PREDICTIVE CLAIM SCORES FOR DYNAMIC MULTI-PRODUCT RISK CLASSIFICATION IN INSURANCE

BY

ROBERT MATTHIJS VERSCHUREN

ABSTRACT

It has become standard practice in the non-life insurance industry to employ generalized linear models (GLMs) for insurance pricing. However, these GLMs traditionally work only with *a priori* characteristics of policyholders, while nowadays we increasingly have *a posteriori* information of individual customers available across multiple product categories. In this paper, we therefore develop a framework to capture this *a posteriori* information over several product lines using a dynamic claim score. More specifically, we extend the bonus-malus-panel model of Boucher and Inoussa (2014) and Boucher and Pigeon (2018) to include claim scores from other product categories and to allow for nonlinear effects of these scores. The application of the proposed multi-product framework to a Dutch property and casualty insurance portfolio shows that customers' individual claims experience can have a significant impact on the risk classification. Moreover, it indicates that considerably more profits can be gained by accounting for their multi-product claims experience.

KEYWORDS

Multi-product risk profiles, dynamic claim score, bonus-malus systems, generalized additive models, cross-selling potential, insurance pricing.

JEL codes: C23, C52, G22

1. INTRODUCTION

It has become the industry standard in non-life insurance to adopt generalized linear models (GLMs) for determining the premium rate structure. Traditionally, these rate structures are based only on *a priori* characteristics of policyholders and do not account for any information available *a posteriori*.

Astin Bulletin 51(1), 1-25. doi:10.1017/asb.2020.34

In addition, customers often hold multiple policies across different product categories, while insurers tend to focus on policies in a single line of business when designing their premia. However, a lot of individual heterogeneity is typically unaccounted for in these *a priori* univariate rate structures, which may (partially) be captured by information observed *a posteriori* and from other product lines.

Several methods have been introduced in the literature to account for this form of heterogeneity. Common shocks and copulas, for instance, can induce correlation between claims from different product categories by introducing common processes and a dependence structure to the marginal claim processes of the same customer, respectively (see, e.g., Bermúdez and Karlis, 2011; Shi and Valdez, 2014). Vector GLMs, multivariate integer-valued autoregressive processes, and multivariate decision trees, on the other hand, can allow for this correlation by directly describing the vector of claims of a customer (see, e.g., Yee and Hastie, 2003; Bermúdez *et al.*, 2018; Quan and Valdez, 2018). Multivariate random effects and multivariate credibility can additionally enable a dynamic correction of the *a priori* rate structure by absorbing any variation not already accounted for by the covariates in GLMs (see, e.g., Englund *et al.*, 2008, 2009; Pechon *et al.*, 2018; Barseghyan *et al.*, 2020). However, Lemaire (1998) argues that past claiming behavior is one of the most important determinants of future claim counts and that a bonus-malus system (BMS) is therefore more intuitive for this correction. In contrast to the random effect and credibility models, the timing of past claims is now explicitly accounted for in these systems through a claim score as a special case of a Markov process with a finite number of states (Kaas *et al.*, 2008). As such, BMSs pose a commercially attractive form of experience rating where only the current level of the score matters instead of the entire claims history.

Despite the appealing framework, these claim scores have primarily been used for designing rate structures for a single product in a cross-sectional setting. Many authors consider BMSs for automobile insurance, for instance, to adjust the given static premium without accounting for any information from other product lines (see, e.g., Pinquet, 1997; Denuit *et al.*, 2007; Tzougas *et al.*, 2014, 2018). A more dynamic approach is followed by Boucher and Inoussa (2014), who argue that it is no longer consistent to use this two-step approach in case of panel or longitudinal data and suggest to estimate the *a posteriori* rate structure in a single step. Boucher and Pigeon (2018) further develop the resulting BMS-panel model and, for practical reasons, consider linear effects for the levels of the claim score.

While the BMS-panel model deals with past claiming behavior in a longitudinal setup, it has thus far only focused on linear effects and a single product. The contribution of this paper is therefore twofold. It extends this BMS-panel model by, on the one hand, allowing the claim scores to affect the rate structure of other product lines and, on the other hand, incorporating a spline for their effects using a generalized additive model (GAM). In addition, a piecewise linear simplification of this spline is considered to accommodate an

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

interpretable rate structure in practice with more flexibility than the pure linear specification. This, in turn, allows us to account for information observed *a posteriori* and from other product lines in our rate structures, to identify the cross-selling potential of customers and to investigate the relation between past claiming behavior across different product categories.

The remainder of this paper is organized as follows. In Section 2, we briefly highlight the concepts behind the industry standard of GLMs, describe the novel extension of the BMS-panel model, and additionally discuss how to determine the optimal claim score. While Section 3 describes the Dutch property and casualty insurance portfolio and comments on the exact optimization procedure, we apply this methodology in Section 4 and elaborate on the results. The final section concludes this paper with a discussion of the most important findings and implications.

2. Modeling framework

2.1. Static *a priori* risk classification

Among non-life insurers, there has been a long tradition of adopting statistical techniques to construct their *a priori* rate structure. These insurers are typically interested in predicting the total claim amount L relative to the exposure to risk e in the form of a risk premium. Technically, this risk premium π is defined as:

$$\pi = \mathbb{E}\left[\frac{L}{e}\right] = \mathbb{E}\left[\frac{N}{e} \times \mathbb{E}\left[\frac{L}{N}\middle| N > 0\right]\right] = \mathbb{E}\left[F\right] \times \mathbb{E}\left[S\right],$$

with N, $S = L/N$, and $F = N/e$ the number of claims, the average claim severity, and the claim frequency, respectively (Antonio and Valdez, 2012; Henckaerts *et al.*, 2020). While we can allow the claim frequencies and severities to interact, it is in general common practice to model these two components independently (see, e.g., Czado *et al.*, 2012; Garrido *et al.*, 2016).

Insurers now traditionally adopt the framework of GLMs for both components to properly estimate these risk premia (Nelder and Wedderburn, 1972). In this framework, we assume that the claim frequency or severity $Y_{i,t}$ is independently distributed for each subject i and period t according to some distribution from the exponential family. The mean predictor $\eta_{i,t}$ in the GLM is then given by:

$$\eta_{i,t} = g\left(\mu_{i,t}\right) = X_{i,t}'\beta \quad \text{for} \quad i = 1, \ldots, M, \quad t = 1, \ldots, T_i,$$

with $\mu_{i,t} = \mathbb{E}\left[Y_{i,t}|X_{i,t}\right]$ the conditional expectation of $Y_{i,t}$ and β a parameter vector containing the effects of the risk factors $X_{i,t}$. In practice, we typically assume a Poisson and Gamma distribution for the claim counts and sizes, respectively, while a logarithmic link function $g(\cdot)$ is commonly adopted to accommodate a multiplicative rate structure (Haberman and Renshaw, 1996).

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

While these GLMs have become the industry standard over the last decades, they lead to a static form of risk classification that only takes *a priori* information of policyholders into account. However, the longitudinal setup in this paper allows us to easily incorporate any *a posteriori* information in the mean predictor to account for past claiming behavior. By additionally including the claims experience from other product categories, we obtain a framework for dynamic multi-product risk classification.

2.2. Dynamic *a posteriori* risk classification

While the cross-sectional model assumes independence between both subjects and periods, we can account for dependencies between periods in the longitudinal setting. As in the BMS-panel model of Boucher and Inoussa (2014), past claiming behavior is summarized by a single claim score, defined recursively as:

$$\ell_{i,t+1} = \min\left(\max\left(\ell_{i,t} + e_{i,t}\mathbb{1}\left(N_{i,t}=0\right) - \Psi\frac{N_{i,t}}{e_{i,t}}, 1\right), s\right),$$

with initial value $\ell_{i,0} = \ell_0$ and indicator function $\mathbb{1}\left(N_{i,t}=0\right)$ that equals one for a period without any claims and zero otherwise. The parameters Ψ, s, and ℓ_0 denote a jump parameter, the maximum score, and the initial score for new policyholders without any experience yet, respectively. The lowest score as well as the jump size after a claim-free period are both fixed at one, since we can already capture their effect indirectly through Ψ and s. We additionally introduce the exposure of risk $e_{i,t}$ into this claim score to account for policies with exposures of less than an entire year. Moreover, the score $\ell_{i,t}$ is an indication of the *a posteriori* risk in a policy, since policyholders who claim more (less) frequently and are thus more (less) risky will receive lower (higher) scores. As such, this type of score corresponds to a BMS with transition rules $+1/-\Psi$, entry level ℓ_0, and maximum level s.

With this claim score, we can directly incorporate past claiming behavior in a longitudinal GLM, even for multiple products. The intuition behind this is that the claim score serves as a relevant predictor for future claiming behavior, rather than an *ex post* punishment (reward) for (no) claims in the past. The linear predictor for product category (c) of this multi-product claim score model is given by:

$$\eta_{i,t}^{(c)} = g^{(c)}\left(\mu_{i,t}^{(c)}\right) = X_{i,t}^{(c)\prime}\beta^{(c)} + \sum_{j=1}^{C} f_j^{(c)}\left(\ell_{i,t}^{(j)}\right) \quad \text{for} \quad c = 1, \ldots, C, \tag{2.1}$$

where $f_j^{(c)}(\cdot)$ represents some function. By additionally requiring that $f_j^{(c)}(\ell_{i,t}^{(c)}) = 0$ whenever $\ell_{i,t}^{(c)} = \ell_0$ or unknown, we can account for policyholders without any *a posteriori* information (yet) and for customers holding only a subset of all available products. In turn, the *a priori* risk premia are fully determined

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

by the policyholder's risk characteristics and any effects of the past claiming behavior, if any, are multiplicative to these premia in case of a logarithmic link function.

While Boucher and Inoussa (2014) and Boucher and Pigeon (2018) consider a logarithmic function, the transformations $f_j^{(c)}(\cdot)$ can be taken in a much more general way. The natural cubic splines used in GAMs, for instance, can already capture nonlinear as well as linear effects of the claim score on the response variable (Hastie and Tibshirani, 1986). These splines are additionally optimal among all twice continuously differentiable functions when minimizing the penalized deviance and can easily be constructed by B-splines (Hastie *et al.*, 2009; Ohlsson and Johansson, 2010). More importantly, we can express the multi-product claim score model as a GAM with these splines, where, given the claim score specification, parameters can be estimated straightforwardly by penalized maximum likelihood and with standard statistical software (see Appendix B). However, a piecewise linearly segmented rate structure is preferable from a practical perspective. We therefore adopt both natural cubic and linear splines in this paper to allow for nonlinear effects of the claim score and to benefit from the existing framework for GAMs. In turn, the resulting multi-product claim score model allows us to dynamically classify policyholders' risk profile based on their experience in multiple product categories.

2.3. Optimality of rate structure

With the multi-product claim score model, we can formulate a rate structure based on the *a priori* characteristics of policyholders and their past claiming behavior across product categories. Depending on our choice for the claim score parameters (Ψ, s, ℓ_0), different premium rates will result from the estimated model. Moreover, since a lot of different parameter combinations, and thus rate structures, are possible in this framework, a criterion is required to assess their performance. While typical statistical goodness-of-fit measures such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are based on maximized likelihoods, we are more interested in the discriminatory power of the predicted premia from a practitioner's point of view. Our aim in this context is to best identify and distinguish between risky and safe customers. A well-known approach for assessing the discriminatory power is based on the Lorenz curve and the Gini index.

The Lorenz curve has first been introduced by Lorenz (1905) in the field of welfare economics as a statistical tool to compare two distributions. In case of perfect alignment between the two distributions, the Lorenz curve reduces to the 45-degree line, or the line of equality. Similarly, the greater the discrepancy between the two distributions, the further the Lorenz curve is away from this line of equality. The Gini index is defined as twice the distance between this Lorenz curve and the line of equality and thus represents a measure of inequality (Gini, 1912). More importantly, in the context of insurance rate making, the

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

Lorenz curve and the corresponding Gini index can also be adopted as a measure of risk discrimination (see, e.g., Frees *et al.*, 2014; Henckaerts *et al.*, 2020). To find the Lorenz curve in practice, we can use the following three steps:

(i) Construct the relativity $R_j = P_j^{\mathrm{A}}/P_j^{\mathrm{B}}$ for each policy $j = 1, \ldots, H$, where P_j^{B} denotes the risk premium of a benchmark model and P_j^{A} the risk premium of an alternative model;
(ii) Order the policies by the relativities R_j from lowest to highest;
(iii) Calculate

$$\left(\hat{F}_P(\omega), \hat{F}_L(\omega)\right) = \left(\frac{\sum_{j=1}^{H} P_j^{\mathrm{B}} \times \mathbb{1}\left\{F_H(R_j) \leq \omega\right\}}{\sum_{j=1}^{H} P_j^{\mathrm{B}}}, \frac{\sum_{j=1}^{H} L_j \times \mathbb{1}\left\{F_H(R_j) \leq \omega\right\}}{\sum_{j=1}^{H} L_j}\right)$$

as a function of $\omega \in [0, 1]$, where L_j denotes the actual observed claim amount of policy j and $F_H(\cdot)$ the empirical cumulative distribution function of the relativities R_j.

In turn, this ordered Lorenz curve $\left\{\left(\hat{F}_P(\omega), \hat{F}_L(\omega)\right) : \omega \in [0, 1]\right\}$ leads to the empirical Gini index, given by the expression:

$$\widehat{\mathrm{Gini}} = 1 - \sum_{j=0}^{H-1}\left(\hat{F}_P\left(F_H(R_{j+1})\right) - \hat{F}_P\left(F_H(R_j)\right)\right)\left(\hat{F}_L\left(F_H(R_{j+1})\right) + \hat{F}_L\left(F_H(R_j)\right)\right)$$

from the trapezoidal rule where $R_0 = 0$. Its asymptotic covariance matrix can be consistently estimated as:

$$\hat{\Sigma}_{\mathrm{Gini}} = \frac{4}{\hat{\mu}_L^2 \hat{\mu}_P^2}\left(4\hat{\Sigma}_h + \frac{\hat{\mu}_h^2}{\hat{\mu}_L^2}\hat{\Sigma}_L + \frac{\hat{\mu}_h^2}{\hat{\mu}_P^2}\hat{\Sigma}_P - \frac{4\hat{\mu}_h}{\hat{\mu}_L}\hat{\Sigma}_{hL} - \frac{4\hat{\mu}_h}{\hat{\mu}_P}\hat{\Sigma}_{hP} + \frac{2\hat{\mu}_h^2}{\hat{\mu}_L\hat{\mu}_P}\hat{\Sigma}_{LP}\right),$$

with $h_j = \frac{1}{2}(\mu_L P_j^{\mathrm{B}} F_L(R_j) + L_j \mu_P[1 - F_P(R_j)])$ for $j = 1, \ldots, H$, and using moment-based estimators for all the means and covariances of L, P, and h (Frees *et al.*, 2011).

Depending on our choice for the benchmark model, there are two versions of the Gini index, namely the simple Gini index and the ratio Gini index. If we simply assume a constant benchmark premium for every policy without any risk discrimination, or $P_j^{\mathrm{B}} = 1$, we are calculating the simple Lorenz curve and the total degree of risk discrimination of the alternative model. However, we often have an existing framework or benchmark premium rate in place, such as a standard GLM, that we would like to improve. In these cases, it makes more sense to not calculate the total degree of risk discrimination but to compare the risk classification resulting from the alternative model to that from this benchmark model. Frees *et al.* (2014) describe a mini-max strategy for determining which model leads to the best risk classification. They calculate the ratio Gini coefficient for every combination of alternative and benchmark model and select the benchmark model that minimizes the maximal coefficient.

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

The intuition behind this is that the model that minimizes the maximal Gini coefficient is the least vulnerable to alternative specifications. The use of this ratio Gini index has an additional practical advantage, since we can directly relate it to the profit potential of the alternative rate structure over the benchmark structure. If we take P_j^{B} to be the risk premia of a benchmark GLM, then the ratio Gini coefficient (divided by two) quantifies how much more profitable it is, on average, to use, for instance, the multi-product claim score model due to the different ordering of risks. In other words, the ratio Gini index enables us to identify which design of the claim score is most profitable and optimal in the sense of risk discrimination as opposed to the benchmark GLM in non-life insurance.

3. Data and empirical considerations

3.1. Property and casualty insurance

To illustrate the implications of the multi-product claim score model in practice, we apply this model framework to real-world non-life insurance claims. More specifically, we analyze a Dutch property and casualty insurance portfolio containing 183,690 policies on general liability, 264,348 on home contents, 111,018 on home, and 363,573 on travel insurance from 2012 up to and including 2018. These policies generally have a duration of 1 year and need to be renewed annually, but policyholders may enter or leave the portfolio at any moment. For each of these four products, we consider a different set of explanatory variables that are known to be used to construct premium rates in the Netherlands. For more details on the exact covariates used for each product category in this paper, see Tables A.1–A.5 in Appendix A.

From the individual claim counts and sizes in this portfolio, we can clearly see excess zeros and negative skewness for each product category in Figure C.1 in supplementary Appendix C. Note that, for illustrative purposes, the claim severities are shown on a logarithmic scale and that the excess zeros in the claim counts seem to indicate that a negative binomial (NB) or zero-inflated (ZI) distribution is more appropriate than a Poisson distribution. More importantly, in Table 1, we show how many policyholders also own other products and in Table 2 how many claims have occurred in each product line. For general liability insurance, for instance, we observe 34,407 customers, of which 5660 [16.45%] own only that product, 13,285 [38.61%] exactly two products, 11,898 [34.58%] exactly three products, and 3564 [10.36%] all four products. Out of the 3520 general liability claims in total, 511 [14.52%] are filed by customers holding only that product, 1192 [33.86%] by those holding exactly two products, 1290 [36.65%] by those holding exactly three products, and 527 [14.97%] by those holding all four products. Moreover, in 339, 147, and 12 cases, the customers holding exactly two, three, and four products, respectively, have also filed at least one claim on the other product(s). In other words, we see that quite a lot of customers have general liability, home contents, and/or

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

TABLE 1

NUMBER OF POLICYHOLDERS IN EACH INSURANCE PRODUCT CATEGORY WITH PERCENTAGES IN SQUARE BRACKETS.

	Products owned				
Policyholders	1	2	3	4	Total
General liability	5660 [16.45%]	13,285 [38.61%]	11,898 [34.58%]	3564 [10.36%]	34,407 [100.00%]
Home contents	10,272 [22.88%]	18,333 [40.84%]	12,724 [28.34%]	3564 [7.94%]	44,893 [100.00%]
Home	688 [3.44%]	5612 [28.08%]	10,119 [50.64%]	3564 [17.84%]	19,983 [100.00%]
Travel	73,412 [89.01%]	2042 [2.48%]	3461 [4.20%]	3564 [4.32%]	82,479 [100.00%]

TABLE 2

NUMBER OF CLAIMS IN EACH INSURANCE PRODUCT CATEGORY WITH PERCENTAGES IN SQUARE BRACKETS AND TAKING INTO ACCOUNT THAT AT LEAST ONE CLAIM HAS ALSO BEEN FILED IN THE OTHER CATEGORIES IN PARENTHESES.

	Products owned									
Claims	1		2		3		4		Total	
General liability	511	(0)	1192	(339)	1290	(147)	527	(12)	3520	(498)
	[14.52%	(0.00%)]	[33.86%	(68.07%)]	[36.65%	(29.52%)]	[14.97%	(2.41%)]	[100.00%	(100.00%)]
Home contents	1546	(0)	3820	(870)	3508	(204)	1297	(16)	10,171	(1090)
	[15.20%	(0.00%)]	[37.56%	(79.82%)]	[34.49%	(18.72%)]	[12.75%	(1.47%)]	[100.00%	(100.00%)]
Home	63	(0)	1285	(522)	2506	(174)	1127	(16)	4981	(712)
	[1.26%	(0.00%)]	[25.80%	(73.31%)]	[50.31%	(24.44%)]	[22.63%	(2.25%)]	[100.00%	(100.00%)]
Travel	9148	(0)	296	(59)	645	(35)	683	(16)	10,772	(110)
	[84.92%	(0.00%)]	[2.75%	(53.64%)]	[5.99%	(31.82%)]	[6.34%	(14.55%)]	[100.00%	(100.00%)]

home insurance, but that there is relatively little overlap with travel insurance. In addition, given that a customer files a claim, we observe a slight yet non-negligible tendency of customers holding exactly two products to have claims in both product categories, while this diminishes as the customer owns more products.

3.2. Optimization methodology

Using the Dutch insurance portfolio, we estimate the multi-product claim score model developed in this paper. We model the claim frequencies and severities independently and assume a Poisson (P), NB, and ZI Poisson (ZIP) distribution for the claim counts and a Gamma (G), Inverse-Gaussian (IG) and Pareto (P) distribution for the claim sizes, both with a logarithmic link function. Moreover, we consider an ordinary GLM for the claim severities and apply the multi-product framework to the claim frequencies, with explanatory variables given in Appendix A for each product category without any interactions. We additionally allow the risk factors to affect the logit of the ZI parameter of the ZIP distribution but include the claim scores only in the Poisson component to avoid identification issues and for consistent interpretation of the marginal effects. While these assumptions can be relaxed, they are adopted nonetheless since they correspond to standard practices in the non-life insurance industry.

Under these assumptions, we apply the multi-product framework to the property and casualty insurance data. Given the claim score parameters, we estimate this framework by penalized maximum likelihood, or penalized iteratively reweighted least squares (PIRLS), which we describe in detail in Appendix B and can be performed efficiently in `R` with the package `mgcv` developed by Wood (2006). However, rather than letting the smoothing penalty determine the number of parameters for the B-splines, we employ $k = 4$ parameters for all of them to sufficiently account for nonlinearities using the $k - 1 = 3$ effective degrees of freedom. We additionally adopt these splines with default knot locations and replace their centering constraint by the *a priori* constraint introduced earlier that $f_j(\ell_{i,t}) = 0$ whenever $\ell_{i,t} = \ell_0$ or unknown, equivalent to having no *a posteriori* information. This, in turn, allows us to exploit all information available, both on customers with or without any *a posteriori* information and with or without all products. Finally, we compare the multi-product framework developed in this paper to the case of linear claim score effects resembling the BMS-panel model of Boucher and Inoussa (2014) and Boucher and Pigeon (2018) to assess the value of our extension.

The above methodology assumes known claim score parameters, while in practice these are unknown as well. We therefore determine the optimal parameters independently for each product by a grid search in terms of the ratio Gini index with a standard GLM as benchmark. More specifically, we estimate the claim score model for each product separately on training data from the period of 2012 up to and including 2017 and select the parameters

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

TABLE 3

ABBREVIATIONS FOR COMBINED FREQUENCY $(X) \in \{P, NB, ZIP\}$ AND SEVERITY $(Y) \in \{G, IG, P\}$ MODELS.

Abbreviation	Frequency Severity model (X)	Severity model (Y)
GLM-$(X)(Y)$	(X) GLM	(Y) GLM
GAM-$(X)(Y)$-One	(X) one-product claim score GAM	(Y) GLM
GLM-$(X)(Y)$-One	(X) one-product claim score GLM	(Y) GLM
GAM-$(X)(Y)$-Multi	(X) multi-product claim score GAM	(Y) GLM
GLM-$(X)(Y)$-Multi	(X) multi-product claim score GLM	(Y) GLM
GAM-$(X)(Y)$-Multi-PL	(X) multi-product claim score piecewise linear GAM	(Y) GLM

that lead to the best ratio Gini index for test data from 2018. We additionally impose the restriction that each claim score level, after truncation toward ℓ_0, must contain at least 0.01% of the training set's exposure to avoid parameter combinations that can lead to unobserved levels. While we consider the values $\{1, 2, \ldots, s-1\}$ for Ψ and $\{3, 4, \ldots, 25\}$ for s, we implement the set $\{2, 3, \ldots, s-1\}$ for ℓ_0 in case policyholders have no prior claims experience.

However, in practice, policyholders often switch between insurers or have been a customer at the insurer previously, meaning that they do in fact have claims experience prior to our data. Fortunately, we have access to the claims history at the insurer from 2005 up to and including 2011, albeit not to policyholders' risk characteristics. All the claims filed, if any, by 12,361, 13,210, 5460, and 13,112 customers are therefore available for general liability, home contents, home and travel insurance in this period, respectively. As a proxy to the unobserved prior claims experience, we can already construct the claim score and use its level at the end of this 7-year period to initialize the claim score for existing policyholders. Given the optimal claim score parameters, we can then specify and estimate the multi-product framework. As such, the multi-product claim score model remains tractable and allows us to incorporate past claiming behavior across product categories for insurance pricing.

4. APPLICATIONS IN NON-LIFE INSURANCE

4.1. Static risk profiles

Based on the Dutch insurance portfolio and the methodology described earlier, we explore how well a standard GLM can describe insurance data. While we adopt the abbreviations in Table 3 henceforth, we show the out-of-sample errors for the estimated GLMs in Figure C.2. Moreover, Table C.1 depicts the maximal ratio Gini coefficients for the static GLMs with the Poisson, NB, and ZIP distribution for the claim frequencies and the Gamma, IG, and Pareto distribution for the claim severities.

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

From the prediction errors, it is apparent that the magnitude of the out-of-sample errors can differ substantially across product categories. For home contents insurance, for instance, we obtain the largest errors, whereas the predicted premia for general liability insurance appear to be much closer to the realized expenses than for the other product lines. More importantly, we find that the prediction errors are distributed largely the same regardless of the model specification. This is additionally supported by the parameter estimates which appear roughly the same for the three-frequency and the three-severity components and are available upon request. The distribution underlying the static GLMs therefore does not seem to really affect the prediction errors or to matter that much for the goodness of fit.

However, in terms of the ratio Gini index, we do find a substantial impact of the distribution underlying the static GLMs. Based on the minimax strategy, we find that the static GLM-NBG or GLM-NBP is the least vulnerable to alternative model choices for general liability or home contents and home insurance, respectively, while the static GLM-ZIPP is the least vulnerable for travel insurance. These results thus seem to indicate that the NB (Pareto) distribution is generally more profitable to adopt for the frequency (severity) component than the Poisson or ZIP (Gamma or IG) distribution. A straightforward selection procedure based on both the AIC and BIC values in Table C.2 additionally confirms these findings.

4.2. Dynamic risk profiles

While the standard GLM is a form of static risk classification, we can also create dynamic risk profiles from the claims experience of individual policyholders. Using the claim score introduced earlier, whose range and hence meaning are reversed compared to Boucher and Inoussa (2014) and Boucher and Pigeon (2018), we first account for this claims experience on a single product and optimize the parameters (Ψ, s, ℓ_0) for each product category separately. The resulting one-product models consider Poisson, NB, and ZIP distributed claim frequencies, Gamma, IG, and Pareto distributed claim severities and use either GAM or GLM specifications. Moreover, from the optimal claim score parameters in Table 4, we see that the claim frequency distribution is not very important. More specifically, there can be substantial differences between the GAM and GLM specifications, but the one-product GAMs only appear to distinguish between Pareto and non-Pareto distributed claim severities. The one-product GLMs, on the other hand, can lead to different but still roughly similar parameters.

In contrast to the one-product models, we can also account for the claims experience across multiple product lines. As such, we extend the one-product models by incorporating the claim score on the other products of the policyholder, if any, given the optimized claim score parameters. We display the

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

TABLE 4

OPTIMAL CLAIM SCORE PARAMETERS FOR EACH PRODUCT CATEGORY WITH DYNAMIC UNIVARIATE RISK CLASSIFICATION, WHERE BOLD PARAMETER COMBINATIONS CORRESPOND TO THE LOWEST MAXIMAL RATIO GINI COEFFICIENT IN TABLE C.3.

	Optimal claim score parameters (Ψ, s, ℓ_0)							
	General liability		Home contents		Home		Travel	
Frequency model	GAM	GLM	GAM	GLM	GAM	GLM	GAM	GLM
GAM/GLM-PG-One	(2, 5, 2)	(2, 6, 1)	(13, 14, 13)	(18, 19, 1)	(2, 7, 1)	(2, 7, 1)	(2, 3, 2)	(2, 3, 2)
GAM/GLM-PIG-One	(2, 5, 2)	**(2, 8, 1)**	(13, 14, 13)	(18, 19, 1)	(2, 7, 1)	(2, 7, 1)	(2, 3, 2)	(2, 3, 2)
GAM/GLM-PP-One	(10, 14, 4)	(2, 3, 1)	(14, 25, 13)	(13, 25, 14)	(12, 17, 13)	(16, 21, 1)	(2, 5, 3)	(2, 5, 3)
GAM/GLM-NBG-One	(2, 5, 2)	(21, 24, 6)	(13, 14, 13)	(18, 19, 1)	(2, 7, 1)	(2, 7, 1)	(2, 3, 2)	(2, 3, 2)
GAM/GLM-NBIG-One	(2, 5, 2)	(2, 8, 1)	(13, 14, 13)	(18, 19, 1)	(2, 7, 1)	(2, 7, 1)	(2, 3, 2)	(2, 3, 2)
GAM/GLM-NBP-One	(10, 14, 4)	(2, 3, 1)	(14, 25, 13)	**(13, 25, 14)**	(12, 17, 13)	**(15, 20, 1)**	(2, 5, 3)	(2, 5, 3)
GAM/GLM-ZIPG-One	(2, 5, 2)	(18, 21, 5)	(13, 14, 13)	(13, 15, 14)	(2, 7, 1)	(2, 7, 1)	(2, 3, 2)	(2, 3, 2)
GAM/GLM-ZIPIG-One	(2, 5, 2)	(17, 20, 5)	(13, 14, 13)	(13, 15, 14)	(2, 7, 1)	(2, 7, 1)	**(2, 3, 2)**	(2, 3, 2)
GAM/GLM-ZIPP-One	(10, 14, 4)	(3, 7, 3)	(14, 25, 13)	(12, 25, 13)	(12, 17, 13)	(16, 21, 1)	(2, 5, 3)	(2, 5, 3)

effects of these claim scores on the linear predictor and the corresponding discounts/surcharges offered to the insured in Figure 1 for each product category separately, where the multi-product GAMs and GLMs lead to almost the same claim score effects and we therefore only show those for the PG-Multi specifications. Figure 2 and Tables C.4 and C.5 additionally present the maximal ratio Gini coefficients and AIC/BIC values when we include these dynamic (multivariate) risk profiles, with all parameter estimates available upon request. The results from the one-product models are provided in Figure C.3 and Table C.3 as well, but since they lead to similar conclusions they are omitted from the text here.

From the claim score effects in Figure 1, we observe that in general policyholders who claim more (less) frequently are more (less) risky. However, we do find some exceptions to this rule in case of the one- and multi-product GAMs. For general liability (home contents) insurance in Figure 1(a) (1(b)), for instance, we find that policyholders with the highest or lowest claim score receive a relatively large (small) discount or surcharge, respectively, but that those with a score between these two extremes receive approximately the same small (large) discount (surcharge). This relation is far less clear and appears even more complicated for the claim scores on other products. We, for instance, find that policyholders merely possessing home contents and/or travel insurance are almost always associated with more risk in the multi-product GAMs. This, in turn, implies that insurers should not target customers holding these products with cross-selling offers since we expect these customers to receive low claim scores or claim relatively often on the other products. Note that the relatively large confidence bands for home contents and travel insurance result from a lack of policyholders with these claim scores, since most policyholders claim very few, more claim score levels lead to sparser distributions and there is relatively little overlap from travel insurance with the other products. The cubic spline approach thus seems more representative of the uncertainty of the claim score effects than the linear approach.

Interestingly enough, the multi-product GLMs do not indicate these subtleties in the claim score due to a lack of exposure in customers owning multiple products and simultaneously having low claim scores. More explicitly, since the effects of the claim scores in these multi-product GLMs are linear, they are essentially a weighted average of all the observed claim scores and do not reflect the actual non-monotonicities in the average observed claim frequencies in the histograms in Figure 1. However, in Tables 1 and 2, we see that most policyholders do not claim (at all) and, as a result, end up with high claim scores. The estimates for the linear claim scores are therefore dominated by policyholders with high claim scores and seem a rather poor linear approximation of the cubic claim scores that is primarily appropriate for the good risks. The flexibility of the one- and multi-product GAMs, on the other hand, allows us to adjust for this lack of exposure by employing multiple cubic splines instead of forcing a single linear relation for all claim scores. This, in turn, enables the cubic claim scores to largely capture the non-monotonic claim frequencies. As such, the

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

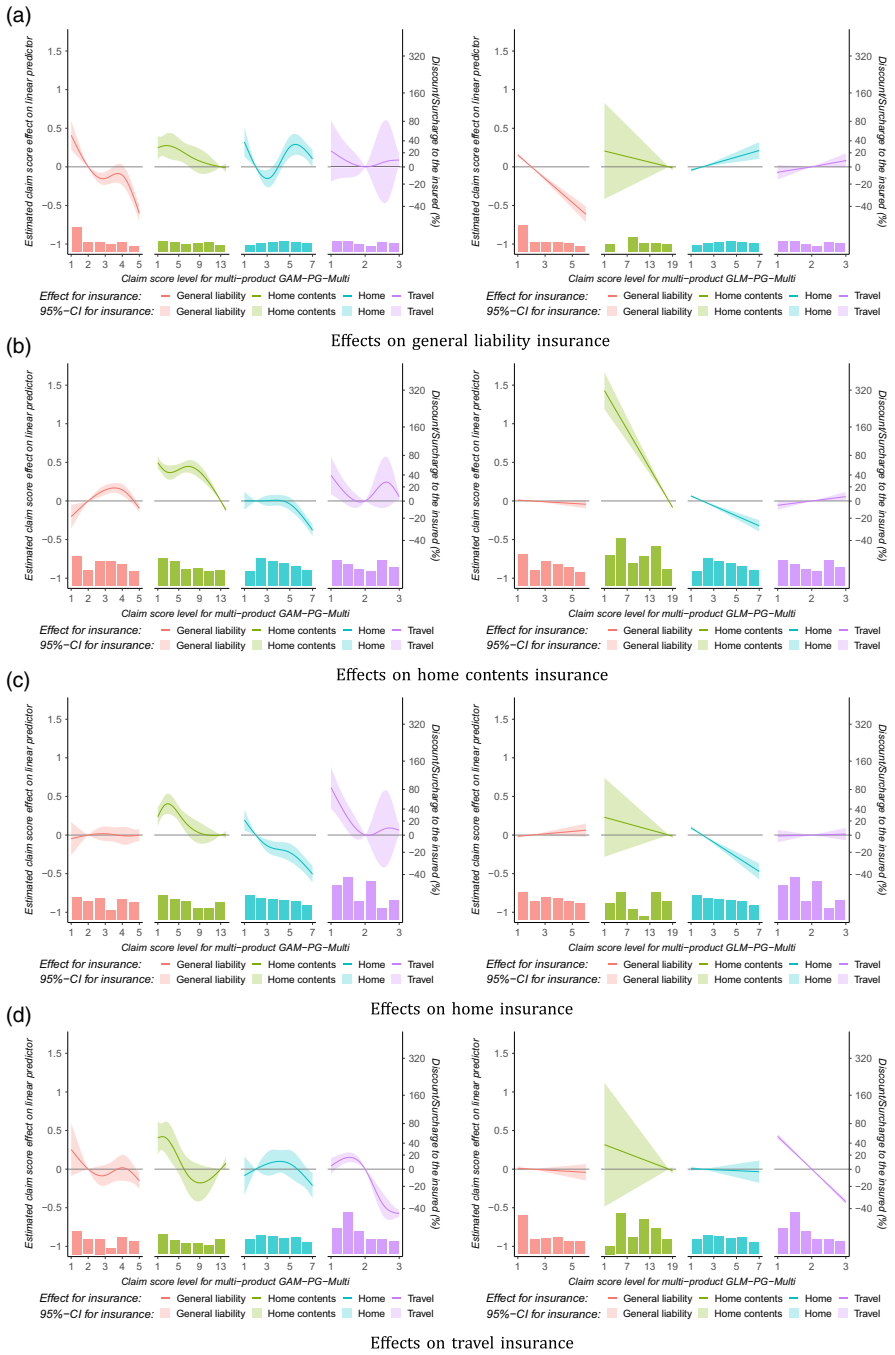

FIGURE 1: Estimated claim score effects on linear predictor and corresponding discounts/surcharges to the insured with 95% confidence intervals and histograms of average observed claim frequencies of general liability (panel (a)), home contents (panel (b)), home (panel (c)), and travel insurance (panel (d)) for multi-product GAM-PG-Multi (left) and GLM-PG-Multi (right).

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

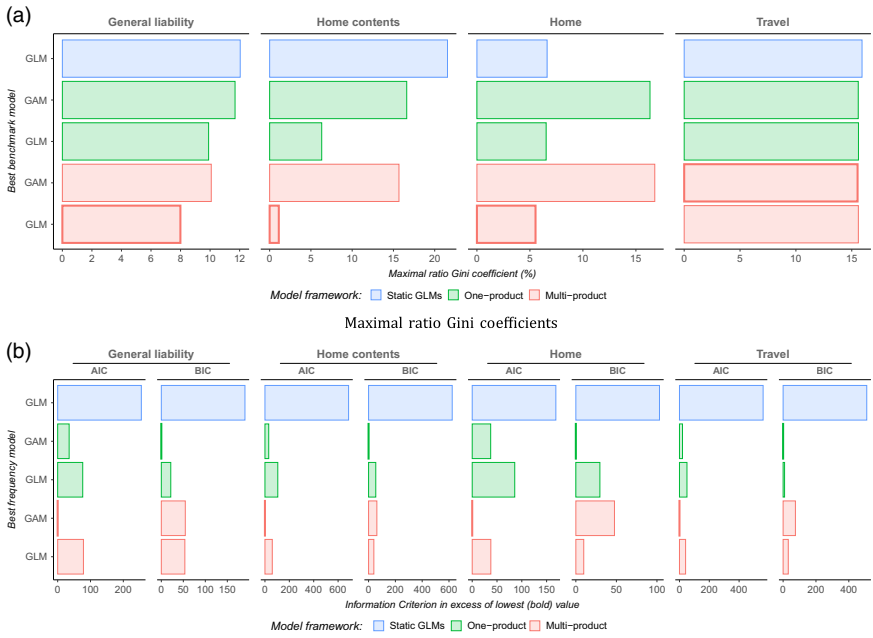

FIGURE 2: Maximal ratio Gini coefficients (panel (a)) and Akaike/Bayesian information criteria in excess of lowest (bold) value (panel (b)) for the best benchmark or frequency model in every model framework and for each product category with dynamic multivariate risk classification. Recall that a lower maximal coefficient (AIC/BIC value) implies a (statistically) less vulnerable and thus better benchmark (frequency) model, and for more details on these coefficients (values), see Table C.4 (C.5).

effects of the multi-product GLMs seem a result of misspecification, whereas the effects of the multi-product GAMs a data-driven result.

Surprisingly, the mini-max strategy of the ratio Gini coefficients largely favors the linear claim score effects over the cubic claim score effects. Figure 2(a), for instance, shows that the linear, rather than cubic, specification is generally the least vulnerable to alternative rate structures, except for travel insurance. Moreover, the claims experience in other product lines appears to be useful for risk classification for every product category. However, based on the AIC and BIC values in Figure 2(b), the cubic, rather than linear, specification systematically leads to better statistical performance. The AIC values further indicate that the claims experience in other product lines consistently improves the performance, while it seems sufficient to only account for the claims experience in the own product category based on the BIC values. Nonetheless, the mini-max strategy indicates that accounting for a customer's one- or multi-product claims experience can already lead to 3% up to 17% more profits, on average, than a static GLM. A standard likelihood ratio (LR) test additionally shows in Table C.5 that the multi-product (one-product) model significantly outperforms the one-product model (static GLM) in-sample for all non-ZIP (most) cubic and some (most) linear specifications and each product category.

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

While these conflicting in- and out-of-sample results may seem surprising at first sight, they can primarily be ascribed to two factors. Firstly, even though the linear specification is favored by the mini-max strategy due to its larger discounts and surcharges, its effects appear to be misspecified for the majority of claim score levels. The cubic specification is therefore preferred based on the AIC and BIC values, since its effects are more representative of the average claim frequencies observed. Secondly, because most policyholders do not claim (at all) and thus end up with high claim scores, there is a huge excess of zeros in the insurance portfolio, in particular when considering multiple product categories. For home contents insurance, for instance, we observe relatively few policyholders with claims in multiple product categories, and we thus also observe little variation in the claim scores on the other products. In case of travel insurance, we do observe a large pool of policyholders, but few of them actually hold multiple products. As a result, it may be statistically sufficient to merely incorporate the claims experience in the own product category, even though accounting for the claims experience in other product categories always appears to improve the rate structure's profitability.

4.3. Piecewise linear simplification

While most claim scores in the multi-product GAMs lead to an intuitive and decreasing relation with respect to a customer's risk, some are less straightforward and more complicated. However, in practice, insurers must explain and justify their premia, and they thus highly prefer intuitive and interpretable premium rates. As a consequence, it makes more sense from a practical perspective to consider a rate structure segmented into piecewise linear components using linear, rather than cubic, splines. We therefore implement this multi-product piecewise linear GAM for Poisson, NB, and ZIP distributed claim frequencies and Gamma, IG, and Pareto distributed claim severities. Moreover, we present the piecewise linear effects on the linear predictor and the corresponding discounts/surcharges offered to the insured resulting from these claim scores in Figure 3 and show the maximal ratio Gini coefficients when including these piecewise linear specifications in Table C.6.

From the claim score effects in Figure 3, we observe approximately the same patterns and subtleties for the piecewise linear GAMs as those for the cubic GAMs. In general, we again expect customers with lower claim scores on a certain product to be associated with more risk on that same product and that customers who merely possess home contents and/or travel insurance are associated with more risk for all other product categories. In terms of cross-selling opportunities, this also means that insurers should, for instance, not target customers holding these products with cross-selling offers. Note that we only show the claim score effects for the piecewise linear GAM-PG-Multi-PL in Figure 3, since the other specifications lead to almost the same relations and that these effects are based on the optimal claim score parameters for the

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

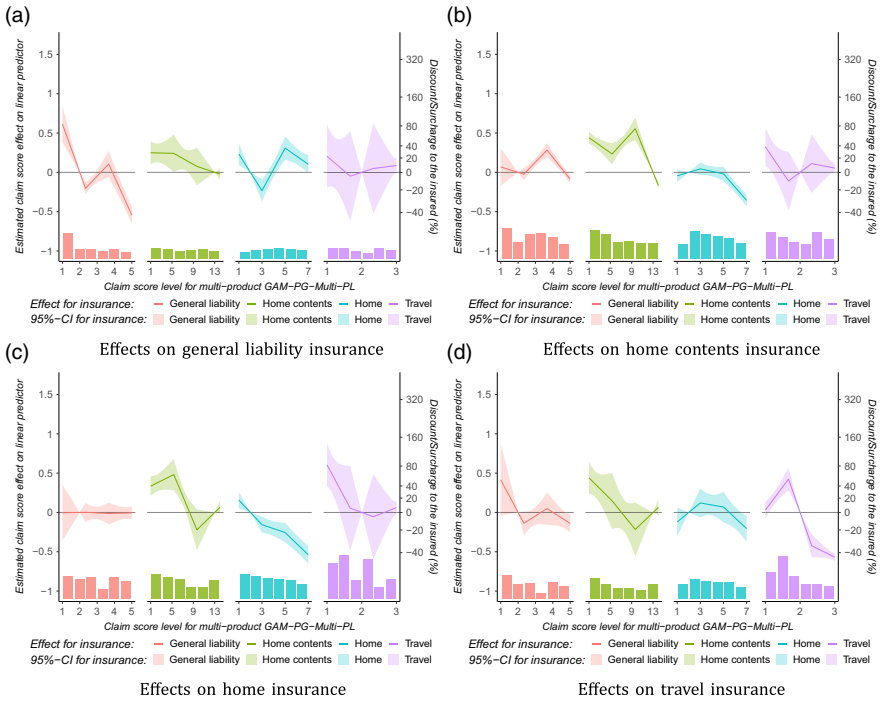

FIGURE 3: Estimated claim score effects on linear predictor and corresponding discounts/surcharges to the insured with 95% confidence intervals and histograms of average observed claim frequencies of general liability (panel (a)), home contents (panel (b)), home (panel (c)), and travel insurance (panel (d)) for multi-product piecewise linear GAM-PG-Multi-PL.

cubic GAMs. The piecewise linear specifications can, of course, be optimized separately as well, but this leads to similar results. The resulting piecewise linear splines therefore simply seem a piecewise linear simplification of the cubic splines and the non-monotonicities in the claim scores indeed a data-driven result that the linear claim scores in the multi-product GLMs are unable to capture. As such, the piecewise linear GAMs represent a more intuitive and interpretable version of the cubic GAM for insurers to adopt in practice but retain the possibility to identify the cross-selling potential of customers across property and casualty insurance.

Despite their promising potential, the piecewise linear GAMs do not consistently outperform their pure linear counterparts in our mini-max strategy. In Table C.6, the multi-product piecewise linear GAM is, for instance, the least vulnerable to alternative rate structures for general liability insurance and only marginally more vulnerable than the multi-product cubic GAM for travel insurance. However, the multi-product GLM remains the least vulnerable for home contents and home insurance, even though the piecewise linear GAM is preferred for these products based on the AIC values in Table C.2.

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

Nonetheless, these piecewise linear GAMs mostly still seem, on average, much less vulnerable to alternative model choices than the static GLMs. We additionally already found in the previous section that the multi-product model significantly outperforms the one-product model in all non-ZIP (all) cubic GAM, and some (all) GLM cases in-sample based on a straightforward LR test (AIC selection procedure). Even though these piecewise linear GAMs are not always optimal in our mini-max strategy, they therefore do seem promising to consider for practitioners in the non-life insurance industry.

The claims experience of customers thus appears to be an important determinant for individual risk classification and it can be very profitable to account for this experience in our premia. Moreover, it seems that accounting for multi-product claims experience is always optimal and becomes more effective in case of a portfolio with relatively more policyholders having claims on multiple products and with more overlap between different product categories. Regardless of these conditions, however, the multi-product piecewise linear GAM seems particularly interesting for its intuitive and interpretable use in practice, and its ability to detect the cross-selling potential of existing individual customers.

5. Conclusion

In this paper, we have presented and applied a multi-product framework for dynamic insurance pricing on the level of individual policyholders. While the industry standard of a GLM typically considers only *a priori* information of policyholders, we have included the *a posteriori* claims experience of customers across multiple product categories in a predictive claim score. As such, we have extended the BMS-panel model of Boucher and Inoussa (2014) and Boucher and Pigeon (2018) by, on the one hand, incorporating the claims experience from multiple product lines and, on the other hand, allowing the respective claim scores to have a nonlinear effect on the (logarithm of the) premium rate structure. Moreover, we have considered both a natural cubic and linear spline for the effects of these claim scores to embed our novel multi-product framework into a GAM and benefit from its existing framework.

In our application of this multi-product framework, we considered a Dutch property and casualty insurance portfolio, including general liability, home contents, home and travel insurance. Using this portfolio, we compared the industry standard of a GLM with cubic splines for the claim scores and linear effects similar to the BMS-panel model. This led to the finding that accounting for a customer's claims experience can be very profitable and substantially outperforms a static GLM based on a mini-max strategy of ratio Gini coefficients. This mini-max strategy generally also favored the linear effects more than the cubic splines in terms of profit potential, even though the linear effects appeared to be dominated by the good risks and thus misspecified for all other

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

risks. The effects from the cubic splines, on the other hand, were actually a data-driven result that yielded more representative confidence bounds in case of claim score levels with little exposure and were preferred based on an AIC and BIC selection procedure. A piecewise linear simplification of the cubic spline supported this claim by resulting in almost the same claim score effects and identifying similar cross-selling opportunities that the linear specification was unable to detect. More importantly, however, our results indicated that it is in fact always optimal or most profitable to account for a customer's claims experience from all product lines and that it becomes more effective in case of a portfolio with relatively more policyholders having claims on multiple products and with more overlap between different product categories. As such, the multi-product framework presented in this paper, and in particular the piecewise linear GAMs for their intuitive and interpretable rate structures, seem promising for practitioners in the non-life insurance industry to implement in their dynamic pricing strategies.

While the focus of this paper has primarily been on separate effects for each claim score, it is also possible to include interaction effects of all these scores. However, a more interesting avenue for future research is to consider a single multidimensional spline in the multi-product GAM for all the separate claim scores combined. This, in turn, may be able to expose complex dependencies between the claim scores of different product categories and enhance the multi-product risk profiles. Alternatively, future research can refine the piecewise linear simplification of the cubic spline using one of the binning strategies mentioned in Henckaerts *et al.* (2018) to improve the profitability of the multi-product piecewise linear GAM. Finally, since the non-monotonic splines can complicate a direct commercial application, a decreasing monotonicity restriction may lead to a more intuitively appealing and feasible framework for non-life insurers to adopt in practice, while simultaneously adjusting their effects.

Acknowledgments

The author gratefully acknowledges financial support from VIVAT. Any errors made or views expressed in this paper are the responsibility of the author alone. In addition, the author would like to thank the editor and two anonymous referees for their valuable comments on a previous version of this paper.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2020.34.

REFERENCES

ANTONIO, K. and VALDEZ, E.A. (2012) Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, **96**(2), 187–224.

BARSEGHYAN, L., MOLINARI, F., MORRIS, D.S. and TEITELBAUM, J.C. (2020) The cost of legal restrictions on experience rating. *Journal of Empirical Legal Studies*, **17**(1), 38–70.

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

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

BOUCHER, J.-P. and INOUSSA, R. (2014) A posteriori ratemaking with panel data. *ASTIN Bulletin*, **44**(3), 587–612.

BOUCHER, J.-P. and PIGEON, M. (2018) *A claim score for dynamic claim counts modeling.* Working Paper, December 2018. Available online at https://arxiv.org/abs/1812.06157.

CZADO, C., KASTENMEIER, R., BRECHMANN, E.C. and MIN, A. (2012) A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, **2012**(4), 278–305.

DENUIT, M., MARÉCHAL, X., PITREBOIS, S. and WALHIN, J.-F. (2007) *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems.* New York: Wiley.

ENGLUND, M., GUILLÉN, M., GUSTAFSSON, J., NIELSEN, L.H. and NIELSEN, J.P. (2008) Multivariate latent risk: A credibility approach. *ASTIN Bulletin*, **38**(1), 137–146.

ENGLUND, M., GUSTAFSSON, J., NIELSEN, J.P. and THURING, F. (2009) Multidimensional credibility with time effects: An application to commercial business lines. *The Journal of Risk and Insurance*, **76**(2), 443–453.

FREES, E.W., MEYERS, G. and CUMMINGS, A.D. (2011) Summarizing insurance scores using a Gini index. *Journal of the American Statistical Association*, **106**(495), 1085–1098.

FREES, E.W., MEYERS, G. and CUMMINGS, A.D. (2014) Insurance ratemaking and a Gini index. *The Journal of Risk and Insurance*, **81**(2), 335–366.

GARRIDO, J., GENEST, C. and SCHULZ, J. (2016) Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, **70**, 205–215.

GINI, C. (1912) *Variabilità e Mutabilità Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche.* Bologna: Cuppini.

HABERMAN, S. and RENSHAW, A.E. (1996) Generalized linear models and actuarial science. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **45**(4), 407–436.

HASTIE, T. and TIBSHIRANI, R. (1986) Generalized additive models. *Statistical Science*, **1**(3), 297–310.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

HENCKAERTS, R., ANTONIO, K., CLIJSTERS, M. and VERBELEN, R. (2018) A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, **2018**(8), 681–705.

HENCKAERTS, R., CÔTÉ, M.-P., ANTONIO, K. and VERBELEN, R. (2020) Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 1–31.

KAAS, R., GOOVAERTS, M., DHAENE, J. and DENUIT, M. (2008) *Modern Actuarial Risk Theory: Using R.* Berlin, Heidelberg: Springer.

LEMAIRE, J. (1998) Bonus-malus systems: The European and Asian approach to merit-rating. *North American Actuarial Journal*, **2**(1), 26–38.

LORENZ, M.O. (1905) Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, **9**(70), 209–219.

NELDER, J.A. and WEDDERBURN, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384.

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

OHLSSON, E. and JOHANSSON, B. (2010) *Non-Life Insurance Pricing with Generalized Linear Models*. Berlin, Heidelberg: Springer.
PECHON, F., TRUFIN, J. and DENUIT, M. (2018) Multivariate modelling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin*, **48**(3), 969–993.
PINQUET, J. (1997) Allowance for cost of claims in bonus-malus systems. *ASTIN Bulletin*, **27**(1), 33–57.
QUAN, Z. and VALDEZ, E.A. (2018) Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, **6**(1), 377–407.
SHI, P. and VALDEZ, E.A. (2014) Multivariate Negative Binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, **55**(1), 18–29.
TZOUGAS, G., VRONTOS, S. and FRANGOS, N. (2014) Optimal bonus-malus systems using finite mixture models. *ASTIN Bulletin*, **44**(2), 417–444.
TZOUGAS, G., VRONTOS, S. and FRANGOS, N. (2018) Bonus-Malus Systems with two-component mixture models arising from different parametric families. *North American Actuarial Journal*, **22**(1), 55–91.
WOOD, S.N. (2006) *Generalized Additive Models: An Introduction with R*. New York: Chapman and Hall/CRC.
YEE, T.W. and HASTIE, T.J. (2003) Reduced-rank vector generalized linear models. *Statistical Modeling*, **3**(1), 15–41.

ROBERT MATTHIJS VERSCHUREN
Amsterdam School of Economics
University of Amsterdam
Roetersstraat 11, 1018 WB, Amsterdam
The Netherlands
E-Mail: r.m.verschuren@uva.nl

APPENDIX A. RISK FACTORS FOR PROPERTY AND CASUALTY INSURANCE

TABLE A.1

DESCRIPTION OF THE KEY VARIABLES USED FOR PROPERTY AND CASUALTY INSURANCE.

Variable	Values	Description
`Count`	Integer	The number of claims filed by the policyholder
`Size`	Continuous	The size of the claim in euros
`Exposure`	Continuous	The exposure to risk in years

TABLE A.2

DESCRIPTION OF THE RISK FACTORS USED FOR GENERAL LIABILITY INSURANCE.

Risk factor	Values	Description
`FamilySituation`	4 categories	Type of family situation

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

TABLE A.3

DESCRIPTION OF THE RISK FACTORS USED FOR HOME CONTENTS INSURANCE.

Risk factor	Values	Description
`ProductType`	3 categories	Type of product
`Year`	6 categories	Calendar year
`Age`	Continuous	Age of the policyholder in years
`BuildingType`	7 categories	Type of building of the home
`RoofType`	2 categories	Type of roof of the home
`FloorSpace`	Continuous	Total floor area of the home in thousands of square meters
`HomeOwner`	3 categories	Whether the policyholder owns or rents its home
`Residence`	5 categories	Residential area of the policyholder
`Urban`	8 categories	Degree of urbanization at home address
`GlassCoverage`	2 categories	Whether the policyholder has glass coverage

TABLE A.4

DESCRIPTION OF THE RISK FACTORS USED FOR HOME INSURANCE.

Risk factor	Values	Description
`ProductType`	3 categories	Type of product
`Year`	6 categories	Calendar year
`Age`	Continuous	Age of the policyholder in years
`FamilySituation`	5 categories	Type of family situation
`BuildingType`	6 categories	Type of building of the home
`RoofType`	2 categories	Type of roof of the home
`Capacity`	Continuous	Total capacity of the home in thousands of cubic meters
`ConstructionYear`	6 categories	Construction year of the home
`Residence`	5 categories	Residential area of the policyholder
`Urban`	8 categories	Degree of urbanization at home address
`GlassCoverage`	2 categories	Whether the policyholder has glass coverage

TABLE A.5

DESCRIPTION OF THE RISK FACTORS USED FOR TRAVEL INSURANCE.

Risk factor	Values	Description
`Region`	3 categories	Regional area covered
`Age`	Continuous	Age of the policyholder in years
`FamilySituation`	5 categories	Type of family situation
`WinterCoverage`	2 categories	Whether the policyholder has winter sport coverage
`MoneyCoverage`	2 categories	Whether the policyholder has money coverage
`VehicleCoverage`	2 categories	Whether the policyholder has vehicle coverage
`MedicalCoverage`	2 categories	Whether the policyholder has medical coverage
`AccidentCoverage`	2 categories	Whether the policyholder has accident coverage
`CancelCoverage`	2 categories	Whether the policyholder has cancellation coverage

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

APPENDIX B. ESTIMATION IN MULTI-PRODUCT CLAIM SCORE MODEL

Essential to the multi-product claim score model developed in this paper is the assumption that the response variable is independently distributed according to some member of the exponential family. If we denote by $Y_{i,t}$ this response for subject i in period t, then its density $p(\cdot)$ can be written as:

$$p(y_{i,t}|\vartheta_{i,t},\varphi) = h(y_{i,t},w_{i,t},\varphi)\exp\left(\frac{w_{i,t}}{\varphi}\big(\vartheta_{i,t}y_{i,t} - A(\vartheta_{i,t})\big)\right) \quad \text{for} \quad i=1,\ldots,M,\ t=1,\ldots,T_i,$$

where $\vartheta_{i,t}$ denotes a distribution parameter, φ a dispersion parameter, $w_{i,t}$ a known weight that is typically set to one or the exposure to risk, $h(\cdot,\cdot,\cdot)$ a known function, and $A(\cdot)$ a known twice continuously differentiable function. The function $A(\cdot)$ is related to the mean $\mu_{i,t}$ and covariance $\Sigma_{i,t}$ of $Y_{i,t}$ through:

$$\mu = \mathbb{E}[Y] = A'(\vartheta) \quad \text{and} \quad \Sigma = \mathbb{V}[Y] = \frac{\varphi}{w}v(\mu),$$

where $v(\mu) = A''(\vartheta) = A''(A'^{-1}(\mu))$ is called the variance function (Ohlsson and Johansson, 2010). As such, it is sufficient to only consider a model for the mean since this can already completely characterize the entire distribution of the response variable.

When considering the multi-product claim score model for the mean equation, inference can be performed straightforwardly by penalized maximum likelihood. Suppose the linear predictor in this model is given by Equation (2.1) with penalized cubic regression splines for the transformations $f_j^{(c)}(\cdot)$. If we consider $(m+1)$-th order B-splines with k parameters and $k+m+1$ knots in a certain interval $[x_1, x_{k+m+1}]$ for the basis of some set of regression splines, then we can represent them by:

$$f_j(x) = \sum_{h=1}^{k}\gamma_h B_h^m(x) \quad \text{for} \quad j=1,\ldots,C,$$

with $m=2$ for cubic splines and $m=0$ for linear splines. The B-spline basis functions $B_h^m(\cdot)$ in this representation are defined recursively as:

$$B_h^m(x) = \frac{x-x_h}{x_{h+m+1}-x_h}B_h^{m-1}(x) + \frac{x_{h+m+2}-x}{x_{h+m+2}-x_{h+1}}B_{h+1}^{m-1}(x) \quad \text{for} \quad h=1,\ldots,k,$$

with initial value $B_h^{-1}(x) = \mathbb{1}(x_h \le x < x_{h+1})$ and where we have omitted the superscripts (c) for the sake of simplicity (Wood, 2006). The knots themselves are typically located at the quantiles of the observations or spaced evenly over the range of the observations for cubic or linear splines, respectively (see, e.g., Wood, 2006). The smooths $f_j(\cdot)$ are usually subject to an additional centering constraint to ensure identification of the mean equation and it is commonly assumed that all its elements sum to zero. As a result, one degree of freedom in the splines is lost due to this identification restriction and $k-1$ effectively remain. The penalized log-likelihood function is now defined as:

$$\begin{aligned}\ell_p(\delta,\varphi,\lambda|y) &= \ell(\delta,\varphi|y) - \frac{1}{2}\sum_{j=1}^{C}\lambda_j\int f_j''(x)^2\mathrm{d}x \\ &= \sum_{i=1}^{M}\sum_{t=1}^{T_i}\left[\log\big(h(y_{i,t},w_{i,t},\varphi)\big) + \frac{w_{i,t}}{\varphi}\big(\vartheta_{i,t}y_{i,t} - A(\vartheta_{i,t})\big)\right] - \frac{1}{2}\sum_{j=1}^{C}\lambda_j\delta' S_j\delta, \quad \text{(B1)}\end{aligned}$$

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press

with $\delta = (\beta, \gamma)$ and where the distribution parameters $\vartheta_{i,t}$ depend on the parameters δ through the linear predictor, λ_j denotes the penalty or smoothing parameter for the j-th regression spline, and S_j a matrix of known coefficients $\tilde{S}_j$ padded with zeros such that $\delta' S_j \delta = \gamma' \tilde{S}_j \gamma$. Note that the first expression in Equation (B1), or $\ell(\cdot)$, actually represents the ordinary log-likelihood function of the model and that the multi-product claim score model can therefore be seen as a penalized GLM in terms of optimization. Maximization of this penalized log-likelihood in terms of the parameters δ given the penalties λ_j leads to the set of $K + \sum_{j=1}^{C} k_j$ normal equations given by:

$$\frac{1}{\varphi} \sum_{i=1}^{M} \sum_{t=1}^{T_i} w_{i,t} \frac{y_{i,t} - \mu_{i,t}}{v(\mu_{i,t}) g'(\mu_{i,t})} X_{i,t} - \sum_{j=1}^{C} \lambda_j S_j \delta = 0,$$

where K denotes the dimension of β and φ is usually omitted since we can incorporate its effect into the penalties. In practice, the smoothing parameters are of course unknown as well and are usually estimated by generalized cross-validation or unbiased risk estimation (see, e.g., Wood, 2006).

It is clear that these normal equations do not lead to an analytical solution for our unknown parameters and that we need to find a numerical solution to them. One way to numerically solve these equations is by the Newton–Raphson method that relies on the gradient of the normal equations with respect to δ, or the Hessian matrix of the (penalized) log-likelihood function. However, a more popular approach for numerically solving these equations in the context of GLMs is called the Fisher scoring method. This method applies the same iterative procedure as the Newton–Raphson method, but now uses the Fisher information matrix $\mathcal{I}(\cdot)$, rather than the Hessian matrix. The Fisher scoring method is therefore characterized by:

$$\delta^{(n+1)} = \delta^{(n)} + \mathcal{I}^{-1}(\delta^{(n)}) J(\delta^{(n)}),$$

with $J(\cdot)$ the Jacobian matrix of the (penalized) log-likelihood function, or the normal equations. Formally, this information matrix is given by the expectation of the negative Hessian matrix, or:

$$\mathcal{I}(\delta) = \mathbb{E}[-H(\delta)] = \frac{1}{\varphi} \sum_{i=1}^{M} \sum_{t=1}^{T_i} \frac{w_{i,t}}{v(\mu_{i,t}) g'(\mu_{i,t})^2} X_{i,t} X_{i,t}' - \sum_{j=1}^{C} \lambda_j S_j,$$

where φ is typically omitted again. The advantages of using this matrix are that it is slightly easier to implement in practice and, by definition, always remains positive definite (Ohlsson and Johansson, 2010). The Hessian matrix, on the other hand, is not necessarily positive definite unless we are already close to convergence. The Fisher scoring method therefore typically leads to more stable convergence than the Newton–Raphson method, whereas the latter method is considered faster. In the context of (penalized) GLMs, Fisher's iterative procedure is also known as PIRLS and can easily be implemented in, for instance, `R` with the package `mgcv` developed by Wood (2006). As such, the multi-product claim score model can heavily benefit from the framework of GLMs and GAMs and rely on existing statistical theory and software for inference.

https://doi.org/10.1017/asb.2020.34 Published online by Cambridge University Press