A Credibility Index Approach for Effective *a Posteriori* Ratemaking with Large Insurance Portfolios

Sebastián F. Calcetero V., Andrei L. Badescu, X. Sheldon Lin
Department of Statistical Sciences
University of Toronto
Toronto, Ontario
*sebastian.calcetero@mail.utoronto.ca

November 12, 2022

Abstract

Credibility, experience rating and more recently the so-called *a posteriori* ratemaking in insurance consists in the determination of premiums that account for both the policyholders' attributes and their claim history. The models designed for such purposes are known as credibility models and fall under the same framework of Bayesian inference in statistics. Most of the data-driven models used for this task are mathematically intractable due to their complex structure, and therefore credibility premiums must be obtained via numerical methods e.g simulation via Markov Chain Monte Carlo. However, such methods are computationally expensive and even prohibitive for large portfolios when these must be applied at the policyholder level. In addition, these computations are "black-box" procedures for actuaries as there is no clear expression showing how the claim history of policyholders is used to upgrade their premiums. In this paper we address these challenges and propose a methodology to derive a closed-form expression to compute credibility premiums for any given Bayesian model. We do so by introducing a credibility index, that works as an efficient summary statistic of the claim history of a policyholder, and illustrate how it can be used as the main input to approximate any credibility formula. The closed-form solution can be used to reduce the computational burden of *a posteriori* ratemaking for large portfolios via the same idea of surrogate modeling, and also provides transparent way of computing premiums from which practical interpretations and risk assessments can be performed.

Keywords— Credibility, Ratemaking, Bayesian, Regression, Experience Rating

1   Introduction

Credibility, experience rating, Bonus Malus Systems (BMS) or more recently the so-called *a posteriori* ratemaking is a fundamental area in actuarial science that enables actuaries to upgrade premiums according to the policyholder's actual experience. Classical credibility theory, as formulated by Bulhman (Bühlmann and Gisler (2005)), is based on the theory of Bayesian inference. The current knowledge on the risk behavior of a policyholder plays the role of the *a priori* information, and together with the actual experience on the claims, leads to a posterior knowledge of the true risk behavior of such policyholder, from which the upgraded premiums can be computed. The nature of the Bayesian credibility model is selected depending on the goals of the modelling

question and most importantly on the behavior of the data, that can be of great complexity due to the size of the portfolio and the hetereogenity. That being said, numerous modeling approaches have been explored in the actuarial literature with the main goal being the calculation of upgraded premiums.

As discussed by Norberg (2004), actuaries have mostly approached the problem of *a posteriori* ratemaking using simplistic assumptions under the class of Bayesian conjugate prior families, and mostly using the Bühlmann linear credibility formula and its variations (for e.g Bühlmann -Straub). The latter provides an approximation to the predictive mean via a weighted average of the claim history's sample mean and the manual premium, and so it provides a non-parametric approach to credibility with a clear interpretation and inexpensive computations. However, this formula is extremely restrictive for *a posteriori* ratemaking applications, as it provides values for only the predictive mean and not for other quantities that the actuary might be interested in (e.g variance, quantiles, probabilities). Furthermore, the resulting premium using Bühlmann's formula is not a data-driven approach, as it relies on strong assumptions, such as linearity and lacks compatibility with different model structures (e.g multivariate models, regression models, heavy tail models). Indeed, it is known that such an expression is exact only for a limited modelling framework (Jewell (1974)), and therefore provides very poor performance when the true predictive mean is nonlinear (Gómez-Déniz and Calderín-Ojeda (2014)).

Nowadays, with the increased availability of micro-level information in insurance data and the development of much more complex insurance products (e.g telematics based insurance, BMS systems) it is necessary to use tailored made models that account for the heterogeneity of the data sets, while still fulfilling regulatory requirements and standards. Such complicated structures that may appear in general insurance cannot be tackled simply neither by the traditional conjugate prior credibility framework, nor by the simplistic closed-form approximation models. In addition, the inclusion of policyholder attributes is of extreme relevance, as noted by Ohlsson (2008), Xacur and Garrido (2018), Diao and Weng (2019) and complicates the nature of the credibility model. It is therefore imperative to use general Bayesian modelling approaches that are data-driven and can provide a more flexible framework from which several quantities of interest can be derived for applications in experience ratemaking, claim reserving and risk management.

The general Bayesian workflow has not been very popular in insurance credibility applications due to the challenges encountered while attempting to obtain closed-form expressions. Many of the Bayesian models that provide a reasonable fit to real insurance data sets are mathematically intractable and the derivation of credibility premiums relies on computationally expensive numerical approximations, such as simulations via Markov Chain Monte Carlo (MCMC) methods (Xacur and Garrido (2018), Zhang et al. (2018), Ahn et al. (2021)). One of the biggest challenge is dealing with large insurance heterogeneous portfolios, for which it is necessary to perform a prohibitive number of simulations from the posterior predictive distribution of each policyholder. In addition, if a lot of experience is observed, these simulations must be performed repeatedly to update the credibility premiums. A second issue of these analytically intractable Bayesian credibility models is the "black-box" approach used to obtain the credibility premiums. These quantities are obtained via numerical approximations that makes the upgrading premium system to lack practical interpretation, thus becoming unappealing to be used by practitioners as results are not explainable to clients or regulators.

There has been a limited number of studies that were performed to approximate the Bayesian premiums in general insurance. Li et al. (2021) tackled this problem by approximating a dynamic Bayesian model via a mathematically tractable discrete Hidden Markov Model. In contrast, in the statistical literature, there have been many papers that attempted to solve the Bayesian inference problem of obtaining accurate approximations to the posterior or predictive distribution. Most of

these approaches discuss the problem of high dimensionality of the observed experience, and their solutions are based on summary statistics (see e.g Sisson et al. (2018)), and very few of them have interpretability as a main goal.

Similar ideas to those based on summary statistics can still be used in *a posteriori* ratemaking. The idea in such a context is to find a summary statistic (lower dimension function) of the policyholder's claim history that still provides the same information contained in that history (see for e.g. Sunnåker et al. (2013), Taylor (1977), Virginia (1998), Sundt (1979a)). The use of such statistics has been limited to the application of linear credibility formulas without motivation on how these statistics capture the information contained in the claim history of the policyholders. Therefore, the choice of suitable summary statistics seems to be a relatively under-explored problem in the credibility literature. Sunnåker et al. (2013) and De Vylder and Ballegeer (1979) are among the very few papers in that direction. Joyce and Marjoram (2008) and Blum et al. (2013) have provided a framework for assessing whether adding a new summary statistics to a given set of statistics provides further information, however no official treatment for the construction of such a statistic was provided. In addition, these statistics are designed towards the computation of distribution functions rather than closed-form expressions for insurance premiums and thus are not useful for *a posteriori* ratemaking applications.

Our paper aims to facilitate and promote the use of data-driven Bayesian models for credibility in insurance by proposing an approach that can be used to perform efficient and transparent experience rating on very complex insurance data sets, with a particular emphasis on large insurance portfolios. We propose a very novel methodology that derives a closed-form expression to approximate accurately the credibility premiums resulting from a given Bayesian model. As part of the methodology, we introduce a summary statistic which we call the "credibility index", that can be seen as an approximation of a sufficient statistic of the claim history for each policyholder, and that provides insights on how this experience is consistent with what is expected for the claim behavior of such policyholder. The newly defined index can be used to accurately approximate the rating factors for various policyholders in an easy and interpretive fashion. The aforementioned closed-form expression, that links the credibility index to the credibility premium policy-wise, also allows us to scale up the premium calculation to the portfolio level effectively, in particular when a portfolio is very large. In this situation, we may perform the computation of true credibility premium on a very small number of policies , say 5%, from the portfolio and use the closed-form expression as a surrogate model to extrapolate the credibility premiums on the selected policies to the rest of the portfolio.

The paper is organized as follows. In Section 2 we provide a summary of the Bayesian framework for experience rating purposes, and review some of the main challenges that will be addressed in this paper. Section 3 introduces the credibility index, along with its motivation and properties. Section 4 discusses how the index can be used to construct a closed-form credibility formula that approximates credibility premiums from a Bayesian model. Section 5 provides sufficiency properties of the newly introduced credibility index and discusses how these guarantee the accuracy of the resulting credibility formula. Section 6 discusses the estimation and implementation of the credibility index and the credibility formula in the context of large insurance portfolios. Section 7 displays a simulation study that illustrate the accuracy of the credibility formula under several choices of Bayesian models. Section 8 illustrates our results on a real European automobile data set, while Section 9 concludes the paper.

3

2 The Bayesian Credibility Framework and Issues

The mathematical framework for classical credibility, as introduced by Bühlmann (see e.g Bühlmann and Gisler (2005)), is based on the Bayesian hierarchical model:

$$\mathbf{Y}_n|\Theta = \theta \sim_{iid} f(y|\theta)$$
$$\Theta \sim P(\theta)$$

where $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ and each Y_j, $j = 1, \ldots, n$ is the claim history (either claim size or frequency) in period j of a policyholder, $f(y|\theta)$ is the density function of the *model distribution* and Θ is the *latent variable* that represents the unobservable risk of each policyholder, and has a prior distribution $P(\theta)$. The variables Y_j and Θ can be either univariate or multivariate, however, we shall keep the notation as if they are univariate for reading convenience. This formulation includes more general hierarchical models with several layers that are commonly used in credibility, see e.g Frees and Valdez (2008), Crevecoeur et al. (2022). For instance, a Bayesian model can have two (or more) layers of latent variables as follows:

$$\mathbf{Y}_n|\Theta_1 = \theta_1 \sim_{iid} f(y|\theta_1)$$
$$\Theta_1|\Theta_2 = \theta_2 \sim P(\theta_1|\theta_2)$$
$$\Theta_2 \sim P(\theta_2)$$

Do note that this is still embedded in the general formulation above by considering $\Theta = (\Theta_1, \Theta_2) \sim P(\theta) = P(\theta_1, \theta_2) = P(\theta_2)P(\theta_1|\theta_2)$. Finally, this setup also includes the case of regression-type credibility models, widely used in *a posteriori* ratemaking in insurance, see e.g Tzougas and di Cerchiara (2021), Denuit et al. (2007), Bermúdez and Karlis (2017). For instance, consider a GLMM-type of structure such as:

$$\mathbf{Y}_n|\Theta = \theta \sim_{iid} f(y|\theta)$$
$$\eta(\theta; \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle) = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle + \varepsilon$$
$$\varepsilon \sim P(\varepsilon)$$

where $f(\cdot)$ is the model distribution, usually a member of the exponential family, $\eta(\cdot)$ would be the so-called link function that links the latent variable Θ to a regression on covariates $\boldsymbol{x}$ with coefficients $\boldsymbol{\beta}$, and ε is a random effect incorporated in the regression. To ease the notation, we will only make explicit the dependence on covariates and their parameters when necessary.

Given such a framework, the goal of *a posteriori* ratemaking consists on using the claim history $\mathbf{Y}_n$ to obtain a premium for the $(n+1) - th$ period. To do so, one first requires the predictive distribution of $Y_{n+1}|\mathbf{Y}_n$, which after simple manipulations can be computed as follows

$$f(Y_{n+1}|\mathbf{Y}_n) = \frac{E_\Theta(f(Y_{n+1}|\Theta)\exp(\ell(\mathbf{Y}_n|\Theta)))}{E_\Theta(\exp(\ell(\mathbf{Y}_n|\Theta)))} \tag{1}$$

where E_Θ is an expectation with respect to the prior distribution of Θ and $\ell(\mathbf{Y}_n|\Theta)$ is the conditional log-likelihood of the observed experience $\mathbf{Y}_n$ given by

$$\ell(\mathbf{Y}_n|\Theta) = \sum_{j=1}^n \log(f(Y_j|\Theta)).$$

The predictive distribution describes the risk behavior of a policyholder, and therefore it is the only quantity needed to compute any *premium principle* that the actuary is considering. As defined

by Kaas et al. (2008), a premium principle is an operator Π that assigns a given risk with a non-negative value, $\Pi(Y_{n+1}|\mathbf{Y}_n)$. Many premium principles have been studied in the actuarial literature as described in Dickson (2005) or Radtke et al. (2016). The most well known examples are premiums based on operators defined through conditional expectations of the form $E(\pi(Y_{n+1})|\mathbf{Y}_n)$ for some weight function $\pi(y)$. Table 1 shows some examples.

Premium principle	Weight functions	Premium $\Pi(Y_{n+1}|\mathbf{Y}_n)$
Net Premium	$\pi(y) = y$	$E(\pi(Y_{n+1})|\mathbf{Y}_n)$
Expected value	$\pi(y) = (1+\alpha)y$	$E(\pi(Y_{n+1})|\mathbf{Y}_n)$
Mean value / Utility function	$\pi(y) = \mathcal{U}(Y)$	$E(\pi(Y_{n+1})|\mathbf{Y}_n)$
Variance /Standard deviation	$\pi_1(y) = y, \pi_2(y) = y^2$	$E(\pi_1(Y_{n+1})|\mathbf{Y}_n) + \alpha E(\pi_2(Y_{n+1})|\mathbf{Y}_n)$
Exponential	$\pi(y) = \exp(\alpha y)$	$\log(E(\pi(Y_{n+1})|\mathbf{Y}_n))/\alpha$
Esscher	$\pi_1(y) = y\exp(\alpha y), \pi_2(y) = \exp(\alpha y)$	$E(\pi_1(Y_{n+1})|\mathbf{Y}_n)/E(\pi_2(Y_{n+1})|\mathbf{Y}_n)$

Table 1: Example of Premium Principles

Although this Bayesian framework is in theory sound, its actual implementation in practice is not often feasible, especially when one needs to account for the claim history and policyholder's attributes in a data-driven fashion. Indeed, the computation of equation (1) and any of the conditional expectations on Table 1 could be a challenging task under most practical considerations. It is known that closed-form solutions to these expression can only be obtained under certain choices of the model distribution and the prior, which apply mostly to a limited number of simple parametric models. Furthermore, most of the data-driven setups that commonly provide satisfactory fits to real insurance data sets do not belong to this limited class of models (see e.g Adam et al. (2021), Czado et al. (2012), Yau et al. (2003)), and so performing credibility is of much higher complexity due to mathematical intractability of premium computations.

The most common approach used for approximation of credibility premiums is via Markov Chain Monte Carlo (MCMC) methods along the same lines as Bayesian inference procedures. These methods enable the actuary to draw samples from the posterior and predictive distribution of the fitted model without the need of having a closed-form expression for their distributions, and so the desired expectations can be obtained as the sample average of the simulated quantities. This process can be computationally expensive as requires the underlying Markov chain to be in a stationary state, and this process must be performed individually for each policyholder. The task becomes more difficult when considering policyholders' attributes on a regression-type model, as the computations of expectations must be performed at a granular level by taking into account the heterogeneity of the overall portfolio. That said, when it comes to relatively large portfolios in non-life insurance, the MCMC process becomes almost prohibitive for practical applications because of the computational burden, Ahn et al. (2021).

The aforementioned issues become more challenging when interpretability is at play. This means that the resulting premium not only must be easy to compute, but also the ratemaking process must be transparent, so that policyholders are priced fairly without any type of discrimination. When numerical approximations are used to calculate premiums in a general Bayesian model, the ratemaking process becomes a "black-box" due to the lack of closed-form solutions linking the policyholder's attributes and their claim history. Therefore it is essential to be able to find a solution that reduces the computational burden and that is also interpretable. In the following sections, we propose a methodology that aims to address these challenges.

3 The Introduction of the Credibility Index

We first introduce and motivate the "credibility index", which is a summary statistic of the claim history of a policyholder and measures the likeliness of the claim history of a policyholder under the current knowledge on the risk (i.e without accounting for claim history). This represents the building block that helps us to obtain closed-form expressions for the credibility premiums in the next section.

To motivate the idea behind our general methodology, we start by discussing the main issue encountered when attempting to derive an analytical expression for the credibility premiums - the high dimensionality. Indeed, each of the values of the claim history $\mathbf{Y}_n$ is an input of such function, and so are all the other attributes of the policyholder (see Friedman (1994) for issues regarding the problem of learning a function). An extended approach discussed in the literature to deal with the dimensionality problem is the use of dimensional reduction techniques via summary statistics. This consists on finding a lower dimension function of the claim history and the policyholder's attributes, that provide the same information contained in the whole set $\mathbf{Y}_n$. This is related to the concept of the so-called *approximately sufficient statistics* in the context of Approximate Bayesian Computation (ABC), see Joyce and Marjoram (2008), Sunnåker et al. (2013) and Fearnhead and Prangle (2012), for further references . Indeed, some approaches in the credibility literature rely implicitly on this idea. For instance, Bühlmann-Straub model performs a dimensional reduction from n to one, by utilizing the summary statistic $\sum_{j=1}^{n} \alpha_j Y_j$, with the coefficients α_j as some tuning parameters. The choice of such summary statistics seems to be an under-explored problem in the credibility literature. Very few papers explored the idea of using different types of summary statistics. See for e.g. Künsch (1992) where the sample median of the set $\mathbf{Y}_n$ is used instead of the sample mean, or Yan and Song (2022) that considered a general class of credibility formulas based on linear combinations of estimators of the mean. However, very few of these have discussed the relevance of such summary statistics.

We follow the same idea of dimensionality reduction and propose a new summary statistic based on the policyholder's claim history that we call the credibility index. This accounts for the heterogeneity of the portfolio, while capturing most of the information from its claim history. To motivate our definition, we first show how the posterior predictive distribution depends on the claim history of a policyholder $\mathbf{Y}_n$. Under the assumption that the log-likelihood (as a function of the parameter), and the support of Θ (denoted as $\mathcal{R}_\Theta$) are continuous, the mean value theorem applied to equation (1) shows that the predictive distribution can be written as:

$$f(Y_{n+1}|\mathbf{Y}_n) = \frac{f(Y_{n+1}|\Theta = \tilde{\theta}) \exp\left(\ell(\mathbf{Y}_n|\Theta = \tilde{\theta})\right)}{E_\Theta\left(\exp(\ell(\mathbf{Y}_n|\Theta))\right)} \qquad (2)$$

for some value $\tilde{\theta} \in \mathcal{R}_\Theta$. The previous expression is quite meaningful in our ratemaking context as it resembles the one of an upgrading process where the prior predictive distribution $f(Y_{n+1}|\Theta = \tilde{\theta})$ is multiplied by an upgrading factor that depends on the claim history of the policyholder. Observe in this expression that $\ell(\mathbf{Y}_n|\Theta = \tilde{\theta})$, i.e the conditional log-likelihood evaluated at the value $\tilde{\theta}$ provided by the mean value theorem, is playing the role of a "summary statistic" of the claim history $\mathbf{Y}_n$ of the policyholder, and the posterior predictive distribution depends on the claim history only through a function of this quantity. This last assertion is however not entirely true in general, as the term in the denominator $E_\Theta\left(\exp(\ell(\mathbf{Y}_n|\Theta))\right)$ is also a function of the claim history $\mathbf{Y}_n$, and so it influences the resulting value of the predictive distribution. This term coincides with the scaling factor that appears in the denominator of the computation of the posterior distribution when using the Bayes rule, and thus intuitively does not add much probabilistic meaning to the problem and

it is rather a value to keep the probability well behaved in the $[0, 1]$ scale. With this in mind, we can argue that the term in the numerator which depends only on $\ell(\mathbf{Y}_n|\Theta = \tilde{\theta})$ may be the most relevant quantity on which the claim history of the policyholder affects the predictive distribution, and therefore the premium. Thus, we propose the following definition for our credibility index.

Definition 1 (Credibility index). *The credibility index for a policyholder is defined as the following summary statistic:*

$$\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{j=1}^{n} \log f(Y_j|\Theta = \tilde{\theta})$$

where $\tilde{\theta} \in \mathcal{R}_\Theta$ is the "mean-value" parameter(which is policyholder dependent).

The credibility index is considered to be a function of the claim history $\mathbf{Y}_n$ and not the latent variable and so we change the notation when referring to the credibility index vs a conditional log-likelihood to avoid any confusion. Note that $\tilde{\theta}$ is essentially the one obtained by the mean value theorem and so it is policyholder dependent. This credibility index is constructed from the log of a probability and therefore it is always on the same scale regardless of the nature of the variables that are in consideration. This is a desirable feature when considering models that account for frequency and severity simultaneously.

Interpretation of the index: The credibility index is in essence a conditional log-likelihood of the claim history of a policyholder, therefore we can think of it as a measure of how likely is the claim history of a policyholder. Such likeliness is computed under the parameters of the model as fitted for the whole portfolio and under the mean-value parameter $\tilde{\theta}$ that may be interpreted as the individual "risk" of a policyholder. We remark that such a value $\tilde{\theta}$ is not necessarily a quantity associated to the posterior distribution of Θ.

To better illustrate how the credibility index can be interpreted, we can make an analogy with a traditional likelihood-ratio hypothesis test, in which the null hypothesis can be thought of as "the current premium is accurate (i.e the model captures the risk of the policyholder appropriately)" vs the alternative "the current premiums is not accurate (i.e the model is not capturing the risk the risk of the policyholder appropriately)". Under the null, we would compute the likelihood test-statistic under the current parameters of the model, which is in essence what the credibility index is, and then proceed to compare it against to a critical value (i.e assess whether there is evidence in favor or against the null hypothesis). If the current premiums are accurate for a policyholder, then the observed experience $\mathbf{Y}_n$ should be "likely" as measured by the credibility index, and thus its value would concentrate around the "no rejection" region. However, if the premium is not accurate, then the observed experience will be "unlikely", as measured by the credibility index and thus this will display values closer to the "rejection region". Along these lines, the credibility index can be used as a measure of consistency/discrepancy between what is currently predicted by the model (i.e current premiums) and the actual claim history of a policyholder. The "critical values" or ranges for which the index could be considered to be measuring a "likely" or "unlikely" claim history, and how to modify the premiums if in the latter case, will be explained in more details in later sections.

Accounting for Policyholder attributes: The actual expression of the index is tailored-made to the model distribution and so it accounts for all the modelling considerations that the actuary took into it (e.g policyholder information, heaviness of the tail, dependence structure, etc), in contrast to a "fixed form" summary statistics, such as the sample mean or the sample median that does not account for any of these specifications. Indeed, by being based on a conditional

7

log-likelihood function, the credibility index can be easily computed for most hierarchical models including those that may account for covariates. In the latter case, the model distribution is of the form $f(y|\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle, \theta)$ where the regression function is introduced as another component in the likelihood independently of the value $\tilde{\theta}$, and so the credibility index would be also covariate dependent as:

$$\mathcal{L}(\mathbf{Y}_n; \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle, \tilde{\theta}) = \sum_{j=1}^n \log f(Y_j | \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle, \Theta = \tilde{\theta})$$

Additivity of the actual experience: The credibility index is an additive function of the claim history of a policyholder in the sense that each period of information on the claim history adds a new term to the computation of the credibility index, and so the overall information collected from a policyholder is just the aggregation of each period of information. Thus, the quantity is extremely easy to update when new data comes in. Later on, we will show how this is useful for iterative ratemaking purposes.

Sub-indexes for multivariate models: In insurance applications where policies have multiple coverages, the variable of interest Y_j is a vector of dimension D, which we shall denote as $Y_j = (Y_j^{(1)}, Y_j^{(2)}, \ldots, Y_j^{(D)})$, with $Y_j^{(d)}$ being the d-component of the vector. In such cases, the distribution function $f(Y_j | \Theta)$ is a multivariate conditional distribution. Most of the credibility models in the literature are constructed under a conditional independence structure, that is, the different components of the vector are conditionally independent given the latent variable Θ. In such framework, we can write the model distribution as the product of the individual conditional distributions of the vector, that is: $f(Y_j | \Theta) = \prod_{d=1}^D f(Y_j^{(d)} | \Theta)$. Therefore we could write the credibility index for a policyholder as

$$\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{d=1}^D \mathcal{L}^{(d)}(\mathbf{Y}_n^{(d)}; \tilde{\theta})$$

with

$$\mathcal{L}^{(d)}(\mathbf{Y}_n^{(d)}; \tilde{\theta}) = \sum_{j=1}^n \log f(Y_j^{(d)} | \Theta = \tilde{\theta}) \quad , d = 1, \ldots, D.$$

Therefore, the credibility index can be decomposed into D credibility sub-indexes, each one associated with the claim history of a policyholder for the d-th component only. Each credibility sub-index can then be interpreted individually, along the same lines as the global index, i.e the likeliness of the claim history for the d-th component of the policyholder, and so the likeliness of the whole claim history vector is given by the addition of the individual sub-indexes. Therefore, the credibility index provides a clear interpretation of the claim history for multivariate models and enables the actuary to identify which of the components in the vector drives most of the "likeness or unlikeness" based on the a priory knowledge on the policyholder. We will discuss how this decomposition is useful for further interpretation of the results and better prediction accuracy in next sections.

Accounting for partially observed claim history: Information from claims is usually subjected to policy limits and deductibles (i.e truncation and censoring), and so it is partially observed. Hence, the values Y_j for such types of claims cannot be used directly in traditional summary statistics. Nevertheless, the credibility index can easily account for such type of observations, along the

same lines as these are addressed when constructing a log-likelihood function. For instance, if the value Y_j is a right-censored observation, then its contribution to the credibility index would be the term $\log\Big(1 - F(Y_j|\Theta = \tilde{\theta})\Big)$, instead of the term $\log\Big(f(Y_j|\Theta = \tilde{\theta})\Big)$, where F is the cumulative distribution function (CDF) associated to the model distribution f. The same analogy can be applied for the case of truncation, or for other different types of modifications, such as adjustments for inflation. An extreme case of censoring worth discussing is *missing data*, which can still be accommodated in the computation of the credibility index. Indeed, if one of the D components of Y_j is entirely missing, then we can still account for the log-likelihood of this observation by a log-likelihood of $\log(1) = 0$, which is equivalent to not adding information to the credibility index. This property is of extreme usefulness in multivariate models as will be discussed in the numerical case studies.

4   A Credibility Formula via the Credibility Index

In this section we proceed to illustrate how the credibility index can be used towards the construction of a closed-form credibility formula for a given Bayesian model. To introduce the main idea and our motivation, note that any credibility formula under any premium principle can be generally written as:

$$\Pi(Y_{n+1}|\mathbf{Y}_n) = G_\Pi(\mathcal{O}, \mathbf{Y}_n, n) \tag{3}$$

where $G_\Pi(\cdot)$ is the theoretical or true functional form, linking the claim history of the policyholder $\mathbf{Y}_n$ and the set of model parameters $\mathcal{O}$, to the credibility premium under the premium principle Π, that represents our quantity of interest. The functional form of such a $G_\Pi(\cdot)$ depends entirely on the premium principle and the underlying Bayesian model. For instance, if we use a premium principle that is based on expectation operators, then one can show that the following structure holds for the credibility premium:

$$\Pi(Y_{n+1}|\mathbf{Y}_n) = \Pi(Y_{n+1}) \exp\Big(g(\mathbf{Y}_n, \tilde{\theta}, n)\Big)$$

for an unknown nonlinear functional $g(\cdot)$, that depends on the claim history $\mathbf{Y}_n$, the individual risk as measured in the mean-value parameter $\tilde{\theta}$ and the number of the previously observed periods. Such an expression isolates the effect of the claim history in a portion, that can be seen as a rating factor, and another portion associated only with the parameters of the model $\mathcal{O}$, which in this case is given by the manual premium $\Pi(Y_{n+1})$. As mentioned in the previous section, such an expression is high dimensional in the number of inputs and the idea of introducing the credibility index is to perform a dimensionality reduction for such an expression. Indeed, by replacing the claim history by the credibility index in the previous expression, we get:

$$\Pi(Y_{n+1}|\mathbf{Y}_n) \approx \Pi(Y_{n+1}) \exp\Big(g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)\Big). \tag{4}$$

This expression is an approximation of the credibility premium formula, due to the fact that we are replacing the whole set of information $\mathbf{Y}_n$ with the credibility index $\mathcal{L}(\mathbf{Y}_n; \tilde{\theta})$. We emphasise that by performing this trick, we reduce drastically the dimension of the functional $g(\cdot)$, who now depends only on the credibility index and mean-value parameter $\tilde{\theta}$, who also depends on each individual claim experience $\mathbf{Y}_n$. We also note that the quality of our proposed approximation relies heavily on the capacity of the credibility index to be a good summary statistic of the claim history of a policyholder. Later on we will prove that if the credibility index is a sufficient statistic for the latent variable Θ, then any credibility premium will depend on the experience $\mathbf{Y}_n$ only as a

function of such index, and therefore the above approximation becomes equality. In the case that the index is not sufficient, but still summarizes a great portion of the information in the sample $\mathbf{Y}_n$, such a relationship still holds, but only represents an approximation. These results will be proved mathematically in the next section.

Interpretation: Expression (4) resembles the one of a rate upgrading process where $\Pi(Y_{n+1})$ plays the role of the manual premium (i.e the premium without experience) and the term on the right , $\exp\Big(g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)\Big)$ plays the role of a rating factor that adjusts the manual premium according to the claim history. As the credibility index measures the magnitude of discrepancy between the manual premium and the claim history, then the function $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)$ describes how much of adjustment should be allocated to reach a premium that is "consistent" with such experience. This adjustment is given by a multiplicative effect under the exponential function to avoid negative premiums. If the function $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)$ takes a value of 0, then no adjustment is necessary; if the function $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) > 0$, then the premium must be increased, and finally if the function $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) < 0$, the premium should be reduced. That being said, the rating process becomes transparent to the actuary in the sense that the credibility premium is derived intuitively and can be easily explained to both clients and regulators.

Similarly, the function $g(\cdot)$ provides the associated ranges for which the credibility index can be defined as likely or unlikely with respect to the claim history for a given policyholder. Indeed, taking the preimage under the function $g(\cdot)$, the values of the credibility index such that $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) \approx 0$ can be considered as the values for which the claim history of a policyholder is consistent with the manual premium (e.g a no rejection zone in our hypothesis test analogy). Analogously, the values of the credibility index such that $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) > 0$ can be considered as the values for which the claim history is riskier than that expected (i.e a one-tailed rejection zone), and finally, the values of the credibility index such that $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) < 0$ can be considered as the values for which the claim history is not as risky as expected (i.e the other tailed rejection zone). Hence, the credibility index provides the actuary with a general metric to compare the risk behaviors of policyholders.

Structure of the function $g(\cdot)$**:**   In principle, the function $g(\cdot)$ is an unknown bivariate surface that depends on both the credibility index and the number of past periods, however, we may want to impose a structure to have interpretability, and also a good quality of the approximation. Indeed, one may impose different structures for the function $g(\cdot)$, for instance one can chose

$$g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) = g_1(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta})) + g_2(n),$$

$$g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) = g_1\left(\frac{1}{n}\mathcal{L}(\mathbf{Y}_n; \tilde{\theta})\right) + g_2(n),$$

$$g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) = g_1\left(\frac{1}{n}\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n\right).$$

Similarly, if the model in consideration is multivariate for $Y_j = (Y_j^{(1)}, \ldots, Y_j^{(D)})$, one may consider a function $g(\cdot)$ that measures the effect of the credibility sub-indexes separately, and so it is possible to quantify the importance of the claim history of each component to the resulting credibility premium. For instance, one may consider the structure of the form

$$g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n) = \sum_{d=1}^{D} g_d(\mathcal{L}(\mathbf{Y}_n^{(d)}; \tilde{\theta})) + g_{D+1}(n),$$

and so each function $g_d(\cdot)$ measures the effect/importance of the past experience on $\mathbf{Y}_n^{(d)}$ over the

resulting premium.

One-Step formula version: So far these expressions upgrade the manual premium with the whole observed experience at once, however, in practice it may be desirable to upgrade the premium one period at a time, as soon as this is observed. We do note that this expression can also be used recursively for such upgrades , as long as the parameters of the model remain unchanged. Simple rearrangement of the above credibility formula shows that:

$$\Pi(Y_{n+1}|\mathbf{Y}_n) = \Pi(Y_n|\mathbf{Y}_{n-1}) \exp\Big(g(\mathcal{L}(\mathbf{Y}_n,\tilde{\theta}),n) - g(\mathcal{L}(\mathbf{Y}_{n-1},\tilde{\theta}),n-1)\Big)$$

$$\mathcal{L}(\mathbf{Y}_n;\tilde{\theta}) = \mathcal{L}(\mathbf{Y}_{n-1};\tilde{\theta}) + \log f(Y_n|\tilde{\theta})$$

Therefore, if we want to upgrade the premium based on the $n-1$ observations, $\mathbf{Y}_{n-1}$, we can use the already available credibility index, $\mathcal{L}(\mathbf{Y}_{n-1};\tilde{\theta})$, add the contribution of he new observation via $\mathcal{L}(\mathbf{Y}_n;\tilde{\theta}) = \mathcal{L}(\mathbf{Y}_{n-1};\tilde{\theta}) + \log f(Y_n|\tilde{\theta})$ (recall the additivity of the index), and teh compute the one-step formula above. Thus, the credibility index can also be thought of as a measure that evolves in time (as the claim history of a policyholder). Keeping track of it through time should be good enough for *a posteriori* ratemaking purposes. We do note, however, that this relationship would hold as long as the parameters of both the model distribution and the credibility index don't change too much as new observations come into play. Indeed, if the parameters of the model change, which is the case if the actuary decides to re-fit the model, it is necessary to find a new representation for the function $g(\cdot)$ and the mean-value parameter $\tilde{\theta}$, and therefore it is preferable to use the expression that upgrades premiums all at once, rather than one-step at a time.

Remark 1. *Note that the particular form of expression (4), although motivated by premium principles defined through expectations, may still be accurate for more general premium principles. Nevertheless, we emphasize that any functional form can be used as desired while keeping in mind the reduction in dimension:*

$$\Pi(Y_{n+1}|\mathbf{Y}_n) = G_\Pi(\mathcal{O},\mathbf{Y}_n,n) \approx G_\Pi(\mathcal{O},\mathcal{L}(\mathbf{Y}_n;\tilde{\theta}),n)$$

5 On the Sufficiency of the Credibility Index

The motivation behind our credibility index is to find a "good summary statistic" for the claim history of a policyholder. Such a "good" summary statistics is formally defined by the concept of a sufficient statistic, see Casella and Berger (2021) for more details and Sundt (1979b) for a whole discussion for the relevance on credibility theory. Indeed, if the credibility index is a sufficient statistic, we can assert that the credibility index is properly capturing all the information of the claim history of a policyholder and therefore our credibility formula (4) is exact. In this section, we show that a subtle modification of the credibility index can potentially be a sufficient statistic for Θ (in the Bayesian sense) for many family distributions, including the exponential dispersion family of distributions, which is commonly used for credibility models in insurance.

We first introduce the motivation for what we called a corrected credibility index. If we can decompose equation 1 as $\ell(\mathbf{Y}_n|\Theta) = l_1(\mathbf{Y}_n) + l_2(\mathbf{Y}_n,\Theta)$ for some functions l_1, l_2 (not necessarily probability functions), we can omit the first term that depends only on $\mathbf{Y}_n$, as it simplifies both the numerator and denominator of equation 2. Therefore we can potentially find a simplified form of the credibility index without losing any information. With this in mind, we introduce a corrected version of the credibility index and some of its properties.

Definition 2. *The corrected credibility index denoted as $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta})$ is defined as the "minimal" term in the following additive decomposition of the credibility index:*

$$\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}) = L(\mathbf{Y}_n) + \tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}).$$

The term "minimal" is defined in the sense that if there is a further sub-decomposition of the form $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = L_1(\mathbf{Y}_n) + L_2(\mathbf{Y}_n; \tilde{\theta})$ for some functions L_1, L_2, then $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = L_2(\mathbf{Y}_n; \tilde{\theta})$.

The corrected credibility index may be a sufficient statistic for the latent variable Θ under certain conditions, and thus it may capture all the information of the claim history of a policyholder. Moreover, we can view the credibility index as an "almost" sufficient statistic as it is closely associated to this possibly sufficient statistic. Along those lines, we can conclude that the credibility index on either its original or corrected version contains a large amount of information on the claim history that is useful for deriving credibility premiums.

We now proceed to show some results on the sufficiency of the corrected credibility index. Let $T(\mathbf{Y}_n)$ be a sufficient statistic for the latent variable Θ. By the Fisher–Neyman factorization theorem, see for e.g Casella and Berger (2021), the conditional likelihood function can be factored into two non-negative functions, and so the log-likelihood function can be additively separated into two functions l_1 and l_2 as:

$$\ell(\mathbf{Y}_n | \Theta) = l_1(\mathbf{Y}_n) + l_2(T(\mathbf{Y}_n), \Theta) \quad \forall \mathbf{Y}_n, \Theta.$$

Theorem 1. *Suppose there exists a value $\tilde{\theta} \in R_\Theta$ at which $l_2(T(\mathbf{Y}_n), \tilde{\theta})$ is a one-to-one function when viewed as a function of the sufficient statistic $T(\mathbf{Y}_n)$. Then the corrected credibility index $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta})$ at the value $\tilde{\theta}$, is a sufficient statistic for the latent variable Θ.*

Proof. The decomposition of the log-likelihood above holds for every value of $\theta \in \mathcal{R}_\Theta$, in particular the one in the assumption. When fixing the value of Θ, the log-likelihood function at the left hand side of the decomposition becomes our credibility index, and so at this particular value $\tilde{\theta}$ we must have the following relationship in terms of the corrected credibility index:

$$\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) + L(\mathbf{Y}_n) = l_1(\mathbf{Y}_n) + l_2(T(\mathbf{Y}_n), \tilde{\theta})$$

And so we have:

$$\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = (l_1(\mathbf{Y}_n) - L(\mathbf{Y}_n)) + l_2(T(\mathbf{Y}_n), \tilde{\theta})$$

The last expression provides an additive decomposition of the corrected credibility index, and by construction of its minimality, we must have that:

$$\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = l_2(T(\mathbf{Y}_n), \tilde{\theta})$$

Now, by assumption of $\tilde{\theta}$, the right-hand side is a one-to-one function when viewed as a function of the sufficient statistic $T(\mathbf{Y}_n)$. Therefore, the left-hand side side is one-to-one function of a sufficient statistic and so the credibility index $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta})$ is also a sufficient statistic for the latent variable Θ. Moreover, if $T(\mathbf{Y}_n)$ is also minimal sufficient, then by the one-to-one correspondence, $\tilde{\mathcal{L}}(\mathbf{Y}_n, \tilde{\theta})$ is also minimal sufficient.

$\square$

Corollary 1. *Under the assumptions above, the following holds for any premium principle Π*

$$\Pi(Y_{n+1} | \mathbf{Y}_n) = G_\Pi(\mathcal{O}, \mathbf{Y}_n, n) = G_\Pi(\mathcal{O}, \tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}), n),$$

12

that is, one can replace the claim history vector with the corrected credibility index without any loss of information for a posteriori *ratemaking purposes.*

Proof. As $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta})$ is a sufficient statistic for the latent variable Θ, then the posterior and predictive distributions will depend on the claim history $\mathbf{Y}_n$ only through a function of $\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta})$. The above expression is just this statement written mathematically.

$\square$

Example 1 (Exponential dispersion family). *In the particular case in which the model distribution $f(Y|\Theta)$ is given by a member of the exponential dispersion family of distributions with a dispersion parameter that does not depend on the latent variable (this is a construction widely used for insurance applications, see e.g Wuthrich and Merz (2022)), we have:*

$$f(Y_j|\Theta = \theta) = \exp\left(\frac{\theta S(Y_j) - C(\theta)}{\varphi_j} - Q(Y_j; \varphi)\right)$$

for some functions $S(\cdot), C(\cdot), Q(\cdot)$ and dispersion parameter ϕ. That said, the conditional log-likelihood takes the form:

$$\ell(\mathbf{Y}_n|\theta) = \sum_{j=1}^n \frac{\theta S(Y_j) - C(\theta)}{\varphi_j} - \sum_{j=1}^n Q(Y_j; \varphi_j)$$

And so the credibility index and the corrected credibility index are respectively:

$$\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}) = \tilde{\theta} \sum_{j=1}^n \frac{S(Y_j)}{\varphi_j} - \sum_{j=1}^n \frac{C(\tilde{\theta})}{\varphi_j} - \sum_{j=1}^n Q(Y_j; \varphi_j)$$

$$\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = \tilde{\theta} \sum_{j=1}^n \frac{S(Y_j)}{\varphi_j} - \sum_{j=1}^n \frac{C(\tilde{\theta})}{\varphi_j}$$

It is known that $\sum_{j=1}^n \frac{S(Y_j)}{\varphi_j}$ is the minimal sufficient statistic for the exponential dispersion family and we can observe from above that the function linking the sufficient statistic and the corrected credibility index is a one-to-one function for $\tilde{\theta} \neq 0$. Therefore, by the theorem above, the credibility index is also a minimal sufficient statistic.

For further illustration, it is known that under certain assumptions of the prior distribution, the credibility formula is exact (see Klugman et al. (2012)) in the sense that the predictive mean becomes equal to the linear credibility formula, and so

$$E(Y_{n+1}|\mathbf{Y}_n) = Z_n * \bar{Y}_n + (1 - Z_n) * E(Y_{n+1})$$

where Z_n is the so called credibility factor. With simple algebra, it is easy to see how this formula can be written in the framework of our proposed credibility formula:

$$E(Y_{n+1}|\mathbf{Y}_n) = E(Y_{n+1}) \exp\left(\log\left(Z_n \frac{\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) + nC(\tilde{\theta})}{n\tilde{\theta}} + (1 - Z_n)E(Y_{n+1})\right) - \log E(Y_{n+1})\right)$$

We note that the credibility index is one-dimensional in the sense that it reduces the information of the whole claim history of a policyholder to a single quantity. Therefore, the credibility index can potentially be a sufficient statistic when the latent variable Θ is one-dimensional. In the case of multivariate models for Y_j with D dimensions, this limitation can be weaken in the sense that

13

now the credibility sub-indexes can also play a role to constitute a set of summary statistics that all together may be sufficient for a latent variable Θ. Indeed, the result of theorem 1 extends directly if we consider a multivariate model for Y_j in which each of the components $Y_j^{(d)}$ is associated to at most one component of the latent variable $\tilde{\Theta}^{(d)}$ at a time. As this is a very particular setup, we illustrate this generalization with just an example as follows.

Example 2 (Multivariate exponential dispersion family). *Let's consider now the case of a particular multivariate exponential dispersion family of distributions with dispersion parameters that do not depend on the latent variable (this construction is also widely used for applications in insurance).*

$$f(Y_j | \Theta = \theta) = \exp\left(\sum_{d=1}^{D} \frac{\theta^{(d)} S^{(d)}(Y_j^{(d)}) - C^{(d)}(\theta^{(d)})}{\varphi_j^{(d)}} - \sum_{d=1}^{D} Q^{(d)}(Y_j^{(d)}; \varphi_j^{(d)}) \right)$$

for some functions $S^{(d)}(\cdot), C^{(d)}(\cdot), Q^{(d)}(\cdot)$. Observe that in this particular case the function $C(\theta)$ is split additively on each of the components, which is the case when the components of the vector Y_j are conditionally independent given Θ. The conditional log-likelihood takes the form

$$\ell(\mathbf{Y}_n | \theta) = \sum_{j=1}^{n} \sum_{d=1}^{D} \frac{\theta^{(d)} S^{(d)}(Y_j^{(d)}) - C^{(d)}(\theta^{(d)})}{\varphi_j^{(d)}} - \sum_{j=1}^{n} \sum_{d=1}^{D} Q^{(d)}(Y_j^{(d)}; \varphi_j^{(d)}),$$

And so the credibility index and sub-indexes are respectively

$$\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{d=1}^{D} \tilde{\theta}^{(d)} \sum_{j=1}^{n} \frac{S^{(d)}(Y_j^{(d)})}{\varphi_j^{(d)}} - \sum_{d=1}^{D} \sum_{j=1}^{n} \frac{C^{(d)}(\tilde{\theta}^{(d)})}{\varphi_j^{(d)}} - \sum_{d=1}^{D} \sum_{j=1}^{n} Q^{(d)}(Y_j^{(d)}; \varphi_j^{(d)}),$$

$$\mathcal{L}^{(d)}(\mathbf{Y}_n^{(d)}; \tilde{\theta}^{(d)}) = \tilde{\theta}^{(d)} \sum_{j=1}^{n} \frac{S^{(d)}(Y_j^{(d)})}{\varphi_j^{(d)}} - \sum_{j=1}^{n} \frac{C^{(d)}(\tilde{\theta}^{(d)})}{\varphi_j^{(d)}} - \sum_{j=1}^{n} Q^{(d)}(Y_j^{(d)}; \varphi_j^{(d)}),$$

$$\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{d=1}^{D} \mathcal{L}^{(d)}(\mathbf{Y}_n^{(d)}; \tilde{\theta}^{(d)}).$$

The corrected credibility index and corrected sub-indexes are respectively

$$\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{d=1}^{D} \tilde{\theta}^{(d)} \sum_{j=1}^{n} \frac{S^{(d)}(Y_j^{(d)})}{\varphi_j^{(d)}} - \sum_{d=1}^{D} \sum_{j=1}^{n} \frac{C^{(d)}(\tilde{\theta}^{(d)})}{\varphi_j^{(d)}},$$

$$\tilde{\mathcal{L}}^{(d)}(\mathbf{Y}_n^{(d)}; \tilde{\theta}^{(d)}) = \tilde{\theta}^{(d)} \sum_{j=1}^{n} \frac{S^{(d)}(Y_j^{(d)})}{\varphi_j^{(d)}} - \sum_{j=1}^{n} \frac{C^{(d)}(\tilde{\theta}^{(d)})}{\varphi_j^{(d)}},$$

$$\tilde{\mathcal{L}}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{d=1}^{D} \tilde{\mathcal{L}}^{(d)}(\mathbf{Y}_n^{(d)}; \tilde{\theta}^{(d)}).$$

It is known that the set of statistics $\sum_{j=1}^{n} \frac{S^{(d)}(Y_j^{(d)})}{\varphi_j^{(d)}}$ are the minimal sufficient statistic for this family and we can observe that the function linking each of these sufficient statistics and the associated corrected credibility sub-indexes are one-to-one functions for $\tilde{\theta}^{(d)} \neq 0$. Therefore, by the theorem above, the set of credibility sub-indexes contains also minimal sufficient statistics.

14

We conclude this section by noting that despite the fact that the mean-value parameter $\tilde{\theta}$ was introduced via the mean value theorem, such a construction does not affect the proof of the sufficiency results of the credibility index. This does not imply any contradiction, but rather shows the flexibility on the choice of the values $\tilde{\theta}$.

6 Estimation of the Credibility Formula - A Surrogate Model for Large Portfolios

In this section we further discuss how to find an estimate of the function $G_\Pi(\cdot)$ and the values $\tilde{\theta}$ for each policyholder. A challenging issue that may prohibit accurate premium calculations is related to the size of the insurance portfolios. In order to tackle such a challenge we provide a general framework in which only a minimal portion of the portfolio, say 5%, is used for the estimation under the principle of *surrogate modeling*, which guarantees a straightforward application.

Surrogates models are a way to reduce the burden of obtaining an output that must take several times under different values of the inputs, in particular when the value of such output is computationally expensive. In such a case, a surrogate model provides an approximation of the output function in a more efficient way than the original computational process. In general terms, by using some known values of the inputs and outputs, a surrogate model is trained as an interpolation function over these values and then extrapolated to a new set of inputs to approximate their outputs. Such training and extrapolation is assumed to be computationally inexpensive when compared to the real mechanism generating the outputs, (see Sobester et al. (2008) for further reference on surrogate modeling). Some applications of such models have been explored in actuarial science in Lin and Yang (2020a) and Lin and Yang (2020b). In our setup, the numerical challenge is associated to the computation of the credibility premium itself, which is usually achieved by an approximation method. As our framework provides a closed-form solution by means of only few inputs (i.e. the credibility index), it can be easily used for the construction of an efficient surrogate model that reduces the computational burden. The following steps, that we will further elaborate on, show how our credibility formula can be estimated and used in a surrogate model.

1. Selecting a sub-portfolio of representative policyholders.

2. Estimate credibility premiums using simulation/numerical schemes only on such sub-portfolio.

3. Estimate the parameter $\tilde{\theta}$ and the nonlinear function $G_\Pi(\cdot)$ with the only sub-portfolio using a tailored method for interpolation.

4. Assess the accuracy of the formula and extrapolate it to the rest of the portfolio.

We will now provide further details on how each of the steps is performed in what follows.

6.1 Selecting a representative sub-portfolio

Selecting a representative sub-portfolio is a crucial task towards obtaining accurate results. Such sub-portfolio must be small enough from a computational point of view, but large enough to exhibit similar properties as in the original portfolio, in order to produce a reliable extrapolation. In the context of large portfolios, a sample between 1% to 10% of the whole portfolio may suffice for this purpose. The selection of the representative policyholders is a problem widely addressed in the statistical literature on *population sampling*, and several methodologies have been developed for this purpose (see Chambers and Clark (2012) for further reference on sampling techniques).

Given the particular goal of extrapolation, it is important for the sub-portfolio to be as similar as possible with respect to all the inputs involved for the computation of the credibility index, which are the parameters of the log-likelihood, the number of periods with observed exposure n and the claim history $\mathbf{Y}_n$. Among all the population sampling techniques, we recommend the use of model-assisted sampling methods, that can sample from large populations while accounting for several attributes of such population to be balanced. In this paper, we use the *cube method*, as described in Tillé (2011), which has been also used in other actuarial papers towards the same applications of surrogate models (see for instance Lin and Yang (2020b)).

6.2 Computing credibility premiums for the sub-portfolio

The computation of credibility premiums can be achieved via any numerical scheme, whether is based on MCMC or quadrature methods. This problem is well documented in the literature of computational Bayesian statistics, so we refer the reader to e.g Sisson et al. (2018) for further details. For the sake of completeness, we illustrate here a tailored made setup, that is quite efficient for the computation of premiums defined through the expectation operators, as motivated in Section 2.

Along the lines with the literature on Approximated Bayesian Computation (ABC), we propose to use an *importance sampling* approach in which the so-called *biased distribution* of the posterior distribution of the latent variables is chosen to be the prior distribution, and therefore the expectations from the posterior predictive distribution can be estimated as

$$E(\pi(Y_{n+1})|\mathbf{Y}_n) \approx \hat{E}(\pi(Y_{n+1})|\mathbf{Y}_n) := \frac{\sum_{k=1}^{K} E(\pi(Y_{n+1})|\theta_k) \exp(\ell(\mathbf{Y}_n|\theta_k))}{\sum_{k=1}^{K} \exp(\ell(\mathbf{Y}_n|\theta_k))}, \tag{5}$$

where θ_k are iid samples from the prior distribution $P(\theta)$, and the value K is large enough to guarantee reasonable estimates of the expectation. Note that this simulation setup is easy to be performed, as it is based on generating samples from the prior distributions, rather than from the posterior one. The reader is directed to the literature on importance sampling (Tokdar and Kass (2010)) for more properties of the proposed estimator.

We note the fact that prior distribution $P(\theta)$ does not involve any term associated with the policyholder itself, and it is the same for each one in the portfolio. Therefore, the same drawn from the prior, θ_k, can be used simultaneously to compute the associated terms in the simulations of all the policyholders in the sub-portfolio. Hence, this simulation setup is very efficient in terms of the number of draws required from the prior distribution, and the size of the portfolio.

Finally, once these expectations are estimated, we are in good position to find the associated credibility premium, according to the premium principle that is being considered (see e.g Table 1). Therefore, we end up with a reliable approximation of such premiums, that we denote by $\hat{\Pi}(Y_{n+1}|\mathbf{Y}_n)$ from now and on.

6.3 Approximating the function $G_\Pi(\cdot)$ and the values $\tilde{\theta}$

To estimate $G_\Pi(\cdot)$ and the value $\tilde{\theta}$, we use a non-parametric interpolation method in which the inputs are the set of parameters of the model, the credibility index and the number of periods of claim history of each policyholder. The outputs are the credibility premiums for each policyholder in the sub-portfolio of representative policyholders. We describe here a simple framework for the estimation of such values. We note that the literature on interpolation of general functions is vast (Mastroianni and Milovanovic (2008)), so the description below is a guideline, rather than a step-by-step numerical procedure to be followed.

As mentioned in Section 4, one needs to select a particular functional form for such function $G_\Pi(\cdot)$. For the sake of convenience, we emphasize on the rating factor form, nevertheless any other functional form can be used. With that in mind, our goal is to fit the relationship (4), in which we now have to estimate the function $g(\cdot)$, that depends only on two inputs, the credibility index and the number of previously observed claim periods, n.

Suppose the sub-portfolio consists of M policyholders indexed by i, $i = 1, \ldots, M$. Let's denote by $\hat{\Pi}_i^p$ the credibility premium for the i-th policy holder that was obtained via simulation (i.e the $\hat{\Pi}(Y_{n+1}|\mathbf{Y}_n)$ values). As these were obtained as the sample mean of a large number of simulations, we can argue that $\hat{\Pi}_i^p \approx N(\Pi_i^p, se_i^2)$, with Π_i^p being the credibility premium $\Pi(Y_{n+1}|\mathbf{Y}_n)$ and se_i^2 the standard error of the estimation. Denote by Π_i the manual premium of the i-th policyholder (i.e the premiums $\Pi(Y_{n+1})$ without claim history incorporated) and by $\mathcal{L}_i(\tilde{\theta}_i)$ the credibility index at the value $\tilde{\theta}_i$, for the i-th policyholder in the portfolio. From the chosen functional form we have,

$$\hat{\Pi}_i^p \approx N(\Pi_i^p, se_i^2)$$

$$\log(\Pi_i^p) = \log(\Pi_i) + g(\mathcal{L}_i(\tilde{\theta}_i), n_i)$$

The setup above can be identified as a generalized regression model with Gaussian response and with a log-link function, where the $\hat{\Pi}_i^p$ play the role of the response, $\log(\Pi_i)$ plays the role of an off-set and both $\mathcal{L}_i(\tilde{\theta})$ and n_i play the role of features with a joint non-linear effect. We do note that the chosen structure imposed on the function $g(\cdot)$ (e.g an additive structure) must be accounted in this step. Recall that such structure must be assessed to guarantee enough flexibility and interpretations of the ratemaking process.

The actuary can fit this model by using any statistical learning techniques, such as Generalized Additive Models (GAM), Neural Networks, Trees based methods, etc. We remark that such a model is extremely easy to fit as only two features are involved. We do note however, that the values $\tilde{\theta}_i$ are unknown and therefore the features of such model are not fixed beforehand. Nevertheless, note that the values $\tilde{\theta}_i$ are only motivated to guarantee equality in equation (2), then one may consider these values as another set of parameters that must be tuned for a better fit. We remark that these values are policyholder dependent and therefore they are functions of the set of parameters of the model and the claim history. With this in mind, we consider a certain structure of the form $\tilde{\theta} = h(\mathcal{O}, \mathbf{Y_n})$, for some unknown function h. As these values are not designed to be interpreted on their own, any statistical learning technique can be used to find a reliable approximation of h.

The estimation of $\tilde{\theta}$ must be achieved jointly with the function $g(\cdot)$, and therefore has to be incorporated as a part of the optimization process when fitting the model for $g(\cdot)$. This can be achieved directly in a tailored algorithm specified by the user, or in an iterative fashion. Indeed, for the easiness of the implementations, it might be easier to consider the latter approach to use some of the already implemented methods in software packages. Along these lines, the general scheme for an iterative estimation is shown in algorithm 1.

6.4 Assessment of the fit and extrapolation

Before performing any type of analysis with the fitted formula, the accuracy of the out-of-sample predictive power must be assessed. In what follows, we briefly mention the key points of this task, however the reader is refereed to Sobester et al. (2008) for further details.

To evaluate the accuracy of the fitted formula, one can perform regular goodness of fit assessment on the sample of policyholders along the same lines as in any regression model. This involves residual checking (e.g scatter plots, histogram error concentrated around 0, no biases, etc) and error metrics analysis (e.g high coefficient of determination, small relative errors, etc). If the

17

Algorithm 1 Fitting of $g(\cdot)$ and θ_i

$R^2 \leftarrow 0$

$\tilde{\theta}_i \leftarrow$ Random number $\forall i = 1, \ldots, M$ $\triangleright$ Start with random values for $\tilde{\theta}_i$

while $R^2 \leq$ Tol **do**

    $\mathcal{L}_i(\tilde{\theta}_i) \leftarrow \sum_{j=1}^n \log f(Y_{i,j}|\Theta = \tilde{\theta}_i) \; \forall i = 1, \ldots, M$ $\triangleright$ Compute the credibility index

    $g(\cdot) \leftarrow \arg\min_g \sum_{i=1}^M \left(\hat{\Pi}_i^p - \Pi_i \exp\left(g(\mathcal{L}_i(\tilde{\theta}_i), n_i) \right) \right)^2$ $\triangleright$ Fit the function $g(\cdot)$

    $\tilde{\theta}_i^o \leftarrow \arg\min_{\tilde{\theta}_i} \left(\hat{\Pi}_i^p - \Pi_i \exp\left(g(\mathcal{L}_i(\tilde{\theta}_i), n_i) \right) \right)^2 \; \forall i = 1, \ldots, M$ $\triangleright$ Find pseudo observations $\tilde{\theta}_i$

    $h \leftarrow \arg\min_h \sum_{i=1}^M \left(\tilde{\theta}_i^o - h(\mathcal{O}, \mathbf{Y}_n) \right)^2$ $\triangleright$ Fit the model $h(\cdot)$

    $\tilde{\theta}_i \leftarrow h(\mathcal{O}, \mathbf{Y}_{i,n}) \; \forall i = 1, \ldots, M$ $\triangleright$ Upgrade the values $\tilde{\theta}_i$ using the fitted values of $h(\cdot)$

    $R^2 \leftarrow 1 - \frac{\sum_{i=1}^M \left(\hat{\Pi}_i^p - \Pi_i \exp\left(g(\mathcal{L}_i(\tilde{\theta}), n_i) \right) \right)^2}{\sum_{i=1}^M \left(\hat{\Pi}_i^p - \frac{1}{M} \sum_{i=1}^M \Pi_i \right)^2}$ $\triangleright$ Upgrade current value of R^2

end while

---

formula is not providing a satisfactory performance under this assessment, the chosen structure of the functional form for $G_\Pi(\cdot)$ must be revised.

Similarly, to evaluate the out-of-sample predictive power, one can use similar ideas from the cross-validation literature on predictive modeling. For instance, the portfolio of representative policyholders can be split into train and test samples. The formula is fitted only with the former set and used to predict on the latter set. If the goodness of fit metrics on the train set are almost the same as their counterpart in the test set, the out-of-sample predictive power of the formula is verified. However, if these two differ drastically, the formula is not reliable for extrapolation, and therefore the sample size of representative policyholders must be increased. The latter can be achieved imitatively by increasing the sample size, say 1% at a time, until the out of sample performance is verified.

Once the fitted credibility formula is assessed, the computation of the credibility premiums for the rest of the portfolio is achieved by just evaluating the formula for each policyholder. Hence, this methodology simplifies considerably the work with large portfolios in the sense that the computational burden is reduced to only a minimal portion of the portfolio.

7 A Simulation Study

In this section, we illustrate how our newly defined credibility index can capture efficiently the information of the claim history of a policyholder via a simulation study. We do so by investigating the accuracy of our proposed credibility formula at approximating the value of credibility premiums on different choices of the model-prior distributions and premium principles commonly observed in actuarial modelling. We separate the analysis depending on whether the variable of interest is continuous or discrete. Results are summarized in Tables 2 and 3, in which we display the accuracy of the formula via the coefficient of determination R^2 of the fitted premiums vs the credibility premiums. As our goal in this section is to evaluate the potential predictive power of our credibility formula, we don't use the surrogate modelling approach and use the whole portfolio as the sample in which credibility premiums are computed.

The general setup for the simulation is explained as follows. We generate synthetic portfolios of policyholders and their claim history from the Bayesian models listed in Tables 2 and 3 . To emulate real portfolios in insurance, we generate a heterogeneous portfolio of 50,000 policyholders,

in which the mean of the model distribution is linked to synthetic covariates and a single latent variable, $\log(\mu_i) = \alpha_i + \theta_i$, where μ_i is the mean of the model distribution, α_i is the systematic component that is observable to the actuary (i.e a regression on covariates) and θ_i is the random component associated with the unobserved risk of a policyholder. Any dispersion parameter of the model distribution is assumed to not be random. The value of α_i, and the dispersion parameters were selected to resemble the heterogeneity that is commonly observed in real data sets.

The claim history for each policyholder is generated according to the resulting value μ_i from the previous step and the selected model distribution. For simplicity, we assume we have $n = 5$ periods of claim history for each policyholder.

To compute the credibility premiums, we use the importance sampling algorithm described in Section 6 with 20,000 samples for each policyholder. We estimate our credibility formula using the rating factor functional form with an additive structure for the function $g(\cdot)$. We use a generalized additive model (GAM) to estimate such function $g(\cdot)$, and use a *random forest* structure to link the values $\tilde{\theta}_i$ as a function of the parameters of the model distribution and their claim history. More specifically, we use the values of the systematic component α_i, the mean of the claim history $\bar{\mathbf{Y}}_{n,i}$ and the observed exposure n_i for each policyholders as the features in the random forest. We assess the accuracy of the credibility formula by comparing the estimated premiums vs the credibility premiums using the coefficient of determination R^2 and results are displayed in Tables 2 and 3.

Model	Premium Principle		
	Expected Value	Standard deviation	Exponential
Poisson-Gamma	98.076%	98.075%	98.083%
Poisson-LogNormal	99.513%	99.514%	99.515%
NegBinom-InvGaussian	96.996%	96.403%	99.999%
Logarithmic-Weibull	97.915%	97.872%	98.460%
GammaCount-Weibull	94.927%	94.912%	94.928%
GenPoisson-Lognormal	99.694%	99.654%	99.695%

Table 2: Coefficient of determination R^2 for simulation study on discrete distributions

Model	Premium Principle	
	Expected Value	Standard deviation
Gamma-Gamma	95.144%	95.145%
Lognormal-InvGaussian	94.512%	94.512%
LogLogistic-Lognormal	99.687%	99.688%
InvGaussian-Weibull	99.873%	98.869%
Pareto-Gamma	98.576%	98.583%
Burr-Lognormal	99.145%	99.147%

Table 3: Coefficient of determination R^2 for simulation study on continuous distributions

Results show that the value R^2 is much larger than 95% in practically all scenarios, which implies that the fitted premiums resemble the credibility premiums regardless of the choice of the model distributions and the premium principles. We noticed that the performance of the fitted formula tends to be lower when the portfolio is more heterogeneous with respect to the policyholder attributes and their claim history, however, the fitted formula is still able to reproduce accurately the credibility premiums on such scenarios. Note that some of the premium principles used in Tables

19

2 and 3 are not defined through expectation operators, nevertheless, the rating factor functional form provides a good fit. In fact, an alternative functional form for $G_\Pi(\cdot)$ might provide a better fit, and so there is room for improvement for the results here displayed. As an overall conclusion, we note that the credibility index is capable of summarizing almost all the information of the claim history a policyholder that is requited for experience rating purposes.

Finally, we noted in our implementation that the values $\tilde{\theta}_i$ are not as influential in the final results in comparison to the function $g(\cdot)$. Indeed, the fitted formula provided values of R^2 very similar to the ones here displayed in the very first iteration of the algorithm. Therefore, the tuning of the values $\tilde{\theta}_i$ is not as relevant as the fitting of the function $g(\cdot)$, and an early stop for the algorithm is recommended if the value of R^2 is already large enough or barely changes from one iteration to another.

8   A Numerical Illustration with Real Data

In this section, we illustrate the use of our credibility formula on a real data set provided by an European automobile insurance company with claim frequencies coming from two business lines. The data set contains information of policyholders' contracts from January 2007 to December 2017 and detailed attributes of the policyholder or their automobiles, such as car weight, engine displacement, engine power, fuel type (gasoline or diesel), car age, age of policyholder. The quantity of interest for this application is a policy with two coverage: Third Party Liability insurance (TPL) and Physical Damage (PD). The number of claims in both lines are potentially dependent as the same car accident can lead to claims in both line of business. It is to note that not all policyholders have a fully observed history in the two lines, as policyholders may initially start with contract on one policy and then upgrade to obtain a policy that covers the two of them, or viceversa (see Figure 1). Thus, the claim history for some of the policyholders can be considered as partially observed in the sense that the number of claims for a particular line may not be available for certain periods.

Table 4 and Figure 1 provide some key statistics of our data. Briefly, the data for the number of claims in both lines present similar behavior to stylized facts in insurance: over-dispersion, large amount of 0's and significant correlation between the two line of business. There are in total 184,848 policyholders that constitutes a large portfolio, yet only 41,956 have fully observed exposure in both lines of business simultaneously. We observe in the right hand side Figure 1 that a considerable number of policyholders renew their contract with this company, therefore a long claim history is available to certain policyholders, having up to 7 years of exposure. Therefore, accounting for such observed experience is relevant for ratemaking purposes.

| | | Claim Frequency | | | |
TYPE	Number of Policyholders	Mean	Median	Variance	Correlation
TPL	138,923	0.038	0.000	0.044	0.245
PD	87,517	0.306	0.000	0.512	
TOTAL	184,484	0.162	0.000	0.279	

Table 4: Summary statistics of the dataset

With this in mind, the insurance company is interested in performing *a posteriori* ratemaking to account for the claim history of the policyholders when renewing their contracts. For illustration purposes, we consider the *Exponential* premium principle, with a surcharge of 5% as the rule for which the insurance company sets the actuarial premium.
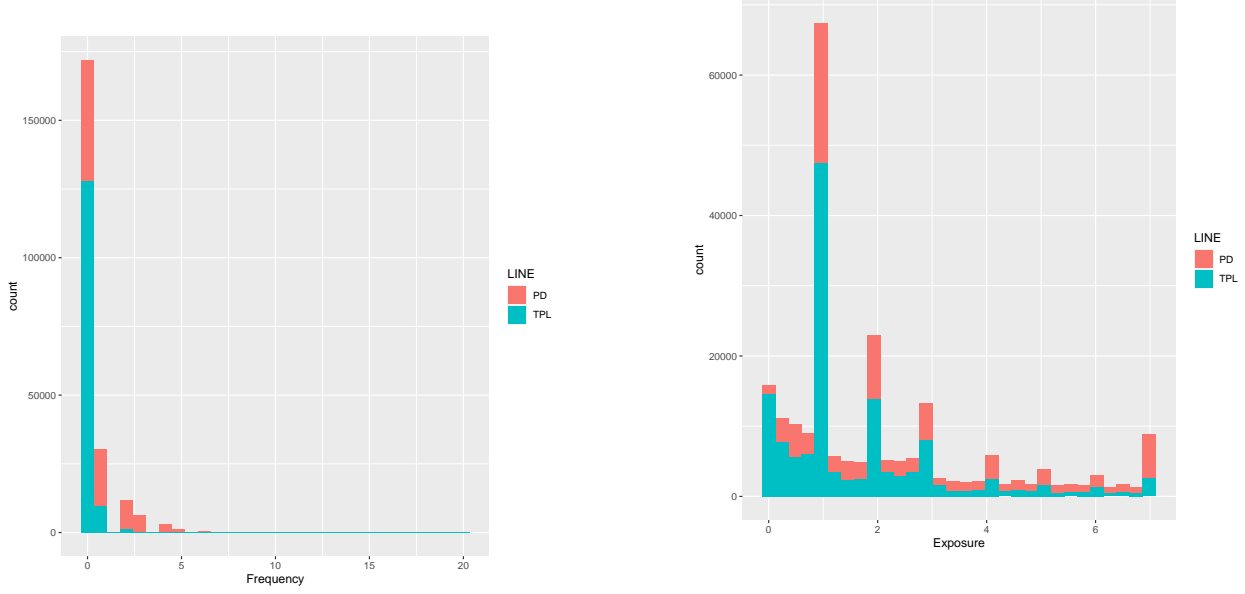
Figure 1: Distribution of claim frequency (left and exposure (right) per claim type

8.1 The Bayesian model and quantities of interest

We illustrate the use our credibility formula starting with a bivariate mixed negative-binomial regression model proposed by Tzougas and di Cerchiara (2021), that is very flexible and it is tailored to deal with over-dispersion and dependent frequencies. Let $Y_j^{(d)}$ be the number of claims from a given policyholder in year j, associated to the dth line of business, with $d = 1$ being PD and $d = 2$ being TPL. Let x be the associate information of covariates, and let $\beta^{(d)}$ be the associated vector of regression coefficients for the d-th coverage. Similarly, let's denote with $\omega^{(d)}$ the time-exposure of the contract for each coverage. We consider the hierarchical model

$$Y_j = \begin{pmatrix} Y_j^{(1)} \\ Y_j^{(2)} \end{pmatrix} \sim_{iid} f(y|\Theta, \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle) = \text{NegBinom}(y^{(1)}; \mu^{(1)}\Theta, r^{(1)}) * \text{NegBinom}(y^{(2)}; \mu^{(2)}\Theta, r^{(2)})$$

where, for $d = 1, 2$

$$\log \mu^{(d)} = \log \omega^{(d)} + \beta_0^{(d)} + \beta_1^{(d)} \text{CarWeight} + \beta_2^{(d)} \text{EngineDisplace} + \beta_3^{(d)} \text{CarAge} + \beta_4^{(d)} \text{Age} + \beta_5^{(d)} \text{EnginePower} + \beta_6^{(d)} \text{Fuel}$$

and $\Theta \sim P(\theta) = \text{InvGauss}(1, \sigma^2)$. We use the notation $\text{NegBinom}(y; \mu, r)$ to denote the probability mass function of a negative binomial with mean μ and dispersion r. Similarly, $\text{InvGauss}(1, \sigma^2)$ denotes an Inverse-Gaussian distribution with mean 1 and variance σ^2.

Observe that this model uses a common random effect for each line of business, which simplifies the task when working with partially observed data. Indeed, even if a policyholder only has observations on only one of the business lines, one can still perform inference on the value of the associated random effect Θ for the two business lines simultaneously.

Note that this model does not have a closed-form expression, neither for the posterior, nor for the predictive distribution, so numerical methods must be used to obtain any quantity of interest (see Tzougas and di Cerchiara (2021) for a detailed discussion). The estimation of the parameters is achieved in R with the aid of generalized linear mixed models with Negative-Binomial response and the fitted parameters are given below:

21

Variable	β_0	β_1	β_2	β_3	β_4	β_5	β_6	r	σ^2
$Y^{(1)}$	-2.78	$-4.48*10^{-5}$	$2.69*10^{-5}$	$-2.94*10^{-2}$	$-4.20*10^{-3}$	$2.24*10^{-3}$	$-8.23*10^{-2}$	0.86	0.58
$Y^{(2)}$	-1.397	$14.19*10^{-5}$	$6.10*10^{-5}$	$-0.20*10^{-2}$	$-7.94*10^{-3}$	$6.06*10^{-3}$	$-22.08*10^{-2}$	0.86	0.58

Table 5: Estimated Parameters of the Bivariate mixed NegBinomial regression model

8.2 Estimation of the credibility formula

In this section proceed to calculate the credibility premium using the exponential principle

$$\Pi(Y_{n+1}|\mathbf{Y_n}) = \frac{1}{0.05}\log\Big(E\Big(\exp\Big(0.05*(Y_{n+1}^{(1)}+Y_{n+1}^{(2)})\Big)|\mathbf{Y_n}\Big)\Big).$$

We proceed by describing the estimation of the credibility formula along the framework of a surrogate model, given the large size of the portfolio. For this application we use the *cube method* (we used the R function `samplecube()`) to obtain a sub-portfolio of 5% the size of the overall portfolio, this consisting of about 9,224 policyholders. In this particular setup, the features that are set to be balanced in the sub-portfolio and matched with those in the overall portfolio are the average number of claims for PD and the average number of claims for TPL, which represent the claim history, and the fitted values of $\mu^{(1)}$ and $\mu^{(2)}$, which represent the policyholder' attributes.

To estimate the credibility premiums we use the importance sampling method described in the estimation section, with 50,000 samples in order to guarantee accuracy. We remark that this step is only necessary to be performed in the sub-portfolio of policyholders identified in the previous step; however we perform the simulation for the entire portfolio for the sake of comparison of the premiums resulting from our formula and the simulated premiums.

As discussed before, the credibility index can be split into further sub-indexes that can be interpreted as summary statistics, one for each line of business in the model. In this case we can identify two credibility sub-indexes, defined as (for $d = 1, 2$)

$$\mathcal{L}^{(d)}(\mathbf{Y}_n; \tilde{\theta}) = \sum_{j=1}^n \log\Big(\text{NegBinom}\Big(Y_j^{(d)}; \mu^{(d)}\exp\Big(\tilde{\theta}\Big), r^{(d)}\Big)\Big).$$

We further use the rating factor functional form, even though the premium principle is not directly given by an expectation operator. By choosing an additive structure for the function $g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)$, our credibility formula is given by

$$\Pi(Y_{n+1}|\mathbf{Y_n}) = \Pi(Y_{n+1})\exp\Big(g_1(\mathcal{L}^{(1)}(\mathbf{Y}_n; \tilde{\theta})) + g_2(\mathcal{L}^{(2)}(\mathbf{Y}_n; \tilde{\theta})) + g_3(n^{(1)}) + g_4(n^{(2)})\Big).$$

The function $g_1(\cdot)$ provides a measure of the effects of the past experience in the number of PD claims in the credibility premium, $g_2(\cdot)$ provides a measure of the effects of the TPL number of claims in the credibility premium, and $g_3(\cdot)$ and $g_4(\cdot)$ provide the effect of the number of periods that the policyholder have been observed for each line of business. We remark the fact that a policyholder may posses only partially observed information in the sense that some of the claim history is associated to only one of the policies and not the two of them simultaneously. Therefore the length of the claim history may differ for each line of business and therefore we consider separate effects for $n^{(1)}$ and $n^{(2)}$.

To perform the estimation of the functions $g_k(\cdot)$ ($k = 1, 2, 3, 4$) and the tuning parameter $\tilde{\theta}_i$ corresponding to the i-th individual, we use the information on only the 5% of the policyholders, as described above. The fitting of the functions $g_j(\cdot)$ is achieved using the R implementation

of a generalized additive model (GAM), with a Gaussian response and taking into account the aforementioned additive structure, in which the $g_k(\cdot)$ are fitted as non-linear functions using natural cubic splines. Similarly, the values of $\tilde{\theta}_i$ are estimated using the structure of a *random forest* in which the prediction function is specified using as features: $\mu^{(1)}$, $\mu^{(2)}$ (which depend on the policyholder attributes), $\bar{Y}^{(1)}$, $\bar{Y}^{(2)}$ (which summarizes the claim history) for each policyholder. This is achieved using the `gbm` package in R with its default routine for calibration of hyper parameters. We choose these models as are flexible and not expensive to train.

The two sets of predictive models for the functions $g_k(\cdot)$ and the structure of the $\tilde{\theta}_i$ are fitted in an iterative fashion as described in the estimation section. We note that the first iteration of the algorithm, in which a constant model for the values $\tilde{\theta}_i$ is used, achieved an error of interpolation of the premiums that is very close to the final iteration of the model illustrated below. That being said, we remark that a single iteration of the algorithm may provide satisfactory results, and if so, it may be preferable to use it directly in order to avoid a bigger computational burden. Moreover, we note that several values $\tilde{\theta}_i$ provided a relatively similar quality in terms of the error of interpolation of the premiums, and thus it might not be necessary to tune these values all the way to the convergence to an optimal solution. Indeed, we recommend an early stopping of the algorithm once the error of interpolation has reached a predetermined tolerance value, or the improvement between iterations is minimal.

8.3   Results and interpretation

The estimated values $\tilde{\theta}_i$ are shown in Figure 2. The figures on the bottom show that the values of $\tilde{\theta}_i$ are highly associated to the manual and credibility premiums (i.e the risk behavior of policyholders). We do note that such association is not perfect in either scenario as the claim history plays a role in the calculation in a non-linear fashion. Nevertheless, recall that the values $\tilde{\theta}_i$ are motivated by the mean value theorem in equation (2), and so their role is mostly for numerical purposes rather than for interpretations for risks assessment. Similarly, the figure on the top-left side shows the distribution of the values $\tilde{\theta}_i$. It can be observed a great variability on these parameters under a bimodal behavior, which can be associated to the risk heterogeneity of the portfolio. Indeed, figure top-right show the distributions of the manual premiums and credibility premiums, and the bimodal behavior is also observable indicating the existence of possibly two major risk classes of policyholders in the portfolio.

To assess how the credibility index is related to the actual risk behavior observed on the claim history of policyholders, we plot it against the standardized claim frequency $\frac{\sum_{j=1}^{n} Y_j^{(d)}}{\sum_{j=1}^{n} \mu_j^{(d)}}$. Such a quantity provides a measure of the observed claim frequency in intuitive fashion (e.g the larger this value, the riskier is the policyholder), while still control by the heterogeneity of policyholder attributes. The graphs for each of the two sub indexes are displayed in Figure 3. It can be seen that the credibility index and the standardized frequency are highly correlated, and therefore the index is strongly associated to the riskiness of policyholders as observed in their claim history. In this case, we get an inverse relationship as the larger the standardized claim frequency, the lower is the value of the credibility sub-index. We do note that the association is not perfect as these two statistics don't capture the same amount of information from the claim history. Indeed, the credibility index provides a wider range of values than the standardized claim frequency, and so can distinguish policyholders' riskiness in a better fashion. For instance, a policyholder with no claims would lead to a standardized claim frequency of 0 regardless of the value of their attributes, however the credibility index will have different values depending on the covariates.

Figure 4 shows the distribution of the credibility sub-index among policyholders. The distribu-
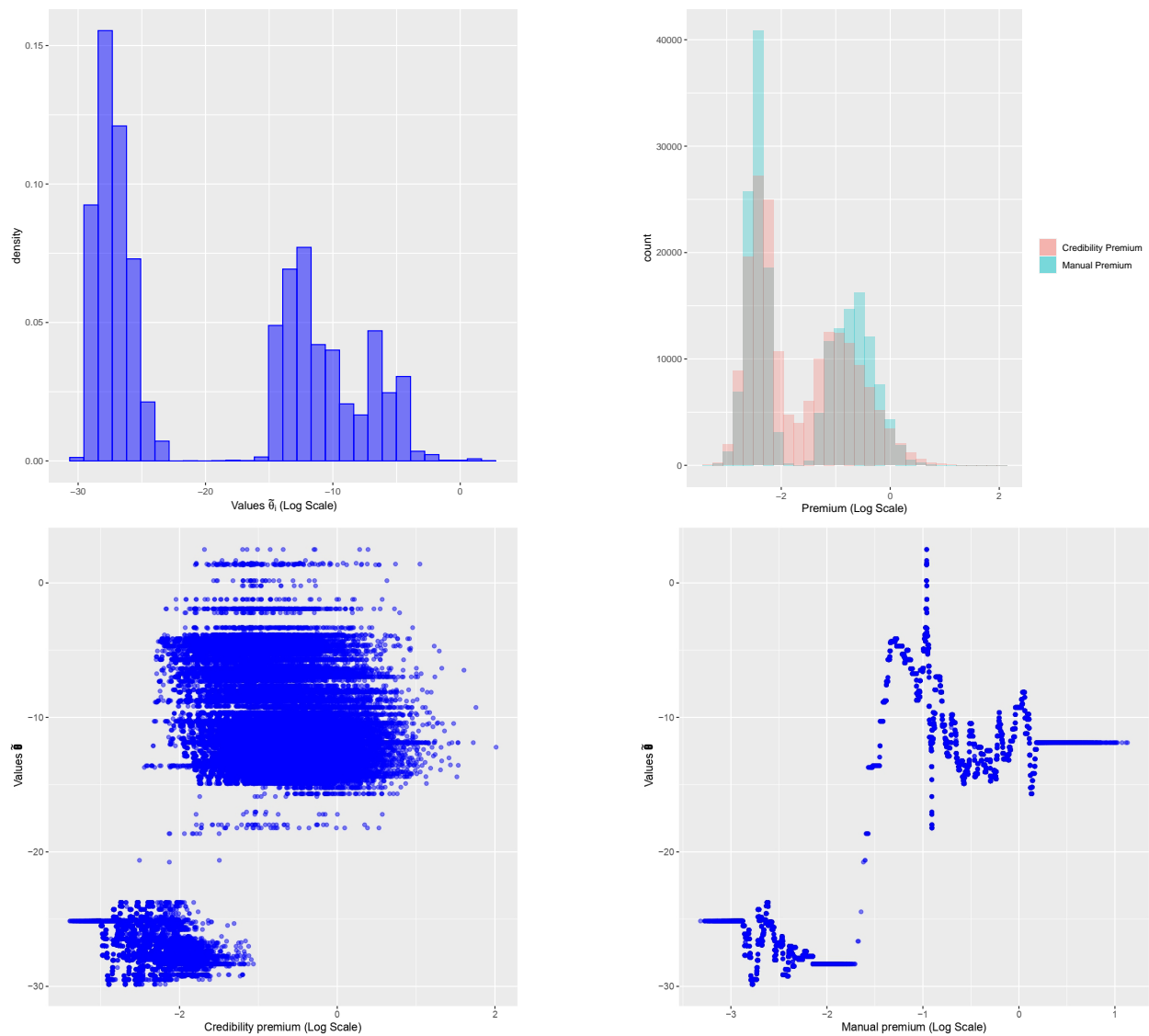
Figure 2: Estimated values of $\tilde{\theta}_i$. Left Histogram /Right Dispersion plot against manual premium.
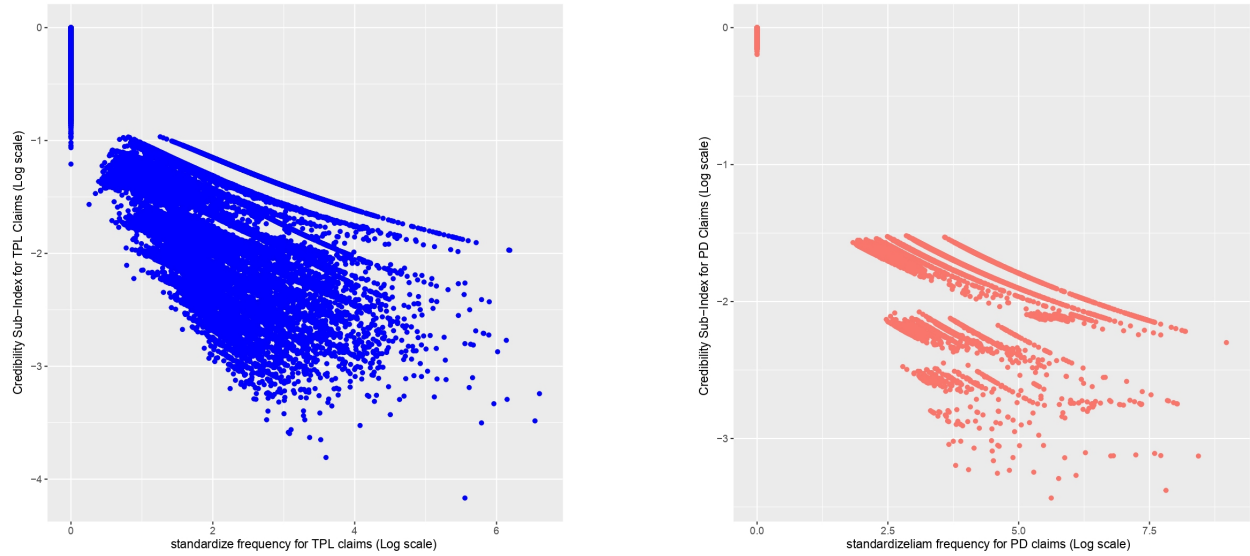
Figure 3: Credibility sub-indexes vs standardized claim frequency. Left PD claims / Right TPL claims.

tion behaves similar to a negative exponential distribution. Note that the peak is around 0 which indicates that the majority of the policyholders tend to have a "low risk" behavior (this is clear as most of them have 0 claims), while relative few tend to have a "risky behavior" as observed in the left tail. Therefore, insurance companies can use the credibility index as measure to assess the observed riskiness of the portfolio vs what was currently expected. Indeed, a distribution of the index that is concentrated towards 0 can be associated to a level of premiums that are not underestimating the risk of the portfolio and therefore there is a good risk assessment of the portfolio, while a distribution with long left-tails is an indicator of an underestimation of the risk of the portfolio that requires attention.

The estimated functions $g_k(\cdot)$ are shown in Figure 5. We obtain a clear interpretation in this context: the greater the credibility indexes, the less is the estimated credibility premium for a particular policyholder. This is consistent with our aforementioned interpretation of the credibility sub-indexes based on Figure 3. In this particular example, for the function $\exp(g_1(\cdot))$, the null-effect is obtained at a value of $\mathcal{L}_i^{(1)}(\tilde{\theta}_i)$ around -13, and around -15 for a value of $\mathcal{L}_i^{(2)}(\tilde{\theta}_i)$ for the function $\exp(g_2(\cdot))$. That is, policyholders with a value less than -13 for the credibility sub-index $\mathcal{L}_i^{(1)}(\tilde{\theta}_i)$ have an associated experience of the number of PD claims to be likely above the expected one, and so must have a revised premium that is larger than the manual premium. Analogously, policyholders with a value larger than -13 for the index $\mathcal{L}_i^{(1)}(\tilde{\theta}_i)$ have an associated experience on the number of PD claims to be below the expected one, and so can have a revised premium that is lower than the manual premium. The same logic applies respectively to the function $\exp(g_2(\cdot))$ with the number of TPL claims and threshold of -15. Similarly, observe that the range of the estimated effect is larger for the function $\exp(g_2(\cdot))$ (i.e from 0.3 to 1.6) than for the function $\exp(g_1(\cdot))$ (i.e from 0.5 to 1.4), thus the experience on the TPL number of claims is more relevant for the determination of the credibility premiums than the experience in the number of PD claims, in the sense that the former can produce a larger deviation in the resulting premium. This is intuitive as TPL claims have a considerable lower frequency than PD claims (see Table 4), and therefore having a TPL claim would impact more on the credibility premium than a PD claim. The two plots
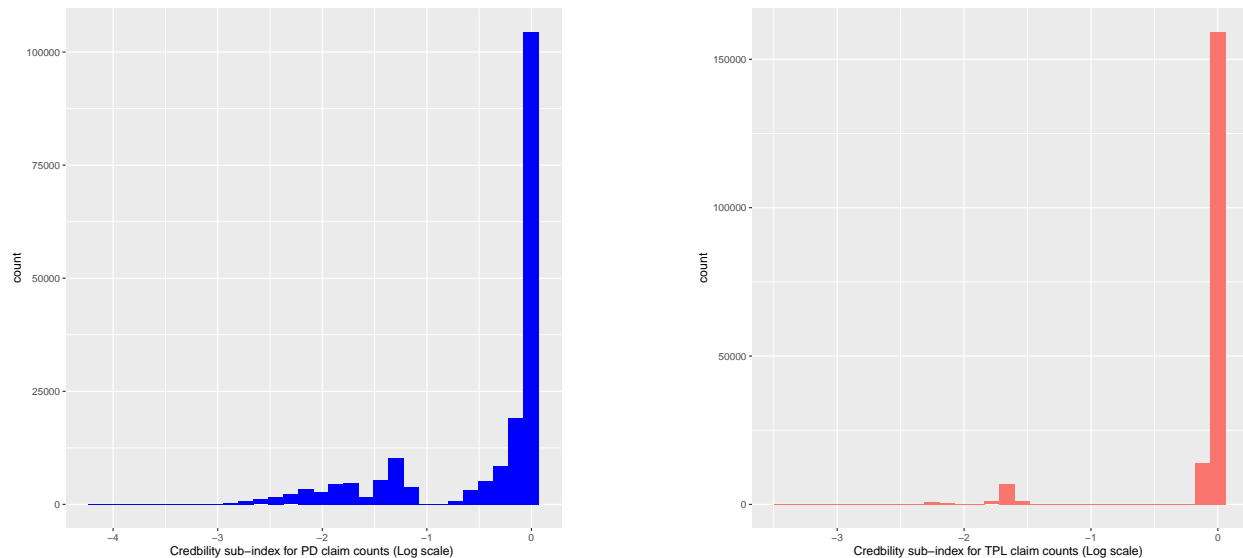
25

Figure 4: Distribution of the credibility sub-indexes. Left is for PD number of claims and Right for TPL number of claims

at the bottom of Figure 5 show the effect of the number of periods of exposure over the resulting credibility premium. The effect of these quantities is not meant to be interpreted directly, however, we can get some insights from these plots. We can observe that the effect from such quantities tends to stabilize as the number of periods increases, which is intuitive as the credibility premium tends to converge to the true premium of a policyholder. Finally, based on the ranges defined above for the overall function $g(\cdot)$, we can use the credibility index to identify three groups of policyholders: less risky than expected, as risky as expected, and more risky than expected. The policyholders falling in the third group could be considered "bad risks" in the sense that their risk is not properly captured in the current premiums and so require extra attention from the remaking process. We do note that this grouping is not aiming to perform a risk classification as it is usually done in ratemaking applications, but rather identifying some riskier than expected policyholders.

8.4 Goodness of fit

We now proceed to assess the quality of our credibility formula as described in Section 6. To do so we have to first verify that our credibility formula has good accuracy at reproducing the credibility premiums, and second we have to verify that it is reliable for extrapolation if only 5% of the portfolio is used (i.e whether 5% is enough or a larger value is necessary),

For the first task, we compare the fitted values of our credibility formula with the credibility premiums obtained form the simulation. The left hand side of Figure 6 shows a dispersion plot of these two premiums. It can be observed that our credibility formula provides premiums that are a good approximation to the credibility premiums since the points lay very close to the 45 degree line. The table on the left hand side of Figure 7 shows some error metrics of the fitted premium vs credibility premiums. We can observe that the interpolation produces an overall value of the coefficient of determination $R^2 = 0.948$, which means that the credibility formula reproduces 95% of the premiums. Moreover, the fitted premiums deviate from the credibility premium only on an average of 3.5%. Similarly the mean error of interpolation is almost 0 implying that there is no bias in the resulting estimation. On another hand, from the perspective of the insurance company,
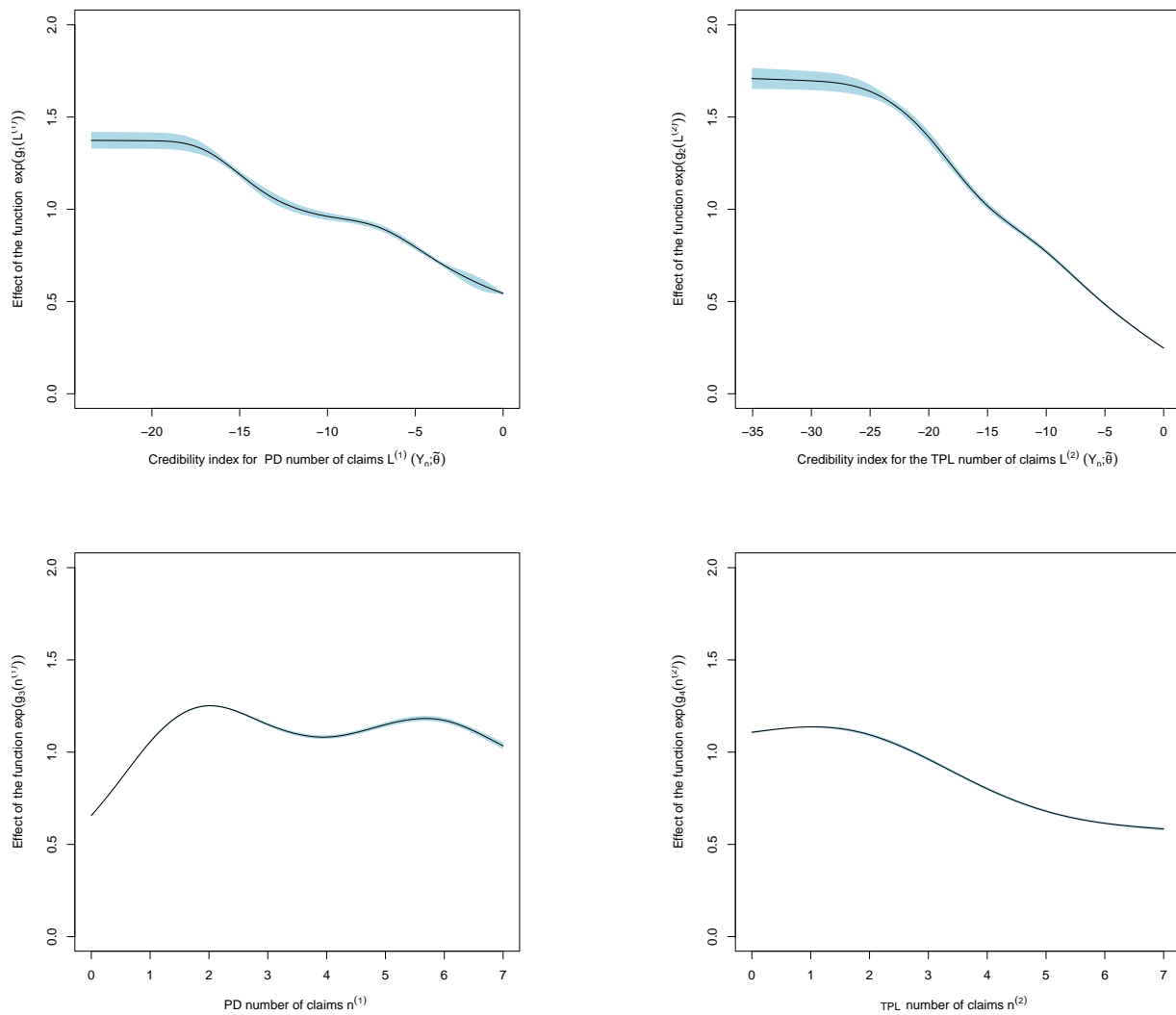
26

Figure 5: Estimation of the functions $g(\cdot)$. Top-left is $g_1(\cdot)$, Top-right is $g_2(\cdot)$, Bottom-left is $g_3(\cdot)$ and Bottom-right is $g_4(\cdot)$

it is desirable that the distribution of the fitted premiums resulting from the credibility formula to behave similar to that obtained from the credibility premiums. Right hand side of Figure 6 shows a QQ plot comparing these two distributions, and it can be observed that the distribution of the fitted premiums is essentially the same to that of the credibility premiums. Therefore, the credibility formula provides an accurate approximation of credibility premiums, and the insurance company can rely on the fitted premiums without having much differences in further portfolio level metrics, such as total premium earned across the portfolio, loss ratios and others.
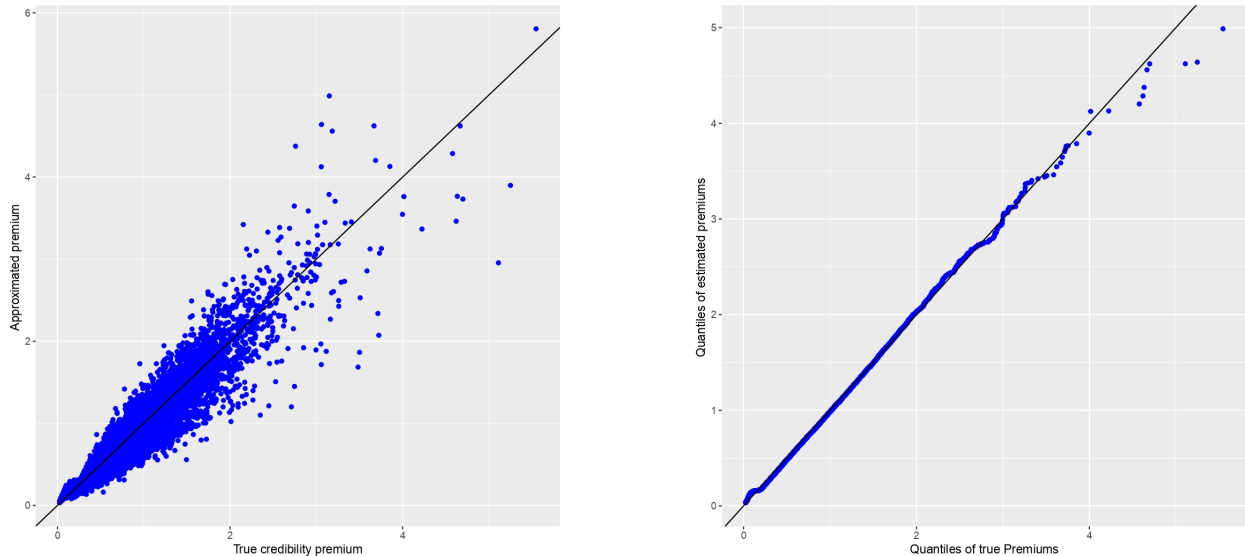

Figure 6: Comparison of approximated premiums vs true premiums. Left dispersion plot and Right QQ-plot.

For the second task, we need to evaluate the "out of sample" behavior of our credibility formula in a test data set, and guarantee that such behavior is the same as in the train data set using "in sample " policyholders. In this particular case of application, we computed the premiums for the 100% of the portfolio, so we can use the 95% that was not in the sample as our test data. The table on the left hand side of Figure 7 shows the error metrics for both the "in sample" and "out of sample" sets of policyholders. Rather than looking at whether these are desirable values for the error metrics as we did before for the first task, we only need to note that the error metrics for the "out of sample" set of policyholders are essentially the same as those that were selected in the sub-portfolio (i.e "in sample"). Therefore the use of only 5% of the portfolio is enough for the credibility formula to consistently extrapolate its predictive power to the 100% portfolio. To assess this claim even further, we also estimate our credibility formula using the 100% of portfolio instead of just the sample of 5%, and compare the two estimations. Indeed, if the fitted values from the two estimations are similar, then the extrapolation using only 5% is reliable. The dispersion plot on the right hand side of Figure 7 compares such two sets of fitted values. It can be observed that the premiums obtained using the selected policyholders are almost the same to the ones using the whole portfolio. Therefore, our credibility formula provides an effective surrogate model to estimate credibility premiums for the portfolio even when only 5% was used for its fitting.

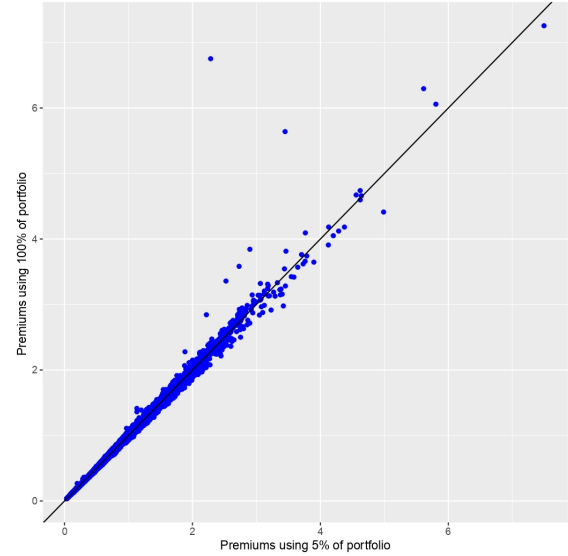| Sub-portfolio	R^2	ME	MAE	MAPE
Out of Sample	0.948	0.0086	0.0351	0.14
In Sample	0.955	0.0088	0.0349	0.14

Figure 7: Assessment of the accuracy of the surrogate model. Left Error Metrics and Right Premiums using 5% vs 100% of the portfolio

8.5 An example with two policyholders

In this subsection we illustrate how the methodology is applied for two policyholders that have similar risk behavior (as quantified by their attributes), but have different claim history. These two policyholders have almost the same predicted mean number of claims, when fitted by the given model, and have a claim history for a total of $n = 4$ years with the insurance company. Their individual attributes are given below.

Policyholder	CarWeight	EngDisp	CarAge	Age	EngPow	Fuel type	Expected Claims
A	1515	1248	1	82	63	Gasoline	0.245 (per year)
B	1475	1360	6	72	55	Gasoline	0.246 (per year)

Table 6: Covariates of the Policyholders

Policyholder A has no record of any claims during those for 4 years, neither in PD nor in TPL, while Policyholder B has two claims, one for PD and one in TPL. Therefore, we can conclude that policyholder A is not as risky as the policyholder B is. Intuitively, the premium should decrease for policyholder A, as there are no claims observed in the period of 4 years, when originally about 4*0.25 = 1 claim is expected on average. On the other hand, the premium should increase considerably for policyholder B as the total number of claims is twice the one expected from the model on such period of time (i.e 2 claims vs 4*0.25=1 claim).

With that in mind, we proceed to further analyze how the credibility index and our credibility formula reflect these two risk behaviors. The credibility index is presented in the Table 7. Observe that the value of $\tilde{\theta}$ is essentially the same for both policyholders as these have essentially the same expected number of claims, as estimated by the mixed Negative Binomial model. However, note the difference in the magnitude of the credibility sub-indexes. Indeed, for policyholder A, which has no claims, the indexes display a relatively minimal value closer to 0, which is associated to a claim history that is less risky than expected, according to our interpretation of Figure 5. On the other hand, for policyholder B, which has several claims, the sub indexes display the opposite

29

behavior with a large negative value which is associated to a claim history that is more risky than expected. Along those lines, the index is indeed capturing the different nature of these two claim histories for these policyholders.

Policyholder	$\tilde{\theta}$	$\mathcal{L}^{(1)}(\mathbf{Y}_n; \tilde{\theta})$	$\mathcal{L}^{(2)}(\mathbf{Y}_n; \tilde{\theta})$
A	$1.61 * 10^{-4}$	$-2.11 * 10^{-5}$	$-1.27 * 10^{-4}$
B	$1.62 * 10^{-4}$	-12.1	-10.3

Table 7: Credibility Sub-Indexes for Policyholders A and B

Now we proceed to look at the estimated credibility premiums in the Table 8. Observe that these are computed as the product of the manual premium times the rating factor given by the credibility formula. The resulting premium behaves consistently with our intuitive analysis. The newly calculated premium will therefore be almost twice as large for policyholder B, while policyholder A will be rewarded with a premium discount, given its no accident history. Note that the claim history is the main driver in the discrepancies in the premiums, and the credibility index is the one capturing such differences.

Policyholder	Manual Premium	$\exp\Big(g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)\Big)$	Credibility Premium
A	0.253	0.854	0.216
B	0.253	1.696	0.429

Table 8: Calculation of credibility premiums

9 Conclusions

Performing accurate experience rating on large insurance portfolios is a challenging task due to two major problems: 1) accounting for the heterogeneity of the policyholders requires flexible, and possibly not mathematically tractable models that can fit complex behaviors in the risk behavior of policyholders and 2) it is necessary to have a great computational power in order to deal with extremely large sized insurance portfolios, especially when no closed-form solutions are available. The first issue can be partially addressed by the use of general Bayesian models beyond the simplistic assumptions commonly used in insurance ratemaking. However, these methods heavily rely on computational techniques which may be affected due to the second problem. Therefore, it is of big importance for actuaries to address effectively the computational issues and to have an effective ratemaking system that is transparent and suited to actuarial standards.

In this paper, we address these aforementioned issues by proposing a methodology in which we compute the credibility premium via the introduction of a credibility index, that is a summary statistics that accounts for how likely it is for a policyholder to experience a certain claim history. The newly proposed credibility formula is obtained as a product between the manual premiums and the actual experience of each policyholder

$$\Pi(Y_{n+1}|\mathbf{Y}_n) = \Pi(Y_{n+1}) \exp\Big(g(\mathcal{L}(\mathbf{Y}_n; \tilde{\theta}), n)\Big).$$

The use of this closed-form expression for the credibility premium enables actuaries to perform ratemaking without the need of relying on cumbersome numerical schemes for any Bayesian model. That is, the actuary will rely on a very efficient formula to approximate credibility premiums for

any policyholder for any given parametric model. In particular, such a closed-form expression can be used along the same lines of surrogate modeling to eliminate the "large portfolio" factor of the equation for ratemaking purposes. Additionally, this expression enables the actuary to provide a transparent picture of the ratemaking process to both clients and regulators, as well as provides a reliable way of performing risk classification among the policyholders.

Future work can explore other setups in which experience rating is defined. For instance, future research should consider the application of the credibility index in the context of evolutionary credibility. So far, the credibility index proposed here considers only the classical approach to credibility in which the seniority of the claims is not of relevance. However, more recent claims may provide a better assessment of the current risk behavior of a policyholder, as the latter may evolve with time. Another future research direction is the construction of other summary statistics that complement the credibility index. Indeed, as motivated in Section 3, the credibility index may not be the only quantity on which the claim history affects the predictive distribution of policyholder. Therefore, it may be possible to improve the accuracy of the credibility formula by considering other type of credibility indexes in the model. Another possible research direction is to consider the development of credibility index that is "non-parametric". The credibility index here proposed is model dependent as rely on a given parametric model, hence it is desirable to develop a credibility index based solely on empirical data-sets.

Finally, it is worth mentioning that the credibility index can find applications in other other areas not related directly to experience rating itself. Indeed, the credibility index can be used on predictive models for claim reserving to account for policyholder's experience. Similarly, the applications of the credibility index can be extended to generality of Bayesian inference. For instance, the idea of a surrogate model for the approximation of expectations can be used in the context of approximate Bayesian computations, particularly in EM algorithms to make them more efficient.

References

Adam, F., Kurnia, A., Purnaba, I., Mangku, I., and Soleh, A. (2021). Modeling the amount of insurance claim using gamma linear mixed model with ar (1) random effect. In *Journal of Physics: Conference Series*, volume 1863, page 012027. IOP Publishing.

Ahn, J. Y., Jeong, H., and Lu, Y. (2021). On the ordering of credibility factors. *Insurance: Mathematics and Economics*, 101:626–638.

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

Blum, M. G., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, 28(2):189–208.

Bühlmann, H. and Gisler, A. (2005). *A course in credibility theory and its applications*, volume 317. Springer.

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Chambers, R. and Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Crevecoeur, J., Robben, J., and Antonio, K. (2022). A hierarchical reserving model for reported non-life insurance claims. *Insurance: Mathematics and Economics*, 104:158–184.

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.

De Vylder, F. and Ballegeer, Y. (1979). A numerical illustration of optimal semilinear credibility. *ASTIN Bulletin: The Journal of the IAA*, 10(2):131–148.

Denuit, M., Maréchal, X., Pitrebois, S., and Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Diao, L. and Weng, C. (2019). Regression tree credibility model. *North American Actuarial Journal*, 23(2):169–196.

Dickson, D. C. M. (2005). *Principles of premium calculation*, page 38–51. International Series on Actuarial Science. Cambridge University Press.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

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

Friedman, J. H. (1994). An overview of predictive learning and function approximation. *From statistics to neural networks*, pages 1–61.

Gómez-Déniz, E. and Calderín-Ojeda, E. (2014). A suitable alternative to the pareto distribution. *Hacettepe Journal of Mathematics and Statistics*, 43(5):843–860.

Jewell, W. S. (1974). Credible means are exact bayesian for exponential families. *ASTIN Bulletin: The Journal of the IAA*, 8(1):77–90.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and bayesian computation. *Statistical applications in genetics and molecular biology*, 7(1).

Kaas, R., Goovaerts, M., Dhaene, J., and Denuit, M. (2008). Premium principles and risk measures. *Modern Actuarial Risk Theory: Using R*, pages 115–134.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. John Wiley & Sons.

Künsch, H. R. (1992). Robust methods for credibility. *ASTIN Bulletin: The Journal of the IAA*, 22(1):33–49.

Li, H., Lu, Y., and Zhu, W. (2021). Dynamic bayesian ratemaking: a markov chain approximation approach. *North American Actuarial Journal*, 25(2):186–205.

Lin, X. S. and Yang, S. (2020a). Efficient dynamic hedging for large variable annuity portfolios with multiple underlying assets. *ASTIN Bulletin: The Journal of the IAA*, 50(3):913–957.

Lin, X. S. and Yang, S. (2020b). Fast and efficient nested simulation for large variable annuity portfolios: A surrogate modeling approach. *Insurance: Mathematics and Economics*, 91:85–103.

Mastroianni, G. and Milovanovic, G. (2008). *Interpolation processes: Basic theory and applications.* Springer Science & Business Media.

Norberg, R. (2004). Credibility theory. *Encyclopedia of Actuarial Science*, 1:398–406.

Ohlsson, E. (2008). Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal*, 2008(4):301–314.

Radtke, M., Schmidt, K. D., and Schnaus, A. (2016). *Handbook on loss reserving.* Springer.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of approximate Bayesian computation.* CRC Press.

Sobester, A., Forrester, A., and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide.* John Wiley & Sons.

Sundt, B. (1979a). On choice of statistics in credibility estimation. *Scandinavian Actuarial Journal*, 1979(2-3):115–123.

Sundt, B. (1979b). On choice of statistics in credibility estimation. *Scandinavian Actuarial Journal*, 1979(2-3):115–123.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., and Dessimoz, C. (2013). Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803.

Taylor, G. C. (1977). Abstract credibility. *Scandinavian Actuarial Journal*, 1977(3):149–168.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: an appraisal. *Survey methodology*, 37(2):215–226.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Tzougas, G. and di Cerchiara, A. P. (2021). The multivariate mixed negative binomial regression model with an application to insurance a posteriori ratemaking. *Insurance: Mathematics and Economics*, 101:602–625.

Virginia, R. (1998). Credibility using a loss function from spline theory. *North American Actuarial Journal*, 2(1):101–111.

Wuthrich, M. V. and Merz, M. (2022). Statistical foundations of actuarial learning and its applications. *Available at SSRN 3822407*.

Xacur, O. A. Q. and Garrido, J. (2018). Bayesian credibility for glms. *Insurance: Mathematics and Economics*, 83:180–189.

Yan, Y. and Song, K.-S. (2022). A general optimal approach to bühlmann credibility theory. *Insurance: Mathematics and Economics*, 104:262–282.

Yau, K., Yip, K., and Yuen, H. (2003). Modelling repeated insurance claim frequency data using the generalized linear mixed model. *Journal of Applied Statistics*, 30(8):857–865.

Zhang, J., Qiu, C., and Wu, X. (2018). Bayesian ratemaking with common effects modeled by mixture of polya tree processes. *Insurance: Mathematics and Economics*, 82:87–94.