HOME AND MOTOR INSURANCE JOINED AT A HOUSEHOLD LEVEL USING MULTIVARIATE CREDIBILITY

Florian Pechon; Michel Denuit and Julien Trufin

DISCUSSION PAPER | 2019 / 13

UCLouvain

LIDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

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

Florian Pechon*
Institute of Statistics, Biostatistics and Actuarial Science
Université catholique de Louvain (UCLouvain)
Louvain-la-Neuve, Belgium

Michel Denuit
Institute of Statistics, Biostatistics and Actuarial Science
Université catholique de Louvain (UCLouvain)
Louvain-la-Neuve, Belgium

Julien Trufin
Department of Mathematics
Université Libre de Bruxelles (ULB)
Bruxelles, Belgium

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 accommodates 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 claim frequencies using the reported number of claims in any of the considered products. Some cross-selling opportunities are identified.

Keywords - Multivariate credibility, Poisson-LogNormal, a posteriori claim frequency.

* Correspondence to: Florian Pechon, Université catholique de Louvain, Institute of Statistics, Biostatistics and Actuarial Science, Voie du Roman Pays, 20, Louvain-la-Neuve, Belgium. Email florian.pechon@uclouvain.be

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 informations available at inception of the policy. These informations include covariates about the policyholder himself, 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 Generalized Linear Models (GLMs) and Generalized 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 informations 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 risks. 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 (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 Third-Party Liability (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.

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 customers starting from the household's first policy cancellation until all the policies have been lapsed.

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 another 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 e.g. 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.

One way to induce dependence is by introducing a copula. Credibility theory can be used in conjunction of copulas, see e.g. 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 number of claims using a Bivariate Integer-valued AutoRegressive regression model (BINAR, 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 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 dependences 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. A maximum likelihood approach was used in the estimation. 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 Bonus-Malus system that models the dependence between at-fault and not-at-fault claims in Motor insurance using bivariate credibility. Boucher & Inoussa (2014) proposed a Bonus-Malus 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 expected in a household. Moreover, different kind of claims are observed in Motor insurance. The paper aims at taking into account all the claims' related information about both products in each household to update the household's claim frequencies.

Our approach can be summarized as follows. First we estimate by maximum likelihood the four kind of a priori 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 a hierarchical structure 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 restricted to Motor Insurance. We follow a maximum likelihood approach, implying that we need to specify the distribution of the random effects. Similarly to Pechon *et al.* (2018) and Pechon *et al.* (2019), the random effects are assumed to be Gaussian. Since the random effects are latent, the evaluation of the likelihood 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 dataset, 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 7-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 correlated the claim frequencies and also the number of claims. The results show that the latent risk 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 main contribution of this paper is on the application side. The model shows using a household (or family) approach that the claim frequencies of different products, subscribed by different policyholders are correlated. The model can be considered as an extension of Pechon *et al.* (2019). Indeed, in Pechon *et al.* (2019), only Motor insurance is considered whereas in this paper, Home insurance is included in addition to the two guarantees in Motor insurance, TPL and MD. Also, in this paper the random effects are correlated using a hierarchical structure, allowing more interpretability.

The model can be used to assess the value of a household, using all the available information: Using a univariate approach, one policyholder in one product may appear to be of poor value for the insurer, while using our multivariate approach makes it possible for the insurer to see the whole picture and may in fact realize that the household as a whole is profitable. Consequently, as shown in the applications, the model can also be used to identify cross-selling opportunities. We also show, with the help of numerical illustrations, the importance of identifying which claims in TPL and in MD originate in fact from the same event and should be considered as only one claim triggering both guarantees at the same time.
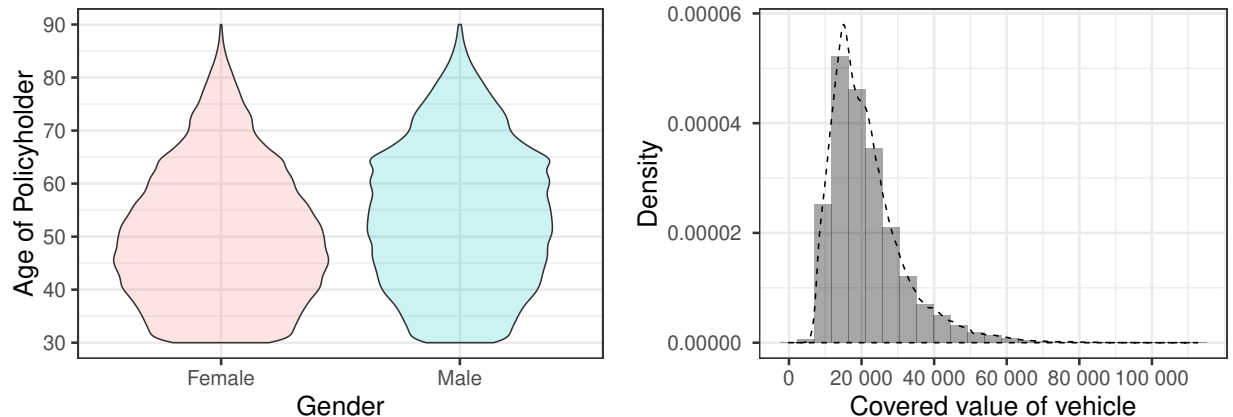
Figure 1: Left: Average age of policyholders split by gender during the years 2011 to 2013. Right: Covered value in Euro of the insured vehicles.

2 Datasets

Let us briefly describe the datasets that will be used to support our analysis. Two detailed datasets relate to portfolios of policies of the same European Insurance company observed during the years 2011-2013. The first dataset is the one used in Pechon *et al.* (2019) and relates to policyholders that 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 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 to 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. Finally, a family identifier allows to match the policyholders that live in the same household. Some statistics about the available covariates are given in Table 1 as well as on Figure 1.

	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 ($\leq$ 3 years)	62.68%		
High	4.10%	Old ($>$ 3 years)	37.32%		

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

The second dataset 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 third-party liability (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 dataset 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 datasets 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.

	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

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

3 Model

3.1 Notations

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

- $N_{h(i),t}^{TPL}$: Number of claims of policyholder i from household h that triggered *only* TPL during year t;

- $N_{h(i),t}^{MD}$: Number of claims of policyholder i from household h that triggered *only* MD during year t;

- $N_{h(i),t}^{MD:TPL}$: Number of claims of policyholder i from household h that triggered both TPL and MD simultaneously during year t;

- $N_{h,t}^{Home}$: Number of claims of household h that triggered Home insurance during year t.

Note that to simplify some expressions, sometimes we will write $N_{h(i),t}^{Home}$ when we actually mean $N_{h,t}^{Home}$ (e.g. in the first assumption later in this section). Clearly, $N_{h(i),t}^{Home} = N_{h,t}^{Home}$ 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 (resp. MD) during year t is $N_{h(i),t}^{TPL} + N_{h(i),t}^{MD:TPL}$ (resp. $N_{h(i),t}^{MD} + N_{h(i),t}^{MD:TPL}$).

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

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

- $N_{h(i),\bullet}^{TPL} = \sum_{t=1}^{T} N_{h(i),t}^{TPL}$;

- $N_{h(i),\bullet}^{MD} = \sum_{t=1}^{T} N_{h(i),t}^{MD}$;

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

- $N_{h,\bullet}^{Home} = \sum_{t=1}^{T} N_{h,t}^{Home}$.

The corresponding a priori claim frequencies are denoted $\lambda_{h(i),\bullet}^{TPL} = \mathrm{E}\left[N_{h(i),\bullet}^{TPL}\right]$, $\lambda_{h(i),\bullet}^{MD} = \mathrm{E}\left[N_{h(i),\bullet}^{MD}\right]$, $\lambda_{h(i),\bullet}^{MD:TPL} = \mathrm{E}\left[N_{h(i),\bullet}^{MD:TPL}\right]$ and $\lambda_{h,\bullet}^{Home} = \mathrm{E}\left[N_{h,\bullet}^{Home}\right]$. 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 $\mathcal{G} = \mathcal{G}_{Motor} \cup \{Home\}$.

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 $\Theta_{h(i)}^{g} = \theta$, the random variables $N_{h(i),1}^{g}, \ldots, N_{h(i),T}^{g}$ 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 $(\Theta_{h(i)}^{g_i}, \Theta_{h(j)}^{g_j}) = (\theta_i, \theta_j)$, 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 $N_{h(j),1}^{g_j}, N_{h(j),2}^{g_j}, \ldots, N_{h(j),T}^{g_j}$ are independent.

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 $\boldsymbol{\Theta}_h = (\Theta_h^{Home}, \Theta_{h(1)}^{TPL}, \Theta_{h(1)}^{MD}, \Theta_{h(1)}^{MD:TPL}, \Theta_{h(2)}^{TPL}, \Theta_{h(2)}^{MD}, \Theta_{h(2)}^{MD:TPL})$ the vector of all random effects. We assume that $\mathrm{E}(\boldsymbol{\Theta}) = \mathbf{1}$. 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 $\mathcal{E}_h^{Household} \sim N\left(-\frac{\varsigma_{Household}^2}{2}, \varsigma_{Household}^2\right)$, $\mathcal{E}_h^{Home} \sim N\left(-\frac{\varsigma_{Home}^2}{2}, \varsigma_{Home}^2\right)$, $\mathcal{E}_h^{Motor} \sim N\left(-\frac{\varsigma_{Motor}^2}{2}, \varsigma_{Motor}^2\right)$, $\mathcal{E}_{h(i)}^{PolId} \sim N\left(-\frac{\varsigma_{PolId}^2}{2}, \varsigma_{PolId}^2\right)$, $\forall g \in \mathcal{G}_{Motor}$ $\mathcal{E}_{h(i)}^{g} \sim N\left(-\frac{\varsigma_g^2}{2}, \varsigma_g^2\right)$.
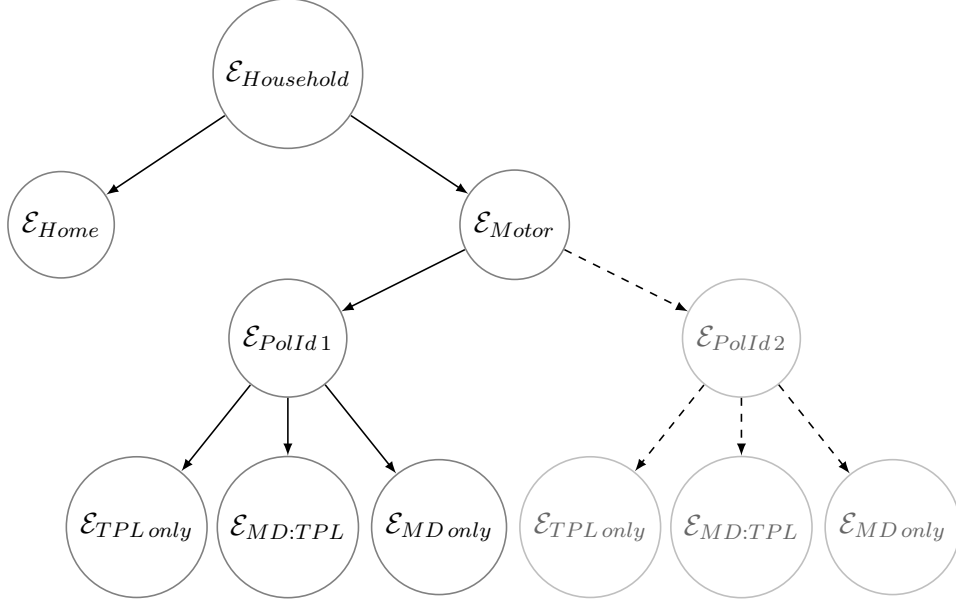
Figure 2: Hierarchical Structure with policyholder effect in Motor insurance.

The random vector $\boldsymbol{\mathcal{E}_h} = (\mathcal{E}_h^{Household}, \mathcal{E}_h^{Home}, \mathcal{E}_h^{Motor}, \mathcal{E}_{h(i)}^{PolId}, \mathcal{E}_{h(i)}^{TPL}, \mathcal{E}_{h(i)}^{MD}, \mathcal{E}_{h(i)}^{MD:TPL})$ has mean vector $E[\exp \boldsymbol{\mathcal{E}_h}] = \mathbf{1}$ and variance-covariance matrix

$$V[\boldsymbol{\mathcal{E}_h}] = diag(\varsigma^2_{Household}, \varsigma^2_{Home}, \varsigma^2_{Motor}, \varsigma^2_{PolId}, \varsigma^2_{TPL}, \varsigma^2_{MD}, \varsigma^2_{MD:TPL}).$$

In short, 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, as in Pechon *et al.* (2019). 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

$$\boldsymbol{\Sigma}_{\log \boldsymbol{\Theta_h}} = \begin{pmatrix} \sigma^2_{Home} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} & \sigma_{HM} \\ \sigma_{HM} & \sigma^2_{TPL} & \sigma_{PolId} & \sigma_{PolId} & & & \\ \sigma_{HM} & \sigma_{PolId} & \sigma^2_{MD} & \sigma_{PolId} & & \sigma_{Motor} & \\ \sigma_{HM} & \sigma_{PolId} & \sigma_{PolId} & \sigma^2_{MD:TPL} & & & \\ \sigma_{HM} & & & & \sigma^2_{TPL} & \sigma_{PolId} & \sigma_{PolId} \\ \sigma_{HM} & & \sigma_{Motor} & & \sigma_{PolId} & \sigma^2_{MD} & \sigma_{PolId} \\ \sigma_{HM} & & & & \sigma_{PolId} & \sigma_{PolId} & \sigma^2_{MD:TPL} \end{pmatrix}$$

where

$$\sigma^2_{Home} = \varsigma^2_{Household} + \varsigma^2_{Home}$$
$$\sigma^2_{TPL} = \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} + \varsigma^2_{TPL}$$
$$\sigma^2_{MD} = \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} + \varsigma^2_{MD}$$
$$\sigma^2_{MD:TPL} = \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} + \varsigma^2_{MD:TPL}$$
$$\sigma_{HM} = \varsigma^2_{Household} \qquad \text{(Household effect)}$$
$$\sigma_{Motor} = \varsigma^2_{Household} + \varsigma^2_{Motor} \qquad \text{(Inter Policyholder)}$$
$$\sigma_{PolId} = \varsigma^2_{Household} + \varsigma^2_{Motor} + \varsigma^2_{PolId} \qquad \text{(Intra Policyholder)}.$$

3.3   Estimation

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

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

Note that if for some household the length of the vector $\boldsymbol{\Theta_h}$ is smaller than seven (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 claim frequency λ to zero yields the correct likelihood.

The computations have been realized in the statistical software R. The estimation was done in a two-part approach. First the *a priori* claim frequencies have been estimated with a Poisson regression using GAMs. As the paper focuses on the a posteriori claim frequencies, the more details about the a priori analyses have been put in the appendix. In a second part, the a priori claim frequencies are fixed and the variance-covariance matrix of the random effects is estimated.

In order to compute the likelihood in (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 Gauss-Hermite quadrature avoids relying on simulation as in Monte-Carlo integration techniques or MCMC in a Bayesian framework. 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 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 optimization process and consequently reduces drastically the computation time.

First, we estimate by maximum likelihood 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 optimizing 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 optimize, convergence is fast.

In a 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 (resp. covariances) are positive. As opposed to Motor insurance, for which we could rely on the results found in Pechon *et al.* (2019), we need to first 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 $\Theta^{TPL}, \Theta^{MD}, \Theta^{MD:TPL}$. 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 equal to 0.1202 (std. error 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.

We display all the estimates at each step in Table 3.

	First Step	Second Step	Third Step (Std. Error)
$\widehat{\varsigma}^2_{Household}$	0.0000	0.0000	0.1220 (0.0156)
$\widehat{\varsigma}^2_{Home}$	0.0000	0.0000	0.1920 (0.0200)
$\widehat{\varsigma}^2_{Motor}$	0.0000	0.1896	0.0824 (0.0235)
$\widehat{\varsigma}^2_{PolId}$	0.2152	0.0256	0.0137 (0.0189)
$\widehat{\varsigma}^2_{TPL}$	0.3417	0.3394	0.3350 (0.0425)
$\widehat{\varsigma}^2_{MD}$	0.1453	0.1485	0.1490 (0.0178)
$\widehat{\varsigma}^2_{MD:TPL}$	0.1063	0.1148	0.1150 (0.0406)

Table 3: Estimates at each step of the optimization process.

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. 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
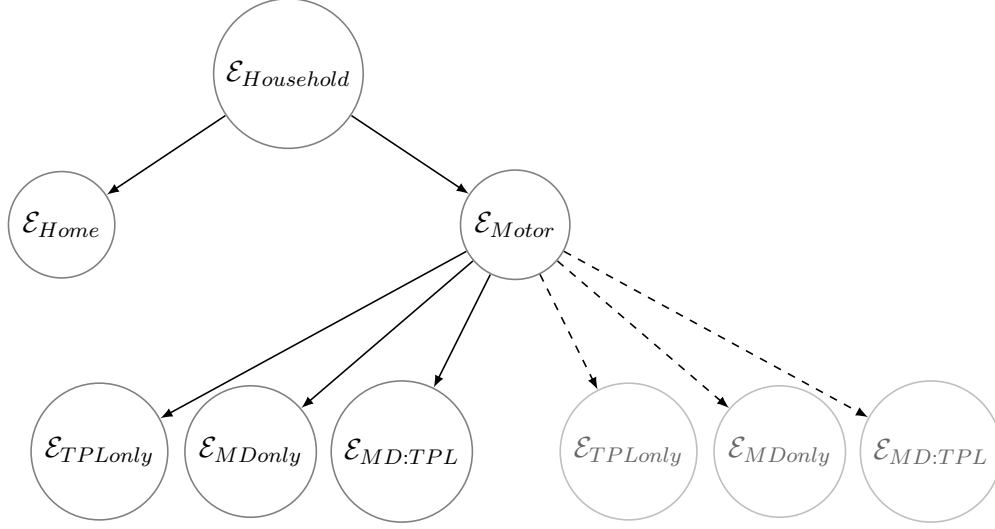
Figure 3: Hierarchical Structure of final model.

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 $\widehat{\varsigma}^2_{Household}$. As noted above, at step 2, one difference with Pechon *et al.* (2019) comes from the dependence structure. If we compare the log-likelihood from both models, we obtain (without some constants involved in the likelihood) $-66\,507.32$ for the model restricted to Motor insurance (step 2) and $-66\,664.46$ for the final model from Pechon *et al.* (2019).

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.

Effect	Estimate	Std.Error
$\widehat{\varsigma}^2_{Household}$	0.1203	0.0173
$\widehat{\varsigma}^2_{Home}$	0.1912	0.0215
$\widehat{\varsigma}^2_{Motor}$	0.0927	0.0208
$\widehat{\varsigma}^2_{PolId}$	0.0000	
$\widehat{\varsigma}^2_{TPL}$	0.3412	0.0173
$\widehat{\varsigma}^2_{MD}$	0.1507	0.0215
$\widehat{\varsigma}^2_{MD:TPL}$	0.1173	0.0208

Table 4: Final Estimates.

This model, without the policyholder effect, amounts to a hierarchical structure as depicted on Figure 3.

Note that given the parametrization using a hierarchical structure, any value within the confidence intervals will yield an appropriate (that is, positive-definite) variance-covariance matrix.

This is one of the advantage of the hierarchical parametrization.

3.4 Implied dependence structure

Now that the model has been estimated by maximum likelihood and that it has even been simplified by removing of the policyholder effect, we can compute the implied dependence structure between these latent factors. We can first 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 will first 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:

$$\mathrm{V}\left[\log \Theta_h^{Home}\right] = \varsigma_{Household}^2 + \varsigma_{Home}^2 \qquad \text{estimated by } 0.3115$$

$$\mathrm{V}\left[\log \Theta_{h(i)}^{TPL}\right] = \varsigma_{Household}^2 + \varsigma_{Motor}^2 + \varsigma_{TPL}^2 \qquad \text{estimated to } 0.5542$$

$$\mathrm{V}\left[\log \Theta_{h(i)}^{MD}\right] = \varsigma_{Household}^2 + \varsigma_{Motor}^2 + \varsigma_{MD}^2 \qquad \text{estimated to } 0.3637$$

$$\mathrm{V}\left[\log \Theta_{h(i)}^{MD:TPL}\right] = \varsigma_{Household}^2 + \varsigma_{Motor}^2 + \varsigma_{MD:TPL}^2 \qquad \text{estimated to } 0.3303.$$

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

$$\mathbb{C}\mathrm{ov}\left[\log \Theta_h^{Home}, \log \Theta_{h(i)}^{TPL}\right] = \mathbb{C}\mathrm{ov}\left[\log \Theta_h^{Home}, \log \Theta_{h(i)}^{MD}\right]$$

$$= \mathbb{C}\mathrm{ov}\left[\log \Theta_h^{Home}, \log \Theta_{h(i)}^{MD:TPL}\right]$$

$$= \varsigma_{Household}^2 \text{ estimated to } 0.1203$$

$$\mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{TPL}, \log \Theta_{h(j)}^{MD}\right] = \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{TPL}, \log \Theta_{h(j)}^{MD:TPL}\right]$$

$$= \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{MD}, \log \Theta_{h(j)}^{MD:TPL}\right]$$

$$= \varsigma_{Household}^2 + \varsigma_{Motor}^2 \text{ estimated to } 0.2130$$

$$\text{and for } i \neq j, \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{TPL}, \log \Theta_{h(j)}^{TPL}\right] = \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{MD}, \log \Theta_{h(j)}^{MD}\right]$$

$$= \mathbb{C}\mathrm{ov}\left[\log \Theta_{h(i)}^{MD:TPL}, \log \Theta_{h(j)}^{MD:TPL}\right]$$

$$= \varsigma_{Household}^2 + \varsigma_{Motor}^2 \text{ estimated to } 0.2130.$$

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 $\boldsymbol{\Theta_h} = (\Theta_h^{Home}, \Theta_{h(1)}^{TPL}, \Theta_{h(1)}^{MD}, \Theta_{h(1)}^{MD:TPL}, \Theta_{h(2)}^{TPL}, \Theta_{h(2)}^{MD}, \Theta_{h(2)}^{MD:TPL})$. We show the estimated variances in Table 6 and the estimated correlations in Table 7.

	$\log \Theta_h^{Home}$	$\log \Theta_{h(1)}^{TPL}$	$\log \Theta_{h(1)}^{MD}$	$\log \Theta_{h(1)}^{MD:TPL}$	$\log \Theta_{h(2)}^{TPL}$	$\log \Theta_{h(2)}^{MD}$	$\log \Theta_{h(2)}^{MD:TPL}$
$\log \Theta_h^{Home}$	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 5: Estimated correlation structure between the random effects on the log scale (i.e. at the score scale).

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

Let us provide some interpretations for the values dispayed in Tables 6 and 7.
The heterogeneity appears to be the smallest in Home insurance. This is very intuitive since the claims related to Motor insurance are strongly linked to the behaviour of the driver, whereas in Home insurance the claims are less the result of choices and decisions made by the policyholder and more related to the architecture of the 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. Nevertheless, one can only take guesses, since these informations are 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

	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

Table 7: Correlation structure between the Lognormal random effects.

for any other period) updates of the claim frequencies (i.e. a posteriori 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 claim frequency corrections. 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 of claim frequencies

As time passes by and the insurer observes the number of claims, we can compute corrections to apply to the a priori 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 are relevant.

Let a household h consist 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 claim frequencies. For instance, at the end of year T, the insurer can then compute the following corrections for the household in Home insurance:

$$
\mathrm{E}\left[\Theta_h^{Home}|\boldsymbol{N_h} = \boldsymbol{n_h}\right]
$$

$$
=\frac{1}{\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right]} \int_{\mathbb{R}_+^7} \theta^{Home} \Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}|\boldsymbol{\Theta} = \boldsymbol{\theta}\right] f_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})d\boldsymbol{\Theta}
$$

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

$$
=\frac{\Pr\left[\boldsymbol{N_h}^{(-Home)} = \boldsymbol{n_h}^{(-Home)}, N_{h,\bullet}^{Home} = n_{h,\bullet}^{Home}+1\right]}{\Pr\left[\boldsymbol{N_h} = \boldsymbol{n_h}\right]}\frac{1+n_{h,\bullet}^{Home}}{\lambda_{h,\bullet}^{Home}}
$$

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 $\boldsymbol{N_h^{(-Home)}}$ (resp. $\boldsymbol{n_h^{(-Home)}}$) is the vector $\boldsymbol{N_h}$ (resp. $\boldsymbol{n_h}$) 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 $\boldsymbol{N_{h(i)}^{(-g)}}$ (resp. $\boldsymbol{n_{h(i)}^{(-g)}}$) is the vector $\boldsymbol{N_h}$ (resp. $\boldsymbol{n_h}$) 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:

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

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

In order to numerically compute these corrections we need to determine the a priori risk profiles of both policyholders in TPL and MD as well as the household's risk profile in Home insurance. For this matter, we bin the predicted a priori 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 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 on 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, 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.

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

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 generalization 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.

On 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 eq. 2).

Similarly, on 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
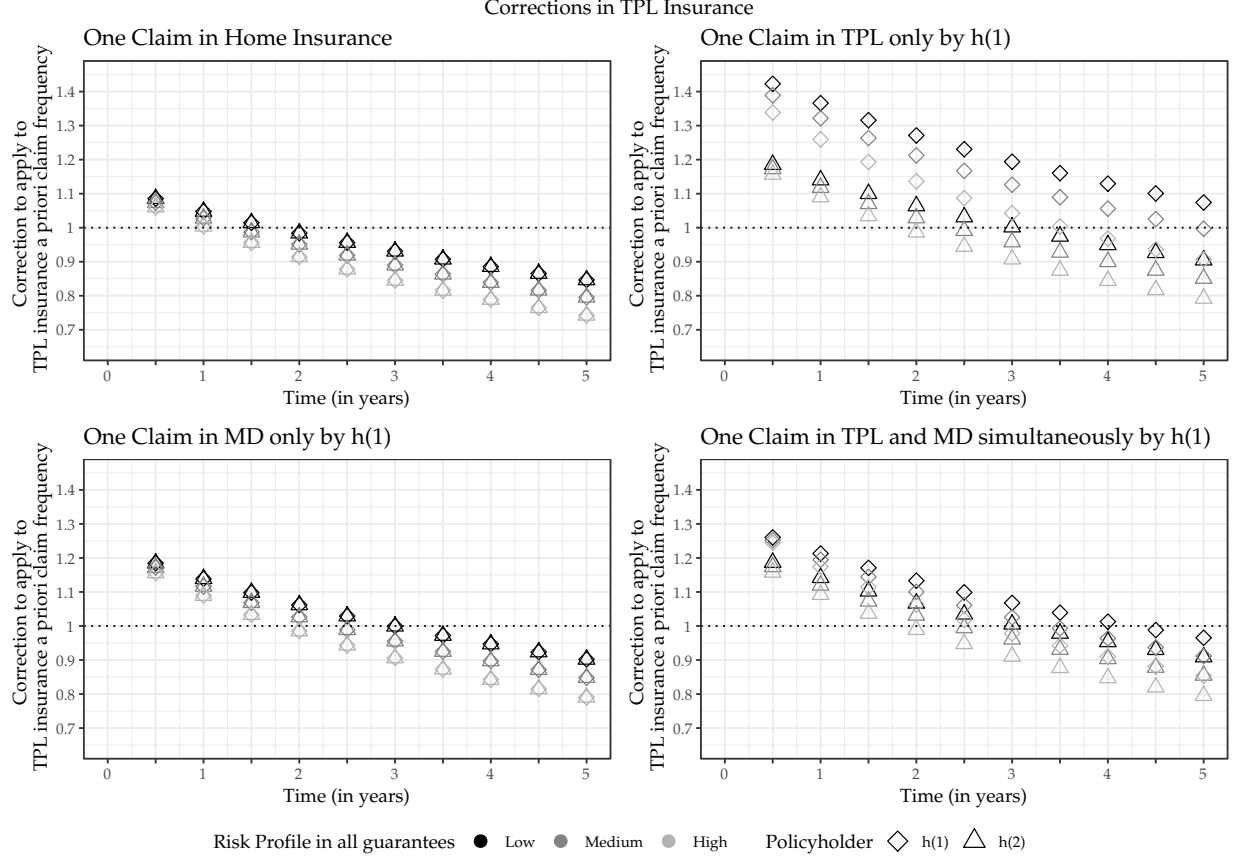
Figure 5: Correction to apply to TPL insurance a priori claim frequencies when one claim was reported.

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 claim frequency than on the TPL a posteriori 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.

Finally, let us analyse the correction factors when a claim was reported in Home insurance (Figure 5 and Figure 6) and compare them to those when no claim was reported (Figure 4). Both in TPL and MD, we observe that after one year, there is a difference of around 10%. As time passes the difference shrinks to around 5% after five years. Consequently, we see that taking into account the reported claims in Home Insurance to compute a posteriori 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 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
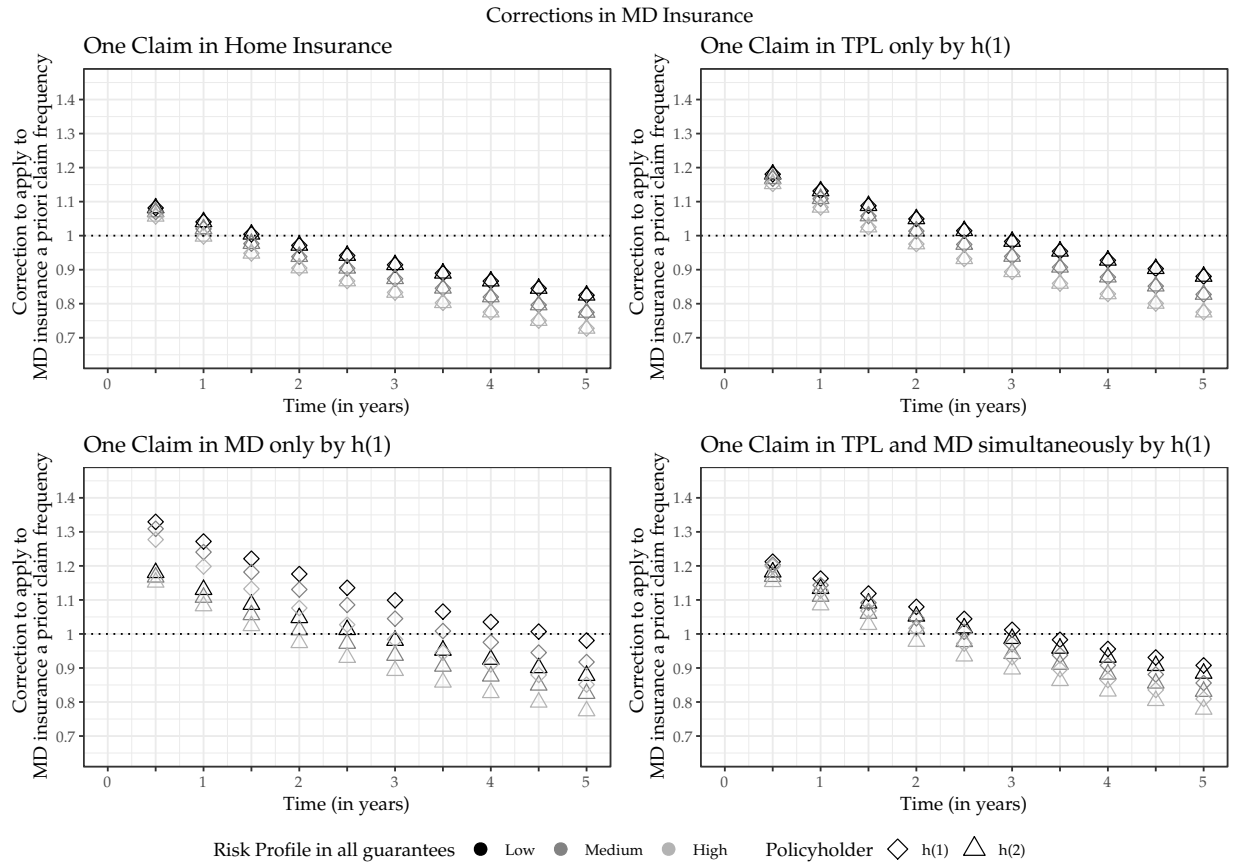
Figure 6: Correction to apply to MD insurance a priori claim frequencies when one claim was reported.
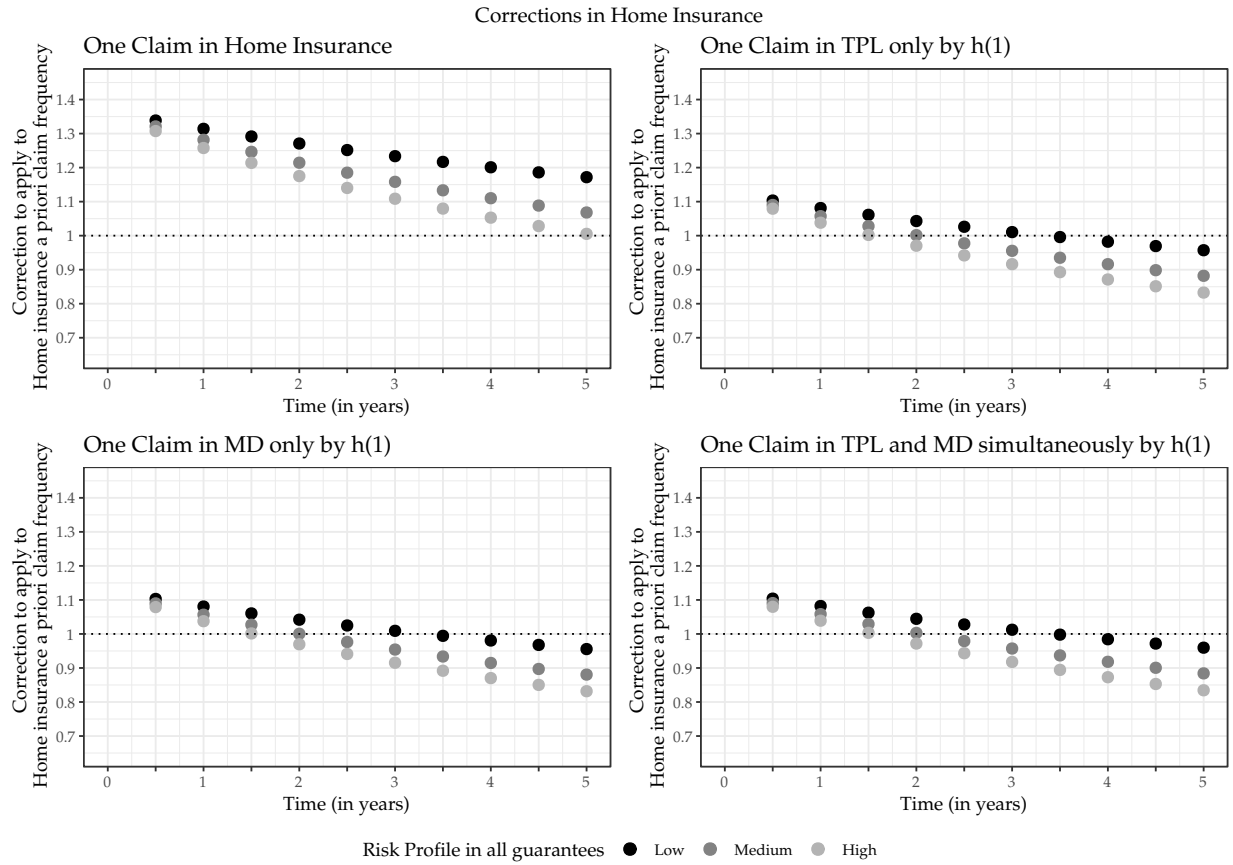
Figure 7: Correction to apply to Home insurance a priori claim frequency when one claim was reported

correction to apply to Home insurance are depicted on 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 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% to 15%). Only a claim in Home insurance yields a correction factor larger than one after five 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.

4.2 Repercussions 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 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 by Table 7;

2. The same variance-covariance matrix as above, except for the off-diagonal terms of the first row (resp. column) are 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 by 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, 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 on 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
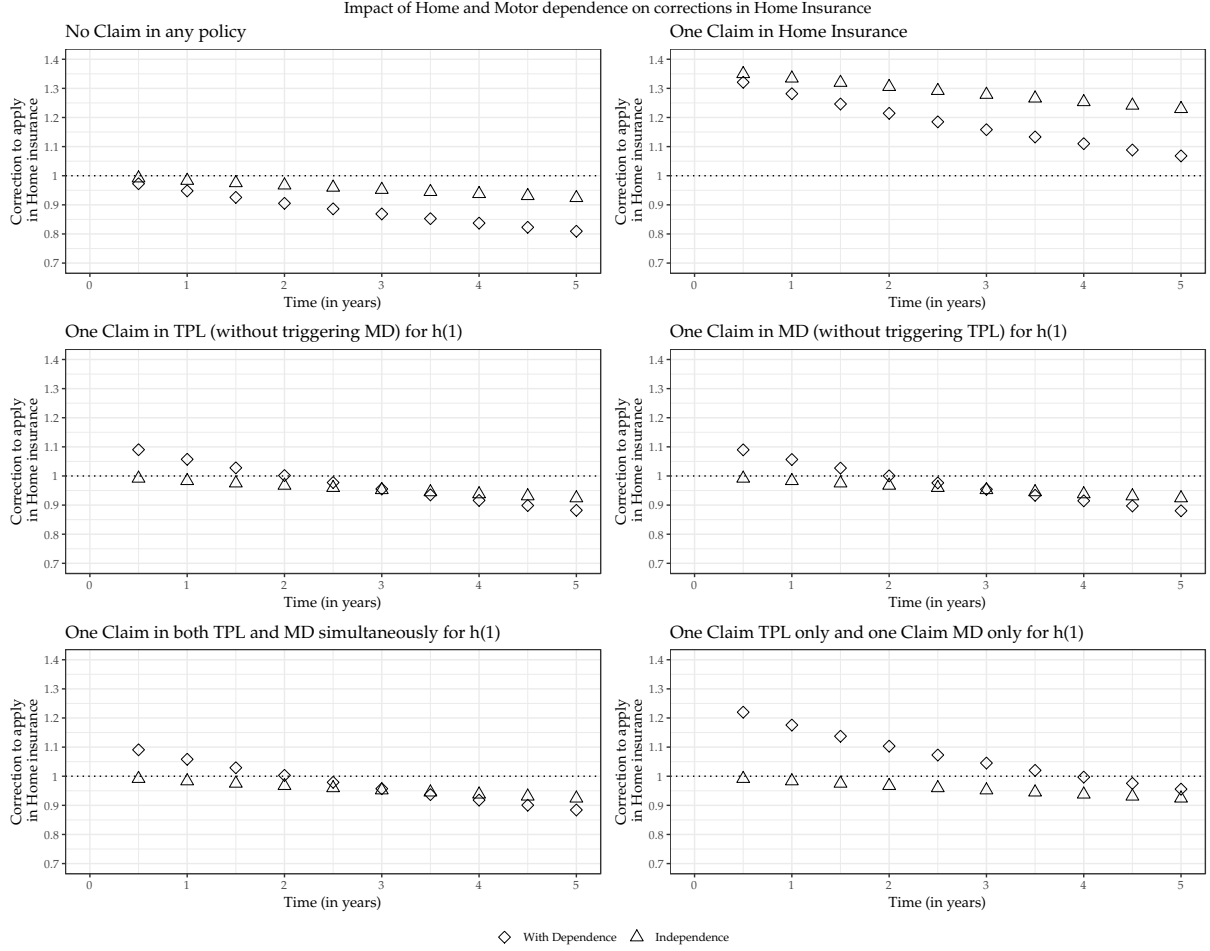
20

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.

falls below one after about 2 years. Note that the difference with the first example (i.e. no claims in any guarantee) is about 5% after five years when considering the dependence. So, correcting the Home insurance a priori 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% to 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 5 of the 6 examples (i.e. when no claim was reported in Home insurance).
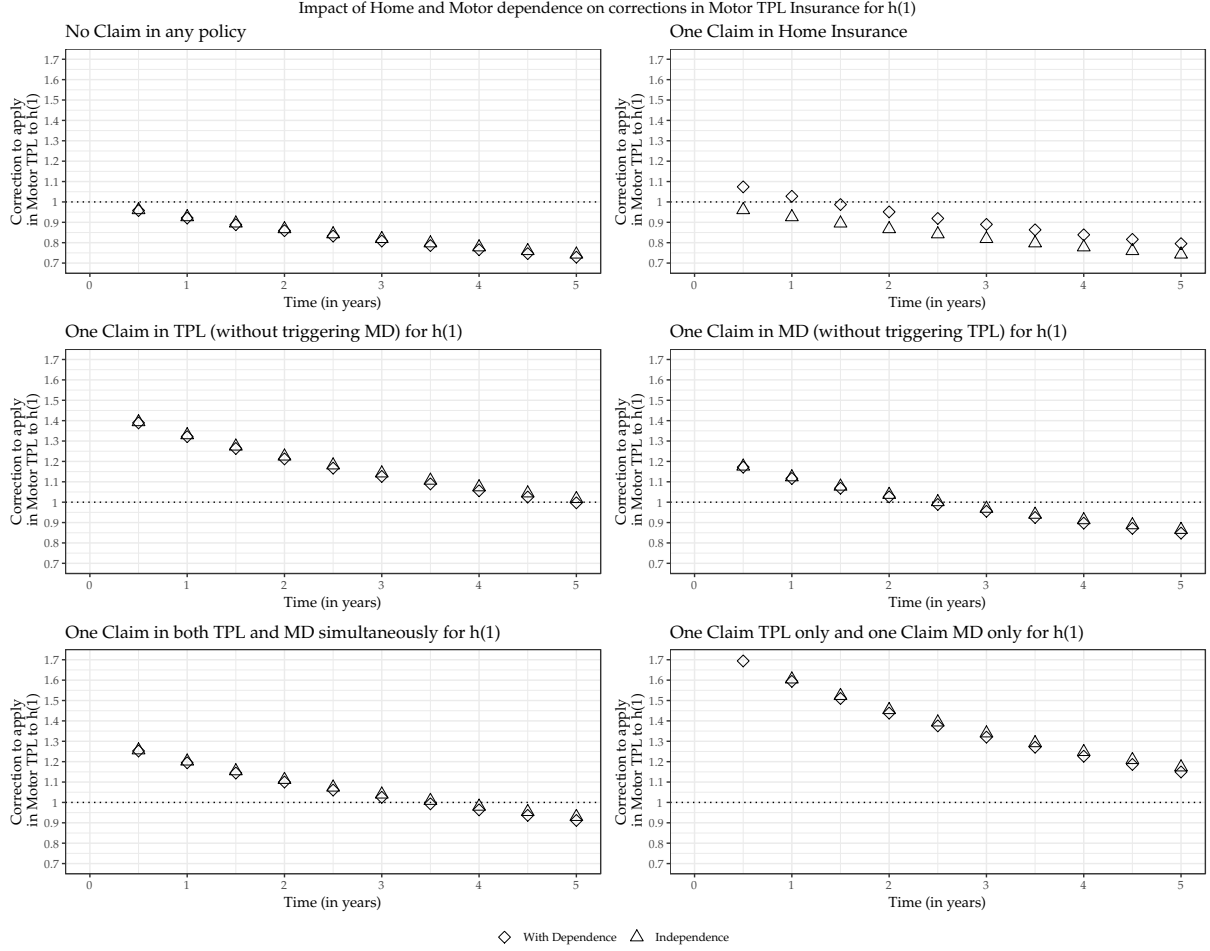
Figure 9: Impact of the dependence of the random effects related to Home and Motor insurance measured by the corrections in Motor TPL for $h(1)$. Six cases are considered, whether a claim occurred or not, and if so, which guarantees were triggered by the claim.

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 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 on Figure 9 for TPL and on 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 (resp. Figure 10) which displays the corrections to apply to $h(1)$ in TPL (resp. 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 on Figure 9 and Figure 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 claim frequencies for two
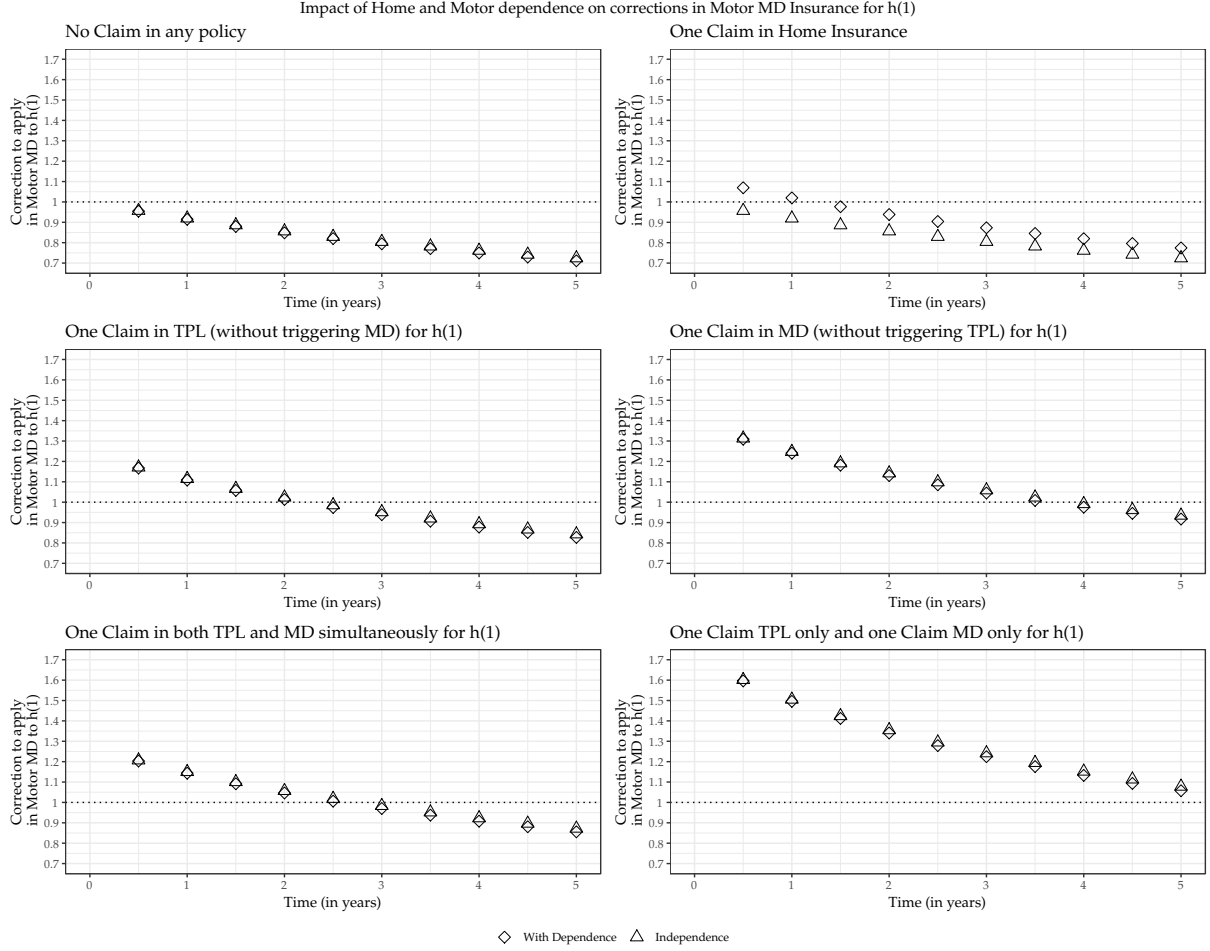
Figure 10: Impact of the dependence of the random effects related to Home and Motor insurance measured by the corrections in Motor MD for $h(1)$. Six cases are considered, whether a claim occurred or not, and if so, which guarantees were triggered by the claim.

policies in Motor insurance (TPL and MD), the a priori 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 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, Figure 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.

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 Conclusion

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 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 parametrization, the hierarchical structure brings some additional interpretability.

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.

Acknowledgements

References

Agresti, Alan. 2003. *Categorical Data Analysis*. Vol. 482. John Wiley & Sons.

Antonio, Katrien, & Zhang, Yanwei. 2014. Nonlinear Mixed Models. *Page 398–424 of:* Frees, Edward W., Derrig, Richard A., & Meyers, Glenn (eds), *Predictive Modeling Applications in Actuarial Science*. International Series on Actuarial Science, vol. 1. Cambridge University Press.

Antonio, Katrien, Guillén, Montserrat, Pérez Martín, AM, *et al.* 2010a. *Multidimensional credibility: a Bayesian analysis of policyholders holding multiple policies*. Tech. rept. Amsterdam School of Economics Research Institute.

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

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

Bermúdez, Lluís, Guillén, Montserrat, & Karlis, Dimitris. 2018. Allowing for time and cross dependence assumptions between claim counts in ratemaking models. *Insurance: Mathematics and Economics*, **83**, 161 – 169.

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

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

Boucher, Jean-Philippe, & Inoussa, Rofick. 2014. A posteriori ratemaking with panel data. *ASTIN Bulletin*, **44**(3), 587–612.

Brockett, Patrick L., Golden, Linda L., Guillen, Montserrat, Nielsen, Jens Perch, Parner, Jan, & Perez-Marin, Ana Maria. 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, Michel, Maréchal, Xavier, Pitrebois, Sandra, & Walhin, Jean-François. 2007. *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons.

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

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

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

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

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

Frees, Edward W, Bolancé, Catalina, Guillen, Montserrat, & Valdez, Emiliano. 2018. Copula modeling of multivariate longitudinal data with dropout. *arXiv preprint arXiv:1810.04567*.

Guillen, Montserrat, Nielsen, Jens Perch, & Pérez-Marín, Ana M. 2008. The ned 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.

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

Kroeze, Karel. 2016. *MultiGHQuad: Multidimensional Gauss-Hermite quadrature*. R package version 1.2.0.

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

Pechon, Florian, Denuit, Michel, & Trufin, Julien. 2019. Multivariate modelling of multiple guarantees in motor insurance of a household. *European Actuarial Journal*, In press.

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

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

Self, Steven G, & Liang, Kung-Yee. 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, Peng, & Yang, Lu. 2018. Pair copula constructions for insurance experience rating. *Journal of the American Statistical Association*, **113**(521), 122–133.

Shi, Peng, Feng, Xiaoping, & Boucher, Jean-Philippe. 2016. Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics*, **10**(2), 834–863.

Tuerlinckx, Francis, Rijmen, Frank, Verbeke, Geert, & De Boeck, Paul. 2006. Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, **59**(2), 225–255.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. 2 edn. Chapman and Hall/CRC.

Appendix

A   A priori model

We aim here to give more details on the a priori model that is used to estimate the a priori claim frequencies in both Motor and Home insurance.

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. After careful variable selection, the final models were as follows:

```
Nclaim_Home ~ offset(log(Expo)) + policy_holder_role_C * risk_classification_cd_C +
contiguity_cd_C+ s(covered_value_building) + s(Longitude, Latitude)

Nclaim_MD ~ offset(log(Expo)) + NewCar + usage_profes + litigation +
gender + s(age, by = gender) + s(covered_value)+s(Longitude, Latitude)

Nclaim_TPL ~ offset(log(Expo)) +NewCar + usage_profes + litigation +
gender + s(age, by = gender) + s(covered_value)+s(Longitude, Latitude)

Nclaim_MDTPL~ offset(log(Expo)) +NewCar+ usage_profes + litigation24+
gender +powerCAT +s(age, by=gender) + s(Longitude, Latitude)
```

Note that the models used in Motor insurance are the same as in Pechon *et al.* (2019).

Due to confidentiality reasons, we cannot display the estimated claim frequencies. We can however show the relative impact of some selected variables.
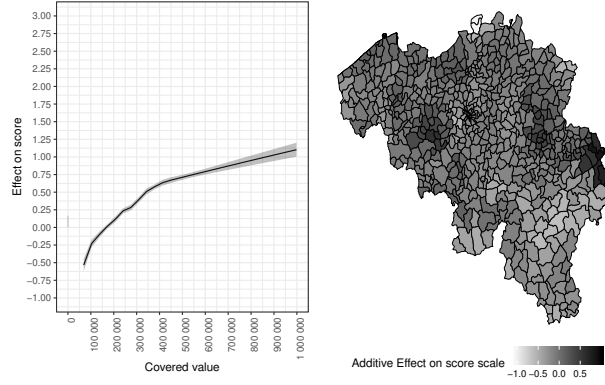
Figure A.1: Left: Effect (on the scale of the score) of covered value on the claim frequency in Home Insurance. Right: Geographic effect (additive on the scale of the score) on the claim frequency in Home Insurance.

For Home insurance, the reference levels are owners covering a house. Tenants appear to have 82% less claims than owners. Apartments appear to be riskier than houses (6%). However, an interaction was found between the variable tenant/owner and the type of home (apartment or house), decreasing the claim frequency for tenants in apartment by an extra 17%. The contiguity of the building was found to be important. As such, contiguous buildings have 20% more claims. Two continuous covariates were included in the a priori model: the covered value of the building as well as the location (represented by latitude and longitude coordinates). The estimated smooth functions of these covariates are depicted on Figure A.1.

For Motor insurance, we show on Figure A.2 the impact of the age of the policyholder, which was estimated for both genders. Also, on Figure A.3, the geographic effect is displayed. Moreover, in what relates the categorical variables, in TPL, policyholders with new cars have about 13.6% less claims and professional usage increases by 37% the claim frequency. Litigation also increased the claim frequency.

In MD, new cars, as opposed to TPL, appeared to have a multiplicative effect of 1.59, while the professional usage increased the claim frequency by 8.5%. Again, the litigation variable showed an increase of the claim frequency when the policyholder once failed to pay its premium in due time.

In MD:TPL, policyholders with new cars also appear to have more claims (multiplicative effect of 1.15), as well as professional usage (1.15) and the litigation variable. Finally, cars with higher power have also an increase of the claim frequency (1.23).
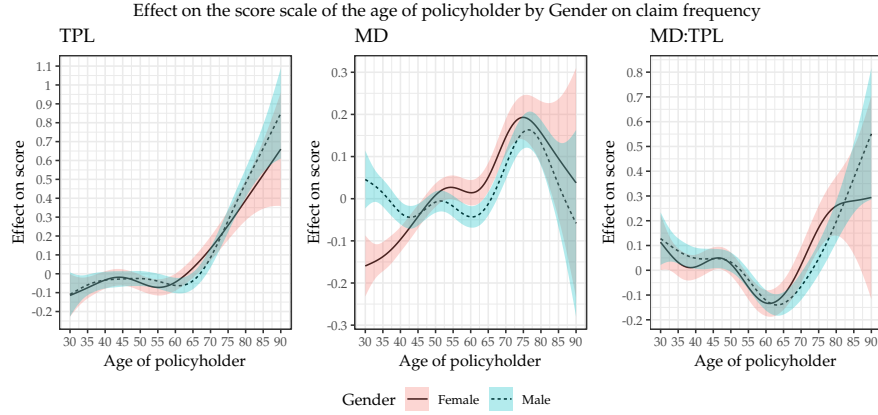
27

Figure A.2: Estimation of the effect of the age on the claim frequency. The effect is shown by gender and by type of claim on the score scale.
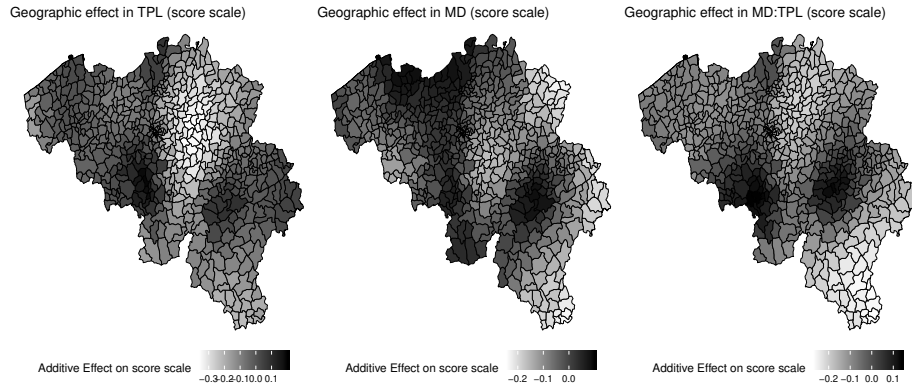

Figure A.3: Geographic effect (additive on the scale of the score) in Motor Insurance for each of the three count variables.