MULTIVARIATE MODELLING OF MULTIPLE GUARANTEES IN MOTOR INSURANCE OF A HOUSEHOLD

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

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

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

Abstract

Actuarial risk classification is usually performed at a guarantee and policyholder level: For each policyholder, the claim frequencies corresponding to each guarantee are modelled in isolation, without accounting for the correlation between the different guarantees and the different policyholders from the same household. However, sometimes, a common event will trigger both guarantees at the same time. Moreover, the claim frequencies for policyholders from the same household appear to be correlated. This paper aims to supplement the standard actuarial approach by combining two guarantees and the policyholders from the household, which allows to refine the prediction on the claim frequencies and account for the common shocks on multiple guarantees. Some possible cross-selling opportunities can also be identified.

1 Introduction and motivation

In motor insurance, ratemaking is usually achieved in two steps. First, the actuaries proceed to the a priori risk classification: Using the characteristics known by the insurer at the time of the subscription of the policy, the actuary partitions the policyholders in order to obtain homogeneous risk classes. Policyholders belonging to the same risk class are then given the same claim frequency estimate. This is routinely done with the help of Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs). These models can be extended with the introduction of random effects. Indeed, some important risk factors are not observed nor known by the insurer at the inception of the contract. This has as consequence that the risk classes obtained by the a priori ratemaking remain heterogeneous. The a posteriori ratemaking aims at adjusting the a priori estimates by using the past claims information, which reveal through time these unobserved, but influential, risk factors. See Denuit *et al.* (2007) for a comprehensive account of these techniques in nonlife insurance.

Multiple guarantees can be underwritten by the policyholders in motor insurance. The most common guarantees in the EU are the Third-Party Liability (TPL) insurance and Material Damage (MD) insurance. The former is compulsory and covers a third-party's loss caused by the insured car. The latter is an optional guarantee that covers the cost of repairing or replacing the insured's own vehicle. The policyholder will typically trigger this guarantee when he is liable for the claim, or couldn't identify the liable person. Generally, the guarantees are sold in packages, and the MD guarantee is part of the 'comprehensive' package.

A priori ratemaking is usually done at a guarantee level, meaning that each guarantee is considered in isolation and independently. However, this is questionnable, as there may be claims that trigger multiple guarantees at the same time. For instance Bermúdez (2009) used a bivariate Poisson to estimate the claim frequencies in TPL and a pool of the other guarantees. The bivariate Poisson was used to estimate the common shocks without having to specify which claims triggered both count variables. Bermúdez & Karlis (2012) extended the bivariate model in Bermúdez (2009), by incorporating a finite mixture in the Poisson regression model. The inclusion of 2-finite random effects allows to capture the overdispersion and the excess of zeros. Also, Bermúdez & Karlis (2011) used a multivariate Poisson regression, as well as zero-inflated Poisson regression, in a Bayesian framework to estimate claim frequencies for three different guarantees. Shi & Valdez (2014) introduced a multivariate Negative Binomial regression model, where each pair of variables has its own covariance structure. This model is then compared to a multivariate model using copulas, joining marginal Negative Binomials. Frees *et al.* (2016) modelled marginally the frequency and severity of several guarantees and then first using copula modelled the dependence between frequency and severity. In a second step the dependence over multiple guarantees was also modelled using a copula. Frees *et al.* (2013) followed a similar approach in a health insurance context by using multivariate two-part regression models to model five different types of expenditures. The frequency is modelled using a binary outcome, indicating if some amount has been paid. The amounts are coupled using a Gaussian copula. These models do not use explicitly the information regarding common shocks, i.e. claims that trigger multiple guarantees at the same time. Nevertheless, since insurers know for each claim which guarantees were triggered, it may be of interest to directly estimate the common shocks frequencies from the data.

A posteriori ratemaking is also generally done at a guarantee level using mixed models. Each policyholder and each guarantee is considered independently. The independence between policyholders from the same household is nevertheless deniable, see for instance Pechon *et al.* (2018) and Shi *et al.* (2016). Moreover, the independence between the random effects related to different guarantees for a given policyholder can also be relaxed. Pinquet (1998) proposed a Bonus-Malus System for different types of claims, by using data related to claims at fault and not at fault with respect to a third party. Englund *et al.* (2008) introduced a multivariate latent risk approach by correlating random effects over multiple guarantees related to the same policyholder. Englund *et al.* (2009) used the model on a dataset related to a Danish insurer to illustrate the improvement in prediction and Thuring (2012) showed that this multivariate credibility model can also be used to find cross-selling opportunities. On aggregated loss, Frees & Wang (2006) used elliptical copulas to model the dependencies between the severities, while the dependence with respect to the count variables (frequencies) was introduced by means of latent correlated factors.

Our approach can be summarized as follows. First, we estimate using the observable characteristics the a priori claim frequencies for three count variables: the two variables counting the claim that trigger only one guarantee, and the variable counting the common shocks. We will assume that each of these count variables follows a Poisson distribution. In a second step, a posteriori ratemaking is done, by the introduction of random effects, leading to a Poisson-mixture. Relaxing the independence between the random effects for a given policyholder as well as for policyholders

living in the same household leads to a multivariate Poisson-mixture. The introduction of the dependence between random effects for a given policyholder is motivated by the fact that the unobserved risk factors influencing the claim frequencies may be the same for multiple guarantees (e.g. regularly driving under dangerous conditions), while the household dependence may come from shared unobserved risk characteristics. So, the dependence between the claim frequencies in TPL and MD is two-sided. Part of the dependence comes from the common shocks and part of the dependence comes from the correlated unobserved risk factors.

The remainder of this paper is organized as follows. In Section 2, we introduce the dataset that will be used to illustrate our methodology. In Section 3, present our model and estimate its parameters on our dataset. In Section 4, we assess the dependence between the claim frequencies in TPL and MD and between policyholders from the same household. In Section 5, we illustrate two applications of our model such as premium corrections and detection of cross-selling opportunities.

2 Dataset

Let us briefly describe the dataset that will be used to support our analysis. The data relate to a portfolio of European policies of both motor TPL and MD insurance observed during the years 2011 to 2013. Only policyholders who have subscribed both guarantees are considered here. Also, in the following, we only consider policyholders aged between 30 and 90 years. Indeed, for the other ages, the number of policyholders that have subscribed both TPL and MD is too small to conduct our present analysis.

For each policyholder we have at our disposal the number of claims triggering each guarantee, as well as the number of claims that trigger both guarantees at the same time. In addition, some characteristics (e.g. age, gender, place of residence) for each policyholder are available. We also know the power, the age, the initial value and the use of the insured car (recall that each policy covers a single vehicle and is associated with a main driver). Furthermore, we also have some information about the contract and whether the premium payment has been split (premiums can be paid annually, semi-annually, quarterly or monthly). The database also contains a litigation variable indicating whether the policyholder has had a failure to pay its premium in due time. Finally, a household identifier allows to determine the policyholders belonging to the same household. On Table 3.1, we show that relative occurrence of types of claims, conditional to the occurrence of a claim.

In the considered dataset, 219 038 households have only one policyholder while 10 982 households contain two policyholders (i.e. 21 964 policyholders belonged to such households).

3 Choice of the model

We consider the two most common guarantees involved in motor insurance, namely TPL insurance and MD insurance. Given the nature of the guarantees, it is possible that a single claim triggers both guarantees at the same time. In fact, a claim will result in three possible cases, depending on which guarantees are triggered. Table 3.1 shows which guarantees are triggered in our dataset, conditional to the occurrence of a claim. A significant proportion of claims (about 20%) trigger both guarantees which hints the need for a model that takes into account common shocks on both guarantees. So, as for each policyholder we have at disposal the number of claims triggering both guarantees, we won't use a bivariate Poisson regression with common shocks as in Bermúdez (2009) for modelling the number of claims that trigger TPL and the number of claims that trigger MD.

Type of claims	Frequency
Only MD	62.0%
Both TPL and MD	19.8%
Only TPL	18.2%

Table 3.1: Relative occurrence according to the type of claims

Let us define the following notations. Let $\mathcal{H}$ be the set of households in the dataset. In a household $h \in \mathcal{H}$ with only one policyholder, we denote by $h(1)$ the corresponding policyholder while in a household $h \in \mathcal{H}$ with two policyholders, we denote by $h(1)$ and $h(2)$ the corresponding policyholders. In this last case, for the ease of the presentation, since in most cases (86%) the age difference between policyholders from the same household is below 15 years, we will designate both policyholders as being spouses. Let us note that since we restrict our analysis to policyholders aged between 30 and 90 years, we cannot be in a "parent-young driver" situation as defined in Pechon *et al.* (2018). So, the "spouse" designation in this paper has to be understood in a broader sense than in Pechon *et al.* (2018).

Furthermore, let us introduce the following claim count variables :

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

This implies that the number of claims for policyholder i from household h that trigger TPL (resp. MD) during year t is $N^{TPL}_{h(i),t} + N^{MD:TPL}_{h(i),t}$ (resp. $N^{MD}_{h(i),t} + N^{MD:TPL}_{h(i),t}$). We denote the corresponding a priori claim frequencies as $\lambda^{TPL}_{h(i),t} = \mathrm{E}\left[N^{TPL}_{h(i),t}\right]$, $\lambda^{MD}_{h(i),t} = \mathrm{E}\left[N^{MD}_{h(i),t}\right]$ and $\lambda^{MD:TPL}_{h(i),t} = \mathrm{E}\left[N^{MD:TPL}_{h(i),t}\right]$.

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

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

The corresponding a priori claim frequencies are denoted $\lambda^{TPL}_{h(i),\bullet} = \mathrm{E}\left[N^{TPL}_{h(i),\bullet}\right]$, $\lambda^{MD}_{h(i),\bullet} = \mathrm{E}\left[N^{MD}_{h(i),\bullet}\right]$ and $\lambda^{MD:TPL}_{h(i),\bullet} = \mathrm{E}\left[N^{MD:TPL}_{h(i),\bullet}\right]$.

The claim frequency analysis that follows can be split into two main parts. First, we perform a marginal analysis using a Poisson GAM regression to account for individual risk profiles. For each count variable and for each policyholder inside the household, we predict the expected number of claims based on the information about the policyholder, his/her car and characteristics about the policy. In a second part, information related to the other guarantee as well as the number of claims filed by the other policyholder from the same household (if any) is included. This dependence between the expected number of claims is introduced by means of a multivariate Poisson-mixture.

In Pechon *et al.* (2018), the authors used a multivariate Poisson-mixture to model the dependencies between claim frequencies in a household for motor TPL insurance. So, in the following, we extend this model to incorporate another guarantee (namely MD) in which common shocks may occur.

3.1 Model

Let $\mathcal{G} := \{TPL, MD, MD:TPL\}$ be the set of types of claims depending on which guarantees are triggered. We introduce a multivariate Poisson mixture model that takes into account the dependence that may exist between the frequencies related to these three count variables for each policyholder as well as the dependence that may exist between the policyholders coming from the same household h. Let us assume the following:

1. $\forall h \in \mathcal{H}$, $\forall i \in h$, $\forall g \in \mathcal{G}$, given $\Theta^g_{h(i)} = \theta$, the random variables $N^g_{h(i),1}, N^g_{h(i),2}, \ldots, N^g_{h(i),T}$ are independent.

2. $\forall h \in \mathcal{H}$, $\forall i, j \in h$, $\forall g_i, g_j \in \mathcal{G}$, 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.

3. The $\boldsymbol{\Theta_h} = (\Theta^{TPL}_{h(1)}, \Theta^{MD}_{h(1)}, \Theta^{MD:TPL}_{h(1)}, \Theta^{TPL}_{h(2)}, \Theta^{MD}_{h(2)}, \Theta^{MD:TPL}_{h(2)})$ (or $\boldsymbol{\Theta_h} = (\Theta^{TPL}_{h(1)}, \Theta^{MD}_{h(1)}, \Theta^{MD:TPL}_{h(1)})$ if there is only one policyholder in the household) are independent and identically distributed following a LogNormal distribution with joint probability density function $f_{\boldsymbol{\Theta_h}}$, $\mathrm{E}[\boldsymbol{\Theta_h}] = \mathbf{1}$ and $\mathrm{V}\left[\log \Theta^g_{h(i)}\right] = \sigma^2_g$ $\forall g \in \mathcal{G}$ and $\forall i \in h$. Furthermore, $\forall h \in \mathcal{H}$, $\forall i \in h$, the correlation matrix of $(\log \Theta^{TPL}_{h(i)}, \log \Theta^{MD}_{h(i)}, \log \Theta^{MD:TPL}_{h(i)})$ is given by

$$\boldsymbol{A} = \begin{pmatrix} 1 & \rho^{TPL,MD} & \rho^{TPL,MD:TPL} \\ \rho^{TPL,MD} & 1 & \rho^{MD,MD:TPL} \\ \rho^{TPL,MD:TPL} & \rho^{MD,MD:TPL} & 1 \end{pmatrix}$$

while $\forall g_i, g_j \in \mathcal{G}$, the correlations $\mathbb{C}\mathrm{orr}\left(\log \Theta^{g_i}_{h(1)}, \log \Theta^{g_j}_{h(2)}\right) := \boldsymbol{B}_{ij}$ for $i, j = 1, 2, 3$ are given by the following matrix

$$\boldsymbol{B} = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{12} & \rho_{22} & \rho_{23} \\ \rho_{13} & \rho_{23} & \rho_{33} \end{pmatrix}$$

So, block $\boldsymbol{A}$ stands for the correlations between the log of the random effects for a policyholder, whereas block $\boldsymbol{B}$ contains the correlations between the log of the random effects of the policyholders from the same household. The variance-covariance matrix of $\log \boldsymbol{\Theta_h}$ is then obtained by

$$\boldsymbol{\Sigma}_{\log \Theta} = diag(\sigma_{TPL}, \sigma_{MD}, \sigma_{MD:TPL}) \times \mathbf{P} \times diag(\sigma_{TPL}, \sigma_{MD}, \sigma_{MD:TPL})$$

where $\mathbf{P}$ is the correlation matrix, given by $\boldsymbol{A}$ for households with only one policyholder, while for households with two policyholders it is given by

$$\mathbf{P} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B} & \boldsymbol{A} \end{pmatrix}.$$

This parametrization implies that the model incorporates 15 parameters, from which 3 correspond to variances, 3 to correlations between random effects for the same policyholder and the remaining 9 to the correlations between random effects for different policyholders from the same household.

Note that we included a parametric hypothesis, namely that the random effects have a LogNormal distribution. This parametric assumption was selected for the same dataset in Pechon *et al.* (2018) after a comparison with Gamma distributions joined by a Gaussian copula using a Vuong test. Since the support of each random effect is $\mathbb{R}^+$, we can write the likelihood as

$$L(\boldsymbol{\Sigma}_{\log \boldsymbol{\Theta}}) = \prod_{h\in\mathcal{H}} \underbrace{\int_0^\infty \cdots \int_0^\infty}_{3\times \#h} \left[\prod_{i\in h} \prod_{g\in\mathcal{G}} \exp\left(-\lambda^g_{h(i),\bullet}\theta^g_{h(i)}\right) \frac{\left(\lambda^g_{h(i),\bullet}\theta^g_{h(i)}\right)^{n^g_{h(i),\bullet}}}{n^g_{h(i),\bullet}!} \right] f_{\boldsymbol{\Theta}_{\boldsymbol{h}}}(\boldsymbol{\theta_h}) d\boldsymbol{\theta_h}$$

where $\#h$ denoted the number of policyholders in the household (1 or 2),
$\boldsymbol{\theta_h} = (\theta^{TPL}_{h(1)}, \theta^{MD}_{h(1)}, \theta^{MD:TPL}_{h(1)})$ for households with only 1 policyholder (#h = 1), or
$\boldsymbol{\theta_h} = (\theta^{TPL}_{h(1)}, \theta^{MD}_{h(1)}, \theta^{MD:TPL}_{h(1)}, \theta^{TPL}_{h(2)}, \theta^{MD}_{h(2)}, \theta^{MD:TPL}_{h(2)})$ for households with two policyholders (#h = 2).

3.2 Estimation of the parameters

Let us now estimate the 15 parameters related to the variance-covariance matrix of the log $\boldsymbol{\Theta}$. We consider the following steps to estimate the parameters by maximum likelihood :

1. Consider each policyholder marginally, without the intra-household correlations (block $\boldsymbol{B}$ is set to 0). This implies that we only have to estimate the 3 parameters from block $\boldsymbol{A}$ and the three variances. We maximise the log-likelihood using the Nelder–Mead algorithm with as initial values the moment estimates (e.g given in Denuit et al., 2007, Section 6.2.7).

2. Fix the six previous estimates to the values found in step 1 and estimate the remaining 9 parameters related to the household dependence (block $\boldsymbol{B}$).

3. All 15 parameters are re-estimated simultaneously using the previous estimates as initial values.

Note that in order to compute the log-likelihood, we need to approximate numerically the integrals. Due to the choice of the density function $f_{\boldsymbol{\Theta}}$, the Gauss-Hermite quadrature can be used, as in Pechon *et al.* (2018). Confidence intervals related to the parameters of block $\boldsymbol{B}$ are all overlapping. As a consequence, we impose a unique correlation parameter ρ in the block $\boldsymbol{B}$ matrix and re-estimate the simplified model that contains only 7 parameters using the three steps approach. A likelihood ratio test is conducted to assess whether the simplified model is significantly different from the full model. The test statistic is 4.0382, implying a p-value of 0.8536. Therefore, the simplified model is chosen.

The final estimates are given in Table 3.2.
These estimates are related to the underlying Normal distribution, which is on the score scale. The variances and correlations of the LogNormals can be obtained thanks to the following Proposition:

Proposition 3.1. *Let $\boldsymbol{X} = (X_1, \ldots, X_q)$ be a random vector obeying the multivariate Normal distribution with variances $V[X_i] = \sigma_i^2$, correlations $\mathbb{C}orr[X_i, X_j] = \rho_{ij}$ (with $\rho_{ii} = 1$) and mean vector $\boldsymbol{\mu} = (-\frac{\sigma_1^2}{2}, \ldots, -\frac{\sigma_q^2}{2})$. Define $Y_i = \exp X_i$. Then $\boldsymbol{Y}$ obeys the multivariate LogNormal distribution with mean vector $\mathbf{1}$,*

$$V[Y_i] = \exp(\sigma_i^2) - 1 \text{ and } \mathbb{C}orr[Y_i, Y_j] = \frac{\exp(\rho_{ij}\sigma_i\sigma_j) - 1}{\sqrt{(\exp(\sigma_i^2) - 1)(\exp(\sigma_j^2) - 1)}}.$$

	Estimate
$\widehat{\mathrm{V}}(\log \Theta^{TPL})$	0.55843
$\widehat{\mathrm{V}}(\log \Theta^{MD})$	0.36473
$\widehat{\mathrm{V}}(\log \Theta^{MD:TPL})$	0.31750
$\widehat{\mathbb{C}\mathrm{orr}}(\log \Theta^{TPL}, \log \Theta^{MD})$	0.52414
$\widehat{\mathbb{C}\mathrm{orr}}(\log \Theta^{TPL}, \log \Theta^{MD:TPL})$	0.69405
$\widehat{\mathbb{C}\mathrm{orr}}(\log \Theta^{MD}, \log \Theta^{MD:TPL})$	0.51272
$\widehat{\rho}$ (Block $\boldsymbol{B}$)	0.44117

Table 3.2: Maximum likelihood estimates of the variances and correlations of the underlying Normal random variables, i.e. $\log \boldsymbol{\Theta}$, where the log is taken on each component of the vector.

	Estimate	Std.Error	p.value	
$\widehat{\mathrm{V}}(\Theta^{TPL})$	0.747932	0.072114	3.34e-25	***
$\widehat{\mathrm{V}}(\Theta^{MD})$	0.440121	0.020956	6.27e-98	***
$\widehat{\mathrm{V}}(\Theta^{MD:TPL})$	0.373689	0.054846	9.53e-12	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta^{TPL}, \Theta^{MD})$	0.465137	0.047079	5.09e-23	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta^{TPL}, \Theta^{MD:TPL})$	0.642053	0.092781	4.51e-12	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta^{MD}, \Theta^{MD:TPL})$	0.470047	0.062446	5.18e-14	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{TPL}, \Theta_2^{TPL})$	0.373520	0.095339	8.94e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{TPL}, \Theta_2^{MD})$	0.383984	0.095567	5.87e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{TPL}, \Theta_2^{MD:TPL})$	0.386138	0.095693	5.46e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{MD}, \Theta_2^{TPL})$	0.383984	0.095567	5.87e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{MD}, \Theta_2^{MD})$	0.396656	0.096654	4.06e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{MD}, \Theta_2^{MD:TPL})$	0.399421	0.097073	3.88e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{MD:TPL}, \Theta_2^{TPL})$	0.386138	0.095693	5.46e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{MD:TPL}, \Theta_2^{MD})$	0.399421	0.097073	3.88e-05	***
$\widehat{\mathbb{C}\mathrm{orr}}(\Theta_1^{MD:TPL}, \Theta_2^{MD:TPL})$	0.402359	0.097667	3.79e-05	***

Table 3.3: Final estimates of the variance-covariance matrix of $\boldsymbol{\Theta}$ along with standard error estimates and p-values.

Using Proposition 3.1, we can deduce the variance-covariance matrix of $\boldsymbol{\Theta}$. The estimates are given in Table 3.3. The standard errors are computed thanks to the Delta Method.

It appears that the heterogeneity is the largest in the TPL case. The correlation between the random effects TPL and $MD:TPL$ for a given policyholder is the largest one, while the correlation between MD and $MD:TPL$ is weaker. Note that the correlations of the random effects of two policyholders from the same household are similar to the ones given in Pechon *et al.* (2018), where only the TPL guarantee is considered.

In the next section, we will analyse the implied correlations between the guarantees TPL and MD.

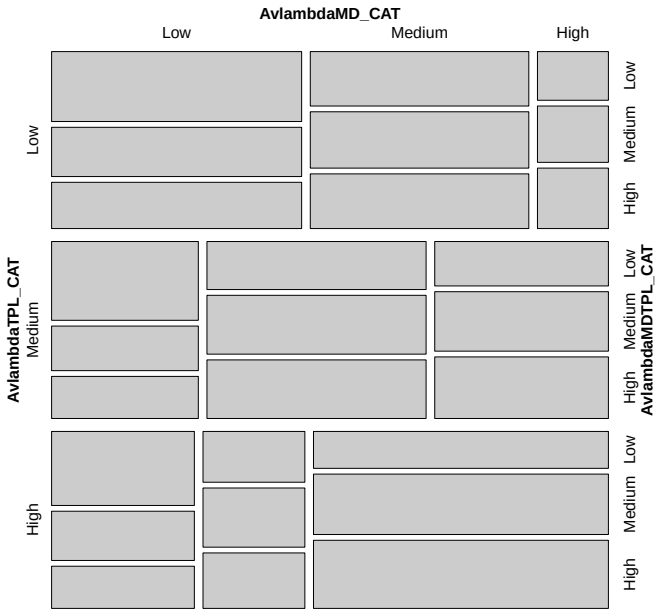

Figure 4.1: Mosaic plot between the categorized a priori claim frequencies $\lambda^{TPL}, \lambda^{MD}, \lambda^{MD:TPL}$.

4 Dependence between claim frequencies in a household

Let us assess the correlation between the claim frequencies in TPL and in MD for a given policyholder. Since these will depend on a priori claim frequencies, let us first discuss the possible dependence between these a priori claim frequencies. To this end, for each of the three a priori claim frequencies, we categorize, based on quantiles, "Low", "Medium", and "High" risk classes. More specifically, we cut the unit interval into three segments of the same length and attribute to each of these segments its median value. So, the "Low" risk profile for instance corresponds to the profiles between quantiles 0 and 1/3, whereas the attributed value corresponds to the quantile 1/6. This categorization is done on each of the three count variables separately.

Then, we can draw a mosaic plot (Figure 4.1), for instance, to visualize the dependence. In case of independence, all the rectangles would be aligned. The dependence appears to be stronger between λ^{TPL} and λ^{MD} than with respect to $\lambda^{MD:TPL}$. In fact, lower (resp. higher) risk profiles in "TPL only" appear associated to lower (resp. higher) risk profiles in "MD only". A similar but weaker association can also be seen with respect to the common shock claim frequencies. This correlation between the a priori claim frequencies partly comes from the fact that the guarantees are all related to the same product (i.e. motor insurance) and consider common risk factors that enter into the a priori model, such as for instance the place of residence, or the age of the policyholder.

In sight of these associations, we can compute the correlations between the claim frequencies in TPL and in MD for a "Medium" risk profile. The medium risk profile is considered as a policyholder with a medium risk profile in each of the three claim frequencies, i.e. median values of $\widehat{\lambda^{TPL}}$, $\widehat{\lambda^{MD}}$ and $\widehat{\lambda^{MD:TPL}}$ (i.e. a medium risk profile in each guarantee).

$$\widehat{\mathbb{C}\mathrm{orr}}\left(\lambda^{TPL}_{h(1)}\Theta^{TPL}_{h(1)} + \lambda^{MD:TPL}_{h(1)}\Theta^{MD:TPL}_{h(1)}, \lambda^{MD}_{h(1)}\Theta^{MD}_{h(1)} + \lambda^{MD:TPL}_{h(1)}\Theta^{MD:TPL}_{h(1)}\right) = 0.6622$$

This correlation can be split into two parts: One part comes from the fact that some claims trig-

ger both guarantees and are accounted by the variable $N^{MD:TPL}$. Another part of this correlation comes from the correlated random effects. Computation shows that about 75% of this correlation comes from the correlated random effects. Similarly, we can compute the correlation between the claim frequencies of two policyholders from the same household. Again, we consider here that both policyholders have a medium risk profile. This gives

$$\begin{aligned}
\widehat{\mathbb{C}\text{orr}}\left(\lambda^{TPL}_{h(1)}\Theta^{TPL}_{h(1)}+\lambda^{MD:TPL}_{h(1)}\Theta^{MD:TPL}_{h(1)},\lambda^{TPL}_{h(2)}\Theta^{TPL}_{h(2)}+\lambda^{MD:TPL}_{h(2)}\Theta^{MD:TPL}_{h(2)}\right) &= 0.4688\\
\widehat{\mathbb{C}\text{orr}}\left(\lambda^{MD}_{h(1)}\Theta^{MD}_{h(1)}+\lambda^{MD:TPL}_{h(1)}\Theta^{MD:TPL}_{h(1)},\lambda^{MD}_{h(2)}\Theta^{MD}_{h(2)}+\lambda^{MD:TPL}_{h(2)}\Theta^{MD:TPL}_{h(2)}\right) &= 0.4875\\
\widehat{\mathbb{C}\text{orr}}\left(\lambda^{TPL}_{h(1)}\Theta^{TPL}_{h(1)}+\lambda^{MD:TPL}_{h(1)}\Theta^{MD:TPL}_{h(1)},\lambda^{MD}_{h(2)}\Theta^{MD}_{h(2)}+\lambda^{MD:TPL}_{h(2)}\Theta^{MD:TPL}_{h(2)}\right) &= 0.4778
\end{aligned}$$

Unlike in the former case, here, only the correlated random effects introduce dependence between the claim frequencies since we assume that no common shock between different policyholders are possible.

Note that these correlations are already adjusted to the known characteristics of the policyholders (which are used to predict the a priori claim frequencies). So these correlations come on top of these possibly correlated a priori characteristics of the policyholders (e.g. place of residence).

5 Insurance applications

The multivariate model can be used for various purposes. We will see that in fact, one can use this model to refine estimated claim frequencies using the household claims history. Indeed, for each policyholder, time will reveal information on the hidden characteristics that are represented by the random effect. Due to the fact that these random effects for policyholders from the same household are correlated, information related to a policyholder will in fact be also relevant for the other policyholder from the same household.

Two main insurance applications will be presented below. First, we will show that using all the household information, we can sharpen the prediction of the claim frequencies. So, for a given household, any claim-free year, or any claim, will affect all the predictions for the claim frequencies of the policyholders of the household under consideration. This also means that households with multiple policyholders can help to improve the predictive power, since there is a higher flow of information.

In a second example, we will show that in a situation where the insurer has only one policyholder in a household, he can use this information to help identifying cross-selling opportunities.

5.1 A posteriori corrections

Let us first assume that a household consists of only one policyholder. We are looking to compute the claim frequencies in TPL (resp. in MD) conditional to the observed number of claims in the past from that policyholder, i.e.

$$\begin{aligned}
&\mathrm{E}\left[\lambda^{TPL}_{h(1),T+1}\Theta^{TPL}_{h(1)}+\lambda^{MD:TPL}_{h(1),T+1}\Theta^{MD:TPL}_{h(1)}|N^{g}_{h(1),\bullet}=n^{g}_{h(1),\bullet}\forall g\in\mathcal{G}\right]\\
&\mathrm{E}\left[\lambda^{MD}_{h(1),T+1}\Theta^{MD}_{h(1)}+\lambda^{MD:TPL}_{h(1),T+1}\Theta^{MD:TPL}_{h(1)}|N^{g}_{h(1),\bullet}=n^{g}_{h(1),\bullet}\forall g\in\mathcal{G}\right].
\end{aligned}$$

The conditional expectations of the random effects can be computed using the Bayes formula and numerical integration

$$\begin{aligned}
&\mathrm{E}\left[\Theta^g_{h(1)}|N^{\tilde{g}}_{h(1),\bullet}=n^{\tilde{g}}_{h(1),\bullet},\forall\tilde{g}\in\mathcal{G}\right]\\
&=\frac{\int_0^\infty\int_0^\infty\int_0^\infty\theta^g_{h(1)}P\left[N^{\tilde{g}}_{h(1),\bullet}=n^{\tilde{g}}_{h(1),\bullet},\forall\tilde{g}\in\mathcal{G}|\Theta^{\tilde{g}}_{h(1)}=\theta^{\tilde{g}}_{h(1)},\forall\tilde{g}\in\mathcal{G}\right]f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_0^\infty\int_0^\infty\int_0^\infty P\left[N^{\tilde{g}}_{h(1),\bullet}=n^{\tilde{g}}_{h(1),\bullet},\forall\tilde{g}\in\mathcal{G}|\Theta^{\tilde{g}}_{h(1)}=\theta^{\tilde{g}}_{h(1)},\forall\tilde{g}\in\mathcal{G}\right]f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}}.
\end{aligned}$$

For the a priori claim frequencies (future and past), we consider three cases. As above, three kinds of risk profiles are constructed, based on quantiles. For instance, for a "Low" risk profile, we will assume that the policyholder's three claim frequencies are the corresponding low claim frequencies. Note that other combinations of risk profiles are possible, but for the ease of the presentation, only these three cases are considered.

We then compute a posteriori expectations for each of these three risk classes and calculate the ratio to the a priori claim frequency. We will call these ratios the correction factors.

As expected, the estimates of the correction factors depicted on Figure 5.1 show that claim-free years will result in a correction below one. This means that in average such policyholders will have lower claim frequencies than the forecast given by their a priori claim frequencies. In contrast, a claim of any kind will increase all the estimates of the claim frequencies. Due to the higher heterogeneity in TPL, the corrections appear to be more severe than in MD. Note that riskier profiles have a less severe correction in case of a claim and stronger decrease in case of claim-free years.

Let us now assume that the household consists of two policyholders. We are interested in the estimates

$$\begin{aligned}
&\mathrm{E}\left[\lambda^{TPL}_{h(i),T+1}\Theta^{TPL}_{h(i)}+\lambda^{MD:TPL}_{h(i),T+1}\Theta^{MD:TPL}_{h(i)}|N^g_{h(j),\bullet}=n^g_{h(j),\bullet}\forall j\in\{1,2\},\forall g\in\mathcal{G}\right]\text{ for }i=1,2,\\
&\mathrm{E}\left[\lambda^{MD}_{h(i),T+1}\Theta^{MD}_{h(i)}+\lambda^{MD:TPL}_{h(i),T+1}\Theta^{MD:TPL}_{h(i)}|N^g_{h(j),\bullet}=n^g_{h(j),\bullet}\forall j\in\{1,2\},\forall g\in\mathcal{G}\right]\text{ for }i=1,2,
\end{aligned}$$

i.e. the conditional expectations of the number of claims triggering TPL (resp. MD). Note that

$$\begin{aligned}
&\mathrm{E}\left[\Theta^g_{h(i)}|N^{\tilde{g}}_{h(j),\bullet}\forall j\in\{1,2\},\forall\tilde{g}\in\mathcal{G}\right]\\
&=\frac{\int_0^\infty\cdots\int_0^\infty\theta^g_{h(i)}P\left[N^{\tilde{g}}_{h(j),\bullet}=n^{\tilde{g}}_{h(j),\bullet}\forall j\in\{1,2\},\forall\tilde{g}\in\mathcal{G}|\Theta^{\tilde{g}}_{h(j),\bullet}\forall j\in\{1,2\},\forall\tilde{g}\in\mathcal{G}\right]f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_0^\infty\cdots\int_0^\infty P\left[N^{\tilde{g}}_{h(j),\bullet}=n^{\tilde{g}}_{h(j),\bullet}\forall j\in\{1,2\},\forall\tilde{g}\in\mathcal{G}|\Theta^{\tilde{g}}_{h(j),\bullet}\forall j\in\{1,2\},\forall\tilde{g}\in\mathcal{G}\right]f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}}.
\end{aligned}$$

Again, we rely on numerical integration to compute the six-dimensional integrals.

The conditional expectation given above can be computed for different number of claims. Here, we only consider a medium risk profile for both policyholders, again to ease the presentation.

On Figure 5.2, we show the ratio of the conditional expectations to the unconditional expectation (i.e. the a priori claim frequencies), when no claim occurred, or when only one claim occurred. In the latter case, three cases are distinguished depending on which guarantees have been triggered. Both the estimates related to the policyholder and its spouse are included on the figure.

As it can be seen, a claim in TPL from the first policyholder, for instance, will have consequences on all estimates for every policyholder from the household. So, a claim will penalize a policyholder but a claim-free year for his or her spouse may in fact lessen this increase. On the other hand, a claim-free household (with at least two policyholders) more rapidly reaches some lower correction levels than in a single policyholder case.

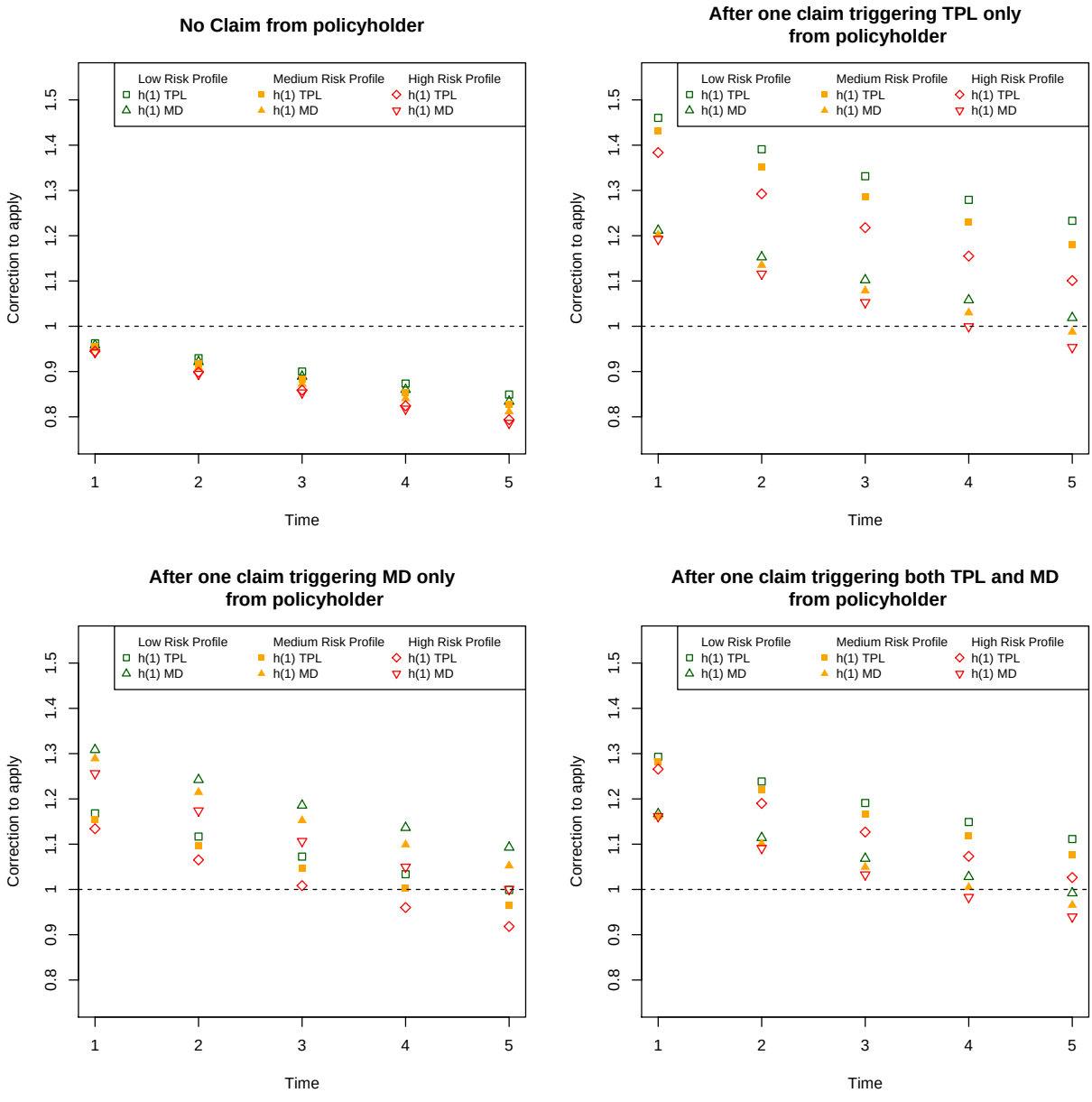

Figure 5.1: Ratio of a posteriori claim frequencies to a priori claim frequencies in TPL and MD in a household with only one policyholder. Four cases are considered depending on whether there was no claim or which guarantees were triggered by the claim.

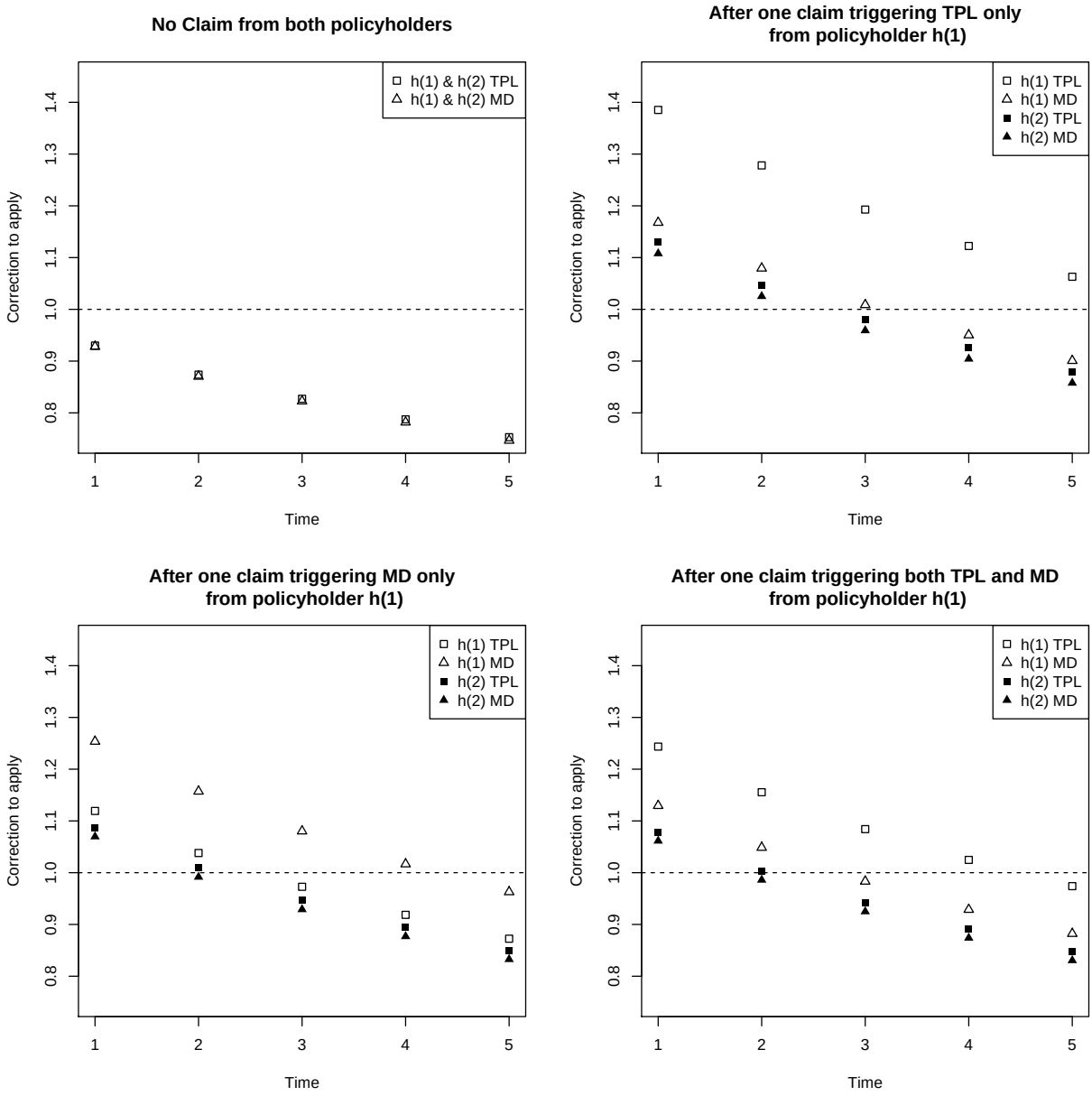

Figure 5.2: Ratio of a posteriori claim frequencies to a priori claim frequencies in TPL and MD in a household with two policyholders. Four cases are considered depending on whether there was a claim and if so, which guarantees were triggered by the claim.

5.2 Cross-Selling

Let us assume that for the last T years, a policyholder has subscribed both the TPL and MD guarantees. Furthermore, let us assume that he is the only policyholder in our database that comes from this household h. We wish to quantify the a posteriori claim frequency correction in TPL and in MD for his or her spouse who is not in the portfolio. This could be useful for instance in a context of detection of cross-selling opportunities.

We will again assume that the policyholder that is already in our database, $h(1)$, has had a constant risk profile for the past T years, i.e. its a priori claim frequency was the same each year of observation. We will assume that $h(1)$ has one of the three aforementioned risk profiles. The other member of the household, the spouse, who does not have a policy at this insurer yet will be called $h(\star)$. First, we will assume that we do not have any claims related information on $h(\star)$. In a second example, we will integrate these specific additional informations.

5.2.1 No claim-related informations on $h(\star)$

Let us first assume that we do not have information related to the number of past claims of the spouse $h(\star)$. We want to assess the correction factor to apply to the spouse, both in TPL and in MD, given the past claims of the policyholder of the household $h(1)$. Since

$$f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta}|N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G})=\frac{P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}|\boldsymbol{\Theta_h}=\boldsymbol{\theta}\right]f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta})}{P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]}$$

where $\boldsymbol{\Theta_h}=(\Theta^{TPL}_{h(1)},\Theta^{MD}_{h(1)},\Theta^{MD:TPL}_{h(1)},\Theta^{TPL}_{h(\star)},\Theta^{MD}_{h(\star)},\Theta^{MD:TPL}_{h(\star)})$, we can deduce the conditional expectations

$$\begin{aligned}
&\mathrm{E}\left[\lambda^{TPL}_{h(\star),T+1}\Theta^{TPL}_{h(\star)}+\lambda^{MD:TPL}_{h(\star),T+1}\Theta^{MD:TPL}_{h(\star)}|N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]\\
&=\lambda^{TPL}_{h(\star),T+1}\frac{\int_0^\infty\cdots\int_0^\infty\theta^{TPL}_{h(\star)}P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}|\boldsymbol{\Theta_h}=\boldsymbol{\theta}\right]f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta})}{P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]}\\
&+\lambda^{MD:TPL}_{h(\star),T+1}\frac{\int_0^\infty\cdots\int_0^\infty\theta^{MD:TPL}_{h(\star)}P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}|\boldsymbol{\Theta_h}=\boldsymbol{\theta}\right]f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta})}{P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]}
\end{aligned}\tag{5.1}$$

and

$$\begin{aligned}
&\mathrm{E}\left[\lambda^{MD}_{h(\star),T+1}\Theta^{MD}_{h(\star)}+\lambda^{MD:TPL}_{h(\star),T+1}\Theta^{MD:TPL}_{h(\star)}|N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]\\
&=\lambda^{MD}_{h(\star),T+1}\frac{\int_0^\infty\cdots\int_0^\infty\theta^{MD}_{h(\star)}P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}|\boldsymbol{\Theta_h}=\boldsymbol{\theta}\right]f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta})}{P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]}\\
&+\lambda^{MD:TPL}_{h(\star),T+1}\frac{\int_0^\infty\cdots\int_0^\infty\theta^{MD:TPL}_{h(\star)}P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}|\boldsymbol{\Theta_h}=\boldsymbol{\theta}\right]f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta})}{P\left[N^g_{h(1),\bullet}=n^g_{h(1),\bullet}\forall g\in\mathcal{G}\right]}.
\end{aligned}\tag{5.2}$$

These expressions can again be computed numerically, e.g. using Gauss quadrature.

On Figure 5.3, we show the correction to apply in TPL (resp. in MD) to the spouse $h(\star)$. We consider three different risk profiles for $h(1)$ and consider as well multiple time histories, ranging

from 1 to 5 years. As could be expected, the correction is below one (i.e. in average, the claim frequency of the spouse will be below the one given by a priori risk classification) when no claims occurred to $h(1)$. In particular, we see that for some riskier profiles for $h(1)$, five claim-free years will in average mean that the spouse seem to have 10% lower claim frequencies with respect to their a priori risk classification, both in TPL and in MD. On the contrary, a claim from $h(1)$ will increase the correction, i.e. the spouse has higher claim frequencies than its a priori risk class. Note however that if only one claim occurred in five years, estimates can in some cases decrease below the unit threshold again. This application may prove to be useful to identify cross-selling opportunities. The difference with Figure 5.2 is that, here, the policyholder $h(\star)$ is considered to have zero exposure.

5.2.2 Integration of claim-related informations on $h(\star)$

Let us now assume that we gather information on the number of past claims of $h(\star)$ has had (at another insurer). This may correspond to the level occupied in a bonus-malus scheme or to the claim experience over the last five years.

We can include that information and again compute the conditional expectation, with respect to its past claims as well as the past claims of policyholder $h(1)$ over the past T years. The difficulty, however, lies in the estimation of the past a priori claim frequencies: We do not know $h(\star)$ risk profile over the past T years, e.g. we do not know which kind of car $h(\star)$ has been driving during the past T years. Additionally, generally the only information related to the past claims that is handed out is only divided in two types of claims: those that triggered TPL and those that triggered MD. In case only one claim was reported, we can deduce that there was not a common shock, i.e. one claim triggering multiple guarantees at the same time. However, in case both kind of claims (TPL and MD) show (at least) one record, one does not know whether this is a common shock, or two separate claims. We can, however, compute the probability that, given that there has been a claim in TPL and in MD, it was a common shock.

Let us now consider the particular case where one claim in TPL and one claim in MD has been observed in the past on $h(\star)$'s side. We do not know whether it is only one claim triggering both guarantees or whether these are two separate claims.

First, we have to assess the probability that it was a common shock. Note that we have to also condition on the claims related information of h(1). We have that

$$\begin{aligned}
&P[N^{MD:TPL}_{h(\star),\bullet} = 1 | N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}] \\
&= \frac{P[N^{MD:TPL}_{h\star,\bullet} = 1, N^{TPL}_{h\star,\bullet} + N^{MD:TPL}_{h\star,\bullet} = 1, N^{MD}_{h\star,\bullet} + N^{MD:TPL}_{h\star,\bullet} = 1, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}]}{P[N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}]} \\
&= \frac{P[N^{MD:TPL}_{h\star,\bullet} = 1, N^{TPL}_{h\star,\bullet} = 0, N^{MD}_{h\star,\bullet} = 0, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}]}{P[N^{TPL}_{h\star,\bullet} + N^{MD:TPL}_{h\star,\bullet} = 1, N^{MD}_{h\star,\bullet} + N^{MD:TPL}_{h\star,\bullet} = 1, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}]} \\
&= \frac{A}{A+B}
\end{aligned}$$

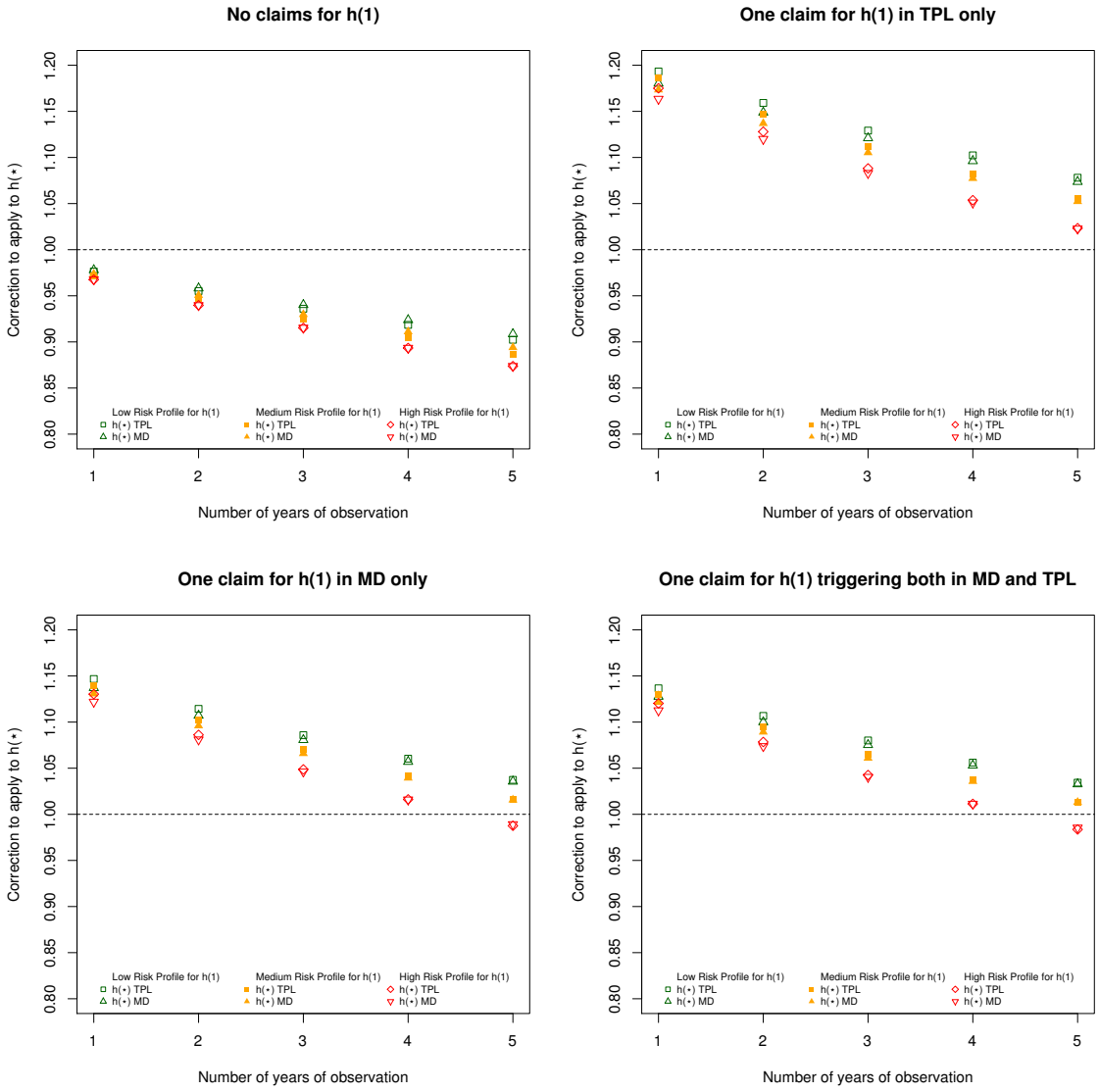

Figure 5.3: Correction to apply to a priori claim frequencies of $h(\star)$ in TPL (resp. in MD) conditional to the number of claims of $h(1)$. No claim related information on $h(\star)$ known.

where

$$\begin{aligned}
A =& P[N^{MD:TPL}_{h\star,\bullet} = 1, N^{TPL}_{h\star,\bullet} = 0, N^{MD}_{h\star,\bullet} = 0, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}] \\
=& \underbrace{\int_0^{+\infty} \cdots \int_0^{+\infty}}_{6\times} P[N^{MD:TPL}_{h\star,\bullet} = 1, N^{TPL}_{h\star,\bullet} = 0, N^{MD}_{h\star,\bullet} = 0, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G} | \boldsymbol{\Theta_h} = \boldsymbol{\theta}] f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
B =& P[N^{MD:TPL}_{h\star,\bullet} = 0, N^{TPL}_{h\star,\bullet} = 1, N^{MD}_{h\star,\bullet} = 1, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}] \\
=& \underbrace{\int_0^{+\infty} \cdots \int_0^{+\infty}}_{6\times} P[N^{MD:TPL}_{h\star,\bullet} = 0, N^{TPL}_{h\star,\bullet} = 1, N^{MD}_{h\star,\bullet} = 1, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G} | \boldsymbol{\Theta_h} = \boldsymbol{\theta}] f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta}) d\boldsymbol{\theta}
\end{aligned}$$

In case more than one claim were reported in TPL and in MD, we have that

$$\begin{aligned}
& P[N^{MD:TPL}_{h(\star),\bullet} = n_3 | N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = n_1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = n_2, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}] \\
=& \frac{P[N^{MD:TPL}_{h(\star),\bullet} = n_3, N^{TPL}_{h(\star),\bullet} = n_1 - n_3, N^{MD}_{h(\star),\bullet} = n_2 - n_3, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}]}{\sum_{n=0}^{min(n_1,n_2)} P[N^{MD:TPL}_{h(\star),\bullet} = n, N^{TPL}_{h(\star),\bullet} = n_1 - n, N^{MD}_{h(\star),\bullet} = n_2 - n, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G}]} \\
=& \frac{A_{n_3}}{\sum_{n=0}^{min(n_1,n_2)} A_n}
\end{aligned}$$

with

$$A_n = \underbrace{\int_0^{+\infty} \cdots \int_0^{+\infty}}_{6\times} P[N^{MD:TPL}_{h\star,\bullet} = n, N^{TPL}_{h\star,\bullet} = n_1 - n, N^{MD}_{h\star,\bullet} = n_2 - n, N^{g}_{h(1),\bullet} = n_g \ \forall g \in \mathcal{G} | \boldsymbol{\Theta_h} = \boldsymbol{\theta}] f_{\boldsymbol{\Theta_h}}(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

We can use numerical integration to compute these probabilities, with as a priori claim frequencies the median values on our portfolio (i.e. we assume a "Medium" risk class for $h(\star)$). This probability will of course depend on the number of years of observation (remember the a priori claim frequencies are aggregated : a medium risk profile over T years will be T times the yearly median claim frequency). The probabilities for different values of T years are given in Table 5.1.

T	1	2	3	4	5
	0.9122	0.8522	0.8077	0.7731	0.7453

Table 5.1: Probability for different number of years of observation T that the two claim records in TPL and MD triggered both guarantees at the same time. "Medium" risk profile assumed for $h(1)$ and for $h(\star)$.

Once the probability that the two claim records in TPL and MD are a common shock has been computed, one can compute the correction to apply to the TPL (resp. MD) a priori claim

frequency. We have that $\forall g \in \mathcal{G}$,

$$
\begin{aligned}
&\mathrm{E}\left[\Theta^{g}_{h(\star)} | N^{\tilde{g}}_{h(1),\bullet} = n^{\tilde{g}}\ \forall \tilde{g} \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1\right] \\
=&\mathrm{E}\left[\Theta^{g}_{h(\star)} | N^{\tilde{g}}_{h(1),\bullet} = n^{\tilde{g}}\ \forall \tilde{g} \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} = 0, N^{MD}_{h(\star),\bullet} = 0, N^{MD:TPL}_{h(\star),\bullet} = 1,\right] \times \\
&P[N^{MD:TPL}_{h(\star),\bullet} = 1 | N^{\tilde{g}}_{h(1),\bullet} = n_{\tilde{g}}\ \forall \tilde{g} \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1] \\
&+ \\
&\mathrm{E}\left[\Theta^{g}_{h(\star)} | N^{\tilde{g}}_{h(1),\bullet} = n^{\tilde{g}}\ \forall \tilde{g} \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} = 1, N^{MD:TPL}_{h(\star),\bullet} = 0,\right] \times \\
&P[N^{MD:TPL}_{h(\star),\bullet} = 0 | N^{\tilde{g}}_{h(1),\bullet} = n_{\tilde{g}}\ \forall \tilde{g} \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1]
\end{aligned}
$$

Using numerical integration and the expression above, we can compute

$$
\begin{aligned}
&\lambda^{TPL}_{h(\star),\bullet}\mathrm{E}\left[\Theta^{TPL}_{h(\star)} | N^{g}_{h(1),\bullet} = n^{g}\ \forall g \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1\right] + \\
&\lambda^{MD:TPL}_{h(\star),\bullet}\mathrm{E}\left[\Theta^{MD:TPL}_{h(\star)} | N^{g}_{h(1),\bullet} = n^{g}\ \forall g \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\lambda^{MD}_{h(\star),\bullet}\mathrm{E}\left[\Theta^{MD}_{h(\star)} | N^{g}_{h(1),\bullet} = n^{g}\ \forall g \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1\right] + \\
&\lambda^{MD:TPL}_{h(\star),\bullet}\mathrm{E}\left[\Theta^{MD:TPL}_{h(\star)} | N^{g}_{h(1),\bullet} = n^{g}\ \forall g \in \mathcal{G}, N^{TPL}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1, N^{MD}_{h(\star),\bullet} + N^{MD:TPL}_{h(\star),\bullet} = 1\right]
\end{aligned}
$$

and compare these a posteriori claim frequencies to the a priori claim frequencies $\lambda^{TPL}_{h(\star),\bullet} + \lambda^{MD:TPL}_{h(\star),\bullet}$ and $\lambda^{MD}_{h(\star),\bullet} + \lambda^{MD:TPL}_{h(\star),\bullet}$.

We illustrate the case where $h(1)$ has been claim-free while $h(\star)$ may have had no claim, or a single claim in TPL or MD as well as the aforementioned case when there is one claim record in both TPL and MD in Figure 5.4. In the very favourable claim-free case, after five years the correction factor drops below 0.75, meaning that the average $h(\star)$ policyholders have a 25% lower claim frequency in both TPL and MD with respect to the a priori claim frequencies. By comparison with Figure 5.1, where we computed to correction to apply to a policyholder without accounting for other policyholders from the household, we see that thanks to $h(1)$ claim-free years, the corrections factor drops from approximately 0.82 to below 0.75. Note that the first three cases considered on Figure 5.4 coincide with the first three cases depicted on Figure 5.2. Only in the last case, a difference appears and comes from the fact that on Figure 5.4, we do not know whether there was only one claim (a double shock) or two separate claims. In the latter case, the correction is expected to be higher than in the former case.

In the situation where $h(\star)$ had a claim triggering MD only, after five years of observation, the estimates are again below one. By comparison with Figure 5.1, we see that the claim-free years of policyholder $h(1)$ helped to decrease the estimates by around 10%.

Acknowledgements

The financial support of the AXA Research Fund through the JRI project "Actuarial dynamic approach of customer in P&C" is gratefully acknowledged. We thank our colleagues from AXA Belgium, especially Arnaud Deltour, Mathieu Lambert, Alexis Platteau, Stanislas Roth and Louise Tilmant for interesting discussions that greatly contributed to the success of this research project.

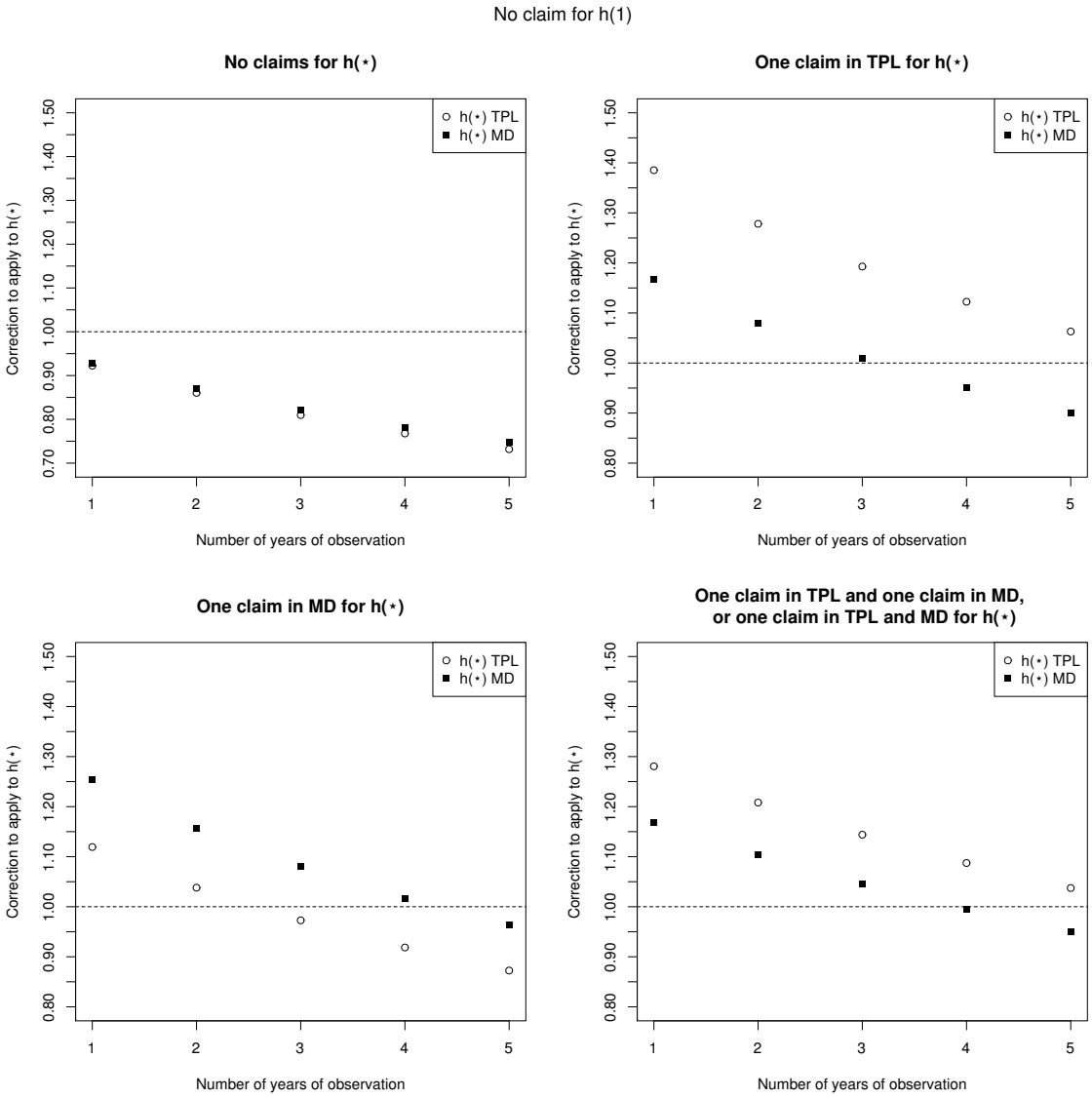

Figure 5.4: Correction to apply to a priori claim frequencies of $h(\star)$ as a function of the number of years of observation when h(1) has had no claim. Four cases are depicted depending on the claims $h(\star)$ may be have had in the past.

Also, we thank our colleagues from the SMCS, the UCL platform for statistical computing, for setting us up a comfortable and efficient working environment.

References

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

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

Bermúdez, Lluís, & Karlis, Dimitris. 2012. A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking. *Computational Statistics & Data Analysis*, **56**(12), 3988–3999.

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.

Englund, Martin, Guillén, Montserrat, Gustafsson, Jim, Nielsen, Lars Hougaard, & Nielsen, Jens Perch. 2008. Multivariate Latent Risk: A Credibility Approach. *ASTIN Bulletin*, **38**(1), 137–146.

Englund, Martin, Gustafsson, Jim, Nielsen, Jens Perch, & Thuring, Fredrik. 2009. Multidimensional credibility with time effects: An application to commercial business lines. *Journal of Risk and Insurance*, **76**(2), 443–453.

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

Frees, Edward W., Jin, Xiaoli, & Lin, Xiao. 2013. Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science*, **7**(2), 258–287.

Frees, Edward W., Lee, Gee, & Yang, Lu. 2016. Multivariate frequency-severity regression models in insurance. *Risks*, **4**(1).

Pechon, Florian, Trufin, Julien, & Denuit, Michel. 2018. Multivariate modelling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin*, 1–25.

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

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

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

Thuring, Fredrik. 2012. A credibility method for profitable cross-selling of insurance products. *Annals of Actuarial Science*, **6**(1), 65–75.