Market-based insurance ratemaking

Pierre-Olivier Goffard, Pierrick Piette, Gareth W. Peters

Market-based insurance ratemaking

Pierre-Olivier Goffard [*1], Pierrick Piette[†2,3], and Gareth W. Peters[‡4]

[1]Université de Strasbourg, Institut de Recherche Mathématique Avancée, Strasbourg, France

[2]Univ Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances (ISFA), Laboratoire SAF EA2429, F-69366, Lyon, France

[3]Seyna, 10 Rue du Faubourg Montmartre 75009 Paris

[4]University of California Santa Barbara, Department of Statistics and Applied Probability, Santa Barbara CA 93106-3110, USA

November 21, 2023

Abstract

This paper introduces a novel method for pricing insurance policies using market data. The approach is designed for scenarios in which the insurance company seeks to enter a new market lacking historical data. The methodology involves an iterative two-step process. First, a suitable parameter is proposed to characterize the underlying risk. Second, the resulting pure premium is linked to the observed commercial premium using an isotonic regression model. To validate the method, comprehensive testing is conducted on synthetic data, followed by its application to a dataset of actual pet insurance rates. To facilitate practical implementation, we have developed an R package called `IsoPriceR`. By addressing the challenge of pricing insurance policies in the absence of historical data, this method contributes to enhancing pricing strategies in emerging markets.

MSC 2010: 62P05, 91G70, 62F15.

Keywords: Insurance Pricing, Bayesian Inference, Approximate Bayesian Computation, Isotonic Regression.

1 Introduction

Modern insurance pricing relies on predictive modeling methods to ensure that premiums reflect, as accurately as possible, the average cost of claims. To achieve this, insurers rely on historical data to train statistical

[*]Email: goffard@unistra.fr.

[†]Email: pierrick.piette@gmail.com.

[‡]Email: garethpeters@ucsb.edu.

learning models and calculate what is called the pure premium. While the foundation of standard actuarial practice often rests on generalized linear models (GLM), see Renshaw [12], the relentless evolution of data science has ushered in a new era where more sophisticated machine learning algorithms are also coming into play, see Blier-Wong et al. [3] and the reference therein. However, a challenge arises when an insurance company enters a new market, lacking historical data on the risks it aims to cover. In this context, conventional predictive modeling tools are failing, leaving insurers at a crossroads, looking for innovative solutions to navigate uncharted territory.

Although an insurance company may lack historical data in a new market, there is an attractive alternative: the ability to observe and analyze market data consisting of rates offered by competitors for similar insurance policies. Our approach leverages this market data to provide insights into the underwritten risk leading to the calculation of insurance premiums.

Consider a scenario in which the underlying risk is represented by a positive random variable, denoted as X. This variable represents the total amount of the claim over the period covered by the insurance policy and is linked to a vector of parameters, $\boldsymbol{\theta} \in \mathbb{R}^d$. Our main objective is to determine the parameter values which best explains the market data, denoted as $\mathcal{D}$. Market data comes from various sources. It can be collected from a multitude of quotes from competitors on the market or extracted from insurance aggregator websites. Our approach uses this market data to derive parameter values, thereby providing a market-aligned pricing strategy.

In order to make inference on the risk profile of the loss process, given by $\boldsymbol{\theta}$, we adopt a Bayesian approach in which the parameter is a random variable having a prior distribution $p(\boldsymbol{\theta})$ that we update using the market data to yield the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. Standard Bayesian techniques are not suitable here, for we cannot write the likelihood function of our data. We do not have actual data points drawn from the random variable X, since it is assumed we are in a setting in which we don't have any historical records or that they are insufficient for practical use. Instead, we will assume that what is observed and available for inference is data that consists of commercial rates offered to customers in the market for equivalent or highly similar insurance products. The parameter θ allows us to compute a pure premium p depending on the coverage brought by the insurance policy. Under the expectation principle, the commercial premium π results from applying a loading f to the pure premium p such that

$$\pi = f(p) > p.$$

The loading function $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is unknown and is estimated via an isotonic regression model. We chose this method because the monotonic relationship between pure and commercial premiums is desirable. Also, market

data is inherently noisy and isotonic regression is more robust to outliers than a simple linear regression for example.

The procedure may be summarized as follows

1. Sample a parameter value θ^*
2. Compute the pure premiums p_i^* for each of the insurance policies $i = 1, \ldots, n$
3. Fit an isotonic regression f^* to learn the relationship between the commercial premia π_i and the pure premia p_i^*
4. Build the 'synthetic' market data $\mathcal{D}^*$ by applying the estimated loading f^* to the pure premium $f^*(p_i)$ for $i = 1, \ldots, n$
5. If the observed and synthetic market data are close enough then we store the parameter value θ^* and the associated loading function f^*.

After iterating the above steps, we get a sequence of parameter-loading function pairs: $(\theta_1^*, f_1^*), (\theta_2^*, f_2^*), \ldots$. This sequence equips us with the means to price our own insurance policies. The problem we tackle is an inverse problem and our solution is inspired from indirect inference methodologies pionneered by Gourieroux et al. [8]. The proposed algorithm resembles Approximate Bayesian Computation (ABC) algorithms described in the book of Sisson et al. [13]. ABC algorithms have found successful applications in a range of actuarial science and risk management problems. We refer the readers to the works of Peters et al. [11], Dean et al. [4], Peters and Sisson [10] and Goffard and Laub [7] for further insights. Isotonic regression, a well-established statistical methodology, see for instance Barlow et al. [1], plays a central role in our approach. Its recent application in actuarial science, as demonstrated by Wüthrich and Ziegel [14], addresses the autocalibration challenges that can arise when pricing insurance contracts using machine learning algorithms.

The rest of the paper is organized as follows. Section 2 describes the risk model used in this study and discusses insurance pricing principles. Section 3 provides a detailed account of the algorithmic procedure. Our method is presented as an Approximate Bayesian Computation (ABC)-type optimization algorithm, which incorporates a simple isotonic regression model. Section 4 presents the results of a simulation study designed to showcase the performance of our method in a controlled environment. Lastly, we apply our algorithm on a dataset made of real-world pet insurance rates in Section 5.

2 Model set up and insurance premium computation

An individual seeks to hedge against a risk X modeled by a positive random variable, over a given period of time, say one year. A common model used for X in property and casualty insurance is given by a compound loss variable

$$X = \sum_{k=1}^{N} U_k,$$

where N is a counting random variable and the U_k's are independent and identically distributed (IID) positive random variables independent from N. The random variable N is the number of occurrences of an event over a given time period (annually), each of these events is associated to a compensation U_k.

2.1 Pure premium computation

An insurance company offers to bear part of this risk $g(X) \leq X$ in exchange for a premium which should compensate the average cost of claim given by

$$p = \mathbb{E}[g(X)],$$

referred to as the pure premium. We consider in this work a function g defined as

$$g(x) = \min(\max(r \cdot x - d, 0), l),$$

where $r \in (0, 1]$ is the coverage rate, $d > 0$ is the deductible and $l > 0$ is the limit. Let us consider a scenario where the risk has a Poisson-lognormal distribution $X \sim \mathsf{Poisson}(\lambda = 3) - \mathsf{LogNorm}(\mu = 0, \sigma = 1)$ and that $n = 100$ insurance coverages are proposed. These are characterized by a rate, a deductible and a limit, set randomly as

$$r_i \sim \mathsf{Unif}([0.5, 1]),\ d_i \sim \mathsf{Unif}([0.5, 6]), \text{ and } l = \infty, \text{ for } i = 1, \ldots, 100.$$

Figure 1 shows the pure premiums

$$p_i = \mathbb{E}(g_i(X)) = \mathbb{E}(\min(\max(r_i \cdot x - d_i, 0), l_i)),\ i = 1, \ldots, 100. \tag{1}$$

as a function of the rates and deductibles.

The pure premium are increasing in the rates and decreasing in the deductible. Note that the pure premiums were estimated via a Crude Monte Carlo simulation method to overcome the lack of explicit formula for the distribution functions of X. In practice, the rate offered to policyholders include a loading to compensate for the variability of the risk and cover the management costs. We describe this loading in the next section.

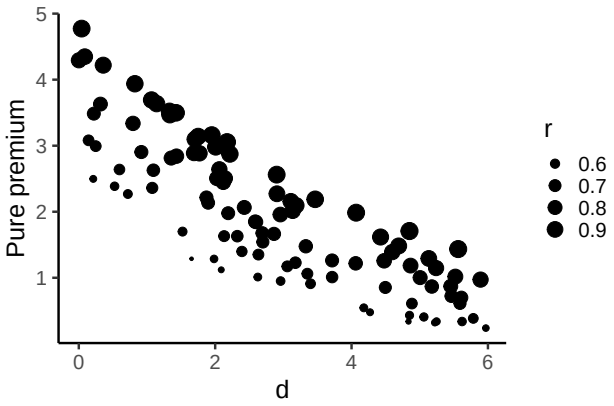


Figure 1: Pure premiums as a function of the rate of coverage (r) and the deductible (d) for a Poisson($\lambda = 3$) – LogNorm($\mu = 0, \sigma = 1$) risk.

2.2 From pure premiums to commercial premiums

Let $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a nondecreasing function, such that

$$\pi = f(p) \geq p.$$

The function f is referred to as the loading function. As the commercial premium is a function of the pure premium then we are applying the expectation premium principle. Other premium principle are also possible like the standard deviation principle discussed in Remark 2.2. A simple loading function is linear in the pure premium as

$$f(x) = (1 + \eta)x,$$

where $\eta > 0$. The loading function used by insurance companies is unknown to us. For instance, take the pure premiums in (1) and apply the following linear loadings

$$\eta_i \sim \text{Unif}([0.5, 2]), \text{ for } i = 1, \ldots, n.$$

The commercial premium then relates to the pure premium as

$$\pi_i = (1 + \eta_i)p_i, \text{ for } i = 1, \ldots, n. \tag{2}$$

Figure 2 displays the commercial premium as a function of the pure premium. In our problem we only observe the commercial premium. When considering a candidate parameter θ of the risk, we can compute the pure premiums but we need to link them to commercial premiums. For that we use isotonic regression which we describe hereafter.

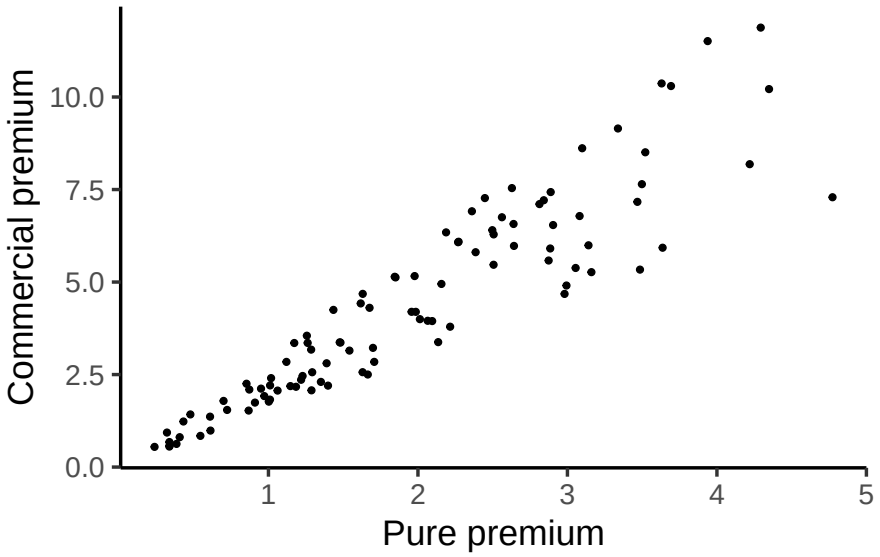


Figure 2: Pure premium as a function of the commercial premium offered by various insurance companies.

2.3 Isotonic regression

Isotonic regression is a statistical technique used for fitting a non-decreasing function to a set of data points. Our datapoints are pairs of pure and commercial premiums $(p_i, \pi_i)_{i=1,\dots,n}$. Assume that the pure premium have been ordered such that $p_i \leq p_j$ for $i \leq j$, isotonic regression seeks a least square fit $\widehat{\pi}_i$ for the π_i's such that $\widehat{\pi}_i \leq \widehat{\pi}_j$ for $p_i \leq p_j$. It reduces to find $\widehat{\pi}_1, \dots, \widehat{\pi}_n$ that minimize

$$\sum_{i=1}^{n} (\widehat{\pi}_i - \pi_i)^2, \text{ subject to } \widehat{\pi}_i \leq \widehat{\pi}_j \text{ whenever } p_i \leq p_j.$$

Since the p_i's fall in a totally ordered space, a simple iterative procedure called the Pool Adjacent Violator Algorithm (PAVA) can be used. Here's a high-level overview of how it works:

1. Initialize the sequence of values to be the same as the data points $\pi_i^* = \pi_i$.

2. Iterate through the sequence and identify "violations," which occur when the current value is greater than the next value, that is

$$\pi_i^* > \pi_{i+1}^* \text{ for some } i = 1, \dots, n.$$

When a violation is found, adjust the values in the associated segment of the sequence to be the average

of the values,

$$\pi_i^* \leftarrow (\pi_i^* + \pi_{i+1}^*)/2,$$

ensuring monotonicity.

3. Repeat Step 2 until no violations are left.

We use the `isoreg` function from *R* to get the fitted values $\widehat{\pi}_i$, $i = 1, \ldots, n$. To complete the isotonic regression task we shall find a function f such that $f(p_i) = \widehat{\pi}_i$. A common choice is a piece-wise constant function that interpolates the $\widehat{\pi}_i$'s. The isotonic fit of the data of Subsections 2.1 and 2.2 is provided on Figure 3.

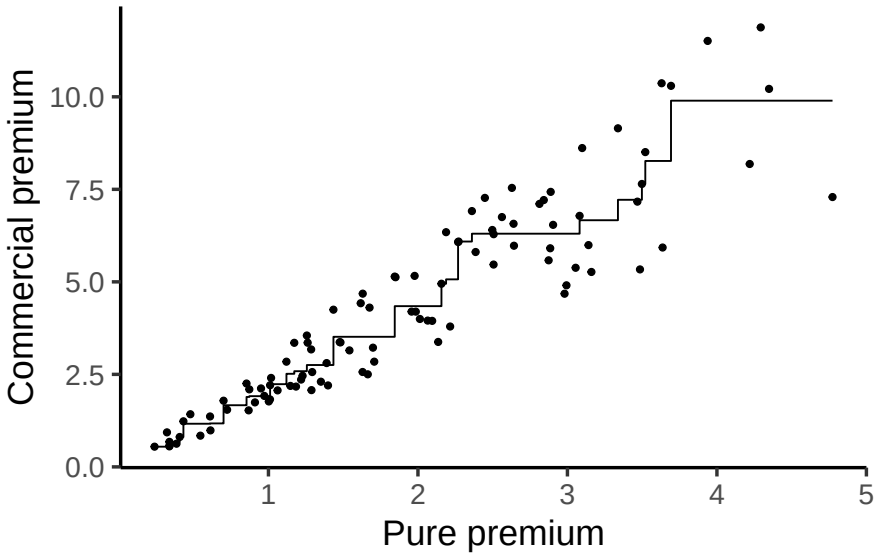


Figure 3: Isotonic link between the pure and commercial premiums.

Remark 2.1. *When looking at Figure 3, one may object that a simple linear regression model could do the job. This impression is partly due to the linear link between pure and commercial premium in (2). Isotonic regression is a non-parametric approach, meaning it doesn't make strong assumptions about the underlying distribution or functional form of the relationship between variables. This can be advantageous when the true relationship is not well represented by a linear model. Furthermore, Isotonic regression is generally more resistant to outliers compared to linear regression due to its piecewise constant nature and the way it enforces monotonicity.*

Remark 2.2. *Our presentation up to now focuses on the expectation premium principle as we try to inform the link f between the commercial premium π and the pure premium $p = \mathbb{E}[g(X)]$. Other premium principles such as the*

standard deviation principle can be considered by slightly adapting the method. Under such principle we have

$$\pi = \mathbb{E}[g(X)] + \eta \cdot \sqrt{\mathbb{V}[g(X)]}. \tag{3}$$

Dividing by $\mathbb{E}[g(X)]$ *on both sides yields*

$$\frac{\pi}{\mathbb{E}[g(X)]} = 1 + \eta \cdot CV[g(X)],$$

where CV is the coefficient of variation. The ratio $\pi/\mathbb{E}[g(X)]$ *is the inverse of the so called loss ratio in practice. It is a widely used indicator of the profitability of an insurance portfolio. The isotonic regression in this context is used to learn the link* f *between the inverse loss ratio and the coefficient of variation as*

$$LR^{-1} = f\{CV[g(X)]\}.$$

More sophisticated premium principles such as the Escher principle or the utility indifference principle are also possible. Premium principles are described at length in actuarial science textbooks such as Dickson [6]. Considering a premium principle instead of another will change the estimation of the risk parameter. It can be cast as a specific instance of model misspecification.

We are now equipped to go over the algorithm to find the parameters θ of the risk X consistent with our commercial rates $\pi_1, \ldots, \pi_n$ of which the isotonic regression is a key ingredient.

3 Market derived insurance ratemaking

The data at hand is a collection of insurance rates $\mathcal{D} = \{\pi_1, \ldots, \pi_n\}$. We assume that these rates were obtained through the following formula

$$\pi_i = f_i\{\mathbb{E}_\theta[g_i(X)]\},\ i = 1, \ldots, n,$$

where the loading function f_i is unknown, the insurance coverage g_i for policy i is known and the risk X is parametrized by an unknown parameter $\theta \in \Theta \subset \mathbb{R}^d$. Our solution alternates between proposing parameter values for the risk and approximating the f_i's using isotonic regression. The learning algorithm is similar to that of ABC algorithms, we simply refine the procedure laid out in the introduction to get an approximation of the posterior distribution $p(\theta|\mathcal{D})$.

We start by setting a prior distribution $p(\theta)$ over the parameter space that we sequentially improve through intermediate distributions characterized by a sequence of tolerance levels $(\epsilon_g)_{g\geq 0}$ that decreases gradually as $\infty = \epsilon_0 > \epsilon_1 > \epsilon_2 > \ldots > 0$. Each intermediate distribution (called a generation) is represented by a cloud of

weighted particles $(\theta_j, w_j)_{j=1,\ldots,K}$. We aproximate each intermediate posterior distribution using a multivariate kernel density estimator (KDE) denoted by $p_{\epsilon_g}(\theta|\mathcal{D})$. The parameters of the algorithm are the number of generations G and the population size K (the number of particles in the cloud).

The algorithm is initialized by setting $\epsilon_0 = \infty$ and $p_{\epsilon_0}(\theta|\mathcal{D}) = p(\theta)$. For generation $g \geq 1$, we hold an intermediate distribution $p_{\epsilon_{g-1}}(\theta|\mathcal{D})$ from which we can sample particles as

$$\theta^* \sim p_{\epsilon_{g-1}}(\theta|\mathcal{D}).$$

We compute the associated pure premium

$$p_i = \mathbb{E}_{\theta^*}[g_i(X)], \text{ for } i = 1,\ldots,n.$$

We wish to ensure that the loss ratios defined by

$$\mathrm{LR}_i = p_i/\pi_i \text{ for } i = 1,\ldots,n,$$

fall in a specific range $\mathrm{LR}_i \in [\mathrm{LR}_{\mathrm{low}}, \mathrm{LR}_{\mathrm{high}}]$. We therefore define the distances

$$d_1^* = \sqrt{\sum_{i=1}^{n}\left[p_i \cdot \mathrm{LR}_{\mathrm{low}}^{-1} - \pi_i\right]_+^2}, \text{ and } d_2^* = \sqrt{\sum_{i=1}^{n}\left[\pi_i - p_i \cdot \mathrm{LR}_{\mathrm{high}}^{-1}\right]_+^2}, \tag{4}$$

where $[x]_+ = \max(x, 0)$ denotes the positive part of x. We then fit the isotonic regression model

$$\pi_i = f(p_i) + e_i, \text{ for } i = 1,\ldots,n,$$

where e_i is an error term that captures the mismatch between the true value of the pure premium and its empirical counterpart estimated by the competitor insurance company using its historical data and the company specific loading function. The commercial premiums based on θ^* and predicted by the isotonic regression model

$$\widehat{\pi}_i = \widehat{f}(p_i), \text{ for } i = 1,\ldots,n,$$

are compared to the observed rate via the distance

$$d_3^* = \sqrt{\sum_{i=1}^{n}(\widehat{\pi}_{i,j} - \pi_i)^2}.$$

that we combine to the distances in (4) to get

$$d^* = d_1^* + d_2^* + d_3^*.$$

If the distance satisfies $d^* < \epsilon_{g-1}$ then we keep the associated particle θ^*. New particles are proposed until we reach N accepted particles denoted by $\theta_1^g, \ldots, \theta_K^g$. We also store the distances $d_1^g, \ldots, d_K^g$. We need to set the next tolerance threshold ϵ_g which is used to calculate the particle weights

$$w_j^g \propto \frac{p(\theta_j^g)}{p_{\epsilon_{g-1}}(\theta)} \mathbb{I}_{d_j^g < \epsilon_{g-1}}, \ j = 1, \ldots, K.$$

The tolerance threshold is chosen so as to maintain a specified effective sample size (ESS) of $K/2$ as in Del Moral et al. [5]. Following Kong et al. [9], the ESS is estimated by $1/\sum_{j=1}^K (w_j^g)^2$. This weighted sample then allow us to update the intermediate distribution as

$$p_{\epsilon_g}(\theta|\mathcal{D}) = \sum_{j=1}^K w_j^g K_H(\theta - \theta_j^g),$$

where K_H is a multivariate KDE with smoothing matrix H. A common choce for the KDE is the multivariate Gaussian kernel with a smoothing marix set to twice the empirical covariance matrix of the cloud of particles $\{\theta_j^g, w_j^g\}$ as in Beaumont et al. [2]. The procedure is summarized in Algorithm 1.

The algorithm's performance is influenced by two key parameters: the number of generations, denoted as G, and the population size, denoted as K. As one would expect, the computational time for the algorithm increases with higher values of both these parameters. Therefore, the choice of suitable values for G and K can be made in consideration of a predetermined computational time budget. An alternative approach to determine the number of generations is to halt the algorithm when the difference between two consecutive tolerance levels falls below a user-defined threshold Δ_ϵ. This adaptive stopping criterion ensures that the algorithm terminates when it reaches a desired level of accuracy. The pure premiums are computed via Monte Carlo simulation. The accuracy depends on the number R of copies of X involved in the Monte Carlo estimations.

In summary, the user must configure several aspects of the algorithm. This includes setting the population size K, determining the number of generations G (or, alternatively, setting Δ_ϵ as the stopping criterion), specifying prior assumptions $p(\theta)$, a corridor of loss ratio to premium $[\mathrm{LR}_{\mathrm{low}}, \mathrm{LR}_{\mathrm{high}}]$ (which can be guided by expert opinions) and selecting the number of Monte Carlo replications R. This last choice has a direct impact on the precision of the calculation of pure premiums.

After the algorithm terminates, it is customary to focus on the last generations of particles for inference. Pointwise estimators are derived from this final set of particles. Two commonly used estimators include the Mean *A Posteriori* (MAP) obtained by averaging the particles in the last cloud and the Mode *A Posteriori* (MODE) which is the mode of the empirical distribution within the final cloud of particles. The forthcoming simulation

Algorithm 1 Population Monte Carlo Approximate Bayesian Computation

set $\epsilon_0 = \infty$ and $p_{\epsilon_0}(\boldsymbol{\theta} \mid \mathcal{D}) = \pi(\boldsymbol{\theta})$
for $g = 1 \rightarrow G$ **do**
  for $j = 1 \rightarrow K$ **do**
    repeat
      generate $\boldsymbol{\theta}^* \sim \widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid \boldsymbol{x})$
      compute $p_i^* = \mathbb{E}_{\theta^*}[g_i(X)]$, for $i = 1, \ldots, n$
      compute $d_1^* = \sqrt{\sum_{i=1}^n \left[p_i \cdot \mathrm{LR}_{\mathrm{low}}^{-1} - \pi_i\right]_+^2}$, and $d_2^* = \sqrt{\sum_{i=1}^n \left[\pi_i - p_i \cdot \mathrm{LR}_{\mathrm{high}}^{-1}\right]_+^2}$
      fit the isotonic regression model $\pi_i = f(p_i^*) + e_i$, for $i = 1, \ldots, n$
      set $\pi_i^* = \widehat{f}(p_i^*)$, for $i = 1, \ldots, n$
      compute $d_3^* = \sqrt{\sum_{i=1}^n (\pi_i^* - \pi_i)^2}$
      compute $d^* = d_1^* + d_2^* + d_3^*$
    until $d^* < \epsilon_g$
    set $\boldsymbol{\theta}_j^g = \boldsymbol{\theta}^*$ and $d_j^g = d^*$
  end for
  find $\epsilon_g \leq \epsilon_{g-1}$ so that $\widehat{\mathrm{ESS}} = \left[\sum_{j=1}^K (w_j^g)^2\right]^{-1} \approx K/2$, where

$$w_j^g \propto \frac{p(\boldsymbol{\theta}_j^g)}{p_{\epsilon_{g-1}}(\boldsymbol{\theta}_j^g \mid \mathcal{D})} \mathbb{I}_{d_j < \epsilon_g}, \quad k = 1, \ldots, K$$

  compute $p_{\epsilon_g}(\boldsymbol{\theta} \mid \mathcal{D}) = \sum_{j=1}^K w_j^g K_{\boldsymbol{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_j^g)$
end for

study, discussed in the following section, is designed to investigate the convergence behavior and to compare the characteristics of the MAP and MODE estimators.

4 Methodology Assessment via Simulation

In this section, we embark on an empirical exploration, seeking to understand how the posterior distribution of the parameters behaves as the sample size n increases. From a theoretical standpoint, it's important to recognize that commercial rates may not capture all the nuances of the underlying claim data. This divergence between the information contained in the commercial rates and the comprehensive claim data implies that the Approximate Bayesian Computation (ABC) posterior does not necessarily align with the true posterior distribution.

This experimentation has been designed to resemble as much as possible the real data situation considered in Section 5. We assume that the risk is given by

$$X = \sum_{i=1}^{N} U_i,$$

where

$$N \sim \mathsf{Poisson}(\lambda = 0.58), \tag{5}$$

and

$$U_i \sim \mathsf{LogNorm}(\mu = 5.75, \sigma = 1),\ i = 1, \ldots, N. \tag{6}$$

The U_i's are IID and independent from N. We suppose that we know the variance parameter σ and we try to draw inference on λ and μ. The parameter values of the claim frequency and severity in (5) and (6) respectively are those infered in Section 5 for the Poisson−LogNorm model using the MAP estimator. The prior distributions are set to independent uniforms for λ and μ as

$$\lambda \sim \mathsf{Unif}([0, 10]),\ \text{and}\ \mu \sim \mathsf{Unif}([-10, 10]).$$

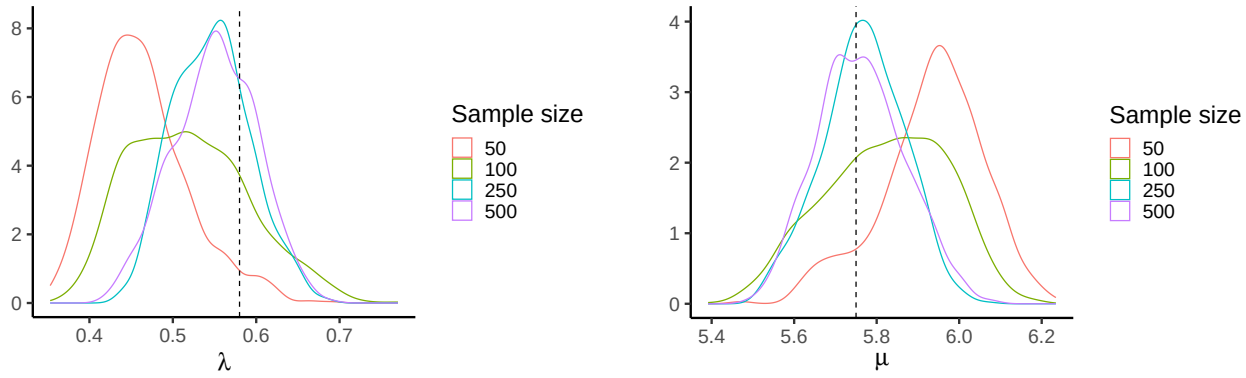
We generate artificial synthetic commercial premiums for this case study according to

$$\pi_i = (1 + \eta_i)\mathbb{E}[g_i(X)] = (1 + \eta_i)\mathbb{E}\{\min[\max(r_i \cdot X - d_i, 0), l_i]\},\ i = 1, \ldots, n,$$

where the premium parameters r, d and l are sampled from that of the real data considered in Section 5, so that the simulated data is as close as possible to the real data. The η_i's are IID from

$$\eta_i \sim \mathsf{Unif}([1.43, 2.5]).$$

which corresponds to loss ratios between 40% and 70%. We further set $\text{LR}_{\text{low}} = 40\%$ and $\text{LR}_{\text{high}} = 70\%$. We consider sample of sizes $50, 100, 250$ and 500. We configure the algorithm with a population size of $1,000$ and use $2,000$ Monte Carlo replications. To ensure the algorithm's efficiency, we set a stopping threshold, requiring that the difference between two consecutive tolerance levels is smaller than 1 for the algorithm to halt. These settings are kept for the analysis of real-world data, as they strike a balanced compromise between accuracy and computing time. We first run our procedure on a single sample of fake data, Figure 4 shows the resulting posterior distributions of λ and μ.



(a) Posterior distributions of λ

(b) Posterior distributions of μ

Figure 4: Posterior distribution of the parameter of the $\text{Poisson}(\lambda = 0.58) - \text{LogNorm}(\mu = 5.75, \sigma = 1)$ model based on synthetic market data of sizes $50, 100, 250$, and 500.

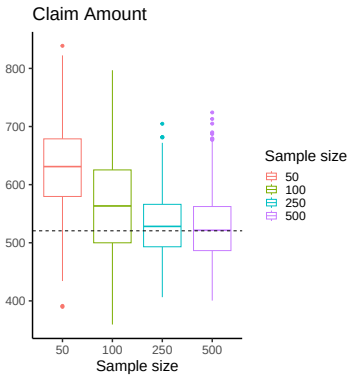
As the sample size increases, the posterior distributions tend to approach the targeted parameter values more closely. Once the parameters have been identified, actuaries often focus on assessing key features of the risk distribution, such as the average claim severity and the probability of no claims. Figure 5 shows the predictive posterior distributions of several metrics, including the average claim amount, the average claim frequency, the probability of no reported claims, the average total claim amount, and the average loss ratio, defined as

$$\overline{\text{LR}} = \frac{1}{n} \sum_{i=1}^{n} \frac{p_i}{\pi_i},$$

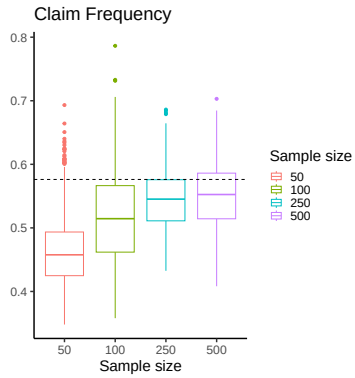
for sample of sizes $50, 100, 250$, and 500.

Irrespective of the sample size, the posterior predictive distribution always include the true value. Note that the true loss ratio is set to the mean of $1/\eta$ where $\eta \sim \text{Unif}([1.43, 2.5])$.

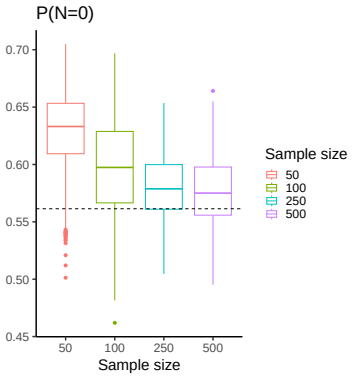
We now repeat 100 identical runs of the experiment. Our goal is to compare the result obtained using our



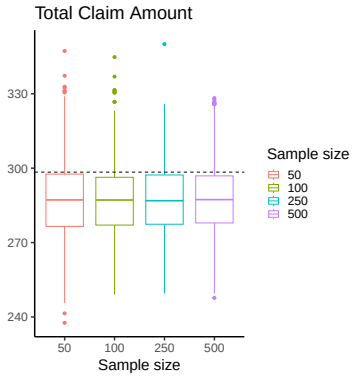
(a) Posterior distribution of $\mathbb{E}(U)$



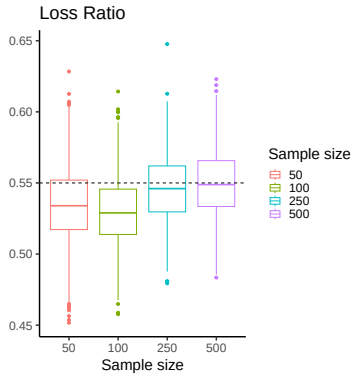
(b) Posterior distribution of $\mathbb{E}(N)$



(c) Posterior distribution of $\mathbb{P}(N = 0)$



(d) Posterior distribution of $\mathbb{E}(X)$



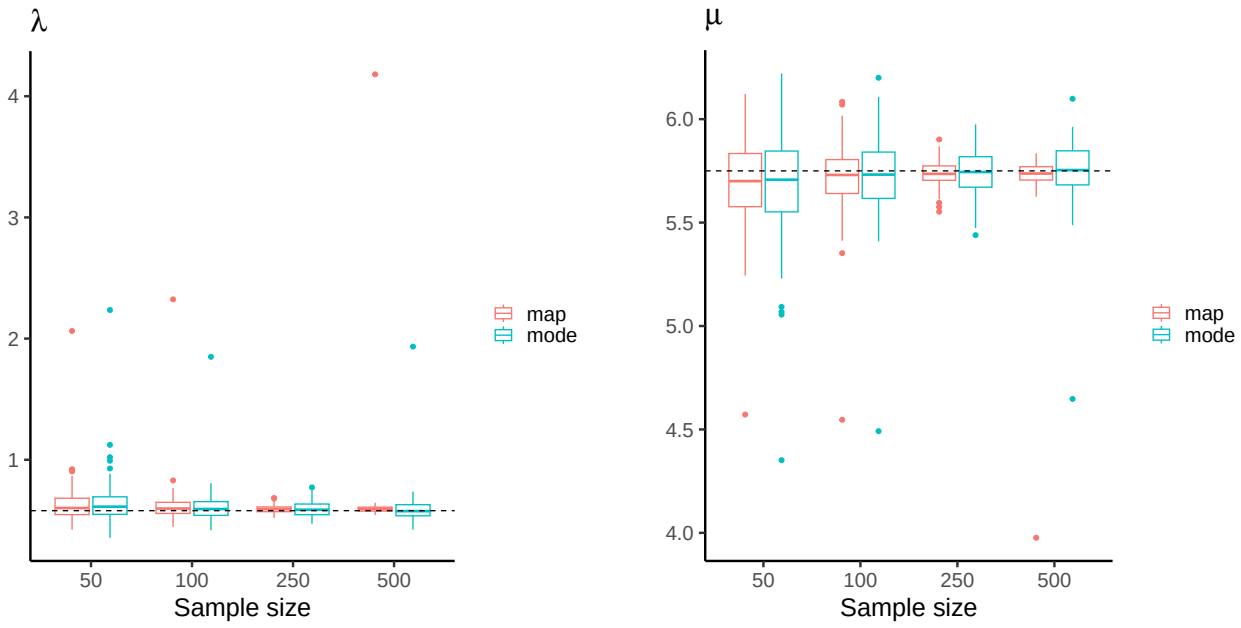
(e) Posterior distribution of LR

Figure 5: Posterior predictive distributions the average claim amount, the average claim frequency, the probability of no reported claims, the average total claim amount and the average loss ratio for the Poisson($\lambda = 0.58$) – LogNorm($\mu = 5.75, \sigma = 1$) model based on synthetic market data of sizes 50, 100, 250, and 500.

two pointwise estimators: the mean *a posteriori* MAP and the mode *a posteriori* MODE. The estimators of the parameters λ and μ are given on Figure 6.

Both of the point-wise estimators seem to converge toward the parameter values that generated the data. The MAP exhibits a better behavior than the MODE as its variability decreases in a notable way as the sample size increases. An increase in the number of particles in the cloud (set to 1,000) would improve the reliability of the posterior distribution and provide a more accurate estimation of the mode

In Figure 7, we present a comparison of key metrics, including the average claim amount, the average claim frequency, the probability of no reported claims, the average total claim amount, and the average loss ratio.



(a) MAP and BP for λ

(b) MAP and BP for μ

Figure 6: MAP and MODE estimators of the parameter of the model $\text{Poisson}(\lambda = 0.58) - \text{LogNorm}(\mu = 5.75, \sigma = 1)$ based on synthetic market data of sizes 50, 100, 250, and 500.

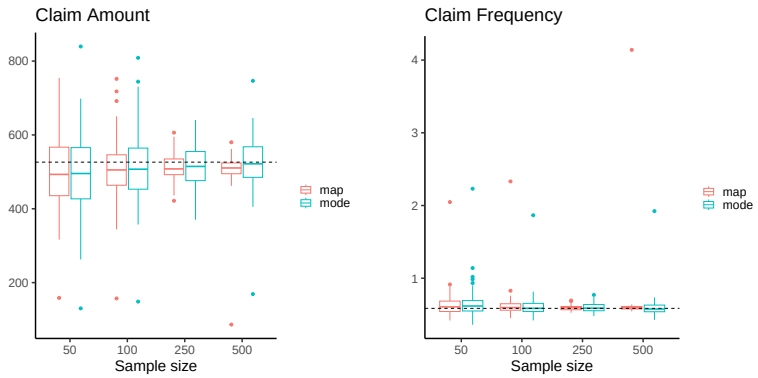
These comparisons are made across different sample sizes, specifically 50, 100, 250 and 500.

Both estimation methods yield satisfactory results in recovering the characteristics of the loss distribution but the use of the MAP yields more reliable estimations.

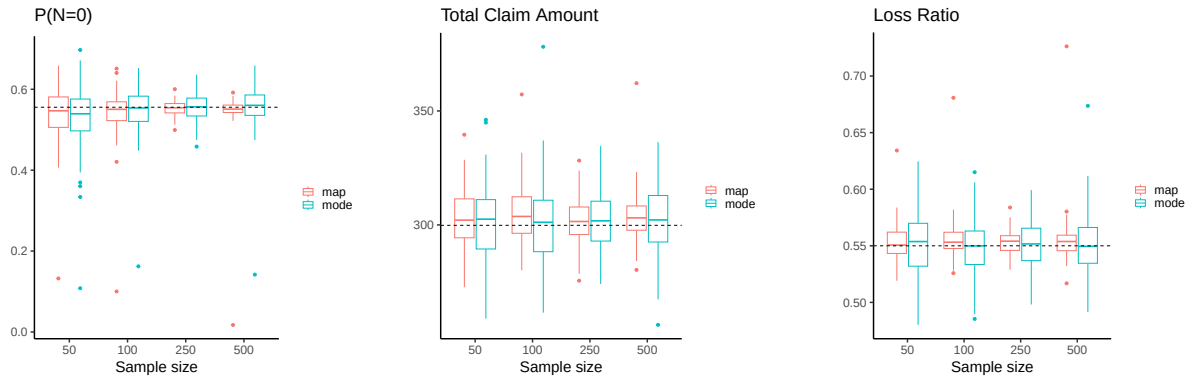
5 Application to the pet insurance market

5.1 Evolution and growth of the pet insurance market

Pet insurance is a product designed to cover the costs of veterinary care for pets. It operates on a similar principle to human health insurance, providing a way for pet owners to manage the financial risks associated with unexpected medical expenses for their animals. Usually the expenses are covered in case of an accident or a disease. Pet owners can choose from different policy options based on their budget and coverage needs. Policies may vary in terms of deductibles (d), coverage limits (l), and reimbursement percentages (r). The cost of premiums can depend on various factors, including the pet's age, breed, health condition, and the level of



(a) Posterior distribution of $\mathbb{E}(U)$ (b) Posterior distribution of $\mathbb{E}(N)$



(c) Posterior distribution of $\mathbb{P}(N = 0)$ (d) Posterior distribution of $\mathbb{E}(X)$ (e) Posterior distribution of LR

Figure 7: MAP and MODE estimator of the features of the $\text{Poisson}(\lambda = 0.58) - \text{LogNorm}(\mu = 5.75, \sigma = 1)$ loss model based on synthetic market data of sizes 50, 100, 250, and 500.

coverage selected.

The pet insurance market has been witnessing significant growth globally, driven by increasing pet ownership (especially with so-called pandemic pets, i.e. animals adopted during 2020 lockdowns), rising veterinary costs and the changing role that a pet plays in a families social structure. This latter factor is also influenced by changing societal views and the increased awareness of the importance of health and welfare of pets, which in turn comes with increased consideration of regular veterinary health checks . In order to offset the cost associated with such expenditures, there has begun to be a broader interest in households purchasing pet insurance.

To date, the adoption and acceptance of pet insurance still varies significantly across regions of the world.

Nordic countries, such as Sweden, have historically had a very high penetration rate with around 70% of pets insured. Some Anglo-Saxon countries (UK and Germany mostly) have seen significant growth in the pet insurance market during the last decades, leading to 30% of penetration rate. Other developed countries, like France, have significantly lower market sizes, with less than 10% of pets that are insured, which suggests high growth potential. The market place for pet insurance in the USA is currently also experiencing sustained growth. According to, MarketWatch guides annual insurance surveys[1], about 44.6%, of pet owners stated they currently have pet insurance in the nationwide survey. Furthermore, the North American Pet Health Insurance Association (NAPHIA) undertook a survey in 2022 on the "State of the Industry Report" and found that more than 4.41 million pets were insured in North America in 2021, up from 3.45 million in 2020. The report also found that $2.84 billion of pet insurance premiums were in force in 2021, a 30.5% increase from 2020.

Primarily, most companies offer pet insurance plans designed to protect dogs and cats. As dog and cat ownership has increased over the last few years, so has the need for pet insurance. As an example in the USA The American Pet Products Association (APPA) found that the increase in pet ownership of dogs and cats between 2017 and 2021 produced and increase in both cat and dog ownership, which further supports the subsequent trends identified by NAPHIA in the following trends in pet insurance uptake demonstrated over time in Table 1.

Year	Cats Insured	Dogs Insured (in millions)
2017	290,000	1.5
2018	348,000	1.8
2019	419,000	2.09
2020	531,000	2.57
2021	727,000	3.25

Table 1: Results on pet insurance uptake for USA from the 2022 survey of the North American Pet Health Insurance Association (NAPHIA) "State of the Industry Report".

This growth continues to spur increases in the capital investments associated with such an insurance line of business:

- in Sweden, Lassie has raised 11m euros in 2022;
- in the UK, ManyPets has raised $350m at a valuation higher than $2bn in 2021;
- in France, Dalma has raised 15m euros in 2022.

[1] https://www.marketwatch.com/guides/pet-insurance/pet-insurance-facts-and-statistics/

Hence, the pet insurance market is becoming more competitive with an increasing number of insurance companies or brokers offering pet insurance policies. To gain new market shares as a new agent, there is a need to propose differentiated products such as new cover mixes without deductible and higher limits.

5.2 Data description

We have collected data on 89 pet health insurance plans offered by various insurers in the French market. These insurance plans are each characterized by specific coverage parameters, including the coverage rate r, the deductible d, and the coverage limit l. The compensation for an annual expense of amount X is calculated as $\min[\max(r \cdot X - d, 0), l]$.

Figure 8 provides a visual overview of the range of insurance coverage options available in the pet insurance market.

We note on Figure 8b that the majority of the insurance coverages do not feature a deductible. We can have a look at the commercial rates offered when $d = 0$ on Figure 9.

As expected, The commercial premiums increase in both the limit and rate of coverage.

5.3 Model fits

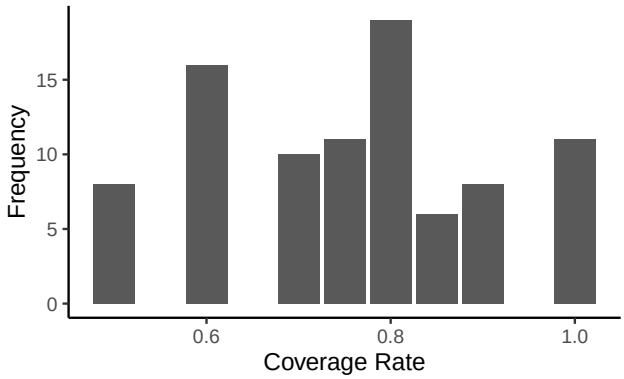
We consider three claim frequency distributions including $\mathsf{Poisson}(\lambda)$, $\mathsf{Bin}(12, p)$ and $\mathsf{Geom}(p)$. The choice of setting the number of trials in the Binomial distribution to 12 aligns with our focus on annual expenses, making it a suitable choice to capture the monthly probability of a claim occurrence. The prior settings for the parameters are as follows:

$$\lambda \sim \mathsf{Unif}([0, 10]),\ p \sim \mathsf{Unif}([0, 1]). \tag{7}$$

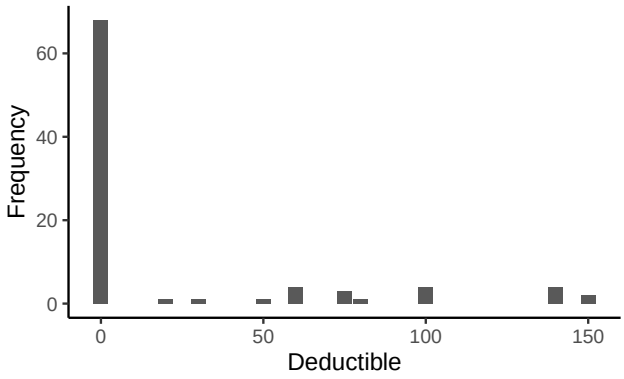
We consider three claim severities distributions including $\mathsf{LogNorm}(\mu = 0, \sigma)$, $\mathsf{LogNorm}(\mu, \sigma = 1)$ and $\mathsf{Gamma}(\alpha, \beta = 1)$. The prior settings over the parameters of the claim sizes distributions are as follows:

$$\mu \sim \mathsf{Unif}([-10, 10]),\ \sigma \sim \mathsf{Unif}([0, 10]),\ \text{and}\ \alpha \sim \mathsf{Unif}([0, 10^5]). \tag{8}$$

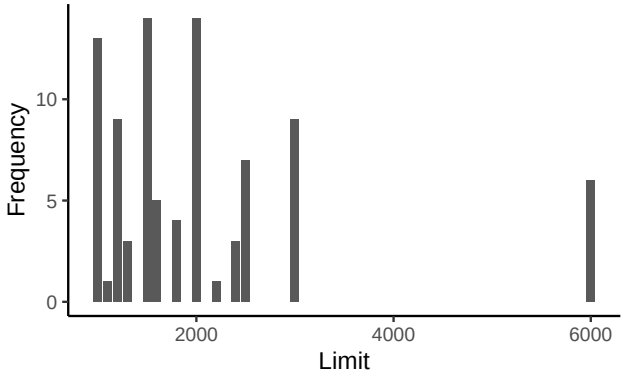
Combining the distributions for the claim frequency and severities makes in total 9 loss models. The population size in the ABC algorithm is set to $N = 1,000$. The pure premiums are computed using $2,000$ Monte Carlo replications. The algorithms stops whenever the difference between two consecutive tolerance levels is lower than $\Delta_\epsilon = 1$. The bounds for the loss ratio corridor are set to $\mathrm{LR}_{\mathrm{low}} = 40\%$ and $\mathrm{LR}_{\mathrm{high}} = 70\%$. The posterior distributions of the parameters for each model models are provided on Figure 10.



(a) Rates of coverage of the insurance policies



(b) Deductible of the insurance policies



(c) Limits of the insurance policies

Figure 8: Overview of the insurance coverage available on the french market.

For all the models, the algorithm updates the prior distribution in an informative way. We note the bimodal posterior distributions for the $\text{Poisson}(\lambda) - \text{Gamma}(\alpha, \beta = 1)$ and $\text{Bin}(12, p) - \text{Gamma}(\alpha, \beta = 1)$ models, see Figure 10g and Figure 10h, which may be problematic if using the MAP as a pointwise estimator. Table 2 provides the tolerance levels (ranked in incresaing order) during the last iteration of the ABC algorithm for the loss models.

The final tolerance level for almost the models lies between 215 and 221 which is higher than the tolerance obtained in the simulation study which was around 55 for 50 data points and 66 for 100 data points. The lack of fit of the $\text{Geom}(p) - \text{LogNorm}(\mu = 0, \sigma)$ is noticeably higher than that of the rest of the models. What we observe here underscores the significant influence of model misspecifications on the accuracy of our analysis. These misspecifications stem from multiple sources, including our assumptions about insurance companies

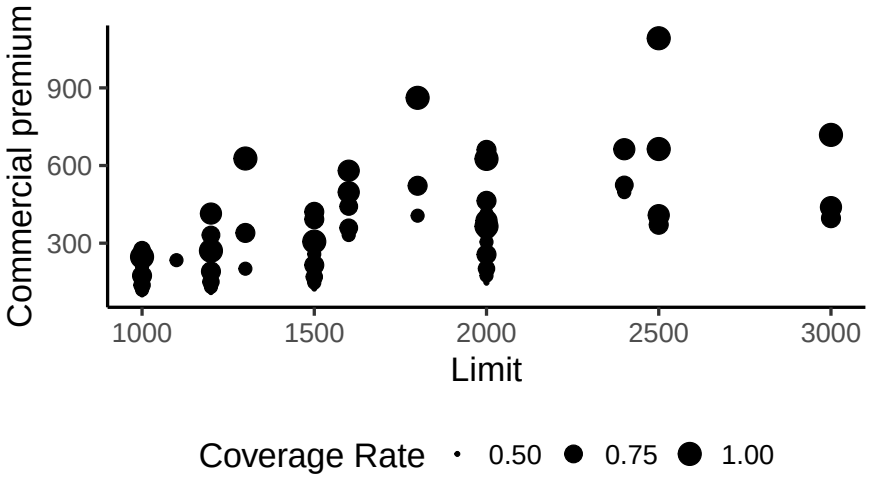


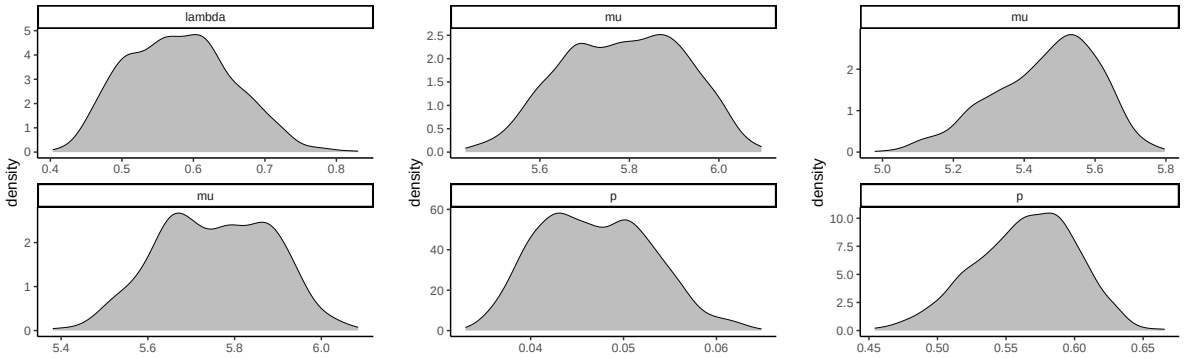
Figure 9: Commercial premiums depending on the rate and limit when no deductible is applied.

Model	ϵ
$\text{Geom}(p) - \text{LogNorm}(\mu, \sigma = 1)$	215.23
$\text{Poisson}(\lambda) - \text{LogNorm}(\mu, \sigma = 1)$	216.10
$\text{Bin}(12, p) - \text{LogNorm}(\mu, \sigma = 1)$	216.24
$\text{Geom}(p) - \text{Gamma}(\alpha, \beta = 1)$	218.92
$\text{Poisson}(\lambda) - \text{LogNorm}(\mu = 0, \sigma)$	219.50
$\text{Bin}(12, p) - \text{LogNorm}(\mu = 0, \sigma)$	219.56
$\text{Bin}(12, p) - \text{Gamma}(\alpha, \beta = 1)$	221.25
$\text{Poisson}(\lambda) - \text{Gamma}(\alpha, \beta = 1)$	221.44
$\text{Geom}(p) - \text{LogNorm}(\mu = 0, \sigma)$	255.03

Table 2: Tolerance level during the last iteration of the ABC algorithm fo each loss model

adhering to the expectation principle for premium calculation and the models employed for claim frequency and claim amounts. Table 3 reports the estimations of the parameters of all the model using the MAP and the MODE.

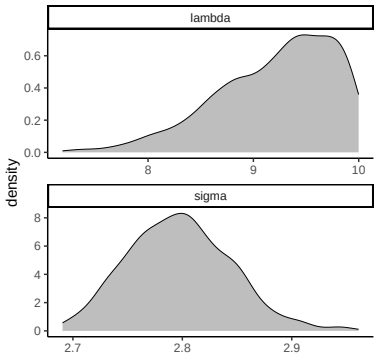
The MAP and MODE estimators continue to exhibit close alignment, albeit to a lesser extent compared to our findings in the simulation study. In the $\text{Poisson}(\lambda) - \text{Gamma}(\alpha, \beta = 1)$ model, the MAP estimator converges toward the first mode of the posterior distribution. However, the MAP estimator for the $\text{Bin}(12, p)$–$\text{Gamma}(\alpha, \beta = 1)$ model positions itself midway between the two modes of the posterior distribution. It's worth highlighting



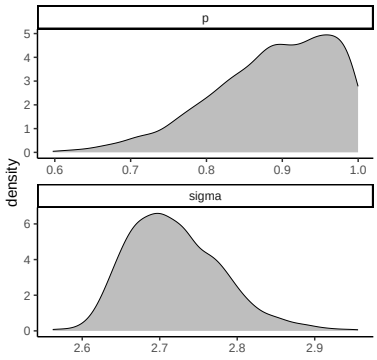
(a) Poisson(λ) – LogNorm($\mu, \sigma = 1$)

(b) Bin($12, p$) – LogNorm($\mu, \sigma = 1$)

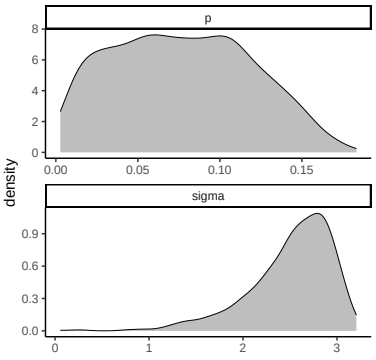
(c) Geom(p) – LogNorm($\mu =, \sigma = 1$)



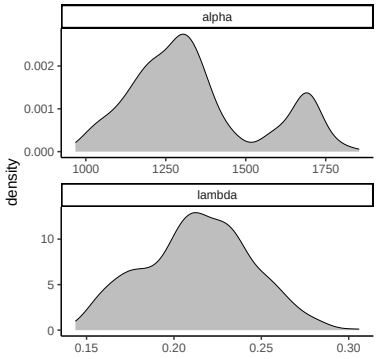
(d) Poisson(λ) – LogNorm($\mu = 0, \sigma$)



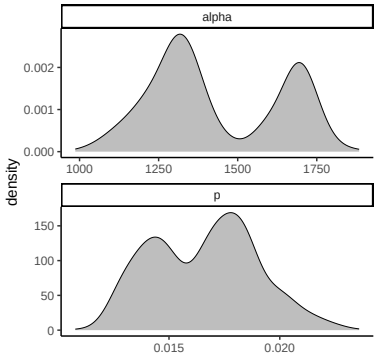
(e) Bin($12, p$) – LogNorm($\mu = 0, \sigma$)



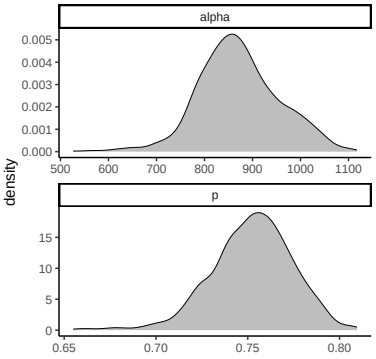
(f) Geom(p) – LogNorm($\mu = 0, \sigma$)



(g) Poisson(λ) – Gamma($\alpha, \beta = 1$)



(h) Bin($12, p$) – Gamma($\alpha, \beta = 1$)



(i) Geom(p) – Gamma($\alpha, \beta = 1$)

Figure 10: Posterior distribution of the parameters of the loss models when fitted to the pet insurance dataset.

Model		MAP	MODE
Poisson(λ) – LogNorm($\mu, \sigma = 1$)	λ	0.58	0.64
	μ	5.75	5.71
Bin($12, p$) – LogNorm($\mu, \sigma = 1$)	p	0.05	0.04
	μ	5.79	5.92
Geom(p) – LogNorm($\mu, \sigma = 1$)	p	0.56	0.55
	μ	5.46	5.43
Poisson(λ) – Gamma($\alpha, \beta = 1$)	λ	0.21	0.26
	α	1341.82	980.25
Bin($12, p$) – Gamma($\alpha, \beta = 1$)	p	0.02	0.02
	α	1432.59	1268.50
Geom(p) – Gamma($\alpha, \beta = 1$)	p	0.75	0.75
	α	871.15	885.37
Poisson(λ) – LogNorm($\mu = 0, \sigma$)	λ	9.21	9.08
	σ	2.80	2.86
Bin($12, p$) – LogNorm($\mu = 0, \sigma$)	p	0.89	0.90
	σ	2.72	2.75
Geom(p) – LogNorm($\mu = 0, \sigma$)	p	0.08	0.01
	σ	2.52	1.64

Table 3: MAP and MODE estimator for the parameters of the loss models.

that the MODE estimation of α in the Poisson(λ) – Gamma($\alpha, \beta = 1$) is lower than one would expect when considering the overall shape of the posterior distribution in Figure 10g.

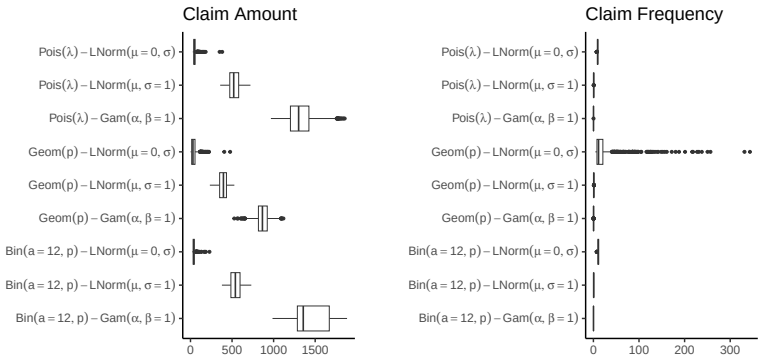
The predictive posterior distributions of the average claim amount, the average claim frequency, the probability of no reported claims, the average total claim amount and the average loss ratio are provided on Figure 11.

Across these models, we note the consistency when estimating the total claim amount, hovering around €300, and the loss ratio, which centers at approximately 70%. However, distinctive perspectives emerge concerning claim frequency and claim sizes, contingent on the choice of model for the latter.

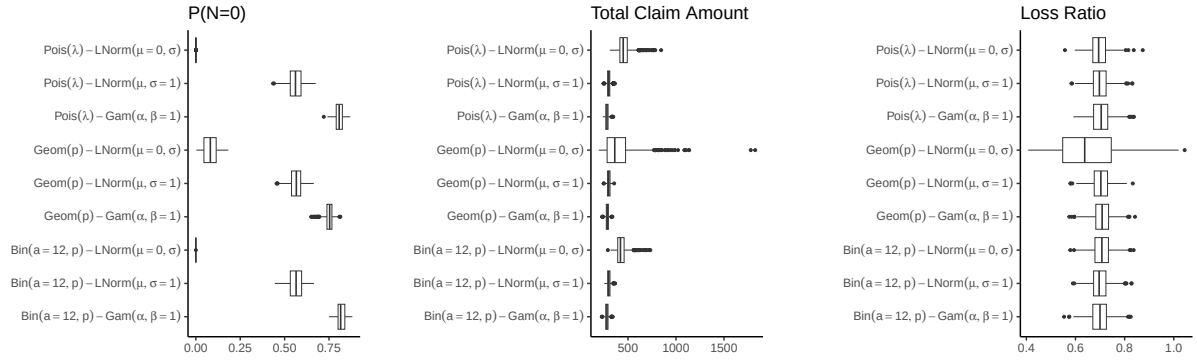
For instance, the use of the Gamma($\alpha, \beta = 1$) distribution results in fewer claims, yet each claim tends to be more severe on average. On the other hand, the Geom(p) – LogNorm($\mu = 0, \sigma$) model stands out by yielding a higher claim frequency, contributing to a more volatile total claim amount and loss ratio.

These observations hold when examining estimations through both the MAP and MODE estimators in Table 4.

Table 5 reports the estimations of the average total claim amounts and the average loss ratio for all the models



(a) Posterior distribution of $\mathbb{E}(U)$ (b) Posterior distribution of $\mathbb{E}(N)$



(c) Posterior distribution of $\mathbb{P}(N = 0)$ (d) Posterior distribution of $\mathbb{E}(X)$ (e) Posterior distribution of LR

Figure 11: Posterior predictive distributions the average claim amount, the average claim frequency, the probability of no reported claims, the average total claim amount and the average loss ratio for the loss models based on the pet insurance market data.

for all models when fitted using the MAP and the MODE.

We estimate the pure premium for each model using the MAP as an estimator of the model parameters and we plot the isotonic regression function for each model to explain the commercial premium on Figure 12.

The $\text{Geom}(p) - \text{LogNorm}(\mu = 0, \sigma)$ model stands out by revealing a lower loss ratio and, consequently, a more pronounced upward trend compared to the other models. To highlight the explanatory power of our methodology, let's focus on the $\text{Poisson}(\lambda)$–$\text{LogNorm}(\mu, \sigma = 1)$ loss model. Note that the choice of the loss model is somewhat arbitrary because the information extracted from the data in Figure 12 is relatively consistent across most of the considered models. In Figure 13, we present a plot that illustrates the relationship between

	$\mathbb{P}(N=0)$		$\mathbb{E}(N)$		$\mathbb{E}(U)$	
Model	MAP	MODE	MAP	MODE	MAP	MODE
Poisson(λ)-LogNorm($\mu, \sigma = 1$)	0.56	0.53	0.58	0.65	519.04	497.03
Bin(12, p)-LogNorm($\mu, \sigma = 1$)	0.56	0.60	0.56	0.50	541.45	625.29
Geom(p)-LogNorm($\mu, \sigma = 1$)	0.56	0.55	0.79	0.83	390.36	380.38
Poisson(λ)-Gamma($\alpha, \beta = 1$)	0.81	0.76	0.22	0.27	1342.06	980.03
Bin(12, p)-Gamma($\alpha, \beta = 1$)	0.82	0.79	0.20	0.23	1432.41	1268.02
Geom(p)-Gamma($\alpha, \beta = 1$)	0.76	0.75	0.34	0.33	871.72	885.33
Poisson(λ)-LogNorm($\mu = 0, \sigma$)	0.00	0.00	9.27	9.04	49.15	43.71
Bin(12, p)-LogNorm($\mu = 0, \sigma$)	0.00	0.00	10.65	10.79	34.72	40.70
Geom(p)-LogNorm($\mu = 0, \sigma$)	0.08	0.01	11.03	90.15	25.30	3.99

Table 4: MAP and MODE estimators of the probability of no claim being reported, the average claim frequency and the average claim amount.

	Loss ratio		$\mathbb{E}(X)$	
Model	MAP	MODE	MAP	MODE
Poisson(λ)-LogNorm($\mu, \sigma = 1$)	0.70	0.75	302.97	320.85
Bin(12, p)-LogNorm($\mu, \sigma = 1$)	0.69	0.71	297.26	308.66
Geom(p)-LogNorm($\mu, \sigma = 1$)	0.71	0.73	301.51	310.84
Poisson(λ)-Gamma($\alpha, \beta = 1$)	0.72	0.68	287.85	266.35
Bin(12, p)-Gamma($\alpha, \beta = 1$)	0.72	0.75	286.72	298.47
Geom(p)-Gamma($\alpha, \beta = 1$)	0.70	0.73	280.51	293.74
Poisson(λ)-LogNorm($\mu = 0, \sigma$)	0.66	0.77	403.51	637.96
Bin(12, p)-LogNorm($\mu = 0, \sigma$)	0.70	0.75	418.14	440.92
Geom(p)-LogNorm($\mu = 0, \sigma$)	0.53	0.86	259.79	348.41

Table 5: MAP and MODE estimators of the average loss ratio and average total claim amounts.

the commercial premium and the pure premium for the Poisson(λ) – LogNorm($\mu, \sigma = 1$) model. Different insurance companies are indicated by distinct colors, providing a visual representation of each company's respective rates.

The accuracy of the loss model fitting enables us to condense the three-dimensional information of the rate of coverage, deductible, and limit into a single metric: the pure premium. Subsequently, isotonic regression unveils the relationship between commercial and pure premiums, providing a link between the two. The distinctions among various players in the pet insurance market come to light through the color-coded points,

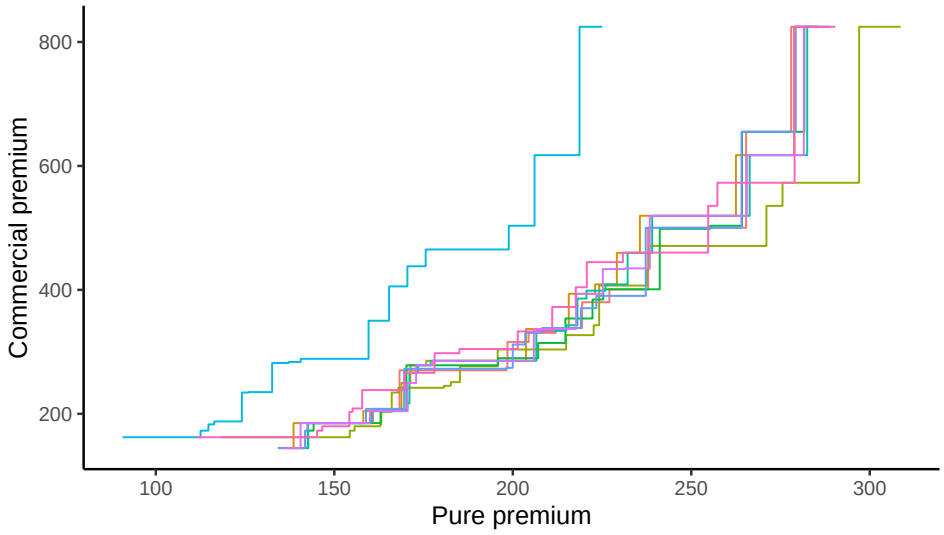


Figure 12: Isotonic link between pure and commercial premium for the different loss models.

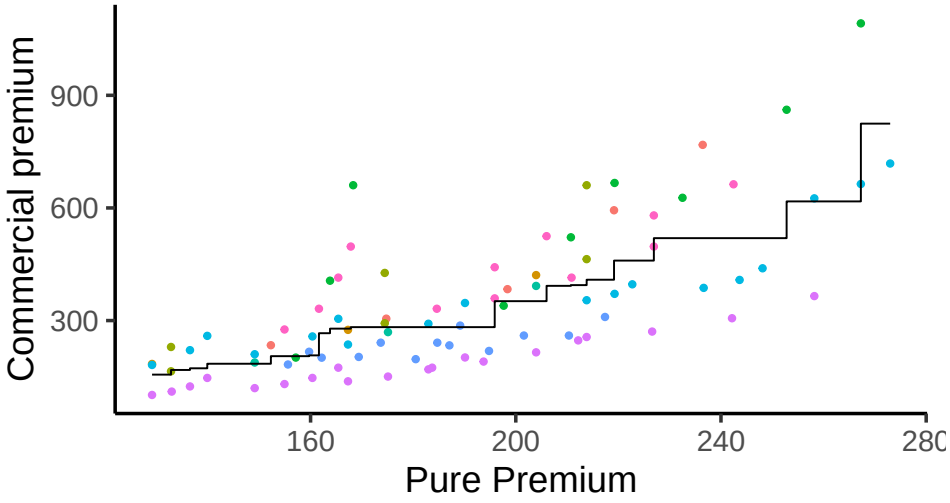


Figure 13: Commercial premium as a function of the pure premium for the $\text{Poisson}(\lambda) - \text{LogNorm}(\mu, \sigma = 1)$ depending on the insurance carrier.

offering insights into the pricing strategies adopted by industry participants.

6 Conclusion

We have developed a robust methodology for risk assessment based on market data. We employ a one-parameter model for the claim frequency and claim size distribution, connecting the pure premium to the commercial premium through an isotonic regression model. This approach optimizes the alignment between commercial and pure premiums while providing a framework for quantifying the associated parameter uncertainty through an Approximate Bayesian Computation algorithm.

The methodology's effectiveness and reliability have been validated within a simulation study and a practical application to an actual pet insurance dataset. This methodology is made accessible to the community through our R package, `IsoPriceR`[2].

While the results are promising, there remain avenues for further research. Future investigations can explore the selection of the most suitable model and consider the integration of historical data when it becomes available. One direction is the development of a credibility framework that combines historical and market data, providing a comprehensive perspective on risk assessment and pricing in emerging markets.

Acknowledgements

Pierre-O's work is conducted within the Research Chair DIALOG under the aegis of the Risk Foundation, an initiative by CNP Assurances.

References

[1] Richard E Barlow, HD Brunk, Daniel J Bartholomew, and James M Bremner. *Statistical inference under order restrictions.(the theory and application of isotonic regression)*. 1972.

[2] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[3] Christopher Blier-Wong, Hélène Cossette, Luc Lamontagne, and Etienne Marceau. Machine learning in p&c insurance: A review for pricing and reserving. *Risks*, 9(1):4, dec 2020. doi: 10.3390/risks9010004.

[4] Thomas A Dean, Sumeetpal S Singh, Ajay Jasra, and Gareth W Peters. Parameter estimation for hidden markov models with intractable likelihoods. *Scandinavian Journal of Statistics*, 41(4):970–987, 2014.

[2] see the market_based_insurance_ratemaking `Github` repository

[5] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and computing*, 22(5):1009–1020, 2012.

[6] D. Dickson. Principles of premium calculation. In *Insurance Risk and Ruin*, pages 38–51. Cambridge University Press, jan 2005. doi: 10.1017/cbo9780511624155.004.

[7] Pierre-Olivier Goffard and Patrick J. Laub. Approximate bayesian computations to fit and compare insurance loss models. *Insurance: Mathematics and Economics*, 100:350–371, sep 2021. doi: 10.1016/j.insmatheco.2021.06.002.

[8] C. Gourieroux, A. Monfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118, dec 1993. doi: 10.1002/jae.3950080507.

[9] Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, mar 1994. doi: 10.1080/01621459.1994.10476469.

[10] Gareth Peters and Scott Sisson. Bayesian inference, monte carlo sampling and operational risk. *Peters GW and Sisson SA (2006)"Bayesian Inference, Monte Carlo Sampling and Operational Risk". Journal of Operational Risk*, 1(3), 2006.

[11] Gareth W Peters, Mario V Wüthrich, and Pavel V Shevchenko. Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance: Mathematics and Economics*, 47(1):36–51, 2010.

[12] Arthur E. Renshaw. Modelling the claims process in the presence of covariates. *ASTIN Bulletin*, 24(2):265–285, 1994. doi: 10.2143/AST.24.2.2005070.

[13] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

[14] Mario V. Wüthrich and Johanna Ziegel. Isotonic recalibration under a low signal-to-noise ratio, 2023.