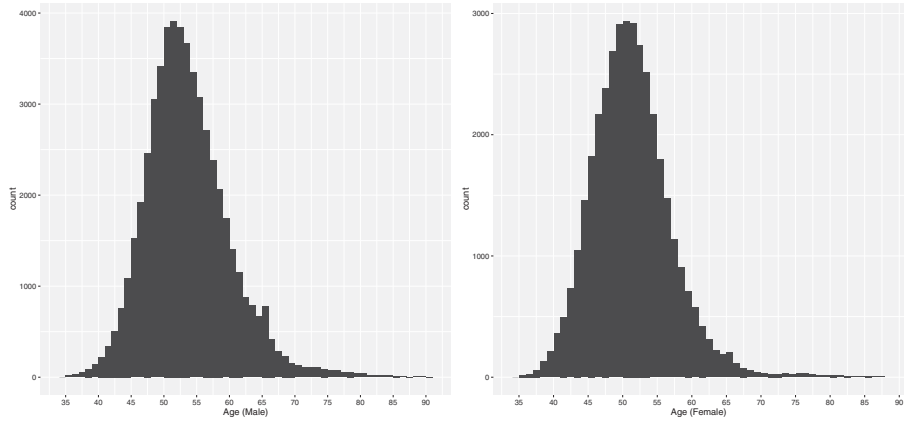

FIGURE 2: Number of male (left)/female (right) policyholders aged a least 35 years in a household comprising at least one policyholder (male or female) aged up to 23 years.

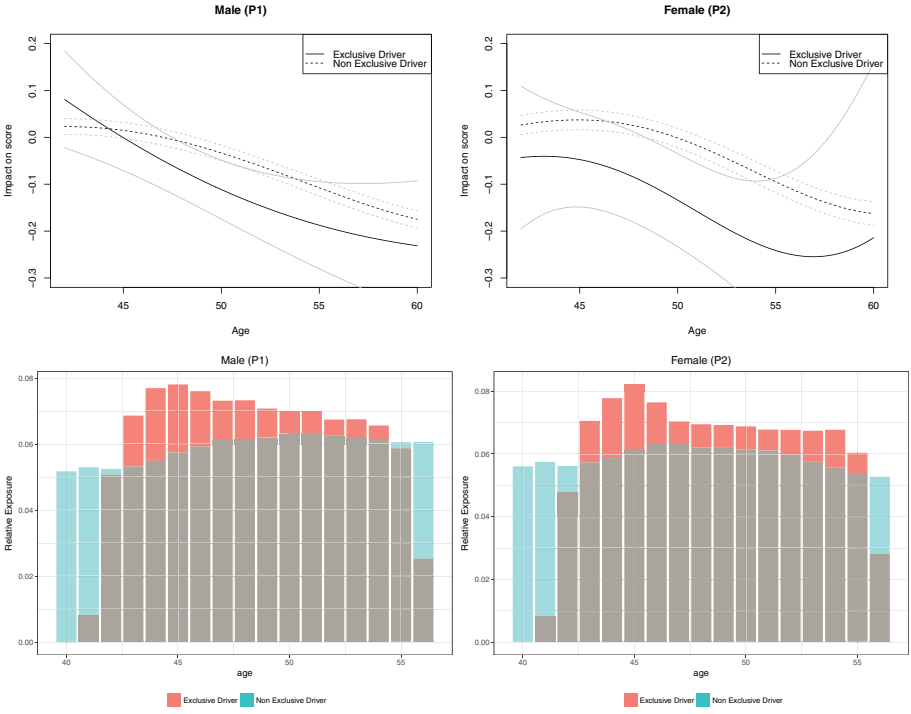

FIGURE 3: Top: Impact of age on claim frequency (on the score scale), for exclusive drivers and for non-exclusive drivers (Left: Male, Right: Female), point estimates and 95% confidence intervals (in lighter lines). Bottom: Distribution of the policyholders' age for exclusive driver (resp. non-exclusive driver) by gender.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

- P4 corresponds to "daughters", i.e., young female drivers, aged at most 23, living with a policyholder from P1 and/or P2.

3. JOINT MODELLING OF PARENTS' CLAIM FREQUENCIES

Let $\mathcal{H}_1$ (resp. $\mathcal{H}_2$) denote the set of all households comprising a member in P1 (resp. P2), i.e., with husband/father (resp. mother/wife) insured. Then, $\mathcal{H}_{12} = \mathcal{H}_1 \cap \mathcal{H}_2$ corresponds to the set of all households with both husband and wife insured. In addition, define the set $\mathcal{H}_{1\setminus 2} = \mathcal{H}_1 \setminus \mathcal{H}_2$ of all households with husband insured, but not his wife, and the set $\mathcal{H}_{2\setminus 1} = \mathcal{H}_2 \setminus \mathcal{H}_1$ of all households with wife insured, but not her husband. Notice that we freely use the terms husband and wife for the ease of exposition, whereas there are now many other forms of cohabitation in addition to marriage, including registered partnership, for instance.

For $h \in \mathcal{H}_1$ (resp. $h \in \mathcal{H}_2$), let N_{ht}^{P1} (resp. N_{ht}^{P2}) be the number of claims filed by husband (resp. wife) during year t. Data are available for years $t = 1, 2, \ldots, T$, with $T = 3$ in our database. For households $h \in \mathcal{H}_{12}$, we observe both N_{ht}^{P1} and N_{ht}^{P2} so that we can study the correlation structure of the spouses' claim frequencies. This is precisely the aim of this section.

Before introducing a joint model for the pairs $(N_{ht}^{P1}, N_{ht}^{P2})$, let us first establish the presence of correlation between these two claim counts. To this end, we work on the basis of the contingency table displayed in Table 1 where we can read the proportions of couples in $\mathcal{H}_{12}$ in each product of risk classes. These risk classes have been created based on quantiles 1/3 and 2/3 on the a priori claim frequencies. We see that the majority of insured couples concentrate along the diagonal (more than 60% of the portfolio), whereas the extreme cases pairing low and high claim frequencies appear to be less common (about 40% of the portfolio). This was expected as the majority of husbands and wives are about the same age and share many characteristics included in the insurance price list.

Using a likelihood ratio test of independence, we clearly reject independence (p-value < 0.001 and $G^2 = 55301.14$). Based on a mixed Poisson construction (see below for a formal definition), we can isolate the correlation not produced by similar observable characteristics. The non-parametric moment estimate for the covariance of the random effects is 0.275956 with confidence interval [0.135659; 0.432074].

Hence, in the following, we use a multivariate Poisson mixture. For more details, we refer the reader, e.g., to Chapters 2 and 6 in Denuit *et al.* (2007). This model is based on the following assumptions (where T denotes the number of observation periods):

1. For $j \in \{1, 2\}$, given $\Theta_h^{Pj} = \theta$, the random variables $N_{h1}^{Pj}, N_{h2}^{Pj}, \ldots, N_{hT}^{Pj}$ are independent, Poisson distributed with respective means $\lambda_{h1}^{Pj}\theta, \lambda_{h2}^{Pj}\theta, \ldots, \lambda_{hT}^{Pj}\theta$.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

TABLE 1

PERCENTAGE OF TOTAL PORTFOLIO EXPOSURE BY RISK PROFILE OF MALE (ROWS) AND FEMALE (COLUMNS) PARENTS.

λ^{P1} \ λ^{P2}	Low	Medium	High
Low	0.2369	0.0828	0.0171
Medium	0.0834	0.1715	0.0828
High	0.0155	0.0829	0.2271

Computed using only households with exactly one male (P1) and one female (P2), regardless if there are any young drivers.

2. Given $(\Theta_h^{P1}, \Theta_h^{P2})$, the random variables $N_{h1}^{P1}, N_{h2}^{P1}, \ldots, N_{hT}^{P1}$ and $N_{h1}^{P2}, N_{h2}^{P2}, \ldots, N_{hT}^{P2}$ are independent.
3. The pairs $(\Theta_h^{P1}, \Theta_h^{P2})$ are independent and identically distributed, with common joint probability density function $f_{\boldsymbol{\Theta}}$, $\mathrm{E}[\Theta_h^{Pj}] = 1$ for $j \in \{1, 2\}$ and variance–covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{\Theta}} = \begin{pmatrix} (\sigma_{\boldsymbol{\Theta}}^{P1})^2 & \sigma_{\boldsymbol{\Theta}}^{P:P} \\ \sigma_{\boldsymbol{\Theta}}^{P:P} & (\sigma_{\boldsymbol{\Theta}}^{P2})^2 \end{pmatrix}.$$

In the remainder of this paper, we also use the correlation coefficient $\rho_{\boldsymbol{\Theta}}^{P:P} = \frac{\sigma_{\boldsymbol{\Theta}}^{P:P}}{\sigma_{\boldsymbol{\Theta}}^{P1}\sigma_{\boldsymbol{\Theta}}^{P2}}$ in addition to the covariance $\sigma_{\boldsymbol{\Theta}}^{P:P}$.

3.1. Bivariate Poisson-LogNormal model

Let us assume that $(\log \Theta_h^{P1}, \log \Theta_h^{P2})$ obeys the bivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}_{\log \boldsymbol{\Theta}}$, where $\mu_j = -\frac{(\sigma_{\boldsymbol{\Theta}}^{Pj})^2}{2}$, $j \in \{1, 2\}$, so that both Θ_h^{P1} and Θ_h^{P2} have unit mean and where

$$\boldsymbol{\Sigma}_{\log \boldsymbol{\Theta}} = \begin{pmatrix} (\sigma_{\log \boldsymbol{\Theta}}^{P1})^2 & \sigma_{\log \boldsymbol{\Theta}}^{P:P} \\ \sigma_{\log \boldsymbol{\Theta}}^{P:P} & (\sigma_{\log \boldsymbol{\Theta}}^{P2})^2 \end{pmatrix}.$$

Let us also introduce the correlation between $\log \Theta_h^{P1}$ and $\log \Theta_h^{P2}$, namely

$$\rho_{\log \boldsymbol{\Theta}}^{P:P} = \frac{\sigma_{\log \boldsymbol{\Theta}}^{P:P}}{\sigma_{\log \boldsymbol{\Theta}}^{P1}\sigma_{\log \boldsymbol{\Theta}}^{P2}}.$$

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

This implies the following variance–covariance matrix for the random effects $(\Theta_h^{P1}, \Theta_h^{P2})$

$$\mathbf{\Sigma}_{\mathbf{\Theta}} = \begin{pmatrix} \exp(\sigma^{P1}_{\log\mathbf{\Theta}})^2 - 1 & \exp\sigma^{P:P}_{\log\mathbf{\Theta}} - 1 \\ \exp\sigma^{P:P}_{\log\mathbf{\Theta}} - 1 & \exp(\sigma^{P2}_{\log\mathbf{\Theta}})^2 - 1 \end{pmatrix}.$$

Also, the correlation between the random effects can be reexpressed as

$$\rho^{P:P}_{\mathbf{\Theta}} = \frac{\exp\sigma^{P:P}_{\log\mathbf{\Theta}} - 1}{\sqrt{(\exp(\sigma^{P1}_{\log\mathbf{\Theta}})^2 - 1)(\exp(\sigma^{P2}_{\log\mathbf{\Theta}})^2 - 1)}}.$$

As the model is fully specified, we can rely on the maximum likelihood approach to estimate both variances and the correlation coefficient which define the variance–covariance matrix of $(\log\Theta_h^{P1}, \log\Theta_h^{P2})$ from which we can thereafter deduce the variance–covariance matrix of $(\Theta_h^{P1}, \Theta_h^{P2})$. Henceforth, let us denote as n_{ht}^{Pj} the realization of N_{ht}^{Pj} recorded in the database and $f_{\Theta^{Pj}}$ the probability density function of Θ_h^{Pj}. The likelihood can be written as

$$\mathcal{L}(\mathbf{\Sigma}) = \mathcal{L}_1 \times \mathcal{L}_2 \times \mathcal{L}_3,$$

where

$$\begin{aligned}
\mathcal{L}_1 &= \prod_{h\in\mathcal{H}_{12}} \mathrm{P}\big[N_{ht}^{Pj} = n_{ht}^{Pj},\ \ t = 1, 2, \ldots, T,\ \ j \in \{1, 2\}\big] \\
&= \prod_{h\in\mathcal{H}_{12}} \int_0^\infty \int_0^\infty \prod_{t=1}^{T} \left(\exp(-\lambda_{ht}^{P1}\theta_1) \frac{(\lambda_{ht}^{P1}\theta_1)^{n_{ht}^{P1}}}{n_{ht}^{P1}!} \exp(-\lambda_{ht}^{P2}\theta_2) \frac{(\lambda_{ht}^{P2}\theta_2)^{n_{ht}^{P2}}}{n_{ht}^{P2}!} \right) \\
&\qquad \times f_{\mathbf{\Theta}}(\theta_1, \theta_2) d\theta_1 d\theta_2, \\
\mathcal{L}_2 &= \prod_{h\in\mathcal{H}_{1\backslash 2}} \mathrm{P}\big[N_{ht}^{P1} = n_{ht}^{P1},\ \ t = 1, 2, \ldots, T\big] \\
&= \prod_{h\in\mathcal{H}_{1\backslash 2}} \int_0^\infty \prod_{t=1}^{T} \exp(-\lambda_{ht}^{P1}\theta_1) \frac{(\lambda_{ht}^{P1}\theta_1)^{n_{ht}^{P1}}}{n_{ht}^{P1}!} f_{\Theta^{P1}}(\theta_1) d\theta_1, \\
\mathcal{L}_3 &= \prod_{h\in\mathcal{H}_{2\backslash 1}} \mathrm{P}\big[N_{ht}^{P2} = n_{ht}^{P2},\ \ t = 1, 2, \ldots, T\big] \\
&= \prod_{h\in\mathcal{H}_{2\backslash 1}} \int_0^\infty \prod_{t=1}^{T} \exp(-\lambda_{ht}^{P2}\theta_2) \frac{(\lambda_{ht}^{P2}\theta_2)^{n_{ht}^{P2}}}{n_{ht}^{P2}!} f_{\Theta^{P2}}(\theta_2) d\theta_2.
\end{aligned}$$

Let $f_{\log\Theta_h^{P1}}$ (resp. $f_{\log\Theta_h^{P2}}$) be the probability density function of the Normal distribution with mean $-(\sigma^{P1}_{\log\mathbf{\Theta}})^2/2$ (resp. $-(\sigma^{P2}_{\log\mathbf{\Theta}})^2/2$) and variance $(\sigma^{P1}_{\log\mathbf{\Theta}})^2$

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

(resp. $(\sigma^{P2}_{\log\boldsymbol{\Theta}})^2$). It is then easy to see that $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ are proportional to

$$\mathcal{L}_1 \propto \prod_{h\in\mathcal{H}_{12}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\lambda^{P1}_{h\bullet}e^{u}-\lambda^{P2}_{h\bullet}e^{v}+un^{P1}_{h\bullet}+vn^{P2}_{h\bullet}} f_{(\log\Theta^{P1}_h,\log\Theta^{P2}_h)}(u,v)dudv,$$

$$\mathcal{L}_2 \propto \prod_{h\in\mathcal{H}_{1\backslash 2}} \int_{-\infty}^{\infty} e^{-\lambda^{P1}_{h\bullet}e^{u}+un^{P1}_{h\bullet}} f_{\log\Theta^{P1}_h}(u)du,$$

$$\mathcal{L}_3 \propto \prod_{h\in\mathcal{H}_{2\backslash 1}} \int_{-\infty}^{\infty} e^{-\lambda^{P2}_{h\bullet}e^{v}+vn^{P2}_{h\bullet}} f_{\log\Theta^{P2}_h}(v)dv,$$

where $\lambda^{P_j}_{h\bullet}=\sum_{t=1}^{T}\lambda^{P_j}_{ht}$ and $n^{P_j}_{h\bullet}=\sum_{t=1}^{T}n^{P_j}_{ht}$ for $j\in\{1,2\}$.

In order to compute the double integrals involved in the log-likelihood, we rely on the Gauss–Hermite quadrature, using the R package MultiGHQuad contributed by Kroeze (2016). The Gauss–Hermite quadrature allows to approximate the integrals of the form $\int_{-\infty}^{\infty} e^{-x^2}f(x)dx$ by linear combinations $\sum_{j=1}^{m}\omega_j f(x_j)$ computed over a grid of $m = k$ (or k^d when the integration is multiple over a domain of dimension d) nodes. The integral thus collapses to a weighted average of m terms. The maximum likelihood estimators can then be found using an optimization algorithm, using as starting values, for instance, the moment estimators. The number of nodes used to approximate these integrals is discussed in Appendix A.1. In the application, we have selected $m = 25$ for $\mathcal{H}_{1\backslash 2}$ and $\mathcal{H}_{2\backslash 1}$ and $m = 25^2$ for $\mathcal{H}_{12}$.

3.2. Bivariate Poisson–Gamma model

In order to challenge the LogNormal assumption, we also consider a different distribution for the random effects. Specifically, we assume here that the random effects Θ^{P1}_h and Θ^{P2}_h are distributed according to the Gamma distributions with unit mean and variances $\mathrm{V}[\Theta^{P1}_h] = a^{P1}$ and $\mathrm{V}[\Theta^{P2}_h] = a^{P2}$. The dependence between Θ^{P1}_h and Θ^{P2}_h is introduced by means of a Gaussian bivariate copula c_R, with correlation parameter R. See, e.g., Denuit *et al.* (2005) for more information about copulas, including the Gaussian one, and associated inference procedures. In case the household consists in only one policyholder, the Poisson–Gamma reduces to the Negative Binomial distribution.

The three variance–covariance parameters are estimated by the maximum likelihood. In the bivariate Poisson–Gamma model with Gaussian copula, the

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

likelihood is proportional to

$$\begin{aligned}
&\mathcal{L}(R, a^{P1}, a^{P2}) \\
&\propto \prod_{h\in\mathcal{H}_{12}} \int_0^\infty \int_0^\infty \exp\big(-\lambda_{h\bullet}^{P1}\theta_1 - \lambda_{h\bullet}^{P2}\theta_2\big)\theta_1^{n_{h\bullet}^{P1}}\theta_2^{n_{h\bullet}^{P2}} c_R(F_{\Theta_h^{P1}}(\theta_1), \\
&\quad F_{\Theta_h^{P2}}(\theta_2))\, f_{\Theta_h^{P1}}(\theta_1)\, f_{\Theta_h^{P2}}(\theta_2) d\theta_1 d\theta_2 \\
&\quad \times \prod_{h\in\mathcal{H}_{1\setminus 2}} \binom{1/a^{P1} + n_{h\bullet}^{P1} - 1}{n_{h\bullet}^{P1}} \left(\frac{\lambda_{h\bullet}^{P1}}{1/a^{P1} + \lambda_{h\bullet}^{P1}}\right)^{n_{h\bullet}^{P1}} \left(\frac{1/a^{P1}}{1/a^{P1} + \lambda_{h\bullet}^{P1}}\right)^{1/a^{P1}} \\
&\quad \times \prod_{h\in\mathcal{H}_{2\setminus 1}} \binom{1/a^{P2} + n_{h\bullet}^{P2} - 1}{n_{h\bullet}^{P2}} \left(\frac{\lambda_{h\bullet}^{P2}}{1/a^{P2} + \lambda_{h\bullet}^{P2}}\right)^{n_{h\bullet}^{P2}} \left(\frac{1/a^{P2}}{1/a^{P2} + \lambda_{h\bullet}^{P2}}\right)^{1/a^{P2}}.
\end{aligned}$$

The numerical evaluation of the double integrals in the log-likelihood can be achieved again by quadrature. However, since the domain of integration is different than in the Poisson-LogNormal case, we will use this time the Gauss–Legendre quadrature, which allows to approximate integrals on the unit interval with the help of the R package mvQuad contributed by Weiser (2016). So, in order to use this quadrature, the double integral is cut into four integrals, which can be reparameterized such that the domain of integration of each of these integrals is $[0, 1] \times [0, 1]$. The Gauss–Legendre quadrature is then applied on each of these four double integrals, with $55^2 = 3,025$ nodes.

3.3. Results

The estimations obtained for both models are displayed in Table 2. Confidence intervals are computed by means of the Delta Method. The nonparametric moment estimates (given e.g., in Denuit *et al.*, 2007, Section 6.2.7) are also given, for the sake of comparison. These estimates have been used as starting values for numerical optimization of the Poisson-LogNormal and Poisson–Gamma likelihoods. The sensitivity of the maximum likelihood estimates with respect to the number of nodes per dimension in the numerical integration is discussed in Appendix A.2.

To choose between the two models, we rely on the Vuong test to assess whether the Poisson-LogNormal model outperforms the Poisson–Gamma one. See Denuit *et al.* (2007) for more details about the Vuong test statistic. The computation of the test statistic yields 12.777 leading to a p-value <0.001. We hereby conclude that the Poisson-LogNormal model outperforms the Poisson–Gamma model. In the following, we continue our analysis with the Poisson-LogNormal specification and we include the children in the household.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

TABLE 2

SUMMARY OF THE ESTIMATES ALONG WITH 95% CONFIDENCE INTERVALS.

	$\widehat{\mathrm{V}}[\Theta^{P1}]$	$\widehat{\mathrm{V}}[\Theta^{P2}]$	$\widehat{\rho}_{\boldsymbol{\Theta}}^{P:P}$
Moment Estimates, Aggregated Data	0.603	0.526	0.490
Moment Estimates, Yearly Data	0.676	0.612	0.370
Bivariate LogNormal Random Effect	0.722	0.670	0.411
95% Confidence Interval	0.718 0.727	0.645 0.695	0.392 0.430
	$\widehat{\mathrm{V}}[\Theta^{P1}]$	$\widehat{\mathrm{V}}[\Theta^{P2}]$	R
Bivariate Gamma Random Effect	0.663	0.588	0.435
95% Confidence Interval	0.661 0.665	0.586 0.591	0.414 0.456

4. INCLUDING CHILDREN IN HOUSEHOLDS

Let $m_{h,3}$ (resp. $m_{h,4}$) be the number of policyholders from P3 (resp. P4) in household h, i.e., the number of sons (resp. daughters) having their own vehicle insured by the company, so that they appear in the database. Further, let $N_{ht}^{P3:j}$, $j = 1, \ldots, m_{h,3}$ (resp. $N_{ht}^{P4:j}$, $j = 1, \ldots, m_{h,4}$) be the number of claims filed by the jth son (resp. daughter) in household h. Notice that we only have information about the individuals who are in the dataset, that is, about those covered by a policy sold by the insurer having providing us with the database. If a person in a household has his or her insurance policy at another insurance company or has no policy, we do not know anything about him or her and hence this person is not considered in the present analysis.

Let us now supplement the model for the numbers of claims filed by the parents with additional assumptions to include the claims filed by their children.

1. For $k \in \{3, 4\}$ and $j \in \{1, \ldots, m_{h,k}\}$, given $\Theta_h^{Pk:j} = \theta$, the random variables $N_{h1}^{Pk:j}, N_{h2}^{Pk:j}, \ldots, N_{hT}^{Pk:j}$ are independent, Poisson distributed with respective means $\lambda_{h1}^{Pk:j}\theta, \lambda_{h2}^{Pk:j}\theta, \ldots, \lambda_{hT}^{Pk:j}\theta$.
2. Given $\Theta_h^{Pk:j}$, $k \in \{3, 4\}$ and $j \in \{1, \ldots, m_{h,k}\}$, the sequences $(N_{h1}^{Pk:j}, N_{h2}^{Pk:j}, \ldots, N_{hT}^{Pk:j})$ are independent for different values of k and j.
3. The random effects $\Theta_h^{Pk:j}$ are LogNormally distributed with the unit mean, and independent for different values of h.
4. Given parents' and children's random effects, the corresponding sequences of yearly numbers of claims are independent.

The dimension of the random effects distribution is equal to the size of the household. More specifically, we denote the variances of the log of the random effects specific to each subpopulation by

$$\mathrm{V}[\log \Theta_h^{Pk}] = (\sigma_{\log \boldsymbol{\Theta}}^{Pk})^2 \text{ for } k \in \{1, 2\}$$

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

and

$$\mathrm{V}[\log \Theta_h^{Pk:j}] = (\sigma_{\log \boldsymbol{\Theta}}^{Pk})^2 \text{ for } k \in \{3,4\} \text{ and } j \in \{1,\dots,m_{h,k}\}.$$

This implies that the variances of the random effects specific to each subpopulation are given by

$$\mathrm{V}[\Theta_h^{Pk}] = \exp((\sigma_{\log \boldsymbol{\Theta}}^{Pk})^2) - 1 := (\sigma_{\boldsymbol{\Theta}}^{Pk})^2 \text{ for } k \in \{1,2\}$$

and

$$\mathrm{V}[\Theta_h^{Pk:j}] = \exp((\sigma_{\log \boldsymbol{\Theta}}^{Pk})^2) - 1 := (\sigma_{\boldsymbol{\Theta}}^{Pk})^2 \text{ for } k \in \{3,4\} \text{ and } j \in \{1,\dots,m_{h,k}\}.$$

Moreover, the correlation matrix between the log of the random effects is assumed to be of the form

$$\boldsymbol{R} = \left(\begin{array}{cc|ccc} 1 & \rho_{\log \boldsymbol{\Theta}}^{P-P} & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \cdots & \rho_{\log \boldsymbol{\Theta}}^{P-CH} \\ \rho_{\log \boldsymbol{\Theta}}^{P-P} & 1 & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \cdots & \rho_{\log \boldsymbol{\Theta}}^{P-CH} \\ \hline \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & 1 & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} & \cdots & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} \\ \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} & 1 & \cdots & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} \\ \cdots & \cdots & \cdots & \cdots & & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} \\ \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{P-CH} & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} & & \rho_{\log \boldsymbol{\Theta}}^{CH-CH} & 1 \end{array}\right),$$

where $\rho_{\log \boldsymbol{\Theta}}^{P-P}$ is the correlation between the log of the two parents' random effects, $\rho_{\log \boldsymbol{\Theta}}^{CH-CH}$ is the correlation between the log of the two children's random effects and $\rho_{\log \boldsymbol{\Theta}}^{P-CH}$ is the correlation between the log of a child and a parent's random effects. The size of this matrix adapts itself to every household, i.e., this matrix will be sized such that each policyholder has a corresponding row (resp. column).

The estimation will be done on the parameters of the multivariate Normal random vector, i.e., on the log scale. Then, we will deduce estimators for the random effects' variance–covariance matrix. In the results hereafter, we will show the estimators of the variance–covariance matrix of the random effects themselves.

There are three parameters to be estimated for the first parents' bloc (one variance for each gender and one correlation), three for the children's bloc (one variance of each gender and one correlation) and one for the parents–children's bloc (one correlation), which sums up to a total of seven parameters.

The log-likelihood is reparameterized so as to change the domain of each parameter to $\mathbb{R}$. In that goal, we take the logarithm of the standard deviations and the logit for the correlations. Then, we proceed along the following three steps:

1. Estimate the "parents' bloc" parameters $\sigma_{\log \boldsymbol{\Theta}}^{P1}$, $\sigma_{\log \boldsymbol{\Theta}}^{P2}$, $\rho_{\log \boldsymbol{\Theta}}^{P-P}$.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

2. Estimate the "children's bloc" parameters $\sigma^{P3}_{\log\boldsymbol{\Theta}}$, $\sigma^{P4}_{\log\boldsymbol{\Theta}}$, $\rho^{CH-CH}_{\log\boldsymbol{\Theta}}$.
3. Estimate the final parameter $\rho^{P-CH}_{\log\boldsymbol{\Theta}}$.

The first step has already been treated previously. The second step only involves policyholders from P3 and P4. Finally, the two previous steps provide initial values for the optimization problem involving the seven parameters.

Let us describe the results obtained for steps 2 and 3. In step 2, both variances are first estimated using moment estimators. These estimates are used as starting values for the maximum likelihood optimization on the children's bloc, along with a correlation coefficient of 0. The optimization run several times with different correlation coefficients as starting value. The following results were found regardless of the starting value for the correlation coefficient: $\widehat{\mathbb{V}}[\Theta^{P3}] = 0.5173634$ with 95% confidence interval [0.4751354, 0.5595914], $\widehat{\mathbb{V}}[\Theta^{P4}] = 0.1428524$ with 95% confidence interval [0.07621682, 0.20948802] and $\widehat{\mathbb{C}\mathrm{orr}}[\Theta^{P3}, \Theta^{P4}] = 0.1727096$ with 95% confidence interval [−1.0257746, 1.3711937]. The latter result shows that the correlation coefficient between children's random effects is not significantly different from 0, so that we set it equal to zero in the following step. Note that the confidence interval is larger than the [−1, 1] interval, which comes from the use of the Delta Method to compute it. In the third step, different starting values were considered for the correlation coefficient $\rho^{P-CH}_{\boldsymbol{\Theta}}$, ranging from −0.5 to 0.5. The optimization found that the maximum of the likelihood was achieved for $\widehat{\rho}^{P-CH}_{\log\boldsymbol{\Theta}} = 0.229569$.

Finally, the optimization run again with all seven parameters, with as initial values the estimates from steps 1–3. The seven parameters characterize the variance–covariance matrix of the logarithm of the random effects (i.e., parameterize the underlying multivariate Normal distribution). Using formulas from Proposition 4.1 given below, we can compute the variance–covariance matrix of the corresponding multivariate LogNormal distribution. Since the correlation between two LogNormally distributed random variables depends on the variances of the underlying Normally distributed random variables, we obtain different correlations for each of the four pairs of parent–child.

Proposition 4.1. *Let* $\boldsymbol{X} = (X_1, \ldots, X_q)$ *be a random vector obeying the multivariate Normal distribution with variances* $V[X_i] = \sigma_i^2$, *correlations* $\mathrm{Corr}[\mathrm{X_i}, \mathrm{X_j}] = \rho_{\mathrm{ij}}$ *(with* $\rho_{ii} = 1$*) and mean vector* $\boldsymbol{\mu} = (-\frac{\sigma_1^2}{2}, \ldots, -\frac{\sigma_q^2}{2})$. *Define* $Y_i = \exp X_i$. *Then,* $\boldsymbol{Y}$ *obeys the multivariate LogNormal distribution with mean vector* $\mathbf{1}$,

$$V[Y_i] = \exp(\sigma_i^2) - 1 \text{ and } \mathrm{Corr}[\mathrm{Y_i}, \mathrm{Y_j}] = \frac{\exp(\rho_{\mathrm{ij}}\sigma_{\mathrm{i}}\sigma_{\mathrm{j}}) - 1}{\sqrt{(\exp(\sigma_{\mathrm{i}}^2) - 1)(\exp(\sigma_{\mathrm{j}}^2) - 1)}}.$$

The final results are shown in Table 3. Let us notice that these results may differ from the previous ones as all the estimators are computed simultaneously. Considering the confidence intervals reported in Table 3, we see that the

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

TABLE 3

ESTIMATED VARIANCE–COVARIANCE PARAMETERS OF THE RANDOM EFFECTS IN THE FINAL MODEL.

	Estimated Variance	Quant. 2.5%	Quant. 97.5%
Θ^{P1}	0.720	0.718	0.723
Θ^{P2}	0.611	0.607	0.615
Θ^{P3}	0.516	0.481	0.551
Θ^{P4}	0.163	0.111	0.216
Pair	Estimated Correlation	Quant. 2.5%	Quant. 97.5%
$(\Theta^{P1}, \Theta^{P2})$	0.430	0.409	0.451
$(\Theta^{P1}, \Theta^{P3})$	0.209	0.175	0.243
$(\Theta^{P1}, \Theta^{P4})$	0.218	0.183	0.254
$(\Theta^{P2}, \Theta^{P3})$	0.212	0.177	0.246
$(\Theta^{P2}, \Theta^{P4})$	0.222	0.185	0.258
$(\Theta^{P3}, \Theta^{P4})$		0	

variances of the random effects significantly differ between the four populations P1–P4. Also, the correlation between husband and wife is significantly larger compared to the correlation between parents and children. The correlations between each pair parent–child now differ because the correlation coefficients not only involve the common Normal correlation but also the marginal variances of the corresponding random effects. They remain nevertheless rather close, in the range 20%–22% in all cases. Notice that the correlation between P3 and P4 has been set to zero. The number of households with at least two young policyholders is too low in our dataset (only 464 households contain multiple young drivers) to be able to estimate the possible existing correlation. However, because we found a 95% confidence interval covering the whole interval $[-1, 1]$, we could have assigned any other value than zero. Because of this limitation of the database, we do not consider examples involving multiple children in the same household in the remainder of this paper.

5. INSURANCE APPLICATIONS

The multivariate modelling inside a household can be useful for various purposes. As we will see, we can use all information about the household's claims to adapt each policyholder's (technical) expected claim frequency. This means that any claim in TPL in the household will change the TPL insurance premium for every member of the household.

Moreover, even in the case where some members of the household do not have an insurance policy with the company, using the multivariate model may help finding interesting cross-selling opportunities. Indeed, the model allows the actuary to find candidates that would appear to be in average less risky than the a priori claim frequency would suggest.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

Finally, one can also use the household's past claims to establish some condition of acceptance for young drivers ensuring that in average the new young policyholders are profitable. This demonstrates the importance of having multiple policyholders from the same household in the portfolio.

5.1. A posteriori corrections

We aim to show how the multivariate model can be used to adapt each individual expected claim frequency using all the household's information. We will start with an example of a household with only two adult policyholders (as in Section 3). The case of a broader household that includes kids (as in Section 4) is considered in the next sections.

We differentiate between three different risk profiles corresponding to the a priori estimated claim frequency: low, medium and high. For P1 and P2, let us make three classes of estimated claim frequencies (on a yearly basis, i.e., as if exposure is 1 year) based on the quantiles 1/3 and 2/3. The claim frequencies labelled as "low", "medium" and "high" will correspond to the quantiles 1/6, 1/2 and 5/6, respectively, which are the medians of each of the three classes. Due to similar characteristics (for example, age, ZIP code), the estimated claim frequencies are related to each other as already noticed in Table 1. This confirms our preliminary analysis conducted on that table.

For the ease of exposition, we assume that this a priori claim frequency is stable for the next 5 years for each policyholder. We also suppose the independence between both random effects, meaning that no correction is applied when the spouse has a claim.

Let us compute the correction to apply to P1, given past claim information of P1 and P2. We get

$$\mathrm{E}\left[\Theta_h^{P1} | N_{h\bullet}^{P1} = n_{h\bullet}^{P1}, N_{h\bullet}^{P2} = n_{h\bullet}^{P2}\right] = \frac{\int_0^\infty \int_0^\infty \theta^{P1} \exp(-\lambda_\bullet^{P1}\theta^{P1} - \lambda_\bullet^{P2}\theta^{P2}) \frac{(\lambda_\bullet^{P1}\theta^{P1})^{n_\bullet^{P1}}}{n_\bullet^{P1}!} \frac{(\lambda_\bullet^{P2}\theta^{P2})^{n_\bullet^{P2}}}{n_\bullet^{P2}!} f_{\boldsymbol{\Theta}}(\theta^{P1}, \theta^{P2}) d\theta^{P1} d\theta^{P2}}{\int_0^\infty \int_0^\infty \exp(-\lambda_\bullet^{P1}\theta^{P1} - \lambda_\bullet^{P2}\theta^{P2}) \frac{(\lambda_\bullet^{P1}\theta^{P1})^{n_\bullet^{P1}}}{n_\bullet^{P1}!} \frac{(\lambda_\bullet^{P2}\theta^{P2})^{n_\bullet^{P2}}}{n_\bullet^{P2}!} f_{\boldsymbol{\Theta}}(\theta^{P1}, \theta^{P2}) d\theta^{P1} d\theta^{P2}}. \tag{5.1}$$

Again, the integrals appearing in (5.1) can be computed using the Gauss–Hermite quadrature. Of course, the correction to apply to P2 is computed in a similar way. As we can see on Figure 4, the a posteriori expected claim frequency of P1 decreases faster than under the independence assumption when both spouses have had no claims. Moreover, in such a case, the decay is even faster when the wife had a higher risk profile. One can also observe that a claim of the wife has consequences on the correction to apply to the husband, even though he had no claim. Finally, when only the husband had a claim, the

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

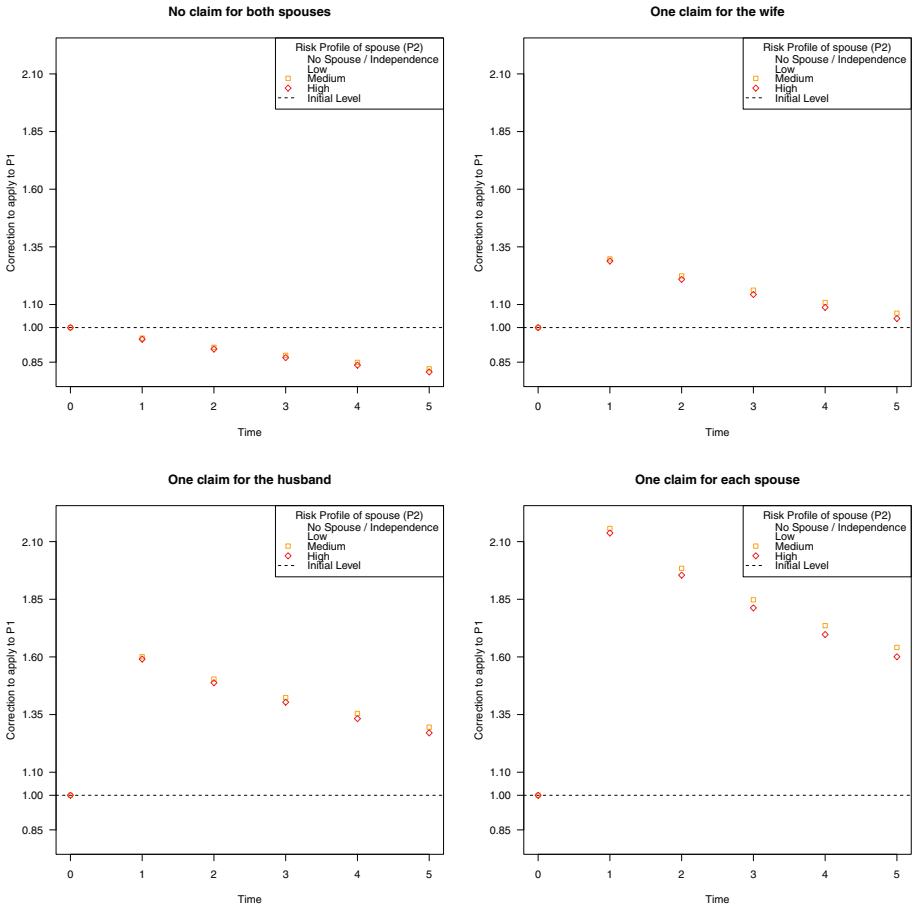

FIGURE 4: Correction to apply to the expected claim frequencies of the husband (low risk profile). Different risk profiles for the wife are assumed (low, medium and high). Four cases are considered: no claim for both spouses, one claim for the wife, one claim for the husband himself and one claim for each spouse. Time is measured in years.

correction is smaller than under the independence hypothesis, followed by a more rapid decay of this correction in the years following the claim.

5.2. Finding cross-selling opportunities

Insurance companies can take advantage of multivariate modelling to find interesting cross-selling opportunities. Since the information about current policyholders inside a household is connected to the potential customers living under that same roof, the company can use that information to compute a (technical) expected claim frequency that can be used to assess the profitability of new customers. This premium, or equivalently the expected cost of the potential future policyholder, may then be used to establish whether this customer would be

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

TABLE 4

EXPECTED VALUE OF Θ_h^{P2} CONDITIONAL TO $N_{h\bullet}^{P1} = n_{h\bullet}^{P1}$, THE NUMBER OF CLAIMS OF THE HUSBAND.

$n_\bullet^{P1}$	0	1	2	3	4
	0.9861	1.2578	1.5950	2.0052	2.4911

profitable. So the multivariate model would help finding cross-selling opportunities. Hereafter, we present two different examples.

5.2.1. *Cross-selling in a household with two adults.* Let us first start with an example with at most two adult policyholders (as in Section 3). The only customer of the household is a 45-year old male who has been in the portfolio for the past 5 years. In the absence of information about past claims recorded by the wife, we can use her husband's past claims to compute her expected claim frequency. Indeed, we can compute the expected value of her random effect, conditional to the information available about her husband's past claims:

$$\begin{aligned}&\mathrm{E}\left[\Theta_h^{P2} | N_{h\bullet}^{P1} = n_{h\bullet}^{P1}\right]\\ &= \frac{\int_0^\infty \int_0^\infty \theta^{P2} \exp(-\lambda_\bullet^{P1}\theta^{P1}) \frac{(\lambda_\bullet^{P1}\theta^{P1})^{n_\bullet^{P1}}}{n_\bullet^{P1}!} f_{\boldsymbol{\Theta}}(\theta^{P1}, \theta^{P2}) d\theta^{P1} d\theta^{P2}}{\int_0^\infty \exp(-\lambda_\bullet^{P1}\theta^{P1}) \frac{(\lambda_\bullet^{P1}\theta^{P1})^{n_\bullet^{P1}}}{n_\bullet^{P1}!} f_\Theta^{P1}(\theta^{P1}) d\theta^{P1}}. \quad (5.2)\end{aligned}$$

Furthermore, let us assume that the husband's estimated claim frequency corresponds to a medium risk profile. Table 4 displays (5.2) for different values of $n_\bullet^{P1}$, i.e., the aggregated number of claims for the husband over the past 5 years. An estimate below 1 indicates that the wife's expected claim frequency is, in average, below the one estimated with the a priori model. Hence, such a potential customer could be targeted by the company.

In case the company can access the information about the number of claims $n_{h\bullet}^{P2}$ of the wife during the last past T^{P2} years, the estimated claim frequency can be computed by using the a priori model aggregated over the past T^{P2} years. In that situation, the conditional expectation is computed as in (5.1). For instance, let us assume that the wife has had an insurance policy during the last 3 years. We assume that similarly to her husband, she has had a medium risk profile in the a priori classification. In that case, Table 5 gives the expected value of Θ_h^{P2} for different different values of $n_{h\bullet}^{P1}$ and $n_{h\bullet}^{P2}$. We can see that when the wife has had no claim in the past 3 years, the estimate of Θ_h^{P2} varies by 25% whether her husband has had none or one claim over the past 5 years. Also, we notice that the a posteriori expected claim frequency is, in average, below the one estimated with the a priori model only in the case where both spouses have had no claim.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

TABLE 5

EXPECTED VALUE OF Θ_h^{P2} CONDITIONAL TO $N_{h\bullet}^{P1} = n_{h\bullet}^{P1}$, THE NUMBER OF CLAIMS OF THE HUSBAND AND TO $N_{h\bullet}^{P2} = n_{h\bullet}^{P2}$, THE NUMBER OF CLAIMS OF THE WIFE.

$n_{h\bullet}^{P1}$ \ $n_{h\bullet}^{P2}$	0	1	2	3	4
0	0.9549	1.5557	2.5045	3.9645	6.1365
1	1.2079	1.9546	3.1177	4.8768	7.4402
2	1.5169	2.4334	3.8389	5.9241	8.8972
3	1.8856	2.9937	4.6631	7.0900	10.4750
4	2.3135	3.6292	5.5744	8.3449	12.1280

5.3. Underwriting rule for young drivers

The multivariate model can also be used for young drivers in a similar way that the one presented in the cross-selling section. In this section, we assume that a young driver wants to get an insurance cover from his or her parents' company. We aim to compare two situations: one in which we only have information about one parent, and the other situation in which we have information about both parents.

Similarly to the cross-selling, we can compute the correction to apply to the young driver, conditional to the claim information of his or her parent(s). We assume that the young driver is a young male (P3). Computations for a young female (P4) are of course similar. Furthermore, both parents are assumed to have a medium risk profile.

When both parents are in the portfolio, we use

$$\mathrm{E}\left[\Theta_h^{P3} | N_{h\bullet}^{P1} = n_{h\bullet}^{P1}, N_{h\bullet}^{P2} = n_{h\bullet}^{P2}\right] = \frac{\int_0^\infty \int_0^\infty \int_0^\infty \theta^{P3} \exp(-\lambda_\bullet^{P1}\theta^{P1} - \lambda_\bullet^{P2}\theta^{P2}) \frac{(\lambda_\bullet^{P1}\theta^{P1})^{n_\bullet^{P1}}}{n_\bullet^{P1}!} \frac{(\lambda_\bullet^{P2}\theta^{P2})^{n_\bullet^{P2}}}{n_\bullet^{P2}!} f_{\boldsymbol{\Theta}}(\theta^{P1}, \theta^{P2}, \theta^{P3}) d\theta^{P1} d\theta^{P2} d\theta^{P3}}{\int_0^\infty \int_0^\infty \exp(-\lambda_\bullet^{P1}\theta^{P1} - \lambda_\bullet^{P2}\theta^{P2}) \frac{(\lambda_\bullet^{P1}\theta^{P1})^{n_\bullet^{P1}}}{n_\bullet^{P1}!} \frac{(\lambda_\bullet^{P2}\theta^{P2})^{n_\bullet^{P2}}}{n_\bullet^{P2}!} f_{\boldsymbol{\Theta}}(\theta^{P1}, \theta^{P2}) d\theta^{P1} d\theta^{P2}},$$

whereas when only the husband ($j = 1$) or the wife ($j = 2$) is in the portfolio, we use

$$\mathrm{E}\left[\Theta_h^{P3} | N_{h\bullet}^{Pj} = n_{h\bullet}^{Pj}\right] = \frac{\int_0^\infty \int_0^\infty \theta^{P3} \exp(-\lambda_\bullet^{Pj}\theta^{Pj}) \frac{(\lambda_\bullet^{Pj}\theta^{Pj})^{n_\bullet^{Pj}}}{n_\bullet^{Pj}!} f_{\boldsymbol{\Theta}}(\theta^{Pj}, \theta^{P3}) d\theta^{Pj} d\theta^{P3}}{\int_0^\infty \exp(-\lambda_\bullet^{Pj}\theta^{Pj}) \frac{(\lambda_\bullet^{Pj}\theta^{Pj})^{n_\bullet^{Pj}}}{n_\bullet^{Pj}!} f_\Theta(\theta^{Pj}) d\theta^{Pj}}.$$

In Figure 5, we can see the conditional expectation of Θ^{P3} in both situations for different past claim scenarios.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

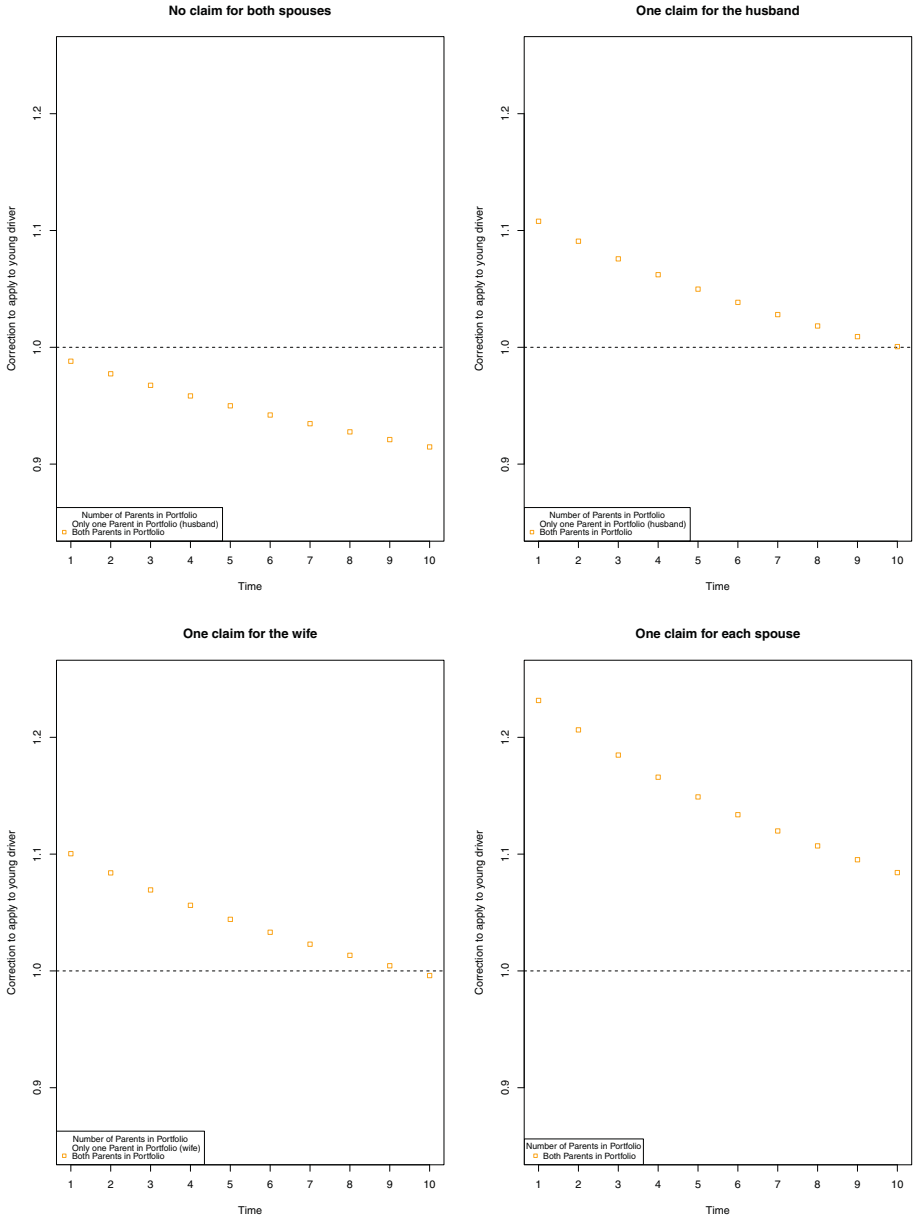

FIGURE 5: Expectation of Θ^{P3} conditional to the number of claims of the father(top)/mother(bottom) (green circle) and of both parents (orange square) throughout time (in years).

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

As we can see, the decay of the estimate of Θ^{P3} is faster when both parents are in the portfolio. We also observe (top-left) that getting an estimate below 0.95 takes 5 years with no claim for both parents, while it takes 10 years with no claim when only the husband is in the portfolio. In addition, we see (top-right and bottom-left) that we need nine claim-free years for both parents for coming back to a level of 1 when a claim occurred while we need more time when only the husband or the wife is in the portfolio (computations show that 20 claim-free years are required).

To sum up, an underwriting rule that would require an a posteriori correction below one would imply no claim for both parents in the observation period or at most one claim in 10 years for both of them. When only one parent is in the portfolio, this rule would require no claim for the parent during the observation period or at most one claim in 21 years.

6. Discussion

In this paper, we have presented an approach that allows to take into account the dependence of the claims' frequency in motor third-party liability insurance of the various policyholders inside a household. It has been shown that the unexplained residual heterogeneity is not independent for members of the same household. The multivariate model is flexible in the sense that it can take into account most households (one or two adult policyholders, with or without young drivers).

The main discovery of this paper is the strength of the positive dependence between policyholders from the same household, showing that a claim from any member of the household will increase in average the estimated a posteriori expected claim frequency for all members of the household. Conversely, for claim-free households, this implies an even lower estimate of the a posteriori claim frequency than in a univariate model.

As this was shown in Section 5, the presented model can be used in practice to use the household's information so as to find cross-selling opportunities, perform underwriting rules for young policyholders or even to correct the expected claim frequencies of policyholders.

From a computational point of view, we note that in order to fasten up the maximum likelihood estimation, one may use a Cholesky decomposition of the multivariate Normal random vector to reparameterize the multiple integrals. Indeed, this will mean that although at each optimization step, the reparameterization is different, the integrand will always include the same density of a centred and reduce multivariate Normal random vector. In this case, it is possible to rely on only one grid in the Gauss quadrature for the whole estimation, instead of computing one grid at each optimization step.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

Of course, the results obtained in this paper refer to the particular data set under study. But as the latter is rather typical for EU markets, we can imagine that similar conclusions would be drawn for other portfolios.

Acknowledgements

Michel Denuit and Florian Pechon gratefully acknowledge the financial support from the contract "Projet d'Actions de Recherche Concertées" No 12/17-045 of the "Communauté française de Belgique", granted by the "Académie universitaire Louvain". Also, the financial support of the AXA Research Fund through the JRI project "Actuarial dynamic approach of customer in P&C" is gratefully acknowledged. We thank our colleagues from AXA Belgium, especially Arnaud Deltour, Mathieu Lambert, Alexis Platteau, Stanislas Roth and Louise Tilmant for interesting discussions that greatly contributed to the success of this research project. Finally, we thank our colleagues from the SMCS for setting us up a comfortable and efficient working environment.

References

Antonio, K., Guillen, M. and Perez Martin, A.M. (2011) Multidimensional credibility: A Bayesian analysis of policyholders holding multiple policies. Working Paper ASE-RI. Available at https://dare.uva.nl/search?identifier=1587a3ef-f8ec-41a7-b61a-ba44d8bfc5ae.

Barseghyan, L., Molinari, F., Morris, D.S. and Teitelbaum, J.C. (2016) Unobserved heterogeneity, experience rating, and insurance demand. Georgetown Law Faculty Publications and Other Works 1127. Available at https://scholarship.law.georgetown.edu/facpub/1127/.

Bühlmann, H. and Gisler, A. (2005) *A Course in Credibility Theory and its Applications*. Berlin: Springer.

Denuit, M., Dhaene, J., Goovaerts, M.J. and Kaas, R. (2005) *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. New York: Wiley.

Denuit, M. and Lang, S. (2004) Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics*, **35**(3), 627–647.

Denuit, M., Marechal, X., Pitrebois, S. and Walhin, J.-F. (2007) *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus–Malus Systems*. New York: Wiley.

Englund, M., Guillen, M., Gustafsson, J., Nielsen, L.H. and Nielsen, J.P. (2008) Multivariate latent risk: A credibility approach. *ASTIN Bulletin*, **38**, 137–146.

Englund, M., Gustafsson, J., Nielsen, L.H. and Nielsen, J.P. (2009) Multidimensional credibility with time effects: An application to commercial business lines. *Journal of Risk and Insurance,* **76**, 443–453.

Fardilha, T., de Lourdes Centeno, M. and Esteves, R. (2016) Tariff systems for fleets of vehicles: A study on the portfolio of Fidelidade. *European Actuarial Journal,* **6**, 331–349.

Frees, E.W.J., Meyers, G. and Cummings, A.D. (2010) Dependent multi-peril ratemaking models. *ASTIN Bulletin,* **40**, 699–726.

Kroeze, K. (2016) MultiGHQuad: Multidimensional Gauss–Hermite quadrature. Available at: http://CRAN.R-project.org/package=MultiGHQuad

Narasimhan, B. (2016) Cubature: Adaptive multivariate integration over hypercubes. Available at: http://cran.r-project.org/package=cubature

Shi, P. (2016) Insurance ratemaking using a copula-based multivariate Tweedie model. *Scandinavian Actuarial Journal,* **2016**, 198–215.

Shi, P., Feng, X. and Boucher, J.P. (2016) Multilevel modeling of insurance claims using copulas. *Annals of Applied Statistics,* **10**, 834–863.

SHI, P. and VALDEZ, E.A. (2014) Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics,* **55**, 18–29.
THURING, F., NIELSEN, J.P., GUILLEN, M. and BOLANCE, C. (2012) Selecting prospects for cross-selling financial products using multivariate credibility. *Expert Systems with Applications,* **39**, 8809–8816.
WEISER, C. (2016) mvQuad: Methods for multivariate quadrature. Available at: http://CRAN.R-project.org/package=mvQuad

FLORIAN PECHON (Corresponding author)
Institute of Statistics
Biostatistics and Actuarial Science
Université catholique de Louvain (UCL)
Louvain-la-Neuve, Belgium
E-Mail: florian.pechon@uclouvain.be

JULIEN TRUFIN
Department of Mathematics
Université Libre de Bruxelles (ULB)
Bruxelles, Belgium
E-Mail: julien.trufin@ulb.ac.be

MICHEL DENUIT
Institute of Statistics, Biostatistics and Actuarial Science
Université catholique de Louvain (UCL)
Louvain-la-Neuve, Belgium
E-Mail: michel.denuit@uclouvain.be

APPENDIX

A.1 Numerical integration

Let us discuss the numerical integration of the bivariate integrals that appear in the likelihood in the bivariate case (i.e., with only parents considered) considered in Section 3.1. We can choose the number of nodes to use, by comparing the Gauss–Hermite approximation with the results obtained with the adaptive multidimensional integration from the R package cubature contributed by Narasimhan (2016). To compute the integral using cubature, the domain $\mathbb{R}^2$ was cut into nine rectangles, on which the integrand was reparameterized so as to have a definite integral on each of the nine subdomains. We then compare the value of the integral to the one obtained with the Gauss–Hermite quadrature for different number of nodes k per dimension. The accuracy of the approximation for the different factors entering the likelihood to be maximized is illustrated on Figure A1, as a function of k and typical values for a priori expected claim frequencies. We can see there that as soon as k reaches 15, the difference between the Gauss–Hermite approximation and the exact value becomes negligible.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

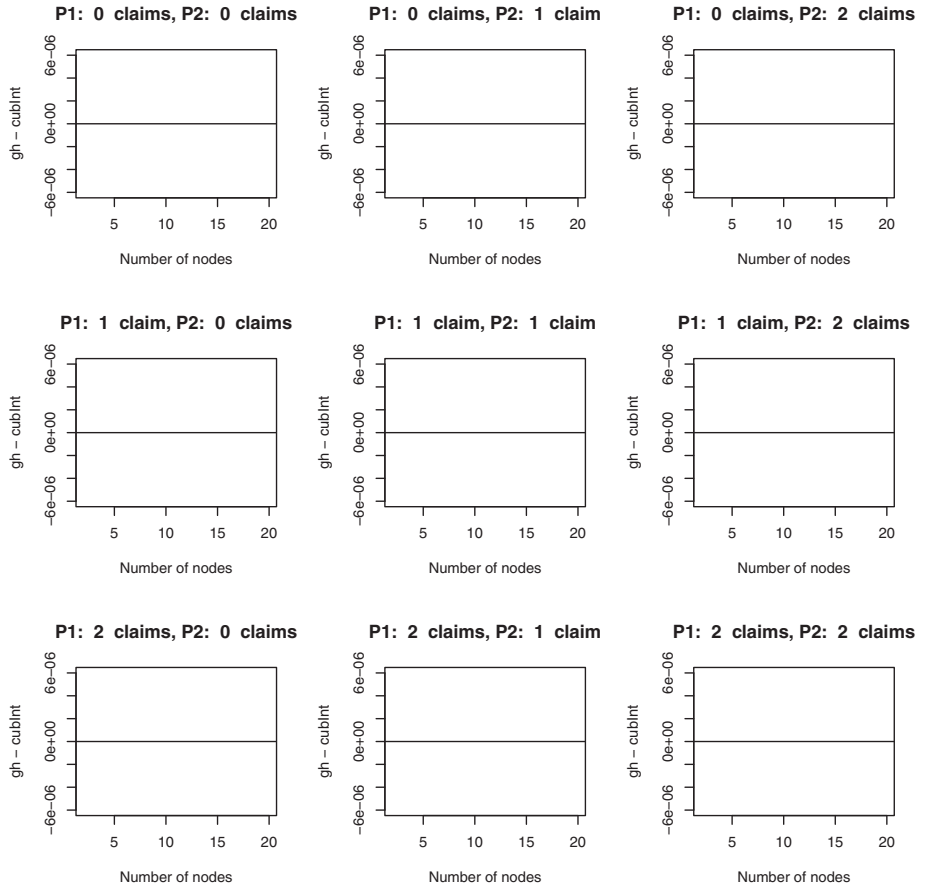

FIGURE A1: Difference between the probabilities approximated by Gauss–Hermite quadrature (denoted as gh) and calculated exactly with adaptive multidimensional integration (denoted as cubInt). From left to right: Number of claims for male varies from 0 to 2. From top to bottom: Number of claims for female varies from 0 to 2. Other parameters were fixed as following: $\lambda_{h\bullet}^{P1} = 0.45$, $\lambda_{h\bullet}^{P2} = 0.36$, $\sigma_{\Theta}^{P1} = 0.8$, $\sigma_{\Theta}^{P2} = 0.7$, $\sigma_{\Theta}^{P:P} = 0.45$.

A.2 Impact of the number of nodes

Let us assess the impact of the number of nodes in the LogNormal and Gamma bivariate cases on the maximum likelihood estimates. To this end, the optimization, in the bivariate case (i.e., with only parents considered) was computed multiple times using 4 to 30 nodes in the LogNormal case, and 10 to 90 by steps of five nodes in the Gamma case. On Figure A2, the estimations of the three parameters are displayed in function of the number of nodes per dimension used to approximate the integral using the Gauss–Hermite quadrature.

It can be seen that there is a very rapid convergence, with stable estimates starting from only 13 nodes. In the Gamma case, Figure A3 displays the estimations as functions of the number of nodes per dimension used to approximate the integral using the Gauss–Legendre quadrature. The graph shows a convergence that appears however to be slower, with stable estimates starting from 55 nodes.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press

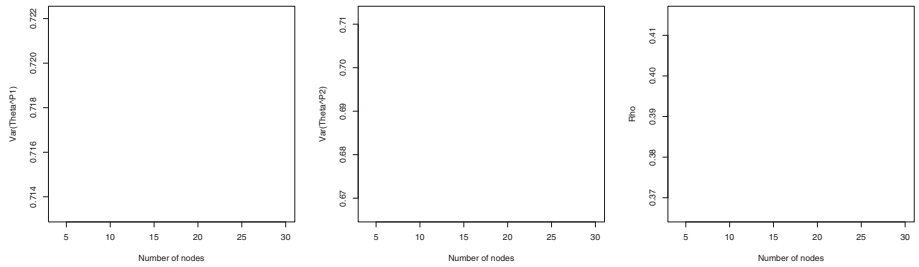

FIGURE A2: Maximum likelihood estimates for the variance of the LogNormal random effects and their correlation in the bivariate LogNormal case for different number of nodes in the Gauss–Hermite quadrature.

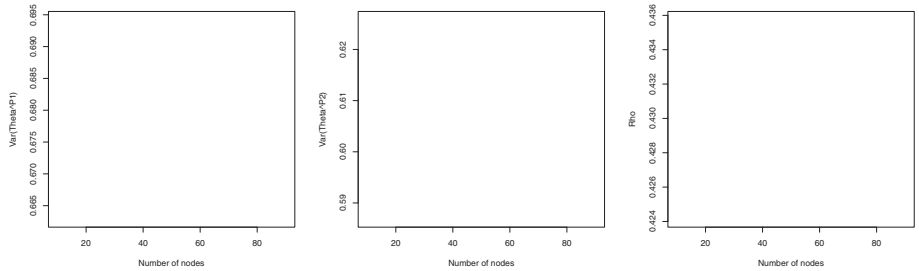

FIGURE A3: MLE for the variance of the Gamma random effects and the dependence parameter R entering the Gaussian copula C_R in the bivariate Gamma case for different number of nodes in the Gauss–Legendre quadrature.

https://doi.org/10.1017/asb.2018.21 Published online by Cambridge University Press