Multidimensional credibility: a Bayesian analysis of policyholders holding multiple contracts
Antonio, K.; Guillén, M.; Pérez Marín, A.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Multidimensional credibility: a Bayesian analysis of policyholders holding multiple contracts

Katrien Antonio* † Montserrat Guillén
Ana–Maria Pérez Marín ‡ §

May 19, 2011

Abstract

Property and casualty actuaries are professional experts in the economic assessment of uncertain events related to non–life insurance products (e.g. fire, liability or motor insurance). For the construction of a fair and reasonable tariff associated with the risks in their portfolio, actuaries have many statistical techniques in their toolbox. In this paper tools for the pricing of multivariate risks are considered. Examples of situations where this problem occurs are numerous; e.g. workers’ compensation schemes where the insurer has information on accidents occurring ‘at work’ as well as ‘not at work’, policyholders having policies in multiple lines of business (e.g. flood, theft etc.) at the same company, or policyholders holding multiple contracts e.g. in a motor insurance context. The latter is the situation we will explore in this paper, using a data set from a European insurance company. The combination of a priori rating (through risk classification based on a

* k.antonio@uva.nl, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam (The Netherlands) and Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven (Belgium).
† Katrien Antonio acknowledges financial support from the The Actuarial Foundation and from NWO through a Veni 2009 grant.
‡ {mguillen, amperez}@ub.edu, University of Barcelona, Diagonal 690, Barcelona, Spain.
§ Ana–Maria Pérez Marín and Montserrat Guillén acknowledge their grant FEDER ECO2010-21787 from the Spanish Ministry of Science.
priori measurable characteristics) and a posteriori rating is considered. A posteriori
the claim experience of a policyholder is taken into account. In a multivariate con-
text the pricing actuary should be able to use claim history from all business lines
or risk components a policyholder is holding within the company. Intuitively, the
number of claims a policyholder has reported on a particular business line or risk
component reveals his general risk proneness and is as such relevant when pric-
ing other lines or components. In contrast to the analytical approaches developed
for multivariate experience rating in the literature, our approach is data–driven
using Bayesian statistics. Empirical evidence has been found of positive depen-
dence between different contracts belonging to the same policy holder, which is an
intuitively appealing result. Focus is on applications of the multivariate Bayesian
experience rating, as well as on graphical representations of the results.

1 Introduction

Pricing is an important activity for property and casualty actuaries. Not all risks are the
same; heterogeneity is omnipresent in the portfolios of general insurance companies.
Pricing risks using specific risk characteristics has a long history in actuarial science. It
is important for the insurer to optimally group the risks in the portfolio so that those
insureds with a similar risk profile pay the same reasonable premium rate. Such is the
idea behind risk classification within the pricing process. A risk classification system
should not only allow insurers to price discriminate their products in a fair and equi-
table manner, but should also be constructed on a sound statistical basis. An overview
of statistical techniques for risk classification is given in Antonio and Valdez (2010).

In this paper an automobile insurance portfolio is considered. Typically, a group
of $N$ policyholders (denoted with $c$ where $c = 1, \ldots, N$) is followed over time $t = 1, \ldots, T_c$. General characteristics of the policyholder (e.g. age, gender) and the insured
vehicle (e.g. age, type of use, make) are available. For each policyholder and each time
period the number of claims and associated claim costs are registered, as is the fraction
of the year that insurance coverage is guaranteed and the insured is exposed to risk (the so-called exposure period). Let $N_{ct}$ be the number of claims from policyholder $c$ in time period $t$, $C_{ctj}$ the cost of the $j$th claim from client $c$ in period $t$ and $E_{ct}$ the exposure of client $c$ in time period $t$. The total cost of claims (or aggregate loss) from client $c$ in period $t$ (denoted with $L_{ct}$) is given by:

$$L_{ct} = \sum_{j=1}^{N_{ct}} C_{ctj}. \quad (1)$$

In order to price this risk, the frequency and severity data are combined into a pure premium $P_{ct}$:

$$P_{ct} = \frac{L_{ct}}{E_{ct}} = \frac{N_{ct}}{E_{ct}} \times \frac{L_{ct}}{N_{ct}} = F_{ct} \times S_{ct}, \quad (2)$$

with $F_{ct}$ the claim frequency per unit of exposure and $S_{ct}$ the so-called severity. Depending on the format of the available data, regression models for $F_{ct}$ and the individual losses $\{C_{ct1}, \ldots, C_{ctN_{ct}}\}$ or $F_{ct}$ and the severity $S_{ct}$ will be specified. Ultimately, the risk $ct$ is priced by applying a premium principle $\pi(.)$ to the random variable $P_{ct}$. In this paper, we essentially focus on the expected value principle which leads us to the net premium:

$$\pi(P_{ct}) = E[P_{ct}] = E[F_{ct}] \times E[S_{ct}], \quad (3)$$

under the (traditional) assumption of independence between claim frequency and severity. Typically, insurers add a ‘risk loading’ to this net premium, to cover items such as administrative expenses, profits, margins for contingencies. If a ‘risk loading’ is added to the net premium, the term ‘gross premium’ is used. Other premium calculation principles may be used, but these are beyond the scope of this paper.

Experience rating is the act of including observed experience in the rating process.
Using experience rating we update the *a priori* premium to an *a posteriori* premium. In an *a priori* rating system only *a priori* correctly measurable and observable rating factors are used. *A posteriori* we also use the reported claim history, since this reveals information about the policyholder which is *a priori* unknown (like aggressiveness, swiftness of reflexes). A standard approach in actuarial practice and literature stipulates that only frequency information is used in an experience rating system. Pinquet (1997) is an exception and applies experience rating to both the frequency and the cost of claims. Following the actuarial literature on experience rating the response variable of interest in this paper is the number of claims. Models for the frequency should be combined with an estimate for the average claim size in order to obtain an estimate for the pure premium (2). For a detailed overview of actuarial models for claim counts (including experience rating), see Denuit et al. (2007). A particular characteristic of our data sample is that the number of claims does not represent ‘number of claimed accidents’ but the ‘number of claimed guarantees’. Indeed, a single accident may affect several guarantees, depending on the insured’s type of coverage. Of course, the same methodology can readily be applied to samples representing ‘number of accidents’ instead of ‘number of affected guarantees’.

Credibility theory is a cornerstone in actuarial mathematics (Hickman and Heacox (1999)) and studies theoretical models for experience rating with latent risk variables. Bühlmann and Gisler (2005) is a recent overview of relevant issues in credibility theory. In the present contribution our interest goes to the concept of multidimensional or multivariate credibility (see Chapter 7 in Bühlmann and Gisler (2005)) which dates back to Jewell (1974) and to Pinquet (1998). This type of experience rating system is relevant when actuaries want to study multivariate risks: for instance, policyholders holding multiple contracts from different business lines (theft, flood damage, …), accidents reported ‘at work’ versus ‘not at work’ in workers’ compensation insurance, or ‘normal claims’ in combination with ‘big claims’ in a specific business line. A multivariate credibility system allows the pricing actuary to use claim history from all available business
lines or risk components when determining the premium for a new business line or a future time period. Dependence between claim behavior in different branches is studied through the introduction of a multivariate latent risk factor, with one component for each line of business or risk component. The amount of dependence will be determined by the distributional assumptions for the response and latent risk factors and the parameter estimates obtained for the multivariate distribution. Intuitively the idea behind a multidimensional credibility model is that the number of claims a policyholder has reported on a particular business line, say $A$, reveals his general risk proneness and is as such relevant when pricing another business line, say $B$. For instance, in the workers’ compensation insurance example, the idea behind a multivariate experience rating system is that people who are ‘dangerous’ in their professional life tend also to take more risks in their free time. Key observation is that – with a multidimensional credibility model – we simultaneously consider observations from different risk components. As such, the data will reveal what can be learned from one risk category with respect to another.

In this paper a data base is studied from a European insurer where policyholders may have just one contract with the company, two contracts or up to 5 contracts. Very often the reason for having many contracts or policies is that households or firms tend to assign the role of policyholder to one person. Multidimensional credibility ideas are suitable in this context. Indeed, the total claim history over all contracts reveals how the household or firm approaches concepts like safety on the road, mechanical check ups on a regular basis, speeding etc. Parameter estimates for the multivariate latent risk distribution will reveal how strong the dependence between different contracts is in this particular example.

Pinquet (1997) and Pinquet (1998) are examples of multidimensional credibility systems. Bühlmann et al. (2003), Englund et al. (2008) and Englund et al. (2009) are more recent contributions of rating systems incorporating data from different business lines on the same policyholder. Our contribution differs from these papers not only by the
kind of application and data sample that is used, but also by the statistical techniques that are used. Our approach is fully Bayesian, whereas research on experience rating has put focus on the theoretical derivation of best linear unbiased credibility premiums, combined with moment estimators for the unknown parameters (as is used in the above mentioned papers). For a hierarchically structured portfolio the merits of a Bayesian experience rating system have been demonstrated in Antonio et al. (2010).

The Bayesian framework has many advantages. Within this general framework standard estimation machinery is available. Nowadays, the implementation of Bayesian models is becoming easier thanks to software packages such as R, WinBugs, OpenBugs and JAGS. A Bayesian approach allows a data driven approach to the problem of multidimensional credibility. Graphical tools are an elegant way to communicate the results obtained from it, as this paper illustrates.

The rest of the paper has been organized as follows. In Section 2 the data set is introduced. Section 3 discusses the stochastic model that we use for experience rating with the multivariate claim counts reported by each policyholder. Model estimation results are in Section 4. Section 5 presents several applications of the Bayesian multivariate credibility model, with an emphasis on graphs representing the experience rating mechanisms. Section 6 concludes.

2 The data

Customers of insurance companies include individuals and firms. In this application we will focus on policy holders having one or more policies underwritten in a European insurance company. We study 48,631 policy holders that have purchased automobile insurance coverage. Data were collected in March 2008 and they include information on customers that had at least one effective policy in December 31, 2005. The data have been registered during 2006 and 2007. During this time, customers may have changed in the sense that they may have canceled (some or all of their) policies or they may have underwritten new ones. New customers entering the company in the
observation period have not been included in the sample. The most important reason for this is that the data have been collected to study customer loyalty and the company did not want to mix customer loyalty with the effects of campaigns attracting new customers. This allows to study the behavior of members of the portfolio with a minimum seniority. However, we are aware that this may bias the sample as new clients are not included.

Automobile insurance covers at least third party liability, but it may also include some extended coverage for bodily injury of the driver. Customers in the sample may have just one contract, 2 contracts or up to 5 contracts. Table 1 gives detailed information on the sample size, the number of policies each policyholder holds and the number of claims reported. From Table 1 one immediately determines that our data sample is characterized by a high average number of claims (per unit of exposure). In light of this observation it is important to note that the data set did not allow to make a distinction between claims at fault and all claims. Therefore, the sample deals with all claims. Besides that, and more important, in this particular data sample an accident does not equal a claim. A single accident may imply several claims, depending on the insured’s type of coverage. According to the company’s definition they count the number of guarantees that were affected. As such, we do not model the ‘number of accidents’ but the ‘number of affected guarantees’. This explains the high claim averages. We see that the average claim frequency gradually increases with the number of policies. A possible explanation for this phenomenon is that policyholders with more contracts more frequently use their cars or motorbikes (for instance for delivery services).

The data in Table 1 should be understood as follows. 34,150 policyholders in the data sample hold exactly one contract at the beginning of the study period. These contracts are exposed to risk during a fraction (between 0 and 1) of the calendar year 2006 and a fraction of the calendar year 2007. This leads to a data base with 68,297 observations, where each observation represents a particular contract in a particular year (2006 or 2007). In total, over all contracts and all years, 46,134 claims have been registered.
Table 1: Sample information: distribution of claim counts for policyholders with 1, . . . , 5 policies.

<table>
<thead>
<tr>
<th>Number of claims</th>
<th>All</th>
<th>1 (68,297 obs.)</th>
<th>2 (39,817 obs.)</th>
<th>3 (14,541 obs.)</th>
<th>4 (4,780 obs.)</th>
<th>5 (3,389 obs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>69.23</td>
<td>69.78</td>
<td>69.65</td>
<td>68.43</td>
<td>66.86</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>14.41</td>
<td>14.31</td>
<td>13.95</td>
<td>14.27</td>
<td>14.58</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7.67</td>
<td>7.44</td>
<td>7.3</td>
<td>8.1</td>
<td>7.35</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.7</td>
<td>3.57</td>
<td>3.6</td>
<td>3.81</td>
<td>4.25</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.98</td>
<td>1.89</td>
<td>2.09</td>
<td>2.2</td>
<td>2.36</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.21</td>
<td>1.16</td>
<td>1.33</td>
<td>1.3</td>
<td>1.56</td>
<td></td>
</tr>
<tr>
<td>&gt; 5</td>
<td>1.8</td>
<td>1.84</td>
<td>2.77</td>
<td>1.9</td>
<td>3.04</td>
<td></td>
</tr>
</tbody>
</table>

# Policy holders | 48,631 | 34,150 | 10,723 | 2,687 | 678 | 393 |
Total Claims     | 89,503 | 46,134 | 26,819 | 10,309 | 3,437 | 2,804 |
Total Exposure (in policy years) | 125,983 | 67,951 | 37,468 | 13,281 | 4,293 | 2,990 |
Mean Claim (Tot. Cl. / Tot. Exp) | 0.71 | 0.68 | 0.72 | 0.78 | 0.80 | 0.94 |

for a total period of exposure of 67,951 years. On average this is 0.68 claims per year of exposure. 10,723 policyholders in the sample hold exactly two contracts. Each of these contracts is exposed to risk during a fraction of calendar year 2006 and 2007, leading to a data base with 39,817 observations (where each policy holder is represented by up to 4 observations, one per contract and per year). For these policyholders, 26,819 claims have been registered for 37,468 years of exposure, which is on average 0.72 claims per year of exposure. Other columns in the Table have a similar interpretation. Histograms and boxplots of the claim distributions corresponding with policyholders having 1 up to 5 contracts are given in Figure 1.

For every policy in the sample, we have information regarding the policyholder (age, gender, address et cetera), the policies he has in force from December 31, 2005 till December 31 2007 (i.e. the issuing date, type of coverage, type of risk, conditions, cancellation date if this occurred). For each policy the claiming behavior registered between December 31st 2005 and December 31 2007 is available. If a policy has been valid only for an interval shorter than those 24 months, we will account for the expo-
Figure 1: Distribution of claim counts for policyholders with 1 up to 5 contracts.

sure time when analyzing the claims. The information on the client is unique for each customer. Table 2 is an overview of the available covariate information which can be used as risk factors when calculating premiums. Due to the limited information we got from the company, we were not able to distinguish different types of coverage. The continuous covariates are calculated with respect to 2005, e.g. a car manufactured in 2005 has vehicle age 0. For the categorical covariates we report the percentage of contracts for which the policyholder is male, lives in a rural area, for which the insured risk is a motorbike, and so on. The corresponding average claim frequency for each category is obtained by dividing the total number of claims reported by the total period of exposure.
Table 2: Covariate information

<table>
<thead>
<tr>
<th>Categorical Covariates</th>
<th>Description</th>
<th>Percentage of contracts</th>
<th>Average Claim Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>75.26</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>23.29</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>1.45</td>
<td>0.85</td>
</tr>
<tr>
<td>Zone</td>
<td>Rural</td>
<td>60.09</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Urban</td>
<td>29.88</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>10.03</td>
<td>0.67</td>
</tr>
<tr>
<td>Marital status</td>
<td>Married</td>
<td>75.16</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Single</td>
<td>19.23</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Unknown, divorced, widow</td>
<td>5.61</td>
<td>0.82</td>
</tr>
<tr>
<td>Type of risk</td>
<td>Motorbike</td>
<td>13.74</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Tourism</td>
<td>78.33</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Van or 4x4</td>
<td>7.93</td>
<td>0.73</td>
</tr>
<tr>
<td>Type of use</td>
<td>Private</td>
<td>99.29</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>0.71</td>
<td>1.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Continuous Covariates</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age of policyholder</td>
<td>18</td>
<td>45</td>
<td>104</td>
</tr>
<tr>
<td>Vehicle age</td>
<td>0</td>
<td>6</td>
<td>78</td>
</tr>
<tr>
<td>Driving license age</td>
<td>0</td>
<td>23</td>
<td>71</td>
</tr>
</tbody>
</table>

3 Model specification

3.1 A random effects Poisson model

A Poisson random effects model is used for experience rating with respect to this portfolio. Our specification is

\[ Y_{c,i,t} \sim \text{POI}(\lambda_{c,i,t}) \text{ with } \lambda_{c,i,t} = e_{c,i,t} \exp (\eta_{c,i,t} + b_{c,i}). \]  

(4)

Here, \( Y_{c,i,t} \) denotes the claims observed in year \( t \) for policy number \( i, i = 1, \ldots, 5 \) on policyholder \( c \). The term \( e_{c,i,t} \) is an exposure variable that gives the length of time during calendar year \( t \) for which the vehicle has insurance coverage. Risk factors are incorporated in terms of explanatory variables. The systematic component \( \eta_{c,i,t} \) is specified.
as follows:

\[ \eta_{c,i,t} = \gamma + \beta_F \times \mathbb{1}_F + \beta_R \times \mathbb{1}_R + \beta_M \times \mathbb{1}_M + \beta_{A1} \times \mathbb{1}_{A1} + \beta_{A2} \times \mathbb{1}_{A2} + \beta_{VA1} \times \mathbb{1}_{VA1}. \] (5)

Hereby the \( \mathbb{1} \)'s represent indicator variables:

- \( \mathbb{1}_F \): 1 if policyholder \( c \) is female, 0 otherwise;
- \( \mathbb{1}_R \): 1 when \( c \) lives in a rural area, 0 otherwise;
- \( \mathbb{1}_M \): 1 when the insured vehicle is a motorbike, 0 otherwise;
- \( \mathbb{1}_{A1} \): 1 when age of \( c \) is \(< 25 \) years, 0 otherwise;
- \( \mathbb{1}_{A2} \): 1 when age of \( c \) is \( \geq 25 \) and \(< 30 \) years, 0 otherwise;
- \( \mathbb{1}_{VA1} \): 1 when vehicle age is \(< 3 \) years, 0 otherwise.

In model specification (4) \( b_{c,i} \) \((i = 1, \ldots, 5)\) is a random effect representing the latent risk inherent to policy \( i \) from policyholder \( c \). Heterogeneity between policyholders is present, since not every relevant risk factor can be measured or observed. The inclusion of random effects allows to take this into account. Moreover, intuitively we would expect observations from different policies belonging to the same policyholder to be dependent. This can be explained by the presence of unobservable household or policyholder effects: e.g. risk averse parents will stimulate their children to drive safely and will pay attention to careful maintenance of the vehicle, et cetera.

Our distributional assumption for the random effects vector \( b_c = (b_{c1}, \ldots, b_{cn_c})' \) is:

\[
b_c \sim N(\mu, \Sigma) = N \begin{pmatrix} -\frac{\sigma_1^2}{2} \ 
\vdots \ 
-\frac{\sigma_n^2}{2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n_c} \\
\sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n_c} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{n^2} 
\end{pmatrix}
\] (6)
Hereby $n_c$ is the number of policies registered on policyholder $c$. As such, $n_c$ is the dimension of the random effects vector. $\Sigma$ is a covariance matrix and therefore symmetric and positive semidefinite. Random effects representing different policyholders are independent.

The specification of the mean vector in (6) is different from what is traditionally used in statistical literature. This specification is often used in actuarial literature because means are unchanged with the introduction of random effects. For example, $E[\exp(b_{c,i})] = \exp(-\frac{\sigma^2}{2} + \frac{\sigma^2}{2}) = 1$. Using the specifications in (4) and (6), the a priori mean, $E[Y_{c,i,t}]$, is therefore given by

$$E[Y_{c,i,t}] := \lambda_{c,i,t}^{\text{prior}} = \epsilon_{c,i,t} \exp(\eta_{c,i,t}).$$

(7)

The a posteriori premium, $E[Y_{c,i,t}|b_{c,i}]$, then becomes

$$E[Y_{c,i,t}|b_{c,i}] = \lambda_{c,i,t}^{\text{prior}} \exp(b_{c,i}).$$

(8)

This specification shows explicitly how an a posteriori correction is made to the a priori premium through the inclusion of the random intercept $b_{c,i}$.

### 3.2 Bayesian analysis

We refer to Antonio and Beirlant (2007) for details on statistical techniques to estimate – on the one hand – the unknown parameters in (5) and (6) and to predict – on the other hand– the random vectors $b_c$. For instance, SAS Proc Nlmixed is readily available for the model specified in (4)–(6). However, the huge sample size (130,824 observations in total) in combination with the high dimensional vector of random effects (up to a 5–variate vector) causes very long computation times in Proc Nlmixed. The use of Bayesian statistics has several advantages for the data set at hand and the model specification. The MCMC simulations converge fast, allowing computations to be done in a reasonable time frame (in contrast to maximum likelihood procedures such as Proc
Multidimensional Credibility – May 2011

Nlmixed. Simulations from the posterior distributions of e.g. (7), (8) and the random effects are readily available and so are simulations from the predictive distribution of future observations. In light of the applications of the multivariate credibility model that will be considered later on, this is a very interesting feature.

The prior distributions used in the Bayesian analysis are selected as follows. For the regression parameters $\gamma$, $\beta_F$, $\beta_R$, $\beta_M$, $\beta_{A1}$, $\beta_{A2}$ and $\beta_{VA1}$ normal priors are used with variance $10^6$. For the the variance components in (6) gamma priors $\Gamma(1, 0.001)$ are specified and for the covariance components $\sigma_{i_1i_2}$ uniform prior on $[-5, 5]$ are used.

The posterior distribution of $b_c$ (given the data) is used to predict the random effects. Due to the structure of the covariance matrix in (6), $b_{c,i}$ will not only depend on history of claims registered for policy $i$ on policyholder $c$ (i.e. $Y_{c,i,t}$), but also on the claims registered for other policies on the same policyholder. This is exactly what we are aiming for with a multivariate credibility model.

4 Model estimation results

Figures 2 and 3 illustrate the mixing and convergence of the chains generated with MCMC sampling. Density estimates and 95% interval estimates of the parameters are given as well.

From the credibility intervals in Figure 2 we conclude that female drivers on average report more claims than male drivers, and so do young drivers and policyholders with new vehicles. Policyholders living in rural areas and policyholders with motorbikes on average report less claims. Estimating the covariance matrix $\Sigma$ using the
Figure 2: Convergence diagnostics for the MCMC updates concerning the parameters $\gamma$, $\beta_F$, $\beta_R$, $\beta_M$, $\beta_{A1−2}$, $\beta_{VA1}$. The figure shows a trace-plot, density, BGR convergence diagnostics and 95% credibility intervals. 2 chains were run, with 4,000 updates for each chain.

The posterior mean of each variance/covariance parameter leads to

$$
\hat{\Sigma} = \begin{pmatrix}
1.54 & 0.39 & 0.36 & 0.44 & 0.42 \\
0.39 & 1.49 & 0.30 & 0.48 & 0.4 \\
0.36 & 0.30 & 1.44 & 0.4 & 0.3 \\
0.44 & 0.48 & 0.4 & 1.59 & 0.27 \\
0.42 & 0.4 & 0.3 & 0.27 & 1.57
\end{pmatrix}.
$$

This is a symmetric and positive definite matrix, and so are the sub-matrices $\hat{\Sigma}_{2\times2}$, $\hat{\Sigma}_{3\times3}$ and $\hat{\Sigma}_{4\times4}$, consisting of the first 2, respectively 3 and 4, rows and columns of $\hat{\Sigma}$. Those sub-matrices are necessary to model policyholders with 2, 3 or 4 policies. We see that random effects corresponding with different policies held by the same policyholder, are positively correlated. This confirms the intuition that policyholders...
who are risky with respect to a certain policy, also tend to be risky with respect to the other policies they may have.

5 A priori premiums and a posteriori updates

In this section applications of the random effects multidimensional credibility model are considered. We primarily illustrate how the insurer can use our statistical model to make a posteriori updates to a priori premiums. Recall from Section 3 that $E[Y_{c,i,t}]$ is used for the a priori premium and $E[Y_{c,i,t}|b_{c,i}]$ for the a posteriori premium. The resulting so-called (theoretical) bonus–malus factor (‘BMF’) is the ratio (a posteriori premium/a priori premium). A BMF $> 1$ indicates that the customer is penalized for his claim behavior, while a BMF $< 1$ implies rewarding customers. See Lemaire (1995) for more details and Antonio et al. (2010) for more examples. Using model specification (4)–(6)

1Please note that policy and client numbers used in this section are introduced by the authors; they are completely independent of the coding used by the company.
the BMF for our model is: \( \exp(b_{c,i}) \). To estimate the BMF we will look at e.g. the mean or median of its posterior distribution (given the data, parameters and other random effects).

### 5.1 Policyholders with 1 contract

34,150 of the 48,631 individuals (i.e. 70%) in our sample hold just one policy. To get an overall picture of the behavior of BMFs, Figure 4 shows the BMF calculated for these policies versus the mean number of claims (per year) registered for the policy. The mean of the posterior distribution of \( \exp(b_{c,1}) \) is used for the BMF. The mean number of claims is obtained as: \( \text{total claims registered in \{2006, 2007\}} / \text{total exposure in \{2006, 2007\}} \). For simplicity, only the policies with full exposure in 2006 and 2007 (i.e. exposure time = 1 in each year) are used in the plot. The latter group represents 33,375 of the 34,150 (i.e. 98%) clients holding 1 policy. Figure 4 reveals that differences among BMFs, for a fixed average number of claims (on the x-axis), are explained by \textit{a priori} differences among policyholders. In the plot we connect low risks, with an \textit{a priori} premium of 0.208, average risks (\textit{a priori} premium of 0.675) and high risks (\textit{a priori} premium of 1.595). The risk characteristics corresponding with these profiles are given in Table 3. Again, these \textit{a priori} premiums are rather high, because we model the number of claimed guarantees, and not the number of claimed accidents. Some numerical illustrations of the BMF behavior are in Table 4. In this Table ‘Exp.’ stands for exposure. Therefore, the reported \textit{a priori} and \textit{a posteriori} premiums concern a period of 1 year of coverage. Figure 4 and Table 4 illustrate that high \textit{a priori} risks are better rewarded for good claiming behavior and less severely penalized for bad behavior, compared to lower \textit{a priori} risks.

In Figure 5 we show the behavior of future \textit{a priori} and \textit{a posteriori} premiums for a selected group of policyholders. The premiums are represented graphically with a boxplot (no whiskers/tails are included to avoid blurring the plot) of their simulated values obtained from the Bayesian analysis. Given data on \( Y_{c,1,t} \) with \( t = 1, 2 \) (being
Figure 4: Policyholders holding 1 policy: BMF versus average number of claims (per year). Right plot is a detail of the left plot.

Table 3: Risk profiles used in Figure 4. Premiums in the Table are calculated for an exposure of 1 year.

<table>
<thead>
<tr>
<th>Name</th>
<th>A priori premium</th>
<th>Female</th>
<th>Rural</th>
<th>Motor</th>
<th>Age &lt; 25 yrs</th>
<th>Age ≥ 25 yrs and &lt; 30 yrs</th>
<th>Veh. Age &lt; 3 yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low risk</td>
<td>0.208</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Average risk</td>
<td>0.675</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>High risk</td>
<td>1.595</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

2007, 2008), for the a priori premium we draw simulations from the posterior distribution of $E[Y_{c,1,3}]$. For the a posteriori premium the posterior distribution of $E[Y_{c,1,3}|b_{c,1}]$ is used. Again we distinguish between low, average and high a priori risks. Above each boxplot the past claim behavior of the corresponding policyholder is mentioned, e.g. ‘2/2 yrs’ means that 2 claims have been reported on a total insured period of 2 years.

5.2 Policyholders with > 1 contract

Similar to what we did in Figure 5, Figures 6 to 8 show boxplots of a priori and a posteriori premiums for selected groups of policyholders, holding 2 up to 4 policies. No il-
Table 4: Numerical illustrations corresponding with Figure 4. All premiums reported in this Table are calculated for an exposure of 1 year (as indicated by the ‘(1)’). ‘Acc. Cl.’ means ‘accumulated number of claims’; it is the total number of claims reported by the policyholder over his exposure period. ‘Acc. Exp.’ is the total exposure period registered for the policyholder.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>L1</td>
<td>0.208 (1)</td>
<td>0.125 (1)</td>
<td>0.60</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>0.292 (1)</td>
<td>1.406</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>0.549 (1)</td>
<td>2.64</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>1.232 (1)</td>
<td>5.925</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>A1</td>
<td>0.675 (1)</td>
<td>0.25 (1)</td>
<td>0.371</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.477 (1)</td>
<td>0.707</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>0.771 (1)</td>
<td>1.142</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A4</td>
<td>1.538 (1)</td>
<td>2.278</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>H1</td>
<td>1.595 (1)</td>
<td>0.362 (1)</td>
<td>0.227</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>H2</td>
<td>0.647 (1)</td>
<td>0.405</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H3</td>
<td>1.011 (1)</td>
<td>0.634</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H4</td>
<td>1.784 (1)</td>
<td>1.118</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Illustration is included of policyholders having 5 policies, but that would be completely similar to the examples discussed below. Figure 6 concerns the group of policyholders having two policies. Boxplots of \(E[Y_{c,1,3}]\) (respectively \(E[Y_{c,1,3}|b_{c,1}]\)) and \(E[Y_{c,2,3}]\) (respectively \(E[Y_{c,2,3}|b_{c,2}]\)) are shown right next to each other (followed by some whitespace to separate two different policyholders). Figure 7 shows policyholders holding 3 policies. Figure 8 those holding 4 policies.

Table 5 gives numerical details for some of the scenarios shown in Figures 6 to 8 (or scenarios not shown in the Figures but present in the data). The comparisons in Table 5 illustrate the impact of the multidimensional credibility model specified in (4)–(6) with positive covariance terms in \(\Sigma\). In (a) (in bold) we compare two policies having the same \(a\) priori premium and the same past claim behavior (3 claims in total over the years 2006 and 2007). The second policy is held by a client having 3 policies realizing a total of 10 claims over 6 insured years (i.e. an average of 1.67 claims per year). The client holding the first policy, on the other hand, has reported a total of 3 claims on 4 insured years (i.e. an average of 0.75 claims per year). Due to the dependence structure in (6) and the positive covariance terms in \(\Sigma\), the bad claim behavior on his other policies
causes the BMF for policy 2 from client \(a_2\) to be higher than the BMF for policy 2 from client \(a_1\) (1.86 versus 1.61). Similar conclusions hold for comparison (b). In comparison (c) bigger premium reductions are realized for a client holding 4 claim–free policies, in contrast with a client holding only 2 claim–free policies. Recall from Section 5.1 that it is important to compare policies with the same \textit{a priori} premium.

To gain insight in the posterior distribution of a vector of random effects used in (4)–(6), Figure 9 is a matrix of scatterplots showing the dependence between the com-
Figure 6: Policyholders holding 2 policies: boxplots of simulations for a priori and a posteriori premiums for a selected group of policyholders.

A priori (red) and a posteriori (grey) premiums

Components of the posterior distribution of $(b_{c,1}, \ldots, b_{c,4})'$. From Table 1 we know that the database contains 687 policyholders having 4 contracts. In this Figure the dots represent couples of posterior means for $(b_{c,i}, b_{c,j})'$, based on 4,000 simulations in the Bayesian model. The positive dependence between the random effects is apparent. Looking at the vector $(b_{c,1}, \ldots, b_{c,5})'$ would reveal a completely similar picture.
Figure 7: Policyholders holding 3 policies: boxplots of simulations for a priori and a posteriori premiums for a selected group of policyholders.

5.3 Summary

Three major applications of the Bayesian multivariate credibility model can be mentioned.

(i) The Bayesian implementation allows to approach the problem of ratemaking in a data–driven way. Graphs like those from the previous Sections help visualizing the riskiness of the portfolio and the behavior of the different a priori risks.
Moreover, parameter estimation is straightforward using the Bayesian toolbox.

(ii) As illustrated in Sections 5.1 and 5.2, we can easily update the risk premium of a client, based on his reported claim history. For instance, for client \( c \) and the next time period \( t = 3 \), the posterior distribution of \( Y_{c,i,3} \), given all parameters in the model and observations \( Y_{c,i,1} \) and \( Y_{c,i,2} \) (with \( j = 1, \ldots, n_c \)), allows to estimate the future number of claims for policy \( i \) from client \( c \). This leads to an updated

---

**Figure 8:** Policyholders holding 4 policies: boxplots of simulations for a priori and a posteriori premiums for a selected group of policyholders.
Table 5: Numerical illustrations corresponding with Figures 6 to 8.

<table>
<thead>
<tr>
<th># policies</th>
<th>client</th>
<th>policy</th>
<th>a priori</th>
<th>a posteriori</th>
<th>BMF</th>
<th>Acc. Cl.</th>
<th>Acc. Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison (a)</strong></td>
<td>2</td>
<td>a1</td>
<td>1</td>
<td>0.675</td>
<td>0.292</td>
<td>0.433</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>a2</td>
<td>1</td>
<td>0.675</td>
<td>0.917</td>
<td>1.359</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>1.085</td>
<td>1.608</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>2.002</td>
<td>2.966</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td><strong>Comparison (b)</strong></td>
<td>2</td>
<td>b1</td>
<td>1</td>
<td>0.989</td>
<td>0.938</td>
<td>0.948</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>1.274</td>
<td>1.289</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>b2</td>
<td>1</td>
<td>0.989</td>
<td>1.032</td>
<td>1.043</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>1.032</td>
<td>1.043</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>2.144</td>
<td>2.168</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>1.27</td>
<td>5.773</td>
<td>4.547</td>
<td>13</td>
</tr>
<tr>
<td><strong>Comparison (c)</strong></td>
<td>2</td>
<td>c1</td>
<td>1</td>
<td>0.869</td>
<td>0.258</td>
<td>0.296</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>0.258</td>
<td>0.296</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>c2</td>
<td>1</td>
<td>0.869</td>
<td>0.217</td>
<td>0.249</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>0.217</td>
<td>0.249</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>0.217</td>
<td>0.249</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>0.217</td>
<td>0.249</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

premium e.g. by multiplying the posterior mean of \(E[Y_{c,i,3}|b_{c,i}]\) with the average severity paid for a claim from this portfolio.

(iii) The Bayesian implementation of the credibility model is especially useful when pricing a new policy on an existing policyholder. For instance, when client \(c\) holds already 2 policies, simulations from \(Y_{c,3,3}\) are obtained by random sampling from the conditional normal distribution of \(b_{c,3}\) given the most recent updates of \(b_{c,1}\), \(b_{c,2}\) and the parameters in \(\Sigma\). This is easily implemented in a Bayesian analysis.

6 Conclusion

This paper presents a statistical approach to multivariate experience rating using reported claim counts. The approach is fully Bayesian and a data–driven alternative for the analytical multivariate credibility schemes that have been presented in actuarial literature. A Bayesian analysis is presented of policyholders holding multiple policies.
within the same European insurance company. The paper illustrates useful graphical tools to visualize the riskiness of the portfolio. Several applications are illustrated, including the pricing of existing policies in a new time period or the pricing of new (extra) policies held by existing policyholders.
References


