Annals of Actuarial Science (2021), **15**, pp. 82–114
doi:10.1017/S1748499520000160

PAPER

Home and Motor insurance joined at a household level using multivariate credibility

Florian Pechon[1]*, Michel Denuit[1] and Julien Trufin[2]

[1]Institute of Statistics, Biostatistics and Actuarial Science Université catholique de Louvain (UCLouvain), Louvain-la-Neuve, Belgium and [2]Department of Mathematics, Université Libre de Bruxelles (ULB), Bruxelles, Belgium
*Corresponding author. E-mail: florian.pechon@uclouvain.be

(Received 20 May 2019; revised 09 May 2020; accepted 22 May 2020; first published online 06 July 2020)

Abstract

Actuarial ratemaking is usually performed at product and guarantee level, meaning that each product and guarantee is considered in isolation. Moreover, independence between policyholders is generally assumed. In this paper, we propose a multivariate Poisson mixture, with random effects correlated using a hierarchical structure, to accommodate for the dependence that may exist between unobserved risk factors across Home and Motor insurance and between policyholders from the same household. The hierarchical structure accounts for the fact that Home insurance covers the whole household, whereas Motor insurance policies are subscribed by specific policyholders within the household. The model allows to periodically correct the a priori expected claim frequencies using the reported number of claims in any of the considered products. Applications show that the impact of the number of claims reported in Motor insurance on the number of claims expected in Home insurance is larger than the other way around. Moreover, an out-of-sample analysis validates an improved predictive power. Also, the model allows to identify more rapidly the riskiest households.

Keywords: Multivariate credibility; Poisson–LogNormal; A posteriori claim frequency

1. Introduction and Motivation

In property and casualty insurance, it is common to perform ratemaking using a two-step approach. First, policyholders are classified by the insurer into risk classes using the information available at inception of the policy. These information include covariates about the policyholder himself or herself, the insured car and/or home. Each risk class contains similar policyholders with respect to the risk they represent and allows to estimate the claim frequency. This risk classification can be routinely done using generalised linear models and generalised additive models (GAMs). The risk classes remain however heterogeneous as important risk factors are not observed and differentiate policyholders belonging to the same a priori risk class. In a second step, the a posteriori ratemaking consists in using credibility theory to correct periodically the estimates to account for this heterogeneity. More specifically, the previous a priori model can be extended thanks to random effects which represent the latent risk factors. As time passes, the number of claims reported every period reveals information about the unobserved risk factors. See Denuit *et al.* (2007) for a comprehensive review of these techniques in nonlife insurance.

In Home insurance, multiple guarantees can be subscribed by policyholders in the EU and cover damage to a building and/or its content by some perils. The guarantees are generally sold in packages, and in some situations, some guarantees may be even compulsory. The policy can cover either the building only, the content only or both. The main guarantees cover damages due to fire, water damage, electricity, broken glass and third-party liability (TPL) (e.g. covering third parties in the event of fire or water damage). Due to the nature of the policy, although one specific policyholder subscribes a Home insurance policy, the whole household is in fact covered.

In Motor insurance, the two most common guarantees that can be bought by policyholders in the EU are TPL insurance and Material Damage (MD) insurance. TPL is compulsory and covers a third-party's loss caused by the insured car. MD is an optional guarantee that covers the cost of repairing or replacing the insured's own vehicle. This guarantee will often be used when the policyholder is liable for the claim, or when no liable person could be identified.

The present paper aims to model claim counts in Home and in Motor insurance, jointly at household level. It is not the first work that combines both products and considers multiple policyholders from the same household. Indeed, Guillen *et al.* (2008) and Brockett *et al.* (2008) moved from a product view to a customer (or household) view and focused on the behaviour of households with multiple policies. Using logistic regression, they analysed how much time the insurer has to retain the customers starting from the household's first policy cancellation until all the policies have been lapsed. Also, Frees (2003) showed, using multivariate credibility in the vein of Jewell (1974), for aggregate loss models, that it is relevant to account for the covariance between different lines of business.

Some works have focused on the dependence arising from the possibility that claims trigger multiple guarantees at the same time. Indeed, in Motor insurance, for instance, it may happen that a single event triggers both TPL and other guarantees (e.g. MD). When only the number of claims in each guarantee is available and it is not possible to know which claims triggered multiple guarantees, one can rely on the bivariate Poisson regression, as, for example, in Bermúdez (2009). The dependence between the number of claims in both guarantees arises from a common latent count variable that counts the claims that triggered both guarantees. This bivariate model has been further extended to more guarantees in Bermúdez & Karlis (2011). The authors used a multivariate Poisson regression as well as a zero-inflation Poisson regression and relied on Bayesian calculations for the estimation.

Other works that focused on the modelling of claim frequencies over multiple guarantees and/or products can be classified in two main categories according to the way claim counts are correlated.

One way to induce dependence is by introducing a copula. Credibility theory can be used in conjunction with copulas, see, for example, Frees & Wang (2005). Frees & Valdez (2008) and Frees *et al.* (2009) used t-copulas to account for the dependencies among claims from different types of coverage in Motor insurance. Frees *et al.* (2018) modelled jointly lapsation and claims in Motor and Home insurance using copulas, by exploiting the idea that claims outcome may be related to the lapsation of a policy. Also, Shi & Valdez (2014) introduced a multivariate negative binomial regression model, where each pair of variables has its own covariance structure. On aggregated loss, Frees & Wang (2006) used elliptical copulas to model the dependencies between the severities, and the dependence between the number of claims was modelled introducing latent correlated factors. Shi *et al.* (2016) introduced dependence between the cost of claims of various types of claims with a Gaussian copula. The copula allows to capture the cross-sectional and temporal dependence among the claims. Shi & Yang (2018) investigate multiple kind of insurance claims using mixed D-vine copula to model the temporal dependence. The result is used to integrate the policyholders' past experience into their future premiums. Bermúdez *et al.* (2018) analysed the time and cross dependence between the number of claims using a bivariate integer-valued autoregressive regression model (see also Karlis & Pedeli 2013). The model allows to account for both the serial correlation arising from multiple observations of the same policyholder over time as well as the correlation between claims in different guarantees.

Another way to introduce dependence is by inclusion of correlated random effects which model the unobserved risk factors that influence the claim frequency. As these latent risk factors may be correlated across guarantees and across policyholders from the household, the random effects are

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

not assumed to be independent amongst the unit household. The introduction of this dependence therefore induces correlation between the claim frequencies which in turn induces correlation between the number of claims. Purcaru & Denuit (2003) analysed what kind of dependence results from the introduction in the claim frequency of correlated random effects. Bermúdez & Karlis (2017) extended the bivariate model in Bermúdez (2009) by inclusion of either one common random effect per policyholder or three correlated random effects, one for each count variable. Pechon *et al.* (2018) used a multivariate Poisson mixture to account for correlated latent risk factors in Motor TPL amongst policyholders from the same household. Also, Pechon *et al.* (2019) used a multivariate Poisson mixture on two guarantees in Motor insurance and on multiple policyholders from the same household. Three count variables were used to model the claim frequencies in both guarantees, depending on which guarantee(s) were triggered and for each count variable a random effect was introduced. The results showed a positive dependence between the random effects related to policyholders from the same household. Also, Antonio *et al.* (2010b) used a multivariate credibility model using a Bayesian approach on a portfolio of fleets of vehicles; see also Antonio *et al.* (2010a). A posteriori ratemaking can also be used to construct a bonus–malus (BM) system. For instance, Pinquet (1998) constructs a BM system that models the dependence between at-fault and not-at-fault claims in Motor insurance using bivariate credibility. Boucher & Inoussa (2014) proposed a BM system for panel data, while differentiating at-fault and not-at-fault claims in Motor insurance.

The aim of this paper is to supplement the existing bibliography on multivariate credibility and present an approach to model the dependencies between claim frequencies in Home and Motor insurance. Specificities of these products are taken into account: in Home insurance, one policyholder will subscribe the policy which will cover the whole household, while in Motor insurance, multiple policies (one per vehicle, each subscribed by the main driver of the car) can be present in a household. Moreover, different kinds of claims are observed in Motor insurance. The paper aims at taking into account all the claim-related information about both products in each household to update the household's expected claim frequencies.

In this paper, we wish to build on Pechon *et al.* (2018) and Pechon *et al.* (2019) and provide a methodology to model the number of claims reported in both Home and Motor insurance at household level. Compared to the two aforementioned papers, we rely here on a hierarchical structure of random effect. Indeed, some latent risk factors may in fact be shared across all policyholders from the household and across all products. Others are maybe specific to Motor insurance, while still being common to all policyholders in Motor insurance. Then, some latent risk factors specific to each policyholder in Motor insurance may exist, and finally each type of claim may also exhibit some latent risk factors. This understanding leads to a hierarchical structure that can be illustrated as in Figure 2. Therefore, this paper is not a mere extension of the two previously cited papers, in which we would only add Home insurance. In fact, a different methodology is used. The hierarchical structure will allow for more interpretation, however, as we will see, at the cost of forcing a positive correlation which is supported by the data used for illustration.

Our approach can be summarised as follows. First, we estimate the four kinds of a priori expected claim frequencies (Home insurance, TPL only, MD only and TPL and MD simultaneously) using the a priori available covariates. We assume that each of these number of claims follows a Poisson distribution. In practice, this can be done using GAMs. In a second step, a posteriori ratemaking is performed with the help of credibility theory. Important risk factors are not observed but are revealed through time by the number of reported claims. Part of these latent risk factors may be shared or correlated across guarantees and products as well as across policyholders from the same household. These dependencies can be accounted for by introducing dependence between the random effects related to the same household. The inclusion of random effects using the hierarchical structure represented in Figure 2 leads to a multivariate Poisson mixture. The hierarchical structure allows to have a well-defined variance–covariance matrix and more interpretable results. In Pechon *et al.* (2019), no hierarchical structure was used, and the model was

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

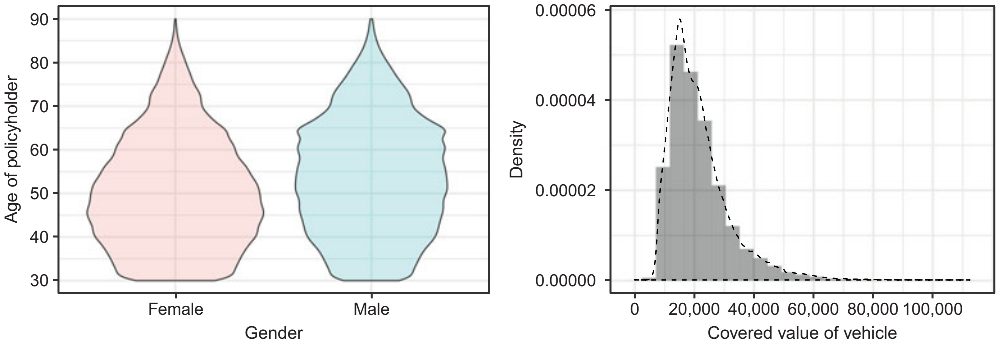

Figure 1. Left: average age of policyholders split by gender during the years 2011–2013. Right: covered value in Euro of the insured vehicles.

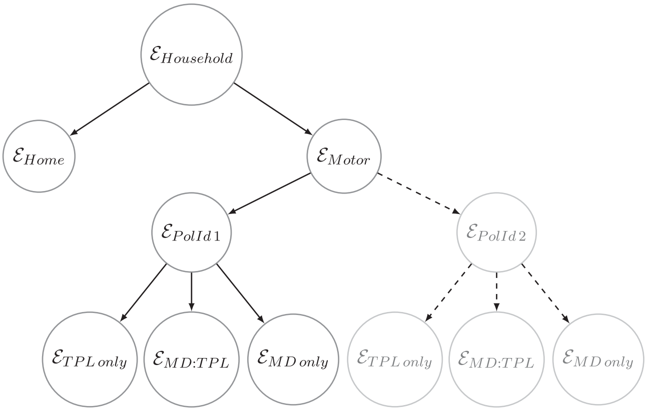

Figure 2. Hierarchical structure with policyholder effect in Motor insurance.

restricted to Motor insurance. Similarly to Pechon *et al.* (2018) and Pechon *et al.* (2019), the random effects are assumed to be Gaussian on the score scale. Parameters are estimated using a likelihood-based inference procedure. The evaluation of the objective function involves the calculation of integrals, which can be computed numerically, for instance using the Gauss–Hermite quadrature. Other possibilities include quasi-likelihood approaches, Monte-Carlo integration or the Laplace method, see Tuerlinckx *et al.* (2006) and Antonio & Zhang (2014) for an extensive review on these methods. As the random effects related to the same household are not assumed to be independent, the contribution of one household to the likelihood involves the computation of as many integrals as there are count variables. On our data set, households had at most two policyholders in Motor insurance and at most one policy in Home insurance, meaning that there are three count variables and three random effects related to each policyholder in Motor insurance, and one count variable and one random effect related to Home insurance. Consequently, the contribution to the likelihood of one household involves at most the computation of a seven-dimensional integral. The Gauss–Hermite quadrature revealed itself to be fast, especially when used in conjunction with the package Rcpp contributed by Eddelbuettel & François (2011) which allows to write chunk of codes in C++ inside an R script enabling a drastic reduction in computational time (up to 30 times faster). Note that this approach is different to a copula approach (e.g. as in Frees & Wang (2006)) which would directly induce dependence between the number of claims or between the claim frequencies, without relying on the introduction of random effects. In our approach, dependence arises from correlated random effects (representing unobserved risk factors) which in turn correlate with the number of claims. The results show that the latent risk

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

factors are correlated across products, guarantees and policyholders from the same household, meaning that the unobserved risk factors may be correlated or shared in the household. Therefore, any claim from the household can be used to predict more accurately the claim frequency in any of the considered guarantees for any policyholder from the household.

The results, which are illustrated in this paper, show that the latent risk factors are positively correlated across both Motor and Home insurance products, guarantees and policyholders from the same household, meaning that the unobserved risk factors may be correlated or shared in the household. The fact that they are positively correlated allows us to use the aforementioned hierarchical structure. Therefore, any claim from the household (in Home or in Motor insurance) can be used to predict more accurately the claim frequency in any of the considered products and guarantees for any policyholder from the household.

A predictive power analysis on out-of-sample data has also been conducted and shows that the multivariate credibility model proposed in this paper improves the predictive power, in comparison for instance to a univariate Poisson mixture. In particular, the model allows to better identify the riskiest households. Therefore, the model can be used to estimate more precisely the expected claim frequencies in a household, using all the available information: using a univariate approach, one policyholder in one product may appear to be a high-risk profile (i.e. a policyholder reporting many claims) for the insurer, while using our multivariate approach makes it possible for the insurer to see the whole picture and may in fact realise that the household as a whole is profitable. Conversely, if a policyholder from a household is reporting many claims in Motor insurance, the model will correct the a priori claim frequencies of the other policyholders from the household, allowing to better identify the riskiest policyholders.

The main contribution of this paper is both methodological and on the application side. The paper suggests a new likelihood-based estimation procedure involving numerical integration in high dimensions with insurance count data. Also, the model shows using a household (or family) approach that the claim frequencies of different products subscribed by different policyholders are correlated. In this paper, the random effects are correlated using a hierarchical structure, allowing more interpretability.

This paper is organised as follows: Section 2 briefly presents the data set used to illustrate the proposed methodology. In Section 3, some notations are introduced, and the model using a hierarchical structure of random effects is described. Then, the model is estimated on the data set, the results are analysed and the correlations are interpreted. Then, in Section 4, some insurance applications illustrate the usefulness of the model. More specifically, a posteriori corrections to both Home and Motor insurance are discussed, and the impact of Home (respectively, Motor) insurance on Motor (respectively, Home) insurance is investigated. In Section 5, cross-validation is used to assess the gain in predictive power of the model. The model is compared to a univariate Poisson–LogNormal mixture model as well as to a Poisson model. Section 6 presents an approach where the estimation of the GAMs and of the variance–covariance matrix of the random effects are cycled. Although the cycling resembles an Expectation–Maximization (EM) algorithm, some characteristics differentiate the two approaches, as for instance, two different objective functions are maximised in the M-step. Section 7 briefly concludes this paper. Additional material is available as online supplements.

2. Data sets

Let us briefly describe the data sets that will be used to support our analysis. We have to stress that, for confidentiality reasons, we are not allowed to show statistics (e.g. mean) on the number of claims. Two detailed data sets relate to portfolios of policies of the same European Insurance company observed during the years 2011–2013. The first data set is the one used in Pechon *et al.* (2019) and relates to policyholders who have subscribed both TPL and MD insurance, in Motor insurance. The policyholders are aged between 30 and 90 years. We have at our disposal the number

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

Table 1. Proportions of policyholders in each level of the categorical variables.

	Usage		Split		Litigation
Private	91.13%	Annually	48.61%	0	97.99%
Professional	8.87%	Monthly	15.79%	2	1.65%
		Semi-annually	22.59%	4	0.36%
		Quarterly	13.01%		
	Power		New car		
Low	95.90%	New (≤3 years)	62.68%		
High	4.10%	Old (>3 years)	37.32%		

Table 2. Number of households according to coverage subscribed in Home and Motor insurance.

	With Home insurance	Without Home insurance	Total
No Motor policies	612,876	0	612,876
One Motor policy	86,321	132,717	219,038
Two Motor policies	5,536	5,446	10,982
Total	704,733	138,163	842,896

of claims triggering each guarantee as well as the number of claims that triggered both guarantees at the same time. Some covariates about the policyholder (e.g. age, gender, place of residence) are available as well as the use of the car. Note that each policy relates to a unique vehicle and is associated with a main driver. Some covariates related to the policy itself are also available, such as whether the payment of the premium has been split. Also, we have access to a variable (litigation) indicating whether the policyholder has had a failure to pay its premium in due time and if so, whether the premium has been paid after a first reminder (level 2) or whether the premium has been paid after a second reminder (level 4), that is, implying an interruption of the coverage in the meantime. Finally, a family identifier allows to match the policyholders who live in the same household. Some statistics about the available covariates are given in Table 1 as well as in Figure 1.

The second data set relates to a portfolio of Home insurance policies of the same European insurance company. A pool of multiple guarantees covering the damages to the building is considered: damage due to fire, water damage, electricity as well as broken glass and TPL (e.g. covering third parties in the event of fire or water damage). These guarantees, all related to the building, have been pooled in order to obtain similar claim frequencies than in Motor insurance. Note that the coverage of the content is not included in the analysis. The data set contains a total of 704,733 households with a Home insurance policy with a total exposure of 1,462,233 years. Some covariates are available, such as the value of the covered building, whether the policyholder is owner (72%) or tenant (28%), whether the building is an apartment (20%) or a house (80%), whether the building is contiguous to other buildings (61%) or not (39%) and finally the ZIP code related to the location of the house. This latter variable can be used to find latitude and longitude coordinates.

The family identifier also allows to match households from both portfolios. More specifically, we can identify which households have subscribed Home insurance and how many Motor insurance policies (covering both TPL and MD) have been bought in the household. Note that this is only possible as long as policyholders buy their insurance policy from the same insurance company, as the data sets relate to only one insurance company.

Among the 704,733 households in the Home insurance portfolio and 230,020 households in the Motor insurance portfolio, 91,857 have a Home insurance policy as well as at least one Motor insurance policy covering TPL and MD. More specific details can be found in Table 2.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

3. Model

3.1. Notations

We define some notations, similar to those in Pechon *et al.* (2019). Let $\mathcal{H}$ be the set of households, and let us introduce the following claim count variables:

- $N^{TPL}_{h(i),t}$: the number of claims of policyholder i from household h that triggered only TPL during year t;
- $N^{MD}_{h(i),t}$: the number of claims of policyholder i from household h that triggered only MD during year t;
- $N^{MD:TPL}_{h(i),t}$: the number of claims of policyholder i from household h that triggered both TPL and MD simultaneously during year t;
- $N^{Home}_{h,t}$: the number of claims of household h that triggered Home insurance during year t.

Note that to simplify some expressions, sometimes we will write $N^{Home}_{h(i),t}$ when we actually mean $N^{Home}_{h,t}$ (e.g. in the first assumption later in this section). Clearly, $N^{Home}_{h(i),t} = N^{Home}_{h,t}$ for every policyholder i belonging to household h.

As in Pechon *et al.* (2019), this implies that the number of claims for policyholder i from household h that triggered TPL (respectively, MD) during year t is $N^{TPL}_{h(i),t} + N^{MD:TPL}_{h(i),t}$ (respectively, $N^{MD}_{h(i),t} + N^{MD:TPL}_{h(i),t}$).

We denote the corresponding a priori expected claim frequencies as $\lambda^{TPL}_{h(i),t} = \mathrm{E}\left[N^{TPL}_{h(i),t}\right]$, $\lambda^{MD}_{h(i),t} = \mathrm{E}\left[N^{MD}_{h(i),t}\right]$, $\lambda^{MD:TPL}_{h(i),t} = \mathrm{E}\left[N^{MD:TPL}_{h(i),t}\right]$ and $\lambda^{Home}_{h,t} = \mathrm{E}\left[N^{Home}_{h,t}\right]$.

We also introduce the aggregated number of claims over the considered time horizon $t = 1, \ldots, T$, namely

- $N^{TPL}_{h(i),\bullet} = \sum_{t=1}^{T} N^{TPL}_{h(i),t}$;
- $N^{MD}_{h(i),\bullet} = \sum_{t=1}^{T} N^{MD}_{h(i),t}$;
- $N^{MD:TPL}_{h(i),\bullet} = \sum_{t=1}^{T} N^{MD:TPL}_{h(i),t}$;
- $N^{Home}_{h,\bullet} = \sum_{t=1}^{T} N^{Home}_{h,t}$.

The corresponding a priori expected claim frequencies are denoted $\lambda^{TPL}_{h(i),\bullet} = \mathrm{E}\left[N^{TPL}_{h(i),\bullet}\right]$, $\lambda^{MD}_{h(i),\bullet} = \mathrm{E}\left[N^{MD}_{h(i),\bullet}\right]$, $\lambda^{MD:TPL}_{h(i),\bullet} = \mathrm{E}\left[N^{MD:TPL}_{h(i),\bullet}\right]$ and $\lambda^{Home}_{h,\bullet} = \mathrm{E}\left[N^{Home}_{h,\bullet}\right]$. In our study, $T = 3$. This means that the policyholders have been observed for at most 3 years. In fact, a variable exposure expresses the amount of time the policyholders were covered during the T years.

Let $\mathcal{G}_{Motor} = \{TPL, MD, MD : TPL\}$ be the set of types of claims in Motor insurance and $\mathcal{G} = \mathcal{G}_{Motor} \cup \{Home\}$ be the set of all claim types considered in this paper.

3.2. Correlation structure

We introduce a Poisson multivariate mixture to allow for dependence between the latent effects which represents all the unobserved risk factors affecting the claim frequency. Let us make the following assumptions:

1. $\forall h \in \mathcal{H}$, $i = 1, 2$, $\forall g \in \mathcal{G}$, given $\Theta^{g}_{h(i)} = \theta$, the random variables $N^{g}_{h(i),1}, \ldots, N^{g}_{h(i),T}$ are independent.
2. $\forall h \in \mathcal{H}$, $i, j = 1, 2$, $\forall g_i, g_j \in \mathcal{G}_{Motor}$ such that $i \neq j$ or $g_i \neq g_j$, given $(\Theta^{g_i}_{h(i)}, \Theta^{g_j}_{h(j)}) = (\theta_i, \theta_j)$, the sequences of random variables $N^{g_i}_{h(i),1}, N^{g_i}_{h(i),2}, \ldots, N^{g_i}_{h(i),T}$ and $N^{g_j}_{h(j),1}, N^{g_j}_{h(j),2}, \ldots, N^{g_j}_{h(j),T}$ are independent.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

3. $\forall h \in \mathcal{H}$, $i = 1, 2$, $\forall g_i \in \mathcal{G}_{Motor}$, given $(\Theta_{h(i)}^{g_i}, \Theta_h^{Home}) = (\theta_i, \theta)$, the sequences of random variables $N_{h(i),1}^{g_i}, N_{h(i),2}^{g_i}, \ldots, N_{h(i),T}^{g_i}$ and $N_{h,1}^{Home}, N_{h,2}^{Home}, \ldots, N_{h,T}^{Home}$ are independent.
4. For a household with two Motor insurance policies and one Home insurance policy, let $\boldsymbol{\Theta_h} = (\Theta_h^{Home}, \Theta_{h(1)}^{TPL}, \Theta_{h(1)}^{MD}, \Theta_{h(1)}^{MD:TPL}, \Theta_{h(2)}^{TPL}, \Theta_{h(2)}^{MD}, \Theta_{h(2)}^{MD:TPL})$ be the vector of all random effects. We assume that $\mathrm{E}(\boldsymbol{\Theta}_h) = \mathbf{1}$. Furthermore, we assume that the random effects can be decomposed using the hierarchical structure described in Figure 2 (with nested random effects):

— $\forall h \in \mathcal{H}$, $\Theta_h^{Home} = \exp\left(\mathcal{E}_h^{Household} + \mathcal{E}_h^{Home}\right)$;

— $\forall h \in \mathcal{H}, \forall i \in h, \forall g \in \mathcal{G}_{Motor}$, $\Theta_{h(i)}^{g} = \exp\left(\mathcal{E}_h^{Household} + \mathcal{E}_h^{Motor} + \mathcal{E}_{h(i)}^{PolId} + \mathcal{E}_{h(i)}^{g}\right)$.

We will assume that $\mathcal{E}_h^{Household} \sim N\left(-\frac{\varsigma_{Household}^2}{2}, \varsigma_{Household}^2\right)$, $\mathcal{E}_h^{Home} \sim N\left(-\frac{\varsigma_{Home}^2}{2}, \varsigma_{Home}^2\right)$, $\mathcal{E}_h^{Motor} \sim N\left(-\frac{\varsigma_{Motor}^2}{2}, \varsigma_{Motor}^2\right)$, $\mathcal{E}_{h(i)}^{PolId} \sim N\left(-\frac{\varsigma_{PolId}^2}{2}, \varsigma_{PolId}^2\right)$, $\forall g \in \mathcal{G}_{Motor}$ $\mathcal{E}_{h(i)}^{g} \sim N\left(-\frac{\varsigma_g^2}{2}, \varsigma_g^2\right)$.

The random vector $\mathcal{E}_{\boldsymbol{h}} = (\mathcal{E}_h^{Household}, \mathcal{E}_h^{Home}, \mathcal{E}_h^{Motor}, \mathcal{E}_{h(i)}^{PolId}, \mathcal{E}_{h(i)}^{TPL}, \mathcal{E}_{h(i)}^{MD}, \mathcal{E}_{h(i)}^{MD:TPL})$ is such that $\mathrm{E}[\exp \mathcal{E}_{\boldsymbol{h}}] = \mathbf{1}$. The variance–covariance matrix of $\mathcal{E}_{\boldsymbol{h}}$ is given by

$$\mathrm{V}\left[\mathcal{E}_{\boldsymbol{h}}\right] = diag(\varsigma_{Household}^2, \varsigma_{Home}^2, \varsigma_{Motor}^2, \varsigma_{PolId}^2, \varsigma_{TPL}^2, \varsigma_{MD}^2, \varsigma_{MD:TPL}^2)$$

The diagonal structure of the matrix $\mathrm{V}\left[\mathcal{E}_{\boldsymbol{h}}\right]$ means that the additive random effects are independent.

Considering Figure 2, $\mathcal{E}_{Household}$ can be understood as the *Household effect*. Indeed, all the shared risk factors across the policyholders from the household and across both products are included in this random effect. Thanks to this random effect, both products are correlated. In Motor insurance, some effects may be common to policyholders from the same household, in addition to the previously cited Household effect. Indeed, the random effect $\mathcal{E}_{Motor}$ induces additional covariance between the claim frequencies in Motor insurance. Finally, some policyholder-related latent factors may exist and influence the claim frequencies; these effects are modelled thanks to $\mathcal{E}_{PolId}$.

Note that the assumptions above that relate to Motor insurance are similar to those made in Pechon *et al.* (2019). The assumption of normality means that the vector of random effects $\boldsymbol{\Theta_h}$ has a multivariate LogNormal distribution with unit mean. In Pechon *et al.* (2018), a Poisson mixture was used to introduce dependence between policyholders from the same household in TPL insurance. The authors compared a Poisson–Gamma, where the Gamma-distributed random effects were correlated using a Gaussian copula, with a Poisson–LogNormal model and concluded that the Poisson–LogNormal model outperformed the Poisson–Gamma model. The model involves a hierarchical structure of random effects representing the latent effects. The variance–covariance matrix of $\log \boldsymbol{\Theta_h}$ (where the log is taken on each component) is given by

$$\boldsymbol{\Sigma}_{\log \boldsymbol{\Theta_h}} = \left(\begin{array}{c|ccc|ccc}
\sigma_{Home}^2 & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} \\
\hline
\sigma_{HM} & \sigma_{TPL}^2 & \sigma_{PolId} & \sigma_{PolId} & & & \\
\sigma_{HM} & \sigma_{PolId} & \sigma_{MD}^2 & \sigma_{PolId} & & \sigma_{Motor} & \\
\sigma_{HM} & \sigma_{PolId} & \sigma_{PolId} & \sigma_{MD:TPL}^2 & & & \\
\hline
\sigma_{HM} & & & & \sigma_{TPL}^2 & \sigma_{PolId} & \sigma_{PolId} \\
\sigma_{HM} & & \sigma_{Motor} & & \sigma_{PolId} & \sigma_{MD}^2 & \sigma_{PolId} \\
\sigma_{HM} & & & & \sigma_{PolId} & \sigma_{PolId} & \sigma_{MD:TPL}^2
\end{array}\right)$$

where

$$\sigma_{Home}^2 = \varsigma_{Household}^2 + \varsigma_{Home}^2$$
$$\sigma_{TPL}^2 = \varsigma_{Household}^2 + \varsigma_{Motor}^2 + \varsigma_{PolId}^2 + \varsigma_{TPL}^2$$

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

$$\begin{aligned}
\sigma^2_{MD} &= \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} + \varsigma^2_{MD} \\
\sigma^2_{MD:TPL} &= \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} + \varsigma^2_{MD:TPL} \\
\sigma_{HM} &= \varsigma^2_{Household} && \text{(Household effect)} \\
\sigma_{Motor} &= \varsigma^2_{Household} + \varsigma^2_{Motor} && \text{(Inter-policyholder)} \\
\sigma_{PolId} &= \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} && \text{(Intra-policyholder)}
\end{aligned}$$

3.3. Estimation

Since the support of each random effect is $\mathbb{R}_+$, we can write the likelihood as

$$L(\boldsymbol{\Sigma}_{\log \boldsymbol{\Theta}}) = \prod_{h \in \mathcal{H}} \int_{\mathbb{R}^7_+} \left[\prod_{i=1}^{2} \prod_{g \in \mathcal{G}} \exp\left(-\lambda^g_{h(i),\bullet} \theta^g_{h(i)}\right) \frac{\left(\lambda^g_{h(i),\bullet} \theta^g_{h(i)}\right)^{n^g_{h(i),\bullet}}}{n^g_{h(i),\bullet}!} \right] f_{\boldsymbol{\Theta}_h}(\boldsymbol{\theta}_h) d\boldsymbol{\theta}_h \tag{1}$$

Note that if for some household the length of the vector $\boldsymbol{\Theta}_h$ is smaller than 7 (i.e. no Home insurance policy or less than two Motor insurance policies), by adopting the convention that $0^0 = 1$ and by setting the corresponding a priori expected claim frequency λ to zero yield the correct likelihood. In such a case, the corresponding integral vanishes.

The computations have been realised in the statistical software R. The estimation was done in a two-part approach.

First, the a priori expected claim frequencies have been estimated with a Poisson regression using GAMs. It has to be noted that no covariates from Motor insurance have been used in the Home insurance regression and vice versa. One reason for this is that some households have only Motor (respectively, Home) insurance, and the covariates related to the missing product would therefore result in missing values, which the GAMs do not handle. This does not prevent us from using common factors, such as the length of time the policyholder or household has been with the company. Such a covariate can be a part of each model as long as the parameters are functionally independent. Four different claims are considered: one in Home insurance and three in Motor insurance. As explained in Section 2, a pool of multiple guarantees in Home insurance is considered, whereas in Motor insurance, three types of claims are considered. The GAMs were chosen to model the claim frequencies using a Poisson regression with a log link function. See Wood (2017) for an introduction to GAMs. In the present study, the mgcv package available in R was used.

The models used in Motor insurance are the same as in Pechon *et al.* (2019). More details about the a priori analyses can be found in Section 6.

In the second part, the a priori expected claim frequencies are fixed, and the variance–covariance matrix of the random effects is estimated by maximising the objective function (1) by holding the a priori expected claim frequencies λ fixed.

Although this two-part approach will be used in this paper, we provide an algorithm in Section 6 that resembles an expectation/conditional maximisation (ECM) procedure. The algorithm proposed in this paper differs from the one proposed in Meng & Rubin (1993) with the fact that we maximise two different objective functions. We stress out that further research is necessary, however, to assess the validity of the proposed algorithm.

In order to compute the likelihood in equation (1), we need to rely on numerical integration. We rely on the multivariate Gauss Hermite quadrature, which is available in R thanks to the package *MultiGHQuad* contributed by Kroeze (2016). The integrand is written in C++ so as to fasten the computation. The C++ integrand can be included in an R function using the Rcpp package (see Eddelbuettel & François 2011). We have a step-by-step approach to estimate the parameters, which we shall describe below. The step-by-step approach allows to break down the estimation in different steps of smaller dimension (i.e. which require less computations) which in turn give

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

Table 3. Estimates at each step of the optimisation process.

	First step	Second step	Third step (s.e.)
$\widehat{\varsigma}^2_{Household}$	0.0000	0.00000	0.128 (0.0161)
$\widehat{\varsigma}^2_{Home}$	0.0000	0.00000	0.185 (0.0204)
$\widehat{\varsigma}^2_{Motor}$	0.0000	0.19038	0.0703 (0.0297)
$\widehat{\varsigma}^2_{PolId}$	0.2414	0.05199	0.0444 (0.0279)
$\widehat{\varsigma}^2_{TPL}$	0.3163	0.31631	0.315 (0.0424)
$\widehat{\varsigma}^2_{MD}$	0.1199	0.12429	0.126 (0.0178)
$\widehat{\varsigma}^2_{MD:TPL}$	0.1225	0.12900	0.133 (0.0411)

us good initial values for the final step that incorporates all the parameters. Having good initial values for this final step allows to have less iterations in the optimisation process and consequently reduces drastically the computation time.

First, we estimate the parameters related to Motor insurance at a policyholder level (the variances $\varsigma^2_{PolId}, \varsigma^2_{TPL}, \varsigma^2_{MD}, \varsigma^2_{MD:TPL}$), the other variances being fixed temporary at zero. As initial values, we rely on the marginal estimates of the variance. To obtain initial values for the parameters $\varsigma^2_{TPL}, \varsigma^2_{MD}, \varsigma^2_{MD:TPL}$, we can estimate the three variances of three different univariate Poisson–LogNormal models. A initial value for ς^2_{PolId} can be given by optimising a trivariate Poisson–LogNormal model with the three previous variances fixed and the covariance (i.e. ς^2_{PolId}) being estimated. Since these models are of low dimension and each of them involves only one parameter to optimise, convergence is fast.

In the second step, the variance ς^2_{Motor} is estimated. In fact, the parameters $\sigma^2_{TPL}, \sigma^2_{MD}, \sigma^2_{MD:TPL}$ are kept constant in this step, and an increase of the parameter ς^2_{Motor} (inducing the dependence between random effects from the same household) is compensated by a decrease of the parameter ς^2_{PolId} (inducing dependence between random effects from the same policyholder). Note that the difference at this point with Pechon *et al.* (2019) comes from the dependence structure. The covariance (induced by Θ_{Motor}) is estimated and is assumed to be the same for any pair of different random effects belonging to the same policyholder, whereas in Pechon *et al.* (2019), each pair of random effects related to the same policyholder had its own covariance parameter.

In the third step, Home insurance is then included. However, the hierarchical structure imposes that the correlations (respectively, covariances) are positive. As opposed to Motor insurance, for which we could rely on the results found in Pechon *et al.* (2019) establishing that the correlation is positive, we first need to assess whether the dependence between Home and Motor insurance is positive before using the hierarchical structure. For this preliminary assessment, we first introduce a single random effect Θ^{Home} and introduce dependence with respect to Motor insurance thanks to a unique covariance parameter between the random variable Θ^{Home} and the random variables Θ^{TPL}, Θ^{MD}, $\Theta^{MD:TPL}$. The variance (i.e. the residual heterogeneity in Home insurance) is estimated marginally and then is fixed at its value to estimate the covariance. We find a covariance equals to 0.1202 (s.e. 0.0174).

Given that the covariance between Home and Motor insurance's random effects is significantly positive, we can rely on the hierarchical structure which imposes a positive dependence and estimate all the parameters simultaneously in a third step by maximising the objective function (3.3) holding the a priori expected claim frequencies λ fixed.

We display all the estimates at each step in Table 3. We note that at each step there is a transfer that occurs: the estimated $\widehat{\varsigma}^2_{PolId}$ from step 1 is split into $\widehat{\varsigma}^2_{PolId}$ and $\widehat{\varsigma}^2_{Motor}$ in step 2. This breakdown can be understood in the following way. The introduction of the variance $\widehat{\varsigma}^2_{Motor}$ induces covariance between policyholders from the same household in Motor insurance, while in step 1, only the random effects of guarantees related to the same policyholder had a strictly covariance.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

Table 4. Final estimates.

Effect	Estimate	s.e
$\widehat{\varsigma}^2_{Household}$	0.1238	0.0175
$\widehat{\varsigma}^2_{Home}$	0.1872	0.0216
$\widehat{\varsigma}^2_{Motor}$	0.1141	0.021
$\widehat{\varsigma}^2_{PolId}$	0.0000	
$\widehat{\varsigma}^2_{TPL}$	0.3213	0.0175
$\widehat{\varsigma}^2_{MD}$	0.1272	0.0216
$\widehat{\varsigma}^2_{MD:TPL}$	0.1336	0.021

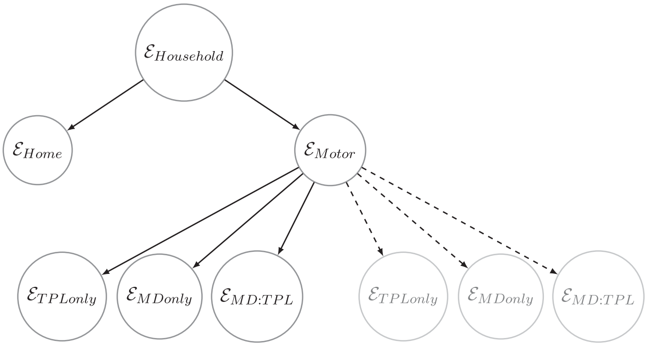

Figure 3. Hierarchical structure of final model.

The increase in $\widehat{\varsigma}^2_{Motor}$ is, however, compensated by the decrease in $\widehat{\varsigma}^2_{PolId}$. The covariance for guarantees related to the same policyholder does not change, while the policyholders from the same household in Motor insurance become correlated. Similarly, from step 2 to step 3, the introduction of the random effect related to the Home induces a decrease of $\widehat{\varsigma}^2_{Motor}$ that is compensated by an increase of $\widehat{\varsigma}^2_{Household}$. The covariances in Motor insurance remain the same than in step 2, however, due to the common variance $\widehat{\varsigma}^2_{Household}$, Home and Motor insurance become correlated, with a covariance equal to $\widehat{\varsigma}^2_{Household}$.

Finally, we observe that the policyholder effect (i.e. ς^2_{PolId}) does not seem to be significantly different from zero. This appears to be coherent with the results found in Pechon *et al.* (2019), where the estimates of the analogue parameters appear to have overlapping confidence intervals. This model is then re-estimated by forcing $\varsigma^2_{PolId} = 0$. The estimates can be found in Table 4.

A likelihood ratio test is conducted to assess whether the policyholder effect is significant (i.e. $\varsigma^2_{PolId} > 0$). The value of the statistic is $t = 0.1866$. Since the test involves a value on the boundary of the domain, Self & Liang (1987) and Agresti (2003) suggest to compute the p-value using $\frac{1}{2}\Pr\left(\chi^2_1 > t\right)$. We obtain a p-value of 0.3329, which confirms that the policyholder effect is not significant. Notice that because of the two-step estimation procedure, this p-value must be considered with caution, but its relatively high value is taken as an evidence supporting the model choice $\varsigma^2_{PolId} = 0$ retained here.

The final model used in the remainder of this paper, without the policyholder effect, amounts to a hierarchical structure as depicted in Figure 3.

Note that given the parametrisation using a hierarchical structure, any value within the confidence intervals will yield an appropriate (i.e. positive-definite) variance–covariance matrix. This is one of the advantages of the hierarchical parametrisation.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

Some sensitivity analysis with respect to the number of nodes per dimension used in the Gauss–Hermite quadrature has been conducted. The results thereof can be found in the Supplementary Material in this paper. It appears that starting at seven nodes per dimension, the estimates stabilise. By choosing $m = 7$ nodes per dimension, for the households with two Motor insurance policies and one Home insurance policy, the number of nodes used to compute the contribution of this household to the likelihood amounts to $m^7 = 7^7 = 823{,}543$. It has to be noted, though, that not all households have these three policies. The contribution to the likelihood of these households therefore implies the evaluation of less integrals.

3.4. Implied dependence structure

Now that the model has been estimated, we can compute the implied dependence structure between these latent factors. We start with the variances, which represent the strength of these latent factors (i.e. the amount of residual heterogeneity in each risk class constructed with the a priori model).

We start with the implied dependence structure of the underlying multivariate normal distribution. Afterwards, the implied dependence structure on the LogNormal scale (i.e. the dependence structure of $\boldsymbol{\Theta_h}$) will be deduced.

For $i = 1, 2$, we can compute the residual heterogeneity:

$$\begin{aligned}
\mathrm{V}\left[\log \Theta_h^{Home}\right] &= \varsigma^2_{Household} + \varsigma^2_{Home} && \text{estimated to 0.3110}\\
\mathrm{V}\left[\log \Theta_{h(i)}^{TPL}\right] &= \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{TPL} && \text{estimated to 0.5592}\\
\mathrm{V}\left[\log \Theta_{h(i)}^{MD}\right] &= \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{MD} && \text{estimated to 0.3651}\\
\mathrm{V}\left[\log \Theta_{h(i)}^{MD:TPL}\right] &= \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{MD:TPL} && \text{estimated to 0.3715}
\end{aligned}$$

As stated above, the hierarchical structure induces positive covariances between the random effects. We can now explicitly compute these. Indeed, we have that for $i, j = 1, 2$

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\log \Theta_h^{Home}, \log \Theta_{h(i)}^{TPL}\right] &= \mathbb{C}\mathrm{ov}\left[\log \Theta_h^{Home}, \log \Theta_{h(i)}^{MD}\right]\\
&= \mathbb{C}\mathrm{ov}\left[\log \Theta_h^{Home}, \log \Theta_{h(i)}^{MD:TPL}\right]\\
&= \varsigma^2_{Household} \text{ estimated to 0.1238}\\
\mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{TPL}, \log \Theta_{h(j)}^{MD}\right] &= \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{TPL}, \log \Theta_{h(j)}^{MD:TPL}\right]\\
&= \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{MD}, \log \Theta_{h(j)}^{MD:TPL}\right]\\
&= \varsigma^2_{Household} + \varsigma^2_{Motor} \text{ estimated to 0.2379}\\
\text{and for } i \neq j, \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{TPL}, \log \Theta_{h(j)}^{TPL}\right] &= \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{MD}, \log \Theta_{h(j)}^{MD}\right]\\
&= \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{MD:TPL}, \log \Theta_{h(j)}^{MD:TPL}\right]\\
&= \varsigma^2_{Household} + \varsigma^2_{Motor} \text{ estimated to 0.2379}
\end{aligned}$$

We can compute the correlations between the log of the random effects. The estimated correlations are shown in Table 5. These correlations are related to the underlying normal distribution, which is on the score scale (i.e. the log scale).

Now that the dependence structure of the multivariate normal random vector is computed, we can deduce the correlation structure between the LogNormal random effects. Let us compute the variances as well as the correlations between the LogNormal distributed random effects

Table 5. Estimated correlation structure between the random effects on the log scale (i.e. at the score scale).

	$\log \Theta_h^{Home}$	$\log \Theta_{h(1)}^{TPL}$	$\log \Theta_{h(1)}^{MD}$	$\log \Theta_{h(1)}^{MD:TPL}$	$\log \Theta_{h(2)}^{TPL}$	$\log \Theta_{h(2)}^{MD}$	$\log \Theta_{h(2)}^{MD:TPL}$
$\log \Theta_h^{Home}$	1.0000	0.2896	0.3574	0.3751	0.2896	0.3574	0.3751
$\log \Theta_{h(1)}^{TPL}$	0.2896	1.0000	0.4745	0.4979	0.3844	0.4745	0.4979
$\log \Theta_{h(1)}^{MD}$	0.3574	0.4745	1.0000	0.6146	0.4745	0.5857	0.6146
$\log \Theta_{h(1)}^{MD:TPL}$	0.3751	0.4979	0.6146	1.0000	0.4979	0.6146	0.6450
$\log \Theta_{h(2)}^{TPL}$	0.2896	0.3844	0.4745	0.4979	1.0000	0.4745	0.4979
$\log \Theta_{h(2)}^{MD}$	0.3574	0.4745	0.5857	0.6146	0.4745	1.0000	0.6146
$\log \Theta_{h(2)}^{MD:TPL}$	0.3751	0.4979	0.6146	0.6450	0.4979	0.6146	1.0000

Table 6. Estimated variances of random effects for $i = 1, 2$.

	Θ_h^{Home}	$\Theta_{h(i)}^{TPL}$	$\Theta_{h(i)}^{MD}$	$\Theta_{h(i)}^{MD:TPL}$
Var	0.3655	0.7406	0.4387	0.3914

Table 7. Estimated correlation structure between the LogNormal random effects.

	Θ_h^{Home}	$\Theta_{h(1)}^{TPL}$	$\Theta_{h(1)}^{MD}$	$\Theta_{h(1)}^{MD:TPL}$	$\Theta_{h(2)}^{TPL}$	$\Theta_{h(2)}^{MD}$	$\Theta_{h(2)}^{MD:TPL}$
Θ_h^{Home}	1.0000	0.2457	0.3193	0.3380	0.2457	0.3193	0.3380
$\Theta_{h(1)}^{TPL}$	0.2457	1.0000	0.4165	0.4410	0.3206	0.4165	0.4410
$\Theta_{h(1)}^{MD}$	0.3193	0.4165	1.0000	0.5730	0.4165	0.5412	0.5730
$\Theta_{h(1)}^{MD:TPL}$	0.3380	0.4410	0.5730	1.0000	0.4410	0.5730	0.6066
$\Theta_{h(2)}^{TPL}$	0.2457	0.3206	0.4165	0.4410	1.0000	0.4165	0.4410
$\Theta_{h(2)}^{MD}$	0.3193	0.4165	0.5412	0.5730	0.4165	1.0000	0.5730
$\Theta_{h(2)}^{MD:TPL}$	0.3380	0.4410	0.5730	0.6066	0.4410	0.5730	1.0000

$$\boldsymbol{\Theta_h} = \left(\Theta_h^{Home}, \Theta_{h(1)}^{TPL}, \Theta_{h(1)}^{MD}, \Theta_{h(1)}^{MD:TPL}, \Theta_{h(2)}^{TPL}, \Theta_{h(2)}^{MD}, \Theta_{h(2)}^{MD:TPL}\right)$$

We show the estimated variances in Table 6 and the estimated correlations in Table 7.

Let us provide some interpretations for the values displayed in Tables 6 and 7.

The heterogeneity appears to be the smallest in Home insurance. This is very intuitive because claims in Motor insurance are strongly related to the way of driving as well as the risk aversion of the policyholder, whereas claims in Home insurance can be more often the result of a external circumstances, such as bad location (e.g. flood zone), or to a badly constructed building (e.g. faulting electrical connections).

The dependence between both Home and Motor insurance materialises in a correlation of about 30%. It appears, however, to be weaker than within guarantees related to Motor insurance. Again, this is intuitive, as less choices are made by the policyholder in Home insurance than in Motor insurance. The correlations in Motor insurance appear to be of similar order of magnitude than those given in Pechon *et al.* (2019), although a different pair appears to have the strongest correlation. By looking at the standard errors of the estimates, we can assume that this could be related to randomness.

Consequently, there appears to be some unobserved risk factor that is affecting the whole household and that is correlating the claim frequencies in both Motor insurance guarantees as well as in Home insurance. This could be a very local geographic effect that the a priori model could not identify using a bivariate function of the latitude and longitude of the place of residence, or even the socio-economic status, as the latitude and longitude variables available in the data set are in fact the centre of each district. Therefore, socio-economic status, often shared in neighbourhoods, could not be integrated in the a priori model and could be part of the unobserved risk factors. Nevertheless, one can only take guesses, since this information is not available and consequently cannot be added to the a priori model.

The fact, however, that the correlation between Motor insurance guarantees is the highest suggests that some other unobserved factor representing the behaviour and risk averseness of the policyholder is hidden in this residual heterogeneity and is also possibly shared among policyholders from the same household.

4. Insurance Applications

In this section, we aim to focus on applications which illustrate the consequence of the dependence between Home and Motor insurance using credibility theory. Some applications related to Motor insurance have been reviewed in Pechon *et al.* (2019). We aim to focus here on the consequence of the dependence between Home and Motor insurance.

In the first application, we will show how the claims experience can be used to produce yearly (or for any other period) updates of the expected claim frequencies (i.e. a posteriori expected claim frequencies). Different scenarios related to the number of claims observed in the different products and guarantees will be discussed.

In the second application, we will show how the newly introduced dependence between the latent factors in Home and Motor insurance impacts the a posteriori corrections. By correction, we mean the conditional expectation of the random effect, that is, the multiplicative correction to apply to the a priori expected claim frequency given the past numbers of claims. More specifically, we will compare the corrections given by our model with the corrections given in case we would assume independence between Home and Motor insurance (but keeping the dependence at household level between the guarantees in Motor insurance).

4.1. A posteriori corrections

As time passes by and the insurer observes the number of claims, we can compute corrections to apply to the a priori expected claim frequency conditional to the observed number of claims thanks to credibility theory. Due to the dependence between the random effects related to the same household, the number of claims of any policyholder from the household as well as of any guarantee is relevant.

Let a household h consists of two policyholders, $h(1)$ and $h(2)$, and let them both have TPL and MD insurance. Moreover, the household also has a Home insurance policy at the same insurance company. We will assume for simplicity that all the policies start at the same date. At any time, the insurer computes for each guarantee the cumulated (i.e. aggregated) number of claims since the contracts inception as well as the cumulated (i.e. aggregated) a priori expected claim frequencies. For instance, at the end of year T, the insurer can then compute the following corrections for the household in Home insurance:

$$\begin{aligned}
&\mathrm{E}\left[\Theta_h^{Home} | \boldsymbol{N_h} = \boldsymbol{n_h}\right] \\
&= \frac{1}{\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right]} \int_{\mathbb{R}_+^7} \theta^{Home} \Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h} | \boldsymbol{\Theta} = \boldsymbol{\theta}\right] f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) d\boldsymbol{\Theta} \\
&= \frac{1}{\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right]} \frac{1 + n_{h,\bullet}^{Home}}{\lambda_{h,\bullet}^{Home}} \int_{\mathbb{R}_+^7} \Pr\left[\boldsymbol{N_h^{(-Home)}} = \boldsymbol{n_h^{(-Home)}},\right. \\
&\qquad \left. N_{h,\bullet}^{Home} = n_{h,\bullet}^{Home} + 1 | \boldsymbol{\Theta} = \boldsymbol{\theta}\right] f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) d\boldsymbol{\Theta} \\
&= \frac{\Pr\left[\boldsymbol{N_h^{(-Home)}} = \boldsymbol{n_h^{(-Home)}}, N_{h,\bullet}^{Home} = n_{h,\bullet}^{Home} + 1\right]}{\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right]} \frac{1 + n_{h,\bullet}^{Home}}{\lambda_{h,\bullet}^{Home}}
\end{aligned}$$

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

where

$$\boldsymbol{N_h} = \left(N_{h,\bullet}^{Home}, N_{h(1),\bullet}^{TPL}, N_{h(1),\bullet}^{MD}, N_{h(1),\bullet}^{MD:TPL}, N_{h(2),\bullet}^{TPL}, N_{h(2),\bullet}^{MD}, N_{h(2),\bullet}^{MD:TPL}\right)$$
$$\boldsymbol{n_h} = \left(n_{h,\bullet}^{Home}, n_{h(1),\bullet}^{TPL}, n_{h(1),\bullet}^{MD}, n_{h(1),\bullet}^{MD:TPL}, n_{h(2),\bullet}^{TPL}, n_{h(2),\bullet}^{MD}, n_{h(2),\bullet}^{MD:TPL}\right)$$

and where $\boldsymbol{N_h}^{(-Home)}$ (respectively, $\boldsymbol{n_h}^{(-Home)}$) is the vector $\boldsymbol{N_h}$ (respectively, $\boldsymbol{n_h}$) without the item related to the Home guarantee.

In Motor insurance, we can compute the conditional expectation of any random effect:

$$\forall g \in \mathcal{G}_{Motor}, \; \mathrm{E}\left[\Theta_{h(i)}^{g} | \boldsymbol{N_h} = \boldsymbol{n_h}\right] = \frac{\Pr\left[\boldsymbol{N_{h(i)}^{(-g)}} = \boldsymbol{n_{h(i)}^{(-g)}}, N_{h(i),\bullet}^{g} = n_{h(i),\bullet}^{g} + 1\right]}{\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right]} \frac{1 + n_{h(i),\bullet}^{g}}{\lambda_{h(i),\bullet}^{g}}$$

where

$$\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right] = \int_{\mathbb{R}_+^7} \prod_{i \in h} \prod_{g \in \mathcal{G}} \exp\left(-\lambda_{h(i),\bullet}^{g} \theta_{h(i)}^{g}\right) \frac{\left(\lambda_{h(i),\bullet}^{g} \theta_{h(i)}^{g}\right)^{n_{h(i),\bullet}^{g}}}{n_{h(i),\bullet}^{g}!} f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta_h}) d\boldsymbol{\theta_h}$$

can be computed numerically, for instance with the Gauss–Hermite quadrature, and where $\boldsymbol{N_{h(i)}^{(-g)}}$ (respectively, $\boldsymbol{n_{h(i)}^{(-g)}}$) is the vector $\boldsymbol{N_h}$ (respectively, $\boldsymbol{n_h}$) without the item related to the guarantee g of policyholder $h(i)$.

Although the modelling in Motor insurance includes three count variables and three random effects to model two guarantees (in order to capture the events that trigger both guarantees at the same time), we can in fact calculate the correction that is applied in TPL and in MD by considering the following linear combinations:

$$\begin{aligned} &\frac{1}{\lambda_{h(i),\bullet}^{TPL} + \lambda_{h(i),\bullet}^{MD:TPL}} \left(\lambda_{h(i),\bullet}^{TPL} \mathrm{E}\left[\Theta_{h(i)}^{TPL} | \boldsymbol{N_h} = \boldsymbol{n_h}\right] + \lambda_{h(i),\bullet}^{MD:TPL} \mathrm{E}\left[\Theta_{h(i)}^{MD:TPL} | \boldsymbol{N_h} = \boldsymbol{n_h}\right]\right) \\ &\frac{1}{\lambda_{h(i),\bullet}^{MD} + \lambda_{h(i),\bullet}^{MD:TPL}} \left(\lambda_{h(i),\bullet}^{MD} \mathrm{E}\left[\Theta_{h(i)}^{MD} | \boldsymbol{N_h} = \boldsymbol{n_h}\right] + \lambda_{h(i),\bullet}^{MD:TPL} \mathrm{E}\left[\Theta_{h(i)}^{MD:TPL} | \boldsymbol{N_h} = \boldsymbol{n_h}\right]\right) \end{aligned} \quad (2)$$

In order to numerically compute these corrections, we need to determine the a priori risk profiles of both policyholders in TPL and in MD as well as the household's risk profile in Home insurance. For this matter, we bin the predicted a priori expected claim frequencies obtained with the GAMs into three categories, using the quantiles 2/6 and 4/6: low-, medium- and high-risk profiles. The three risk profiles are then associated with the numerical values given by the quantiles 1/6, 3/6 and 5/6 (i.e. the corresponding median value of estimated expected claim frequency of each risk class). In the following numerical examples, we will only consider the cases in which the same a priori risk profile is shared across all the policyholders and all the guarantees. Of course, other combinations are possible.

4.1.1. *A posteriori corrections in the claim-free case*

Let us first illustrate the example in which no claim was reported in any guarantee. We show in Figure 4 the corrections to apply in Home and Motor insurance when no claim occurred as time passes. As expected, the correction factor drops as time passes and no claim is reported. The riskier profiles have a stronger correction, in line with the idea that claim-free years are more expected from a safer policyholder than from a riskier policyholder. Note that the corrections are similar in TPL and in MD, although they appear to be slightly stronger in MD. This comes from the fact that the average claim frequency is higher in MD than in TPL, although this effect is, in these examples,

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

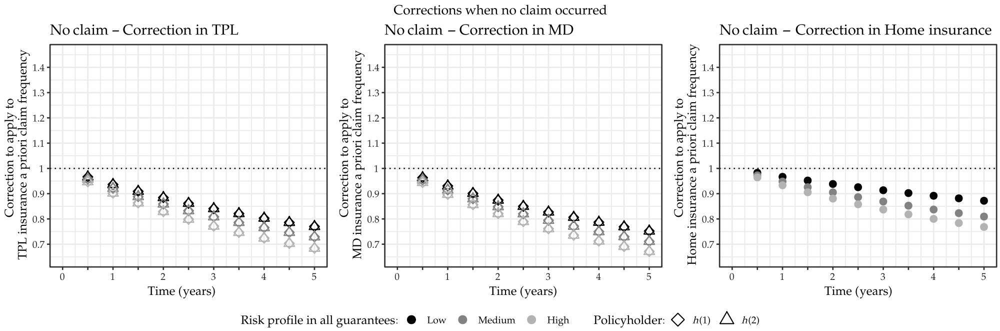

Figure 4. Corrections to apply in TPL (left), MD (middle) and Home insurance (right) when no claim occurred in the household.

partly compensated by the fact that the heterogeneity is greater in TPL than in MD. Also in Home insurance, a correction factor below 90% is observed after 5 claim-free years, meaning that in average, claim-free households (in Motor and Home insurance) with a low-risk profile have about 10% less claims in Home insurance than the claim frequency given by their a priori risk class.

4.1.2. A posteriori corrections in Motor insurance after a claim

In the following, we will first consider that a claim occurred (either in Home or in Motor insurance) and discuss the corrections to apply to the guarantees in Motor insurance. The corrections to apply in Home insurance will be considered later. Note that, even though two policyholders are considered in the household, we can consider in these examples without loss of generalisation that $h(1)$ experiences the claims in Motor insurance and $h(2)$ remains claim-free. Also, note that for Home insurance, it is neither $h(1)$ nor $h(2)$ that experiences the claim, but rather the whole household h.

In Figure 5, the corrections to apply in TPL to both policyholders are depicted. The figures show that a claim in any guarantee will increase the correction factors in all the guarantees and for all the policyholders from the household. We see, however, that the biggest increase is for $h(1)$ in TPL (i.e. the policyholder at fault in the triggered guarantee). We also note that both $h(1)$ and $h(2)$ have the same corrections after a claim of $h(1)$ in MD only. This comes from the fact that the model assumes the same covariance between the different random effects in Motor insurance (regardless if they are related to the same policyholder from the household). This is not the case for the corrections after a claim triggering both TPL and MD, as the correction in TPL explicitly combines both kind of claims: those triggering only TPL and those triggering TPL and MD (see equation (2)).

Similarly, in Figure 6, the corrections to apply in MD to both policyholders are shown. By comparing the corrections in TPL after a claim in TPL (Figure 5) and the corrections in MD after a claim in MD (Figure 6), it appears that we have a stronger correction factor in TPL when a claim is reported in TPL than in MD when a claim in MD is reported. This comes from the greater heterogeneity in TPL (see Table 6) and the lower claim frequencies in TPL which means that a claim is less expected in TPL than in MD. Moreover, the decrease of these correction factors is greater in MD than in TPL, in line with the greater claim frequencies in MD than in TPL.

Furthermore, it appears that a claim triggering both guarantees at the same time has less consequences on the MD a posteriori expected claim frequency than on the TPL a posteriori expected claim frequency. One explanation is that in average the claim frequency related to MD:TPL claims, $\lambda^{MD:TPL}$, only plays for about 28% of the total claim frequency in MD (i.e. $\lambda^{MD} + \lambda^{MD:TPL}$), whereas it is about 55% of the total claim frequency in TPL (i.e. $\lambda^{TPL} + \lambda^{MD:TPL}$). Hence, the correction arising from the conditional expectation of $\Theta^{MD:TPL}$ is less weighted (i.e. in the convex combination (2)) in the MD case than in the TPL case.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

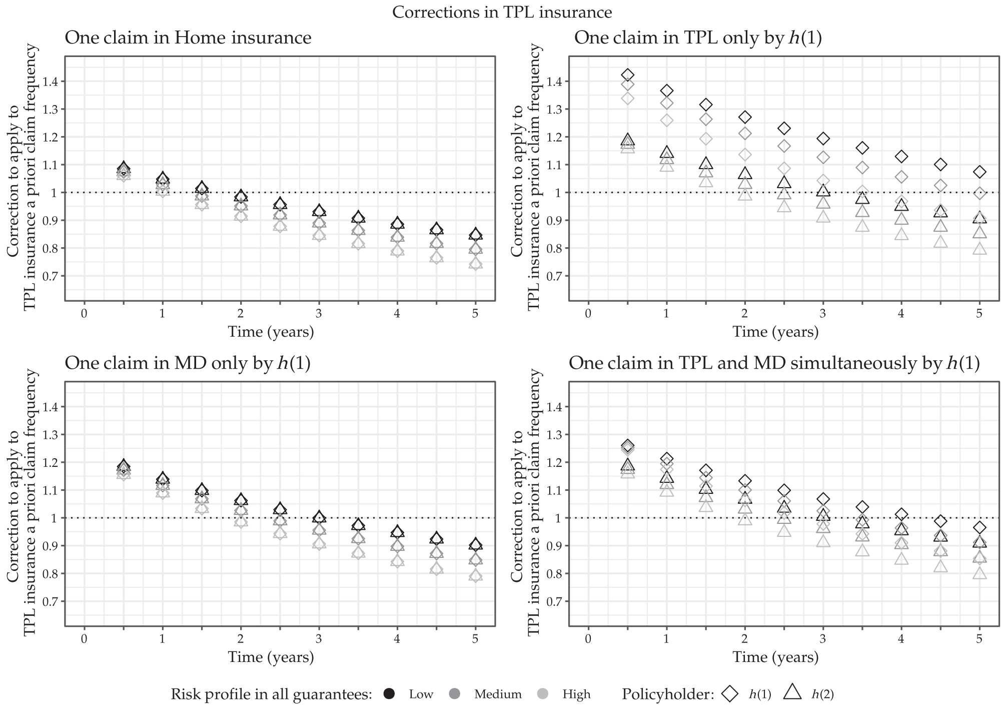

Figure 5. Correction to apply to TPL insurance a priori expected claim frequencies when one claim was reported.

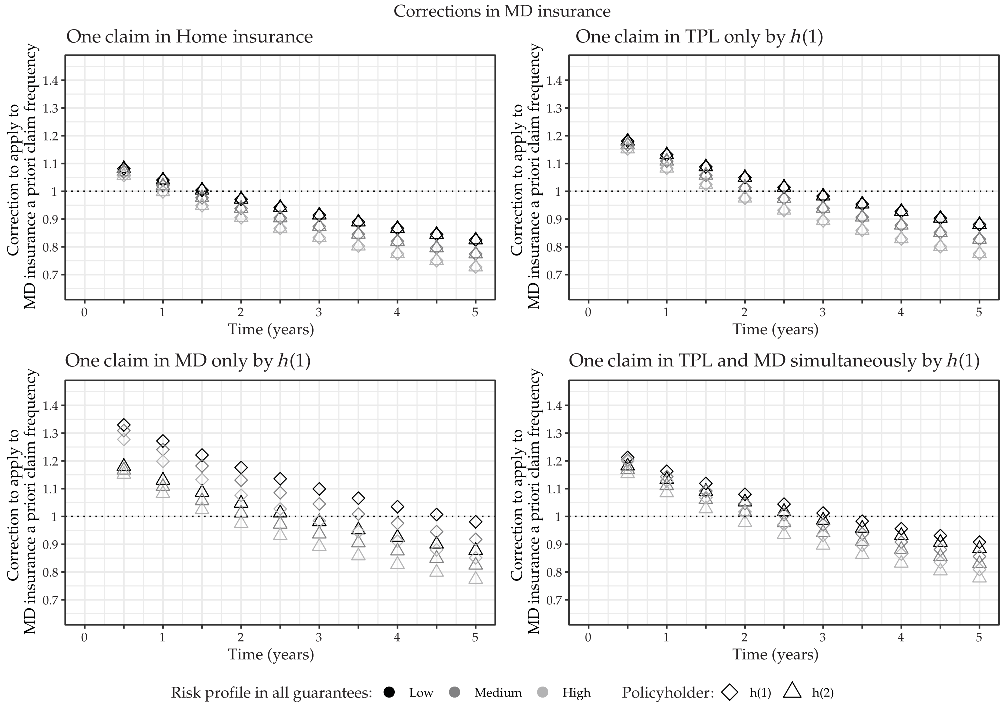

Figure 6. Correction to apply to MD insurance a priori expected claim frequencies when one claim was reported.

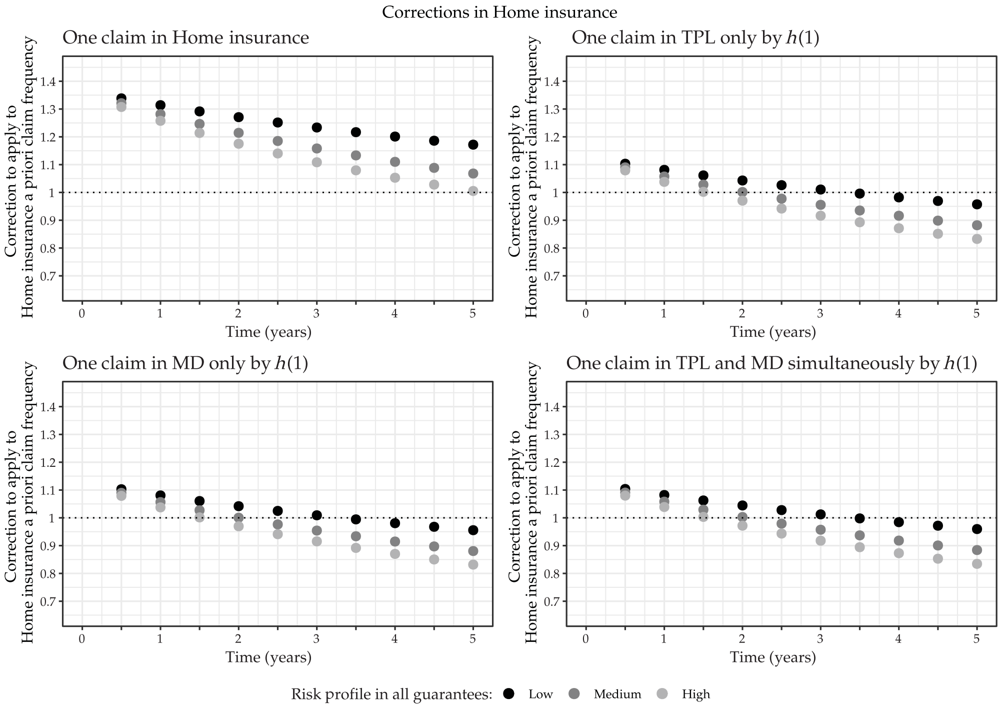

Figure 7. Correction to apply to Home insurance a priori expected claim frequency when one claim was reported.

Finally, let us analyse the correction factors when a claim was reported in Home insurance (Figures 5 and 6) and compare them to those when no claim was reported (Figure 4). Both in TPL and in MD, we observe that after 1 year, there is a difference of around 10%. As time passes, the difference shrinks to around 5% after 5 years. Consequently, we see that taking into account the reported claims in Home insurance to compute a posteriori expected claim frequencies is pertinent.

4.1.3. A posteriori corrections in Home insurance after a claim

We can also compute the corrections to apply to the a priori expected claim frequency in Home insurance as a function of the number of reported claims in both Home and Motor insurance. As before, we assume a household with two policies in Motor insurance as well as one Home insurance. The correction to apply to Home insurance is depicted in Figure 7. Note that, on the contrary to the Motor insurance case, only one correction applies, as the policyholder is assumed to be the whole household in this case.

Symmetrically to the Motor insurance case, a claim in Home insurance appears to increase the most the correction factor to apply to the a priori expected claim frequency in Home insurance. We note, however, that in the long run, differences between lower- and higher-risk profiles appear to be large (about 10–15%). Only a claim in Home insurance yields a correction factor larger than 1 after 5 years. Moreover, note that the correction factors arising from a claim in any guarantee in Motor insurance are all equal, regardless of which guarantee was triggered. The reason of this comes from the common covariance between the Home insurance guarantee and the three random effects related to the three count variables in Motor insurance.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

4.2. Impact of dependence between Home and Motor insurance on corrections

Let us assess on an example in what way the dependence between Motor and Home insurance changes the corrections to apply.

We will consider a household with two policyholders, each holding a policy in Motor insurance, covering both TPL and MD. In addition, the household also holds a Home insurance policy. To ease the presentation, we will consider a unique risk profile for each guarantee and each of the policyholders. More specifically, we will consider the median expected a priori claim frequency for each of the four count variables (i.e. medium-risk profile in the previous example).

In order to measure the impact of the dependence between Home and Motor insurance, we will distinguish two cases:

1. The correlation structure of the random effects is given in Table 7;
2. The same variance–covariance matrix as above, except for the off-diagonal terms of the first row (respectively, column), is set to 0 (i.e. independence between Home and Motor insurance is assumed). The dependence in Motor insurance for policyholders from the same household is therefore kept at there levels given in Table 7.

Six examples are considered: (i) a claim-free case (meaning that in all three policies no claim was reported), (ii) one claim in Home insurance (during the first semester), (iii) one claim from policyholder $h(1)$ in Motor TPL, (iv) one claim from policyholder $h(1)$ in Motor MD, (v) one claim from policyholder $h(1)$ that triggered both TPL and MD simultaneously and (vi) two claims from policyholder $h(1)$: one triggering TPL and the other triggering MD.

4.2.1. Impact on Home insurance

Let us first show the corrections to apply in Home insurance, by discussing the six examples detailed above. The corrections in Motor insurance will be presented later. The corrections are depicted in Figure 8.

In the claim-free case, we see that both Motor policies help decrease the estimate by an extra 10% compared to the independence case. In the second example, where one claim was reported in Home insurance, the correction is above 1, but the dependence allows this correction to be weaker at first. As time goes by, the correction decreases faster than in the independence case thanks to the claim-free years in Motor insurance. In the next three examples, in which one claim was reported in Motor insurance, we observe similar corrections to be applied in Home insurance. We see that even though a claim was reported in Motor insurance, the correction factor in Home insurance falls below 1 after about 2 years. Note that the difference with the first example (i.e. no claims in any guarantee) is about 5% after 5 years when considering the dependence. So, correcting the Home insurance a priori expected claim frequency without considering the Motor insurance experience may yield too advantageous corrections in the considered examples. In the last example, we note the importance of distinguishing when a single event is triggering two guarantees (i.e. only one claim) compared to two events each triggering a different guarantee (i.e. two claims). Indeed, in the former case, the correction factor in Home insurance is 5–10% lower than in the latter case. These differences also exhibit the importance of identifying whether two claim counts are in fact related to a single event (i.e. one claim triggering both guarantees) or whether they are unrelated and should be considered as two separate claims.

Finally, note that as could be expected, the corrections in the independence case are the same in five of the six examples (i.e. when no claim was reported in Home insurance).

4.2.2. Impact on Motor insurance

Let us now assess the corrections to apply in Motor insurance (TPL and MD), by discussing the same six examples as previously. Even though the modelling in Motor insurance involves three

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

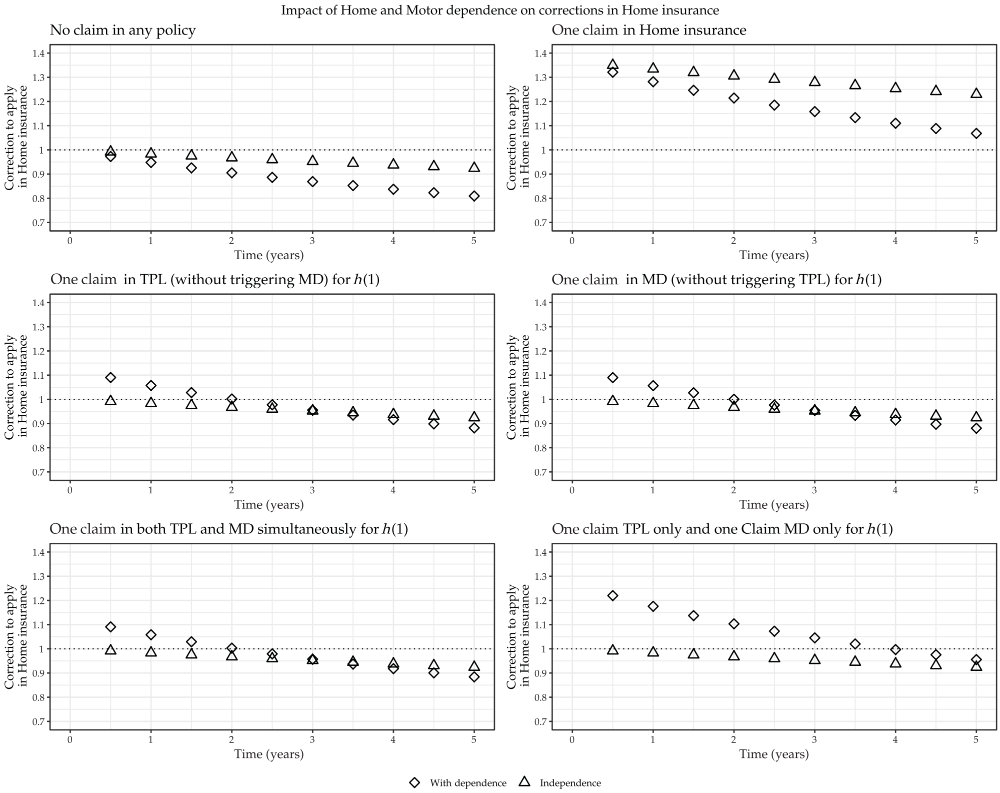

Figure 8. Impact of the dependence of the random effects related to Home and Motor insurance measured by the corrections in Home insurance. Six cases are considered, whether a claim occurred or not, and if so, which guarantees were triggered by the claim.

count variables, we can compute two correction factors, one for TPL and one for MD, using the formulas given in (2). The corrections are shown in Figure 9 for TPL and in Figure 10 for MD. Note that only the corrections for policyholder $h(1)$ (who is the only policyholder among the two policyholders in the household that experiences claims in Motor insurance) are shown. By looking at Figure 9 (respectively, Figure 10) which displays the corrections to apply to $h(1)$ in TPL (respectively, MD), we note that the dependence between Home and Motor insurance does not seem to considerably change the corrections. While the independence case considers that there is independence between Home and Motor insurance, no independence between guarantees in Motor insurance for different policyholders from the same household is assumed. This means that the observed number of claims of any policyholder in Motor insurance remains relevant in both cases displayed in Figures 9 and 10. Let us explain why the additional information related to Home insurance does not change a lot the correction factors. Compared to the aggregated a priori expected claim frequencies for two policies in Motor insurance (TPL and MD), the a priori expected claim frequency in Home insurance is actually small. Actually, it is comparable to the one of one policyholder in Motor TPL, so that it amounts to around 1/7 of the total expected claim frequency in the household. This means that a claim-free year in Home insurance will have a little impact on the correction factors in TPL and MD. Only when a claim occurs in Home insurance, it appears to have an impact on the correction factors in TPL and MD. In that situation, Figures 9 and 10 show that after 5 years there is a difference of about 5% in the correction factors to apply to TPL and MD due to the dependence with Home insurance.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

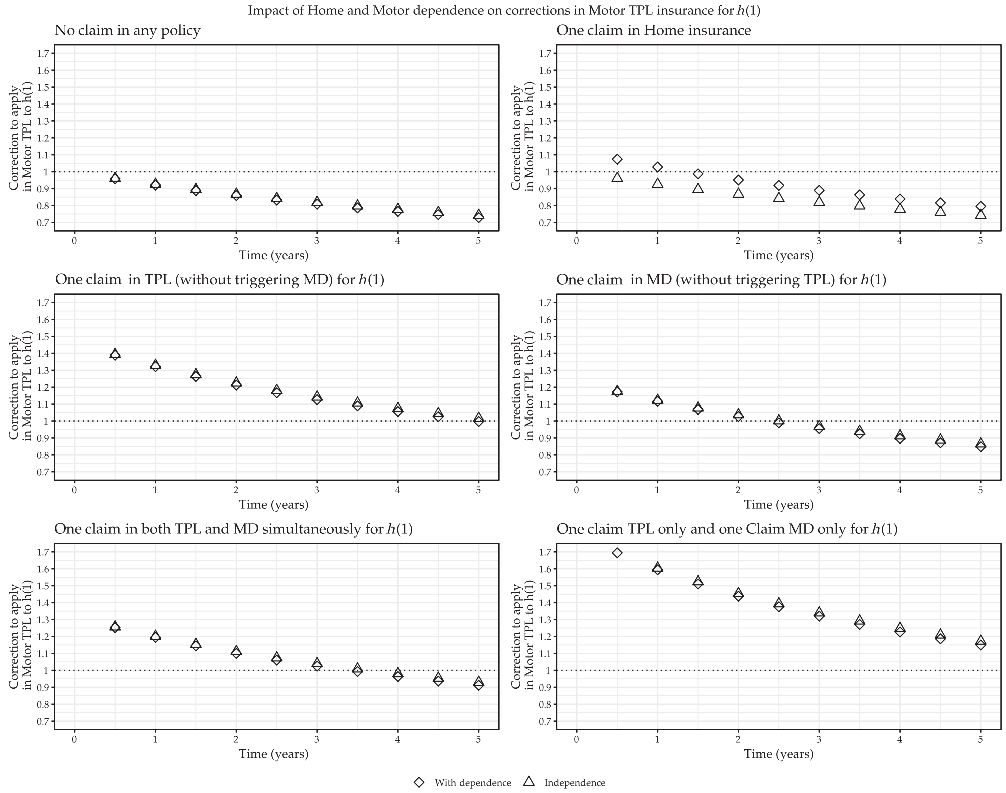

Figure 9. Impact of the dependence of the random effects related to Home and Motor insurance measured by the corrections in Motor TPL for $h(1)$. Six cases are considered, whether a claim occurred or not, and if so, which guarantees were triggered by the claim.

The corrections to apply to $h(2)$ (who does not experience any claims in any of the six cases) exhibit a similar difference between the independence case and the dependence case.

5. Predictive performance on out-of-sample data

It is common practice nowadays to compute the predictive performance on out-of-sample data. Indeed, by measuring the predictive power of a model on data that have not been used in the estimation, it is possible, for instance, to detect overfitting or assess the predictive power of a model.

Here, we only use years 1–2 and predict claim numbers of year 3. Thus, we set $T = 2$. The prediction of each claim frequency given by our model is the product of two factors: the expected a priori claim frequency from the GAM and a *correction factor*, the expectation of the random effect conditional to the reported number of claims from the past, that is, for policyholder i from household h, the expected claim frequency in year $T + 1$ in guarantee $g \in \mathcal{G}$ given past experience of the whole household is given by

$$\widehat{\lambda}^{g}_{h(i),T+1} \mathrm{E}\left[\Theta^{g}_{h(i)} | \mathbf{N}_{\mathbf{h}\bullet} = \mathbf{n}_{\mathbf{h}\bullet}\right]$$

where $\mathbf{n}_{\mathbf{h}\bullet}$ is the aggregated number of claims over the period $t = 1$ to $t = T$. Since the random effect related to policies of the same household is correlated, any number of claim of any policy of the household is predictive of any claim frequency.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

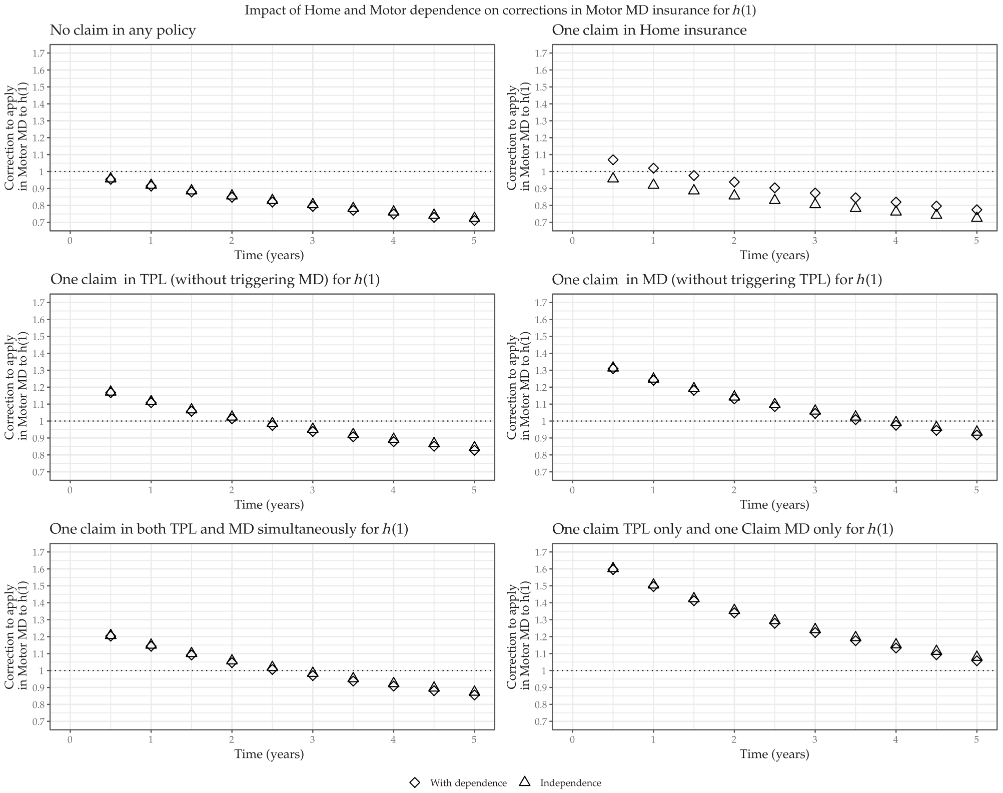

Figure 10. Impact of the dependence of the random effects related to Home and Motor insurance measured by the corrections in Motor MD for $h(1)$. Six cases are considered, whether a claim occurred or not, and if so, which guarantees were triggered by the claim.

We aim to compare our model to two different models. First, we compare our multivariate Poisson mixture to the expected claim frequencies obtained with the GAMs, that is, $\widehat{\lambda}^{g}_{h(i),T+1}$. Also, we wish to compare our multivariate Poisson mixture to a univariate Poisson mixture. Let us put labels on each of the three models, that is,

Model	Description
Model 1	GAMs
Model 2	GAMs + univariate a posteriori correction
Model 3	GAMs + multivariate a posteriori correction

The univariate Poisson mixture (model 2) considers that all the random effects are independent from each other, that is, only the number of claims of the policyholder itself in the considered guarantee is relevant for the correction of the predicted expected a priori claim frequency. In mathematical terms, for policyholder i from household h, the expected claim frequency in year $T+1$ in guarantee $g \in \mathcal{G}$ is given by

$$\widehat{\lambda}^{g}_{h(i),T+1} \mathrm{E}\left[\Theta^{g}_{h(i)} | N^{g}_{h(i),\bullet} = n^{g}_{h(i),\bullet}\right]$$

We will consider that in the univariate case, the variances of the random effects are the same as in the multivariate case: in the variance–covariance matrix, the original diagonal terms are kept while all the off-diagonal terms are put to 0.

Fivefolds have been constructed on the data to perform cross-validation. The folds have been constructed in the following way to ensure correct balancing.

For each household, the number of policies in Motor insurance (0, 1 or 2) and in Home insurance (0 or 1) is determined. Then, for each policy, the number of claims $(0, 1, 2+)$ reported by each policyholder in each guarantee during the two first years (2011 and 2012, i.e. $t = 1, 2$) is determined. Both, the number of policies and the numbers of claims are then crossed to obtain a categorical variable informing which policies the household has subscribed to and how many claims have been reported in each of these policies. The newly created categorical variable is then used to split the households into fivefolds, using stratified sampling. Stratified sampling ensures that all the claimed policies, or all the households with a single Motor insurance policy, are not grouped into a single fold.

Once the households have been split into the fivefolds, the cross-validation involves an estimation part and a prediction part. Let us describe the process for model 3. The process is similar for models 1 and 2.

Let us consider the first fold. In the first step, the multivariate Poisson mixture model (model 3) is estimated on the data that consist of all the folds, except the first fold. The first fold is therefore considered as out-of-sample, as it is not involved in the estimation process.

Next, the claim frequencies related to the household in the first fold are predicted: for each year and for each household, the estimated expected claim frequencies are obtained using the GAMs with the a priori available data.

Then, the observed number of claims of the two first years is used to compute the correction factor to apply to the third year's a priori expected claim frequencies from the household. The predictive performance can then be evaluated on the data from the first fold and related to the third year.

Finally, we can iterate and leave the second fold out of the estimation process and predict the claim frequencies on the second fold. It has to be noted that the 3 years of observations have similar claim frequencies on portfolio level.

We will use two different measures to assess the predictive power of our model. We will rely on the Poisson deviance as well as the loss ratio lift, which, in our context, will be the ratio of the actual number of claims to the expected number of claims. See Henckaerts *et al.* (2019) for more details on these measures.

5.1. Poisson deviance

Let us first use the Poisson deviance. Indeed, one way to measure the predictive power of a model with count data is to compute the Poisson deviance on the out-of-sample data. The Poisson deviance has been computed on each fold on the third year's predicted claim frequencies and observed number of claims.

First, we computed the deviance on Home insurance, Motor TPL and Motor MD separately. In Figure 11, on top, we see that the GAMs alone, as expected, have the worst performance in the three cases, and the multivariate Poisson mixture outperforms the univariate Poisson mixtures. We note, however, that in Home insurance, the difference is small between the univariate and the multivariate credibility models. This difference is, however, the greatest in Motor TPL (which is where the greatest heterogeneity has been found). On the bottom, the three cases are summed to obtain the deviance of the whole portfolio over all insurance products. Using a univariate Poisson mixture allows to take into account part of the heterogeneity and 'improves therefore the predictability. However, the GAMs along with multivariate credibility appear to outperform both models, on every fold, by exploiting the dependence between the latent risk factors.

Let us note that small differences in deviances for insurance type count data (i.e. low claim frequencies) can in reality imply large differences in the predicted claim frequencies. Therefore, we will also use a second method to assess the predictive power of our model by using the lift.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

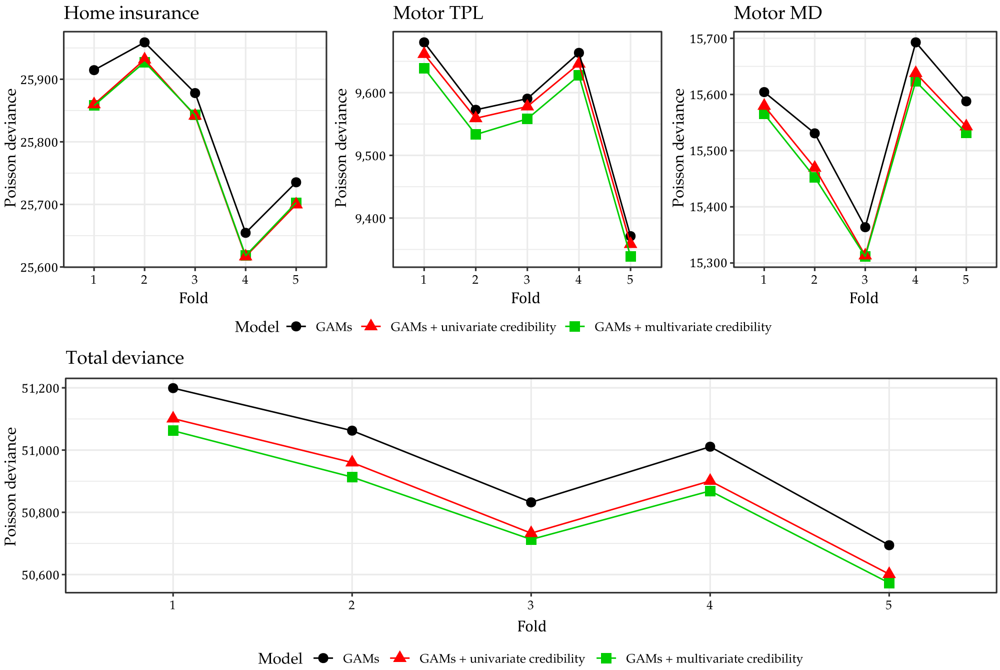

Figure 11. Poisson deviance on out-of-sample data. Three models are considered: GAMs, GAMs corrected by a univariate credibility model and GAMs corrected by the multivariate credibility model. Top: the deviance for Home insurance and Motor TPL and MD is illustrated. Bottom: total deviance (sum of the three deviances).

5.2. Lift

Another way to compare the predictive performance is to rely on the lift. Define the relativity $r^{g}_{h(i)}$ as the ratio of the a posteriori expected claim frequency according to the competing model to that of the benchmark model. If the model does not include any random effect (model 1), then the a posteriori expectation reduces to $\widehat{\lambda}^{g}_{h(i),T+1}$.

Based on these relativities, it is then possible to bin policyholders into groups of similar exposure. Then, on each bin, the loss ratio, that is, the ratio of the sum of the actual number of claims observed in the bin to the sum of the expected number of claims in the bin given by the benchmark model. A loss ratio of 1 for each group means that the benchmark model is perfectly accurate. The loss ratios for each group can be plotted and, if an upward trend is visible, it implies that the competing model outperforms the benchmark model. The intuitive idea is as follows. Since we have sorted the relativities, from smallest to largest, if we have an upward trend, then it means that for small relativities, the loss ratio is below 1. Consequently, since the relativity is small, the competing model is predicting less claims than the benchmark model, and the loss ratio below 1 implies that there are less actual claims than predicted by the benchmark model. Therefore, it could be that competing model is predicting better than the benchmark model. A similar argument holds for larger relativities. See Henckaerts *et al.* (2019) for a comprehensive presentation of the lift.

Let us use the lift methodology in two different comparisons. First, we will compare model 3 (comp) with model 1 (bench). Then, we will compare model 3 (comp) with model 2 (bench). In each comparison, we will consider the three different claim frequencies that have been covered in this paper: claim frequencies in Home insurance, in Motor TPL and in Motor MD.

The computations are done in the third year of fold 1, when the model has been estimated on the four remaining folds (i.e. folds 2–5). Since we are using the third year to compute the lift, we

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

can use the two first years of observations to compute a posteriori corrections for the third year. When comparing model 3 with model 1, the relativities can be written as

$$\begin{cases} r_{h(i)}^{Home} = \dfrac{\lambda_{h,3}^{Home}\mathrm{E}\left[\Theta_h^{Home}|\mathbf{N_{h,1-2}}=\mathbf{n_{h,1-2}}\right]}{\lambda_{h,3}^{Home}} = \mathrm{E}\left[\Theta_h^{Home}|\mathbf{N_{h,1-2}}=\mathbf{n_{h,1-2}}\right] \\ r_{h(i)}^{TPL} = \dfrac{\lambda_{h(i),3}^{TPL}\mathrm{E}\left[\Theta_{h(i)}^{TPL}|\mathbf{N_{h,1-2}}=\mathbf{n_{h,1-2}}\right] + \lambda_{h(i),3}^{MD:TPL}\mathrm{E}\left[\Theta_{h(i)}^{MD:TPL}|\mathbf{N_{h,1-2}}=\mathbf{n_{h,1-2}}\right]}{\lambda_{h(i),3}^{TPL}+\lambda_{h(i),3}^{MD:TPL}} \\ r_{h(i)}^{MD} = \dfrac{\lambda_{h(i),3}^{MD}\mathrm{E}\left[\Theta_{h(i)}^{MD}|\mathbf{N_{h,1-2}}=\mathbf{n_{h,1-2}}\right] + \lambda_{h(i),3}^{MD:TPL}\mathrm{E}\left[\Theta_{h(i)}^{MD:TPL}|\mathbf{N_{h,1-2}}=\mathbf{n_{h,1-2}}\right]}{\lambda_{h(i),3}^{TPL}+\lambda_{h(i),3}^{MD:TPL}} \end{cases}$$

where $\mathbf{n_{h,1-2}}$ is a vector indicating the number of claims observed in every policy of the household during years 1 and 2, that is, in a household h with a Home insurance policy and two Motor insurance policies,

$$\mathbf{n_{h,1-2}} = \sum_{\mathbf{t=1}}^{\mathbf{2}} \left(\mathbf{n_{h,t}^{Home}}, \mathbf{n_{h(1),t}^{TPL}}, \mathbf{n_{h(1),t}^{MD}}, \mathbf{n_{h(1),t}^{MD:TPL}}, \mathbf{n_{h(2),t}^{TPL}}, \mathbf{n_{h(2),t}^{MD}}, \mathbf{n_{h(2),t}^{MD:TPL}}\right)'$$

while when comparing model 3 with model 2, the relativities are given by

$$\begin{cases} r_{h(i)}^{Home} = \dfrac{\lambda_{h,3}^{Home}\mathrm{E}\left[\Theta_h^{Home}|\mathbf{N_h}=\mathbf{n_h}\right]}{\lambda_{h,3}^{Home}\mathrm{E}\left[\Theta_h^{Home}|N_h^{Home}=n_h^{Home}\right]} \\ r_{h(i)}^{TPL} = \dfrac{\lambda_{h(i),3}^{TPL}\mathrm{E}\left[\Theta_{h(i)}^{TPL}|\mathbf{N_h}=\mathbf{n_h}\right] + \lambda_{h(i),3}^{MD:TPL}\mathrm{E}\left[\Theta_{h(i)}^{MD:TPL}|\mathbf{N_h}=\mathbf{n_h}\right]}{\lambda_{h(i),3}^{TPL}\mathrm{E}\left[\Theta_{h(i)}^{TPL}|N_{h,\bullet}^{TPL}=n_h^{TPL}\right] + \lambda_{h(i),3}^{MD:TPL}\mathrm{E}\left[\Theta_{h(i)}^{MD:TPL}|N_{h,\bullet}^{MD:TPL}=n_h^{MD:TPL}\right]} \\ r_{h(i)}^{MD} = \dfrac{\lambda_{h(i),3}^{MD}\mathrm{E}\left[\Theta_{h(i)}^{MD}|\mathbf{N_h}=\mathbf{n_h}\right] + \lambda_{h(i),3}^{MD:TPL}\mathrm{E}\left[\Theta_{h(i)}^{MD:TPL}|\mathbf{N_h}=\mathbf{n_h}\right]}{\lambda_{h(i),3}^{MD}\mathrm{E}\left[\Theta_{h(i)}^{MD}|N_{h,\bullet}^{MD}=n_h^{MD}\right] + \lambda_{h(i),3}^{MD:TPL}\mathrm{E}\left[\Theta_{h(i)}^{MD:TPL}|N_{h,\bullet}^{MD:TPL}=n_h^{MD:TPL}\right]} \end{cases}$$

The policies are then sorted based on their relativity and are binned into 25 risk classes of equal (or almost equal) exposures. On each of these risk classes, the aforementioned loss ratio can be computed, comparing for each risk class the actual number of reported claims with the expected number of reported claims (given by the benchmark model).

The resulting loss ratios are illustrated in Figure 12 in blue. We have used a local polynomial regression to compute the smooth curves. We also computed in dark orange the loss ratios with the competing model (i.e. the multivariate credibility model) in the denominator.

First, let us comment on the plots on the left, which compare our multivariate credibility model (model 3) to the GAMs (model 1). The blue curves suggest an increasing trend, in particular, in the higher risk classes. The higher risk classes imply that the multivariate credibility model yields higher claim frequencies than the GAMs, whereas the high loss ratio observed on this risk classes suggests that the GAMs underestimate clearly the claim frequencies. As explained above, if the blue line is not close to 1, then the benchmark model is not perfectly accurate. Moreover, if an uprising trend is visible, then, in fact, the competing model is outperforming the benchmark model. Therefore, one can say that the multivariate credibility model will at least less underestimate the expected claim frequencies, but could in fact overestimate the expected claim frequencies. This is the reason why we also plotted the orange curve which computes the loss

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

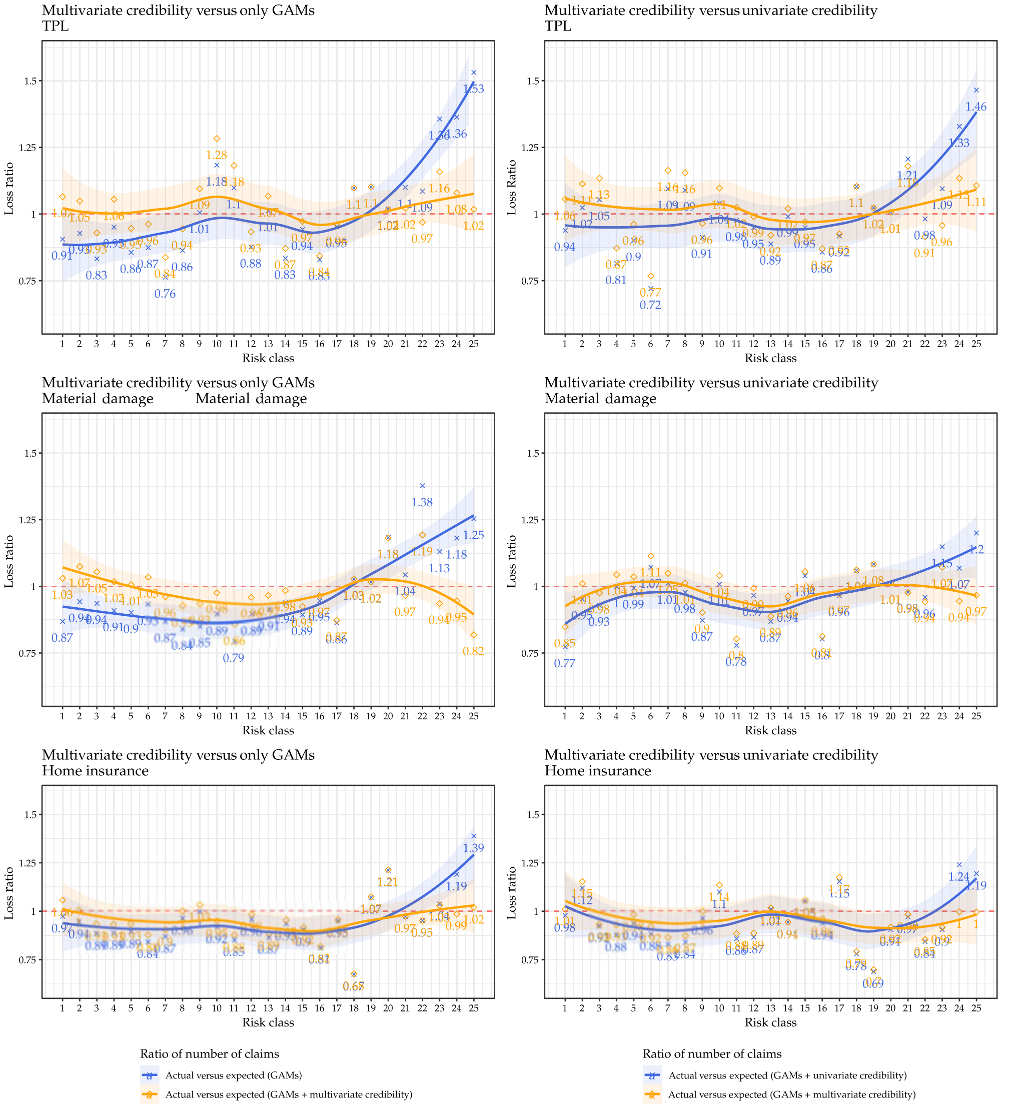

Figure 12. Lift on out-of-sample data. Two comparisons are considered. Left: multivariate credibility is compared to the predictions of the GAMs. Right: multivariate credibility is compared to the predictions of univariate credibility.

ratio between the actual and the expected number of claims (given by the multivariate credibility model). As can be seen, for instance, on the plot on top left, the orange curve is much closer to the horizontal line at level 1. Being close to this red line suggests that the expected number of claims are closer to the actual number of claims (i.e. the model is predicting better on this risk class). It has to be noted that for the orange curve, only the closeness to the horizontal line at level 1 has to be seen (since we are not comparing the predictions with respect to another model). As such, we see that for the three plots on the left, the orange curve is, for most risk classes, closer to the red line than the blue curve is (i.e. the multivariate credibility model outperforms the GAMs on these risk classes). This is especially true for the higher risk classes, where our multivariate credibility model was able to account for the higher latent risk factors.

Let us now compare the multivariate credibility model (model 3) with the univariate credibility model (model 2). In TPL and MD, the blue lines again suggest an increasing trend, especially

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

for the second half of the risk classes. For Home insurance, this upward trend is not as visible. However, when looking at the orange curve, we can see that again, for many classes, the multivariate credibility model is close to predicting the actual number of claims on each risk class (i.e. close to 1). We note that again, there is a big discrepancy in this last risk classes (i.e. the riskiest classes). Indeed, compared to the univariate credibility model, the multivariate credibility is able to better identify the riskiest households by using information related to other products and other policyholders from the household.

6. An alternative to the two-step estimation

6.1. Description of the algorithm

In this paper, model fit is based on a two-step approach. First, the expected a priori claim frequencies are estimated with GAMs using the a priori available information. The expected a priori claim frequencies are then fixed and random effects are introduced. In the second step, the variance–covariance matrix of these random effects is estimated. Therefore, one could say that a limitation of our approach is that parameters in the GAMs could depend on the variance–covariance matrix and could therefore have different estimates, would we have estimated both the parameters of the GAMs and the variance–covariance matrix at the same time.

In this section, we aim to conduct an experiment and show how we can actually cycle both steps. We can summarise the cycling of the two-part approach in the following ways:

1. We compute the conditional expectation of the random effects in the current model. If this is the first time we iterate, these conditional expectations are fixed at 1.
2. The GAMs are estimated, with the conditional expectations found in step 1 put in the offset.
3. The variance–covariance matrix of the random effects is estimated by using the estimated expected a priori claim frequencies from the previous step.
4. Repeat steps 1–3, until convergence is achieved.

The algorithm described resembles to what is known as an ECM algorithm (see Meng & Rubin 1993). Indeed, the E-step corresponds in our situation to conditional expectation computed in step 1. Then, the conditional maximisation step corresponds to both steps 2 and 3. One key difference, however, relates to the fact that at both M-steps (steps 2 and 3 in the described algorithm above), two different objective functions are maximised, while Meng & Rubin (1993) maximise the same objective function at the different M-steps. Indeed, at step 2, the Poisson log-likelihood is maximised, whereas at step 3, the log-likelihood of the multivariate Poisson mixture is maximised.

We stress out that, since this algorithm cannot be considered as an ECM algorithm, further research is necessary to assess its validity. Therefore, this section, along with the practical application of the algorithm that follows, should only be considered as a preliminary experiment and further research could be devoted to this issue.

6.2. Practical application of the algorithm

This algorithm has been implemented on the model presented in this paper. The steps have been repeated five times. We wish now to illustrate the coefficients of the discrete variables in Tables 8–11. The parameters do appear to change a bit after the first iteration and seem to converge rapidly. Note that, for confidentiality reasons, we are not allowed to show the intercept. We can, however, state that the absolute change in the intercept between the first and last iterations was as follows: 0.003 in TPL, −0.018 in MD, −0.024 in MD:TPL and −0.007 in Home insurance.

For the continuous variables, we show, in Figure 13, the estimated effects at iterations 1 and 5 on the log scale. We see that the differences remain small (below 1% in most cases).

Table 8. Estimate of discrete parameters in TPL at each iteration of the cycling algorithm.

Iteration	New car	Prof. usage	Litigation 2	Litigation 4	Female
1	−0.146	0.316	0.365	0.887	0.051
2	−0.153	0.316	0.362	0.890	0.050
3	−0.153	0.316	0.362	0.890	0.050
4	−0.153	0.316	0.362	0.890	0.050
5	−0.153	0.316	0.362	0.890	0.050

Table 9. Estimate of discrete parameters in MD at each iteration of the cycling algorithm.

Iteration	New car	Prof. usage	Litigation 2	Litigation 4	Female
1	0.468	0.082	0.207	0.330	0.065
2	0.457	0.082	0.206	0.323	0.061
3	0.456	0.082	0.206	0.322	0.061
4	0.456	0.082	0.206	0.322	0.061
5	0.456	0.082	0.206	0.322	0.061

Table 10. Estimate of discrete parameters in MD:TPL at each iteration of the cycling algorithm.

Iteration	New car	Prof. usage	Litigations 2–4	Female	Power high
1	0.144	0.142	0.295	0.112	0.211
2	0.137	0.141	0.293	0.111	0.199
3	0.137	0.141	0.293	0.111	0.198
4	0.137	0.141	0.293	0.111	0.198
5	0.137	0.141	0.293	0.111	0.198

Table 11. Estimate of discrete parameters in Home insurance at each iteration of the cycling algorithm.

Iteration	Tenant	Apartment	Contiguity	Tenant:Appartment
1	−1.740	0.062	0.191	−0.181
2	−1.742	0.060	0.192	−0.179
3	−1.742	0.060	0.192	−0.179
4	−1.742	0.060	0.192	−0.179
5	−1.742	0.060	0.192	−0.179

Finally, for the geographic effect, Figures 14 and 15 compare the estimated effects at iterations 1 and 5.

We also comment on the parameters involved in the variance–covariance matrix. The estimates at the different steps are given in Table 12. We note that, similarly to the parameters of the GAMs, the changes occur mainly at the first iteration steps, and convergence is rapid.

To sum up, cycling the different steps produces estimates that rapidly stabilise, with only limited changes between first and last iterations.

7. Discussion

In this paper, we have presented a model that allows the actuary to account for the dependence between two different products which can have multiple guarantees as well as multiple

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

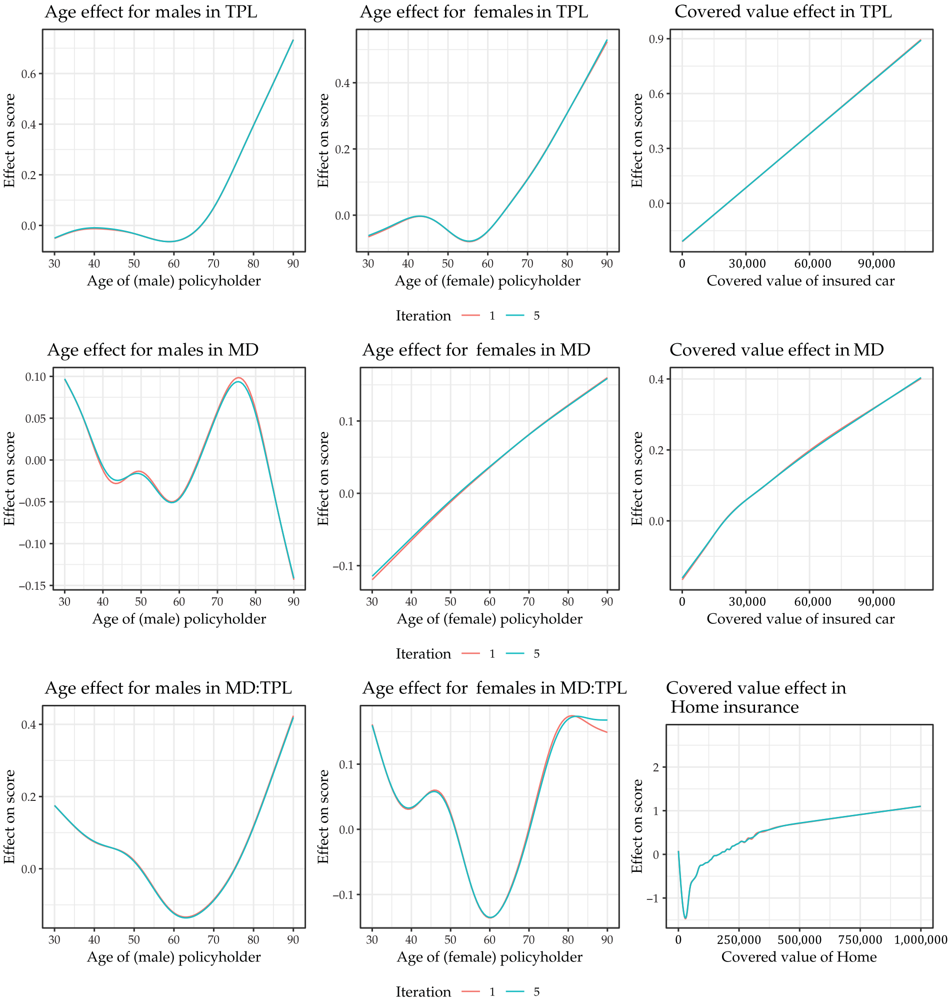

Figure 13. Effect on score at the first iteration and the last iteration (= 5) for the univariate continuous variables.

policyholders from the same household. The model captures the dependencies that exist between unobserved risk factors that influence the claim frequencies. Home insurance is subscribed by one specific policyholder, but in reality, the whole household is covered by the policy. In Motor insurance, each policy relates to a single car and policyholder. The proposed model takes into account these specificities by the introduction of a hierarchical dependence structure of the random effects which model the unobserved risk factors. On top of easing the parametrisation, the hierarchical structure brings some additional interpretability, at the cost of forcing a positive correlation.

The results show that there is a dependence between Home and Motor insurance. Consequently, any claim from any guarantee is relevant to refine the predictions on the claim frequencies on both considered products. The hierarchical structure hints towards the fact that part of the unobserved risk factors are common to the whole household, whereas some are more specific to the product (i.e. Home or Motor insurance), and others are specific to the different guarantees in Motor insurance. The results shown in the applications stress out the benefits for the insurer to have customers who own multiple policies over different lines of business, as it improves its understanding about the customer's risk profile.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

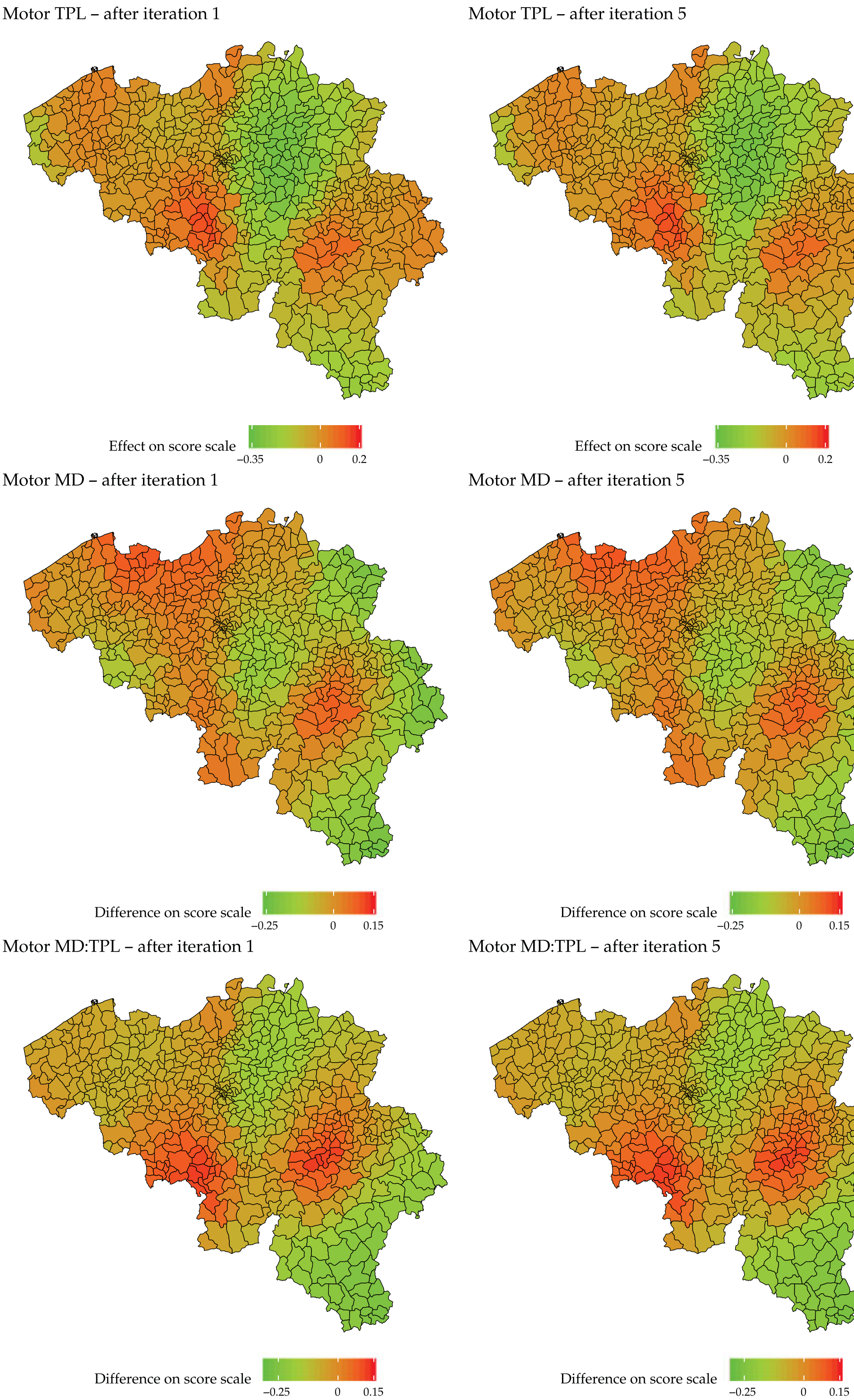
Motor TPL – after iteration 1
Motor TPL – after iteration 5
Effect on score scale
−0.35
0
0.2
Effect on score scale
−0.35
0
0.2
Motor MD – after iteration 1
Motor MD – after iteration 5
Difference on score scale
−0.25
0
0.15
Difference on score scale
−0.25
0
0.15
Motor MD:TPL – after iteration 1
Motor MD:TPL – after iteration 5
Difference on score scale
−0.25
0
0.15
Difference on score scale
−0.25
0
0.15

Figure 14. Effect on score of the geographic effects estimated at the first iteration and the last iteration (= 5) in Motor insurance.

https://doi.org/10.1017/S1748499520000160 Published online by Cambridge University Press

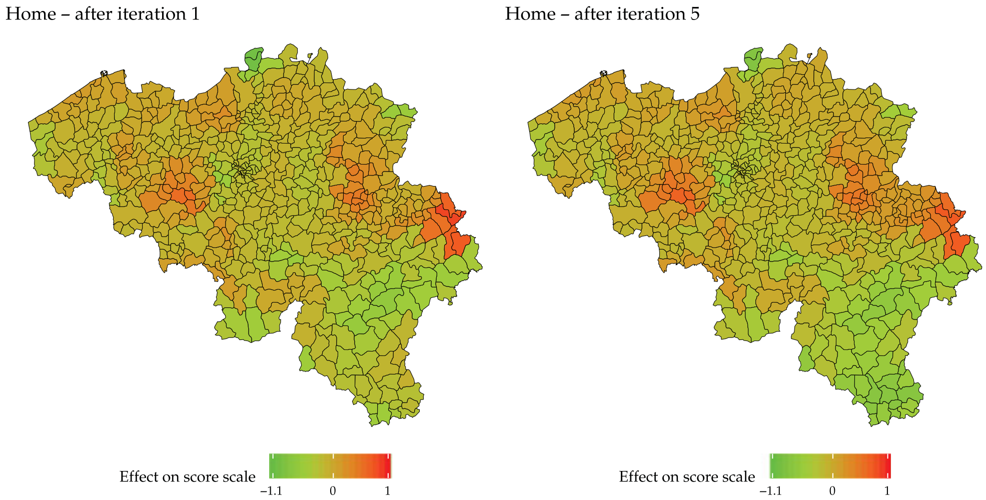

Figure 15. Effect on score of the geographic effect estimated at the first iteration and the last iteration (= 5) in Home insurance.

Table 12. Estimate of the parameters defining the variance–covariance matrix at each iteration of the cycling algorithm.

Iteration	$\widehat{\varsigma}^{2}_{Household}$	$\widehat{\varsigma}^{2}_{Home}$	$\widehat{\varsigma}^{2}_{Motor}$	$\widehat{\varsigma}^{2}_{TPL}$	$\widehat{\varsigma}^{2}_{MD}$	$\widehat{\varsigma}^{2}_{MD:TPL}$
1	0.1238	0.1872	0.1141	0.3213	0.1272	0.1336
2	0.1245	0.1870	0.1139	0.3229	0.1277	0.1348
3	0.1243	0.1872	0.1146	0.3217	0.1275	0.1340
4	0.1243	0.1872	0.1146	0.3217	0.1275	0.1340
5	0.1243	0.1872	0.1146	0.3217	0.1275	0.1340

We need to stress out that one limitation of this paper is that the models only rely on claim frequencies and do not consider the cost of claims. If costs of claims were to be considered, then some specificities would need to be taken into account. For instance, in Motor insurance, two different types of claims should be distinguished (bodily damage and MD), due to the fact that bodily damages are in general way more costly. By considering only the claim frequencies, the models are therefore greatly simplified. Replacing the Poisson distribution with the Tweedie model appears to be a promising way to include costs in the analysis.

The out-of-sample analysis indicated that the multivariate credibility model improves the prediction compared to a univariate mixture. The Poisson deviances suggest that our model outperforms both the univariate Poisson mixture as well as the GAMs. However, the Poisson deviance did not allow us to identify for which risk classes our model is particularly better in terms of predictive power. The analysis with the lift allowed us to identify these and showed that our model is better at identifying the riskiest policyholders (i.e. the policyholders having many claims across all their policies).

Finally, an experiment was conducted by proposing an algorithm that cycles the estimation of the GAMs and of the variance–covariance matrix. The algorithm showed a rapid convergence, while the parameters of the GAMs remained, for the most, close to a 1% deviation to those obtained using the two-part approach. Further work is, however, necessary to assess the validity of the method from a theoretical point of view.

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 and Stanislas Roth 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.

Supplementary Material. For supplementary material referred to in this article, please visit https://doi.org/10.1017/S1748499520000160.

References

Agresti, A. (2003). *Categorical Data Analysis*, vol. **482**. John Wiley & Sons.

Antonio, K. & Zhang, Y. (2014). Nonlinear mixed models. In E.W. Frees, R.A. Derrig & G. Meyers (Eds.), *Predictive Modeling Applications in Actuarial Science* (pp. 398–424). International Series on Actuarial Science, vol. 1. Cambridge University Press.

Antonio, K., Guillén, M., Pérez Martn, A.M., et al. (2010a). Multidimensional credibility: a Bayesian analysis of policyholders holding multiple policies, technical report, Amsterdam School of Economics Research Institute.

Antonio, K., Frees, E.W. & Valdez, E.A. (2010b). A multilevel analysis of intercompany claim counts. *ASTIN Bulletin: The Journal of the IAA*, **40**(1), 151–177.

Bermúdez, L. (2009). A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics*, **44**(1), 135–141.

Bermúdez, L., Guillén, M. & Karlis, D. (2018). Allowing for time and cross dependence assumptions between claim counts in ratemaking models. *Insurance: Mathematics and Economics*, **83**, 161–169.

Bermúdez, L. & Karlis, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, **48**(2), 226–236.

Bermúdez, L. & Karlis, D. (2017). A posteriori ratemaking using bivariate Poisson models. *Scandinavian Actuarial Journal*, **2017**(2), 148–158.

Boucher, J.-P. & Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin*, **44**(3), 587–612.

Brockett, P.L., Golden, L.L., Guillen, M., Nielsen, J.P., Parner, J. & Perez-Marin, A.M. (2008). Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? *Journal of Risk and Insurance*, **75**(3), 713–737.

Denuit, M., Maréchal, X., Pitrebois, S. & Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons.

Eddelbuettel, D. & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**(8), 1–18.

Frees, E.W. (2003). Multivariate credibility for aggregate loss models. *North American Actuarial Journal*, **7**(1), 13–37.

Frees, E.W. & Valdez, E.A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, **103**(484), 1457–1469.

Frees, E.W. & Wang, P. (2005). Credibility using copulas. *North American Actuarial Journal*, **9**(2), 31–48.

Frees, E.W. & Wang, P. (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics*, **38**(2), 360–373.

Frees, E.W., Shi, P. & Valdez, E.A. (2009). Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin*, **39**(1), 165–197.

Frees, E.W., Bolancé, C., Guillen, M. & Valdez, E. (2018). Copula modeling of multivariate longitudinal data with dropout. arXiv preprint arXiv:1810.04567.

Guillen, M., Nielsen, J.P. & Pérez-Marín, A.M. (2008). The need to monitor customer loyalty and business risk in the European insurance industry. *The Geneva Papers on Risk and Insurance – Issues and Practice*, **33**(2), 207–218.

Henckaerts, R., Côté, M.-P., Antonio, K. & Verbelen, R. (2019). Boosting insights in insurance tariff plans with tree-based machine learning. arXiv preprint arXiv:1904.10890.

Jewell, W.S. (1974). Exact multidimensional credibility. *Bulletin of Swiss Association of Actuaries*, **74**, 193–214.

Karlis, D. & Pedeli, X. (2013). Flexible bivariate INAR (1) processes using copulas. *Communications in Statistics-Theory and Methods*, **42**(4), 723–740.

Kroeze, K. (2016). Multighquad: Multidimensional Gauss-Hermite Quadrature. R package version 1.2.0.

Meng, X.-L. & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**(2), 267–278.

Pechon, F., Trufin, J. & Denuit, M. (2018). Multivariate modelling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin*, **48**(3), 969–993.

Pechon, F., Denuit, M. & Trufin, J. (2019). Multivariate modelling of multiple guarantees in motor insurance of a household. *European Actuarial Journal*, **9**(2), 575–602.

Pinquet, J. (1998). Designing optimal bonus-malus systems from different types of claims. *ASTIN Bulletin*, **28**(2), 205–220.

Purcaru, O. & Denuit, M. (2003). Dependence in dynamic claim frequency credibility models. *ASTIN Bulletin*, **33**(1), 23–40.

Self, S.G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**(398), 605–610.

Shi, Peng, & **Valdez, Emiliano A.** (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, **55**(1), 18–29.

Shi, P. & Yang, L. (2018). Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association*, **113**(521), 122–133.

Shi, P., Feng, X. & Boucher, J.-P. (2016). Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics*, **10**(2), 834–863.

Tuerlinckx, F., Rijmen, F., Verbeke, G. & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: a review. *British Journal of Mathematical and Statistical Psychology*, **59**(2), 225–255.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd edition. Chapman and Hall/CRC.

Cite this article: Pechon F, Denuit M and Trufin J (2021). Home and Motor insurance joined at a household level using multivariate credibility. *Annals of Actuarial Science* **15**, 82–114. https://doi.org/10.1017/S1748499520000160