# Online supplement for "Market-based insurance ratemaking: application to pet insurance"

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]Hestialytics, Paris

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

February 6, 2025

The goal of this supplementary material is to discuss the identifiability issue associated to finding the parameters of the risk model. Let $X$ be a random variable that represents the total health expenses over a given time period, say one year, associated to a pet. Our goal is to find the parameter $\theta$ that best explains our market data made of insurance quotes

$$\widetilde{p}_i = f_i\left[\mathbb{E}(g_i(X))\right],\ i = 1,\ldots,n, \tag{1}$$

where $g_i$ are the coverage functions and $f_i$ are the loading function. The coverage functions are known and of the form

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

where $r$ is the rate of coverage, $d$ is the deductible and $l$ is the limit. The loading functions $f_i$ are unknown and will be approximated by a generic function $f$.

In Section 1, we consider the problem of finding $\theta$ if we know the pure premiums

$$p_i = \left[\mathbb{E}(g_i(X))\right],\ i = 1,\ldots,n,$$

---

[*]Email: goffard@unistra.fr.

[†]Email: pierrick.piette@gmail.com.

[‡]Email: garethpeters@ucsb.edu.

which is not a realistic situation in practice. We look into the actual problem, relying on the commercial premiums (1), in Section 2. An isotonic regression model is used to approximate $f$. This choice is compared to a linear fit $f(x) = b \cdot x$. We consider the model

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

where $N \sim \text{Pois}(\lambda = 3)$ and the $U_k$ are IID and lognormally distributed $U \sim \text{LogNormal}(\mu = 0, \sigma = 1)$. Our aim is to estimate the parameters $\theta = (\begin{array}{ccc} \lambda & \mu & \sigma \end{array})$ based on three insurance quotes associated to the following coverage functions

$$g_1 = 0.85 \cdot X, \, g_2 = \max(X - 1.8, 0), \text{ and } g_3 = \min(X, 6).$$

In terms of coverage rate, deductible and limit, it is equivalent to $(r_1 = 0.85, d_1 = 0, l_1 = \infty)$, $(r_2 = 1, d_2 = 1.8, l_2 = \infty)$, and $(r_3 = 1, d_3 = 0, l_3 = 6)$ respectively. The pure premium associated to these coverages are provided in Table 1.

| r | d | l | Pure premium |
|---|---|---|---|
| 0.75 | 0.00 | Inf | 3.71 |
| 1.00 | 1.80 | Inf | 3.40 |
| 1.00 | 0.00 | 6.00 | 3.63 |

Table 1: Pure premium associated to $g_1, g_2$ and $g_3$ of the $\text{Pois}(\lambda = 3) - \text{LogNormal}(\mu = 0, \sigma = 1)$ risk model.

# 1 Optimization problem 1

The first optimization that we consider reads as follows:

**Problem 1.** *Find $\theta \in \Theta \subset \mathbb{R}^d$ to minimize $d\left[p_{1:n}, p_{1:n}^\theta\right]$, where*

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

*are the pure premium associated to $X$ parametrized by $\theta \in \Theta \subset \mathbb{R}^d$ and $d(\cdot, \cdot)$ denotes a distance function over the observation space.*

Wew measure the discrepancy between observed and model-generated pure premiums using the root mean square error (RMSE) defined as

$$\text{RMSE}\left[p_{1:n}, p_{1:n}^\theta\right] = \sqrt{\sum_{i=1}^{n} w_i^{\text{RMSE}}[p_i - p_i^\theta]^2}. \tag{2}$$

The statistical framework is that of minimum distance estimation. We do not have access to the full shape of the data distribution. We must base our inference on specific moments, just as in the generalized method of moments, a popular method among econometricians (see Hansen [1]). The model is identifiable if there exists a unique $\theta^*$ such that

$$\theta^* = \arg\min_{\theta \in \Theta} \text{RMSE}(p_{1:n}, p_{1:n}^\theta). \tag{3}$$

Existence stems from the fact that the parameter space $\Theta$ is compact and the map $\theta \mapsto \text{RMSE}(p_{1:n}, p_{1:n}^\theta)$ is continuous. Uniqueness is more difficult to verify as it depends on the functional $g_i$'s. Given the model for $X$ and the insurance coverages, the pure premium do not have an analytical expression making it difficult to show the convexity of (2). A simple necesssary condition is that the number of parameters must be smaller than $n$, the number of moments considered. We consider here a situation where we have only three pure premiums ($n = 3$) and we know that

$$\text{RMSE}\left[p_{1:n}, p_{1:n}^{\theta^*}\right] = 0,$$

Figure 1 shows the plot of the functions

$$\lambda \mapsto d\left(p_{1:n}, p_{1:n}^\theta\right)\Big|_{\mu=0, \sigma=1}, \, \mu \mapsto d\left(p_{1:n}, p_{1:n}^\theta\right)\Big|_{\sigma=1, \lambda=3}, \text{ and } \sigma \mapsto d\left(p_{1:n}, p_{1:n}^\theta\right)\Big|_{\mu=0, \lambda=3}.$$

By examining the parameters individually, the model appears to be identifiable. However, this does not imply that a unique parametrization can be determined unless two out of the three parameters are fixed. A grid search procedure is not feasible for exploring the three-dimensional parameter space. Instead, we employ the optimization procedure described in Section 3.3 of our paper. The parameters of the algorithm are configured as follows:

$$J = 500, \quad R = 500, \quad \text{and} \quad \epsilon_{\min} = 0.02.$$

Here, $J$ represents the population size of the clouds of particles, $R$ is the number of Monte Carlo replications, and $\epsilon_{\min}$ is the threshold for the tolerance level. The algorithm terminates once the tolerance threshold reaches $\epsilon_{\min}$. The choice of the threshold $\epsilon_{\min} = 0.02$ is based on an estimation of the error in the RMSE resulting from using a Monte Carlo approximation to compute the pure premium. The prior distributions of the parameters are defined as:

$$\lambda \sim \text{Unif}([0, 10]), \quad \mu \sim \text{Unif}([-3, 3]), \quad \text{and} \quad \sigma \sim \text{Unif}([0, 2]).$$

Figure 2 displays the resulting posterior distributions.

If the pure premium could be calculated exactly, our algorithm would yield Dirac distributions concentrated at the true parameter values. However, since we rely on a Monte Carlo approximation, the use of an Approximate

(a) $\lambda \mapsto d\left(p_{1:n}, p_{1:n}^{\theta}\right)\Big|_{\mu=0,\sigma=1}$

(b) $\mu \mapsto d\left(p_{1:n}, p_{1:n}^{\theta}\right)\Big|_{\sigma=1,\lambda=3}$

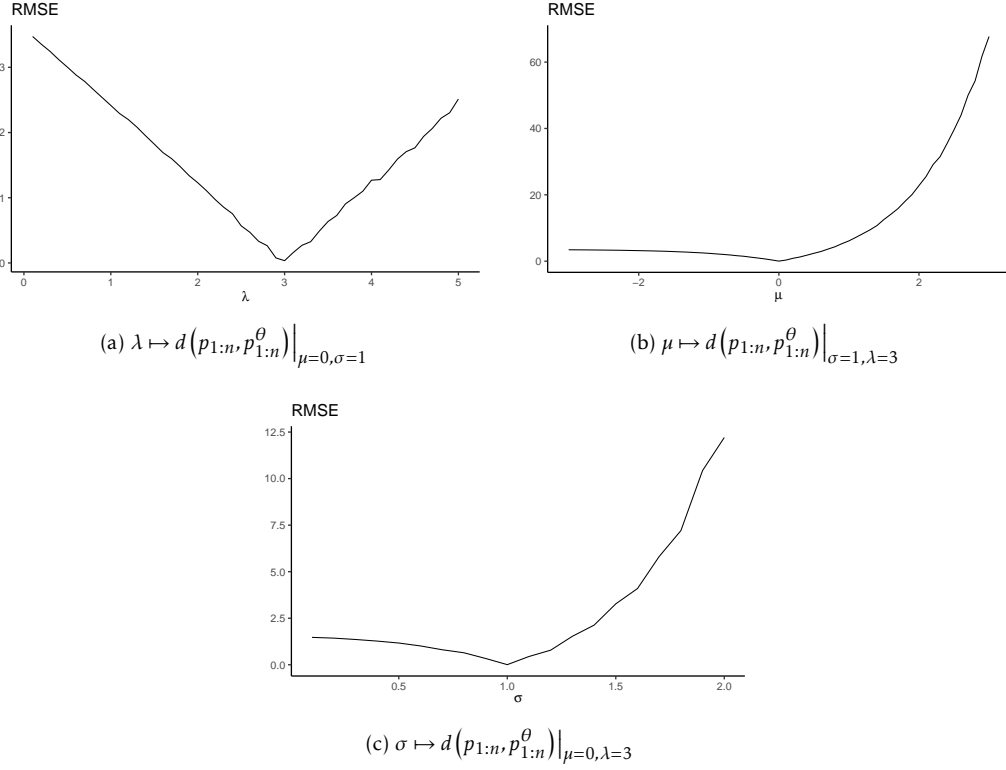(c) $\sigma \mapsto d\left(p_{1:n}, p_{1:n}^{\theta}\right)\Big|_{\mu=0,\lambda=3}$

Figure 1: Plots of the RMSE when looking into each parameter separately.

Bayesian Computation (ABC) procedure is justified, as it incorporates the uncertainty introduced by the Monte Carlo error into the posterior distribution. in view of Figure 2, we conclude that having access to the pure premium enables the identification of the true parameters of the model. The situation becomes more challenging when only the commercial premium is available, as will be discussed in the next section.

## 2 Optimization problem 2

As mentionned earlier, we do not have access to the pure premiums. Instead we have a collection of commercial rates $\widetilde{p}_{1:n} = \{\widetilde{p}_1, \ldots, \widetilde{p}_n\}$ defined as

$$\widetilde{p}_i = f_i(p_i) = f_i\left\{\mathbb{E}_{\theta_0}\left[g_i(X)\right]\right\}, \ i = 1, \ldots, n.$$

It leads us to formulate the following optimization problem:

**Problem 2.** *Find $\theta \in \Theta \subset \mathbb{R}^d$ and $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ to minimize $d\left[\widetilde{p}_{1:n}, f(p_{1:n}^{\theta})\right]$, where the function $f$ is applied*

(a) Posterior distribution of $\lambda$



(b) Posterior distribution of $\mu$
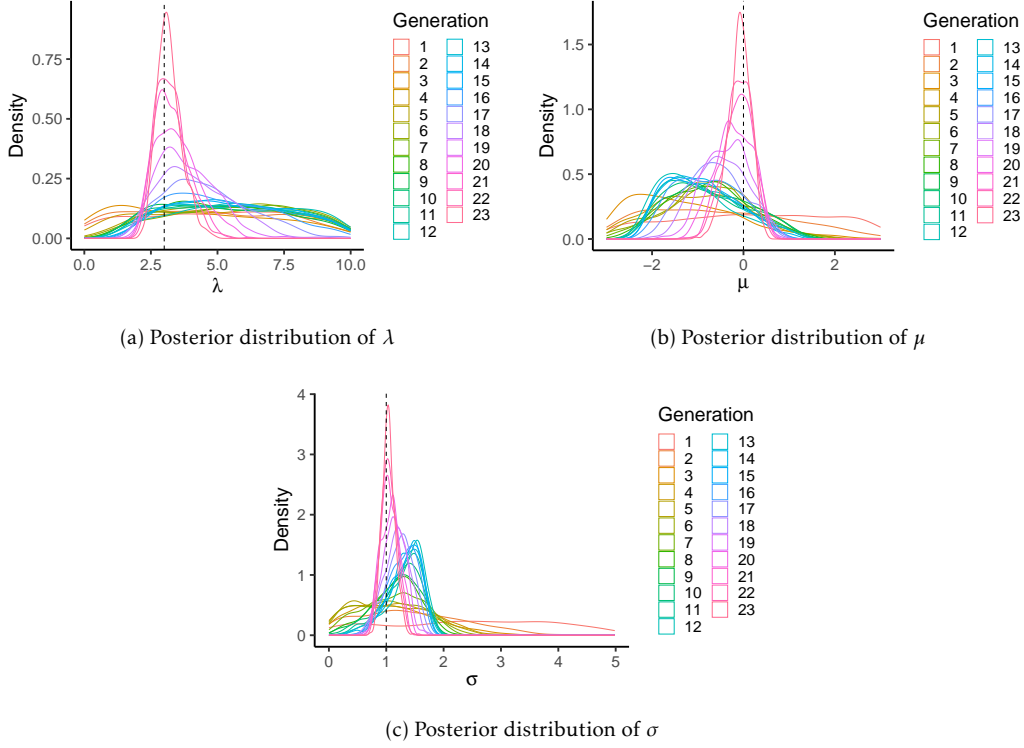


(c) Posterior distribution of $\sigma$

Figure 2: Posterior distributions of the parameters of the $\text{Pois}(\lambda) - \text{LogNormal}(\mu, \sigma)$ risk model.

*elementwise on $p_{1:n}^{\theta}$ and $d(\cdot, \cdot)$ denotes a distance function over the observation space, subject to*

$$\widetilde{p}_i \geq p_i^{\theta}, \text{ and } f(p_i^{\theta}) \geq p_i^{\theta}, \text{for } i = 1, \ldots, n. \tag{4}$$

Our first task is to find a generic function $f$ to represent the safety loading functions $f_i$'s used by the competitors. For this, we use isotonic regression. It is a statistical technique used for fitting a non-decreasing function to a set of data points. The idea is that if two pure premium satisfy $p_i \leq p_j$ then the commercial premium should also verify $\widetilde{p}_i \leq \widetilde{p}_j$. Consider a collection of candidate pure premiums $p_{1:n}^{\theta}$, associated to a candidate risk parameter $\theta$. Our datapoints are therefore pairs of pure and commercial premiums $(p_i^{\theta}, \widetilde{p}_i)_{i=1,\ldots,n}$. Suppose the pure premium have been ordered such that $p_i^{\theta} \leq p_j^{\theta}$ for $i \leq j$, isotonic regression seeks a least square fit $\widetilde{p}_i^{\theta}$ for the $\widetilde{p}_i$'s such that $\widetilde{p}_i^{\theta} \leq \widetilde{p}_j^{\theta}$ for $p_i^{\theta} \leq p_j^{\theta}$. It reduces to find $\widetilde{p}_1^{\theta}, \ldots, \widetilde{p}_n^{\theta}$ that minimize

$$\sum_{i=1}^{n} w_i^{\text{iso}}(\widetilde{p}_i^{\theta} - \widetilde{p}_i)^2, \text{ subject to } \widetilde{p}_i^{\theta} \leq \widetilde{p}_j^{\theta} \text{ whenever } p_i^{\theta} \leq p_j^{\theta},$$

where $\left(w_i^{\text{iso}}\right)_{i=1,\ldots,n}$ denotes the weights associated to each pair $(p_i^{\theta}, \widetilde{p}_i)_{i=1,\ldots,n}$. Since the $p_i^{\theta}$'s fall in a totally

ordered space, a simple iterative procedure called the Pool Adjacent Violators 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 $\widetilde{p}_i^* = \widetilde{p}_i$.

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

$$\widetilde{p}_i^* > \widetilde{p}_{i+1}^* \text{ for some } i = 1,\ldots,n.$$

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

$$\widetilde{p}_i^* \leftarrow (\widetilde{p}_i^* + \widetilde{p}_{i+1}^*)/2,$$

   ensuring monotonicity.

3. Repeat Step 2 until no violations are left.

We use the isoreg function from $R$ to get the fitted values $\widetilde{p}_i^\theta$, $i = 1,\ldots,n$. To complete the isotonic regression task we shall find a function $f$ such that $f(p_i^\theta) = \widetilde{p}_i^\theta$. A common choice is a piece-wise constant function that interpolates the $\widetilde{p}_i^\theta$'s.

We now turn to the definition of a distance. Our starting point to measure the discrepancy between observed and model-generated commercial rates is the root mean square error (RMSE) defined as

$$\text{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right] = \sqrt{\sum_{i=1}^{n} w_i^{\text{RMSE}}[\widetilde{p}_i - f(p_i^\theta)]^2}, \tag{5}$$

for a candidate risk parameter $\theta$ and an isotonic fit $f$. We seek

$$\theta^* = \underset{\theta \in \Theta}{\arg\min} \, \text{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right].$$

The existence of such $\theta^*$ is guaranteed because $\theta \mapsto \text{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right]$ only takes a finite number of values. Indeed, to each $\theta \in \Theta$ is associated a unique permutation $s^\theta \in S_n$, where $S_n$ denotes the set of all the permutations of $\{1,\ldots,n\}$, such that

$$p_{s^\theta(1)}^\theta \leq \ldots \leq p_{s^\theta(n)}^\theta.$$

This permutation $s^\theta$ defines a unique isotonic fit $f$ based on

$$\widetilde{p}_{s^\theta(1)}^\theta \leq \ldots \leq \widetilde{p}_{s^\theta(n)}^\theta,$$

6

leading to a given RMSE value $\mathrm{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right]$. Concretely, for $\theta_1, \theta_2 \in \Theta$, if it holds that $s^{\theta_1} = s^{\theta_2}$ then $\mathrm{RMSE}\left[\widetilde{p}_{1:n}, f(p_{i:n}^{\theta_1})\right] = \mathrm{RMSE}\left[\widetilde{p}_{1:n}, f(p_{i:n}^{\theta_2})\right]$. The application $\theta \mapsto s_n^\theta$ is surjective since $S_n^\Theta = \{s_n^\theta ; \theta \in \Theta\}$ is finite. The fact that $\Theta$ is a continuous space implies that $\theta^*$ cannot be unique. Our problem is an ill-posed inverse problem. Ill-posedness is usually dealt with by adding a regularization to the objective function that one wants to minimize. The ratio of $p/\widetilde{p}$ corresponds to what practitioners would call the expected Loss Ratio (LR). Our solution is based on targeting a given loss ratio. The loss ratio is a standard measure to assess the profitability of insurance lines of business. An insurance company that enters a new market is likely to have insights on the loss ratio relative to this market, for example by having informal discussions with reinsurers, brokers or competitors. These insights may translate into the definition of a lower and upper bound denoted by $\mathrm{LR}_{\mathrm{low}}$ and $\mathrm{LR}_{\mathrm{high}}$, respectively. We can then assume that the loss ratios $\mathrm{LR}_i = p_i/\widetilde{p}_i$, for $i = 1, \dots, n$, should fall in the range $[\mathrm{LR}_{\mathrm{low}}, \mathrm{LR}_{\mathrm{high}}]$, which we refer to as the loss ratio corridor. Assuming that $\mathrm{LR}_{\mathrm{high}} < 1$, we may ensure both constraints in (4) by adding to our distance (5) two regularization terms defined as

$$\mathrm{Reg}_{\mathrm{low}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right) = \sqrt{\sum_{i=1}^{n} w_i^{\mathrm{RMSE}} \left(\widetilde{p}_i - p_i^\theta \cdot \mathrm{LR}_{\mathrm{low}}^{-1}\right)_+^2},$$

and

$$\mathrm{Reg}_{\mathrm{high}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right) = \sqrt{\sum_{i=1}^{n} w_i^{\mathrm{RMSE}} \left(p_i^\theta \cdot \mathrm{LR}_{\mathrm{high}}^{-1} - \widetilde{p}_i\right)_+^2},$$

where $(x)_+ = \max(x, 0)$ denotes the positive part of $x$. The distance we consider within Problem 2 is now given by

$$d\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right] = \mathrm{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right] + \mathrm{Reg}_{\mathrm{low}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right) + \mathrm{Reg}_{\mathrm{high}}\left(\widetilde{p}_{1:n}, p_{1:n}^\theta\right).$$

We take again our three insurance coverages $g_1, g_2$ and $g_3$ and we distinguish two safety loading situations. The well-specified case in Section 2.1 and the misspecified in Section 2.2

## 2.1 Well-specified safety loading

Consider the following loading functions

$$f_1(x) = 1.38 \cdot x, \ f_2(x) = 1.1 \cdot x, \ \text{and} \ f_3(x) = 1.4 \cdot x.$$

The isotonic model is well specified in this case because the pure and commercial premium are ordered in the same way. In that case the isotonic regression exactly interpolates the datapoints as shown on Figure 3.

Figure 4 shows the plot of the $\sigma \mapsto \mathrm{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^\theta)\right]\Big|_{\lambda=3, \mu=0}$ when $f$ is isotonic (Figure 4a) and when $f$ is linear (Figure 4b).
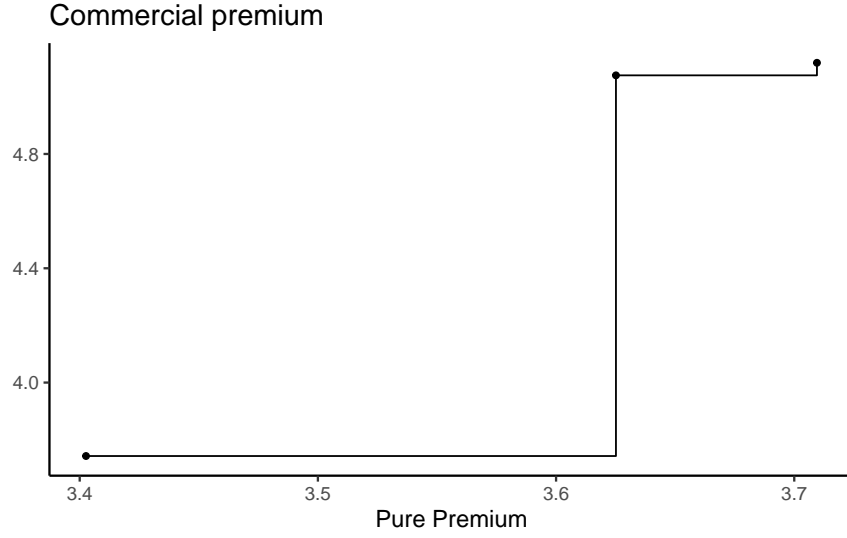
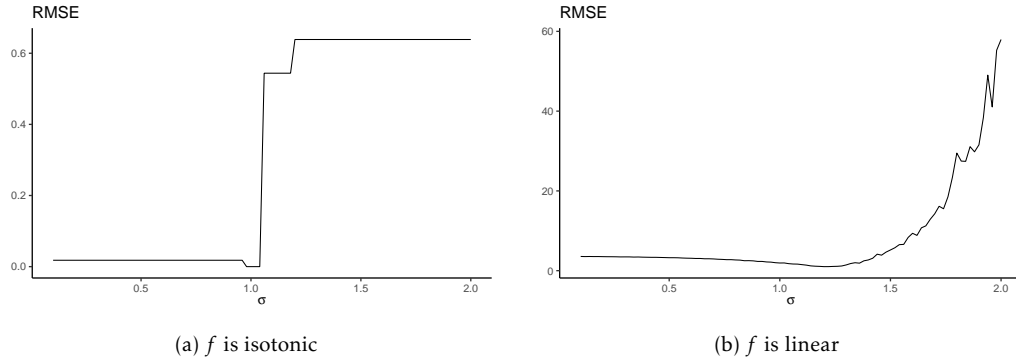Figure 3: Isotonic fit of the pure and commercial premium when the isotonic model is well-specified.



(a) $f$ is isotonic

(b) $f$ is linear

Figure 4: $\sigma \mapsto \text{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^{\theta})\right]\Big|_{\lambda=3,\mu=0}$.

Note that in this example, we do not include the regularization terms. When $f$ is isotonic, the values of $\sigma$ for which the RMSE is zero correspond to the same ordering of the pure premium as that of the true parameter $\sigma = 1$ (see Figure 4a). Consequently, several values of $\sigma$ are optimal, including the true one.

In contrast, when $f$ is linear, the RMSE is minimized for the parameter value that best aligns the commercial premium. Although one value of $\sigma$ seems to stand out, it is notably different from the true value (see Figure 4b). From Figure 4a, one might expect that running our ABC algorithm would produce a uniform posterior distribution.

The parameters of the algorithm are set as follows:

$$J = 2000, \quad R = 500, \quad \text{and} \quad \epsilon_{\min} = 10^{-8}.$$

We define a prior distribution on $\sigma$ as $\sigma \sim \mathsf{Unif}([0, 2])$. The tolerance level is set very low because we are specifically interested in the values of $\sigma$ that produce the correct ordering and an RMSE close to zero. The resulting posterior distribution is shown in Figure 5.
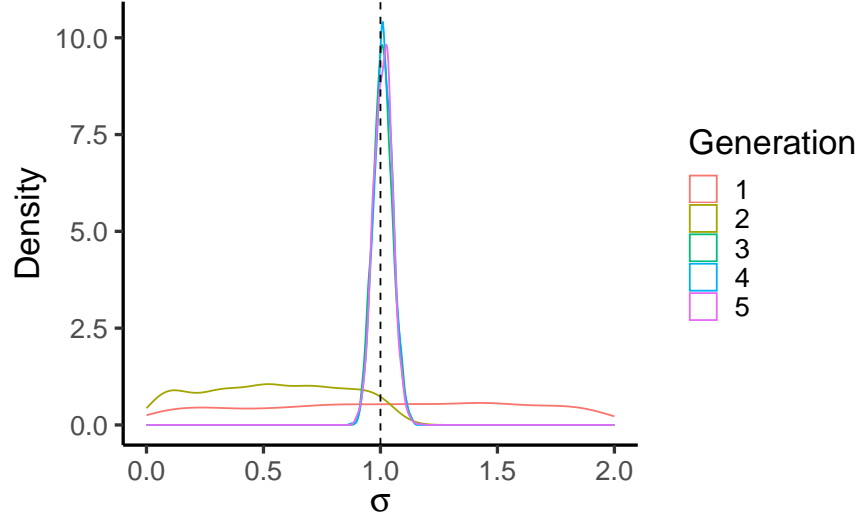


Figure 5: Intermediate posterior distributions for $\sigma$.

The posterior distribution is not entirely uniform, which can be attributed to the Monte Carlo approximation used. This approximation may alter the ordering of the pure premium when approaching the edges of the admissible range of $\sigma$ values. The posterior distribution shown in Figure 5 might appear satisfactory to a naive observer, leading one to argue that regularization is unnecessary in this case.

However, when estimating $\lambda$, $\mu$, and $\sigma$ simultaneously, the parameters can offset each other, and the correct ordering of the pure premium alone does not provide sufficient information to identify the parameters accurately. To address this, we run our ABC algorithm with the following settings:

$$J = 500, \quad R = 500, \quad \text{and} \quad \epsilon_{\min} = 10^{-8}.$$

The prior distributions for the model parameters are defined as:

$$\lambda \sim \mathsf{Unif}([0, 10]), \quad \mu \sim \mathsf{Unif}([-3, 3]), \quad \text{and} \quad \sigma \sim \mathsf{Unif}([0, 2]).$$

(a) Posterior distribution of $\lambda$



(b) Posterior distribution of $\mu$
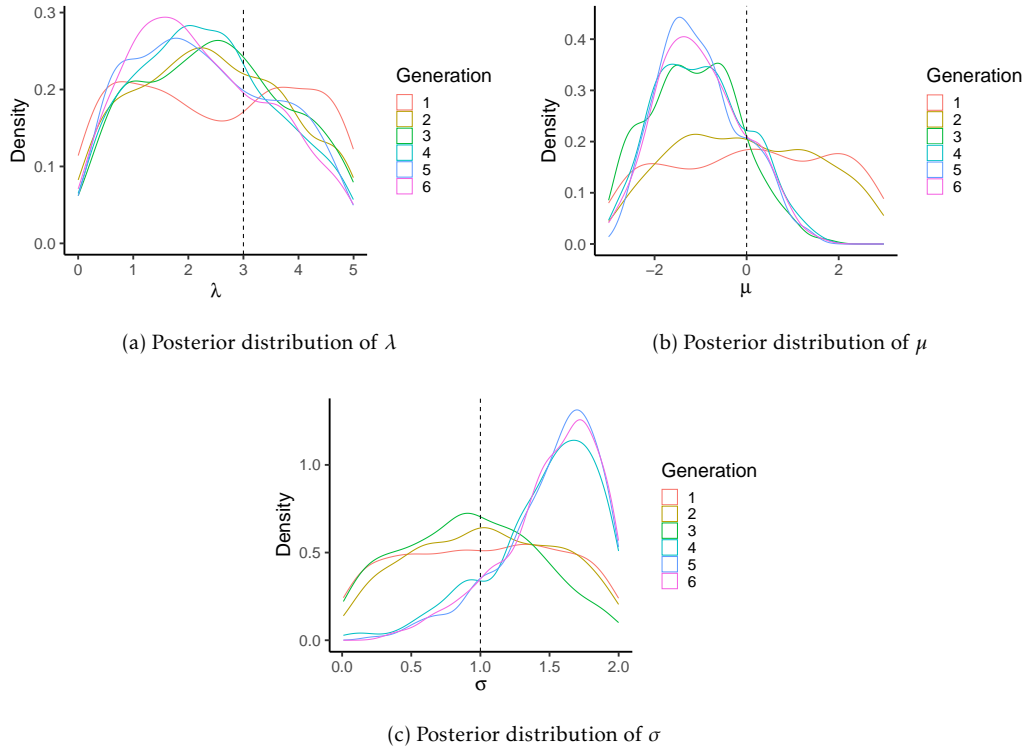


(c) Posterior distribution of $\sigma$

Figure 6: Posterior distributions of the parameters of the $\text{Pois}(\lambda) - \text{LogNormal}(\mu, \sigma)$ risk model.

The resulting posterior distributions are provided in Figure 6.

Too many parameter combinations result in the correct ordering, making it difficult to identify the relevant portions of the parameter space. Regularization is essential to address this issue, as it allows us to exclude irrelevant regions. To this end, we define a loss ratio corridor with:

$$\text{LR}_{\text{low}} = 0.7 \quad \text{and} \quad \text{LR}_{\text{high}} = 0.9.$$

The posterior distributions are shown in Figure 7.

The posterior do not center arround the true value but at least the true values belongs to the set of possible values. The posterior distributions are informative despise the fact that we have only used three commercial rates.

(a) Posterior distribution of $\lambda$

(b) Posterior distribution of $\mu$
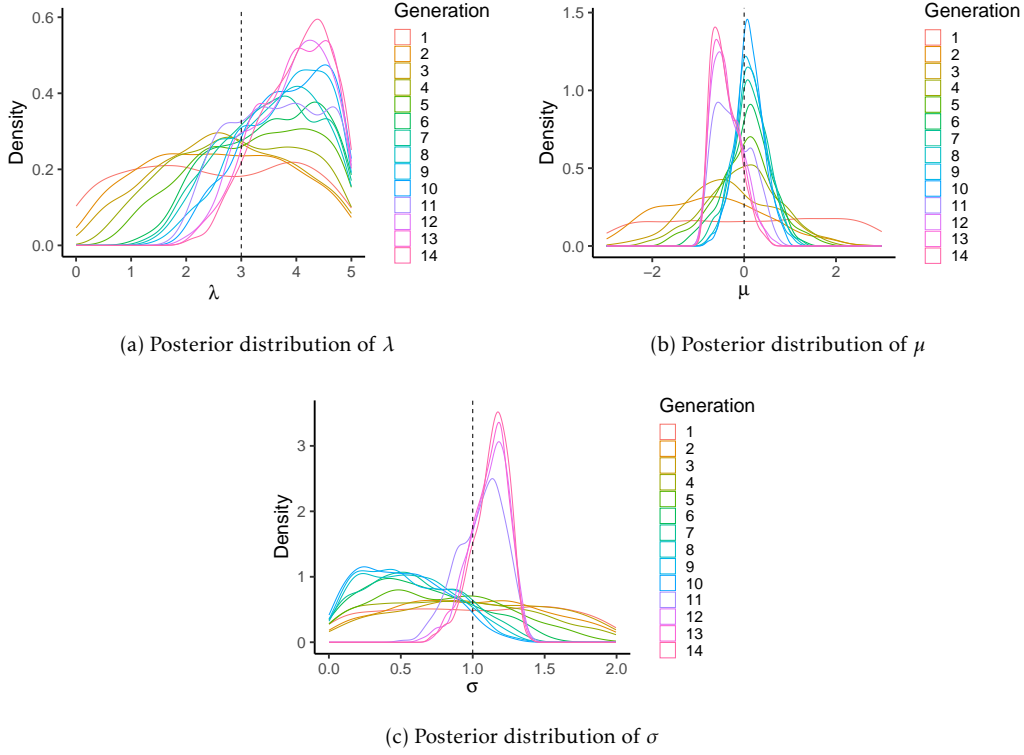
(c) Posterior distribution of $\sigma$

Figure 7: Posterior distributions of the parameters of the $\text{Pois}(\lambda) - \text{LogNormal}(\mu, \sigma)$ risk model.

## 2.2 Misspecified safety loading

Consider the following loading functions:

$$f_1(x) = 1.1 \cdot x, \quad f_2(x) = 1.2 \cdot x, \quad \text{and} \quad f_3(x) = 1.3 \cdot x.$$

In this case, the isotonic model is misspecified because the pure premium and commercial premiums are not ordered consistently. As a result, the isotonic regression approximates the data points through interpolation, as illustrated in Figure 8.

Figure 9 shows the plot of the $\sigma \mapsto \text{RMSE}\left[\widetilde{p}_{1:n}, f(p^{\theta}_{1:n})\right]\Big|_{\lambda=3, \mu=0}$ when $f$ is isotonic (Figure 9a) and when $f$ is linear (Figure 9b).

The isotonic condition does not hold for the true parameter, and the RMSE of the isotonic model is minimized for incorrect values. Similarly, the linear fit is also inaccurate, just like in Figure 3.
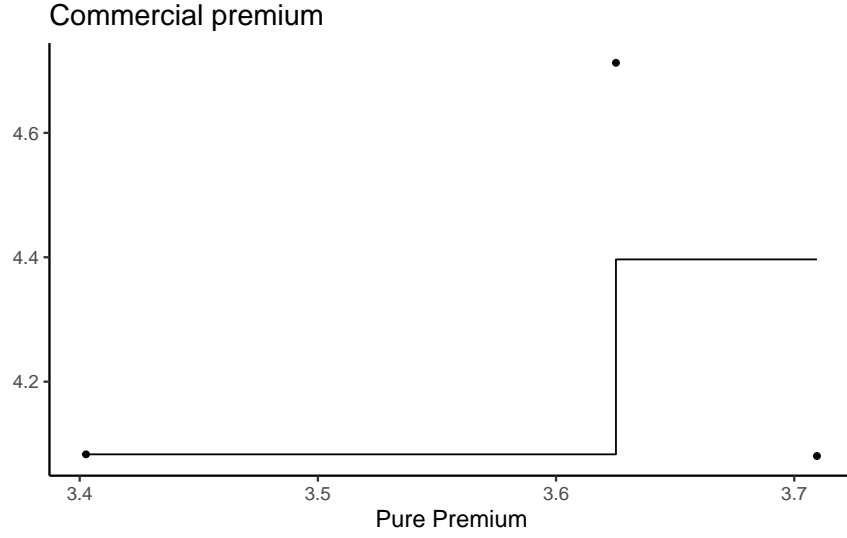
Figure 8: Isotonic fit of the pure and commercial premium when the isotonic model is misspecified.
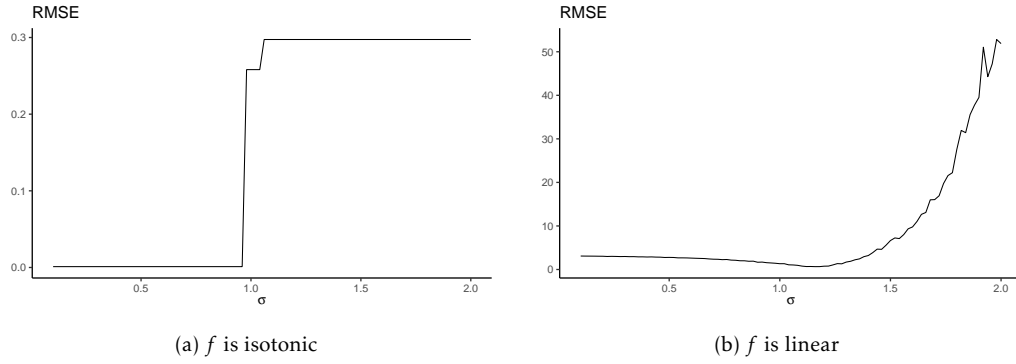


(a) $f$ is isotonic



(b) $f$ is linear

Figure 9: $\sigma \mapsto \mathrm{RMSE}\left[\widetilde{p}_{1:n}, f(p_{1:n}^{\theta})\right]\Big|_{\lambda=3, \mu=0}$.

We run our ABC algorithm with the following settings:

$$J = 500, \quad R = 500, \quad \text{and} \quad \epsilon_{\min} = 0.01.$$

The prior distributions for the model parameters are defined as:

$$\lambda \sim \mathsf{Unif}([0, 10]), \quad \mu \sim \mathsf{Unif}([-3, 3]), \quad \text{and} \quad \sigma \sim \mathsf{Unif}([0, 2]).$$

The resulting posterior distributions are provided in Figure 10. Note that regularization terms are not included at this stage.
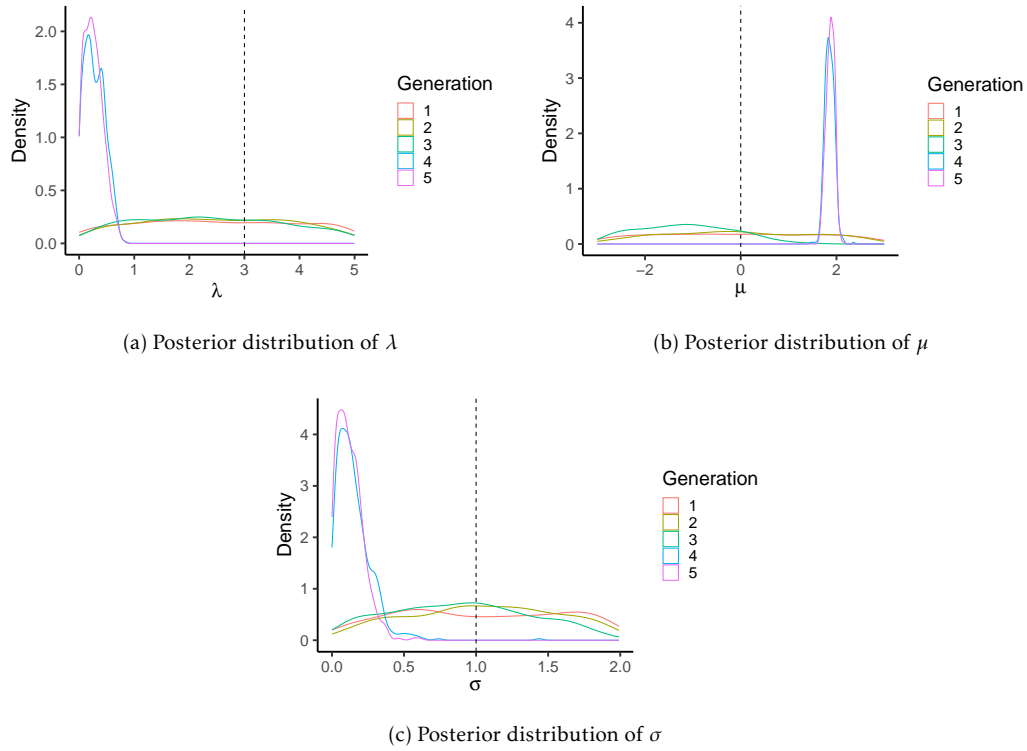
(a) Posterior distribution of $\lambda$

(b) Posterior distribution of $\mu$

(c) Posterior distribution of $\sigma$

Figure 10: Posterior distributions of the parameters of the $\mathrm{Pois}(\lambda) - \mathrm{LogNormal}(\mu, \sigma)$ risk model.

The posterior distribution is useless. We run again our ABC algorithm, this time setting a loss ratio corridor with

$$\mathrm{LR}_{\mathrm{low}} = 0.76, \text{ and } \mathrm{LR}_{\mathrm{high}} = 0.9.$$

The posterior distributions are provided on Figure 11.

The posterior distributions are slightly more informative now. However, using only three rates appears insufficient to achieve reliable results. The simulation study and the real data analysis in our paper consider scenarios with more than 25 rates and only two parameters to estimate.

# References

[1] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029, July 1982. ISSN 0012-9682. doi: 10.2307/1912775.
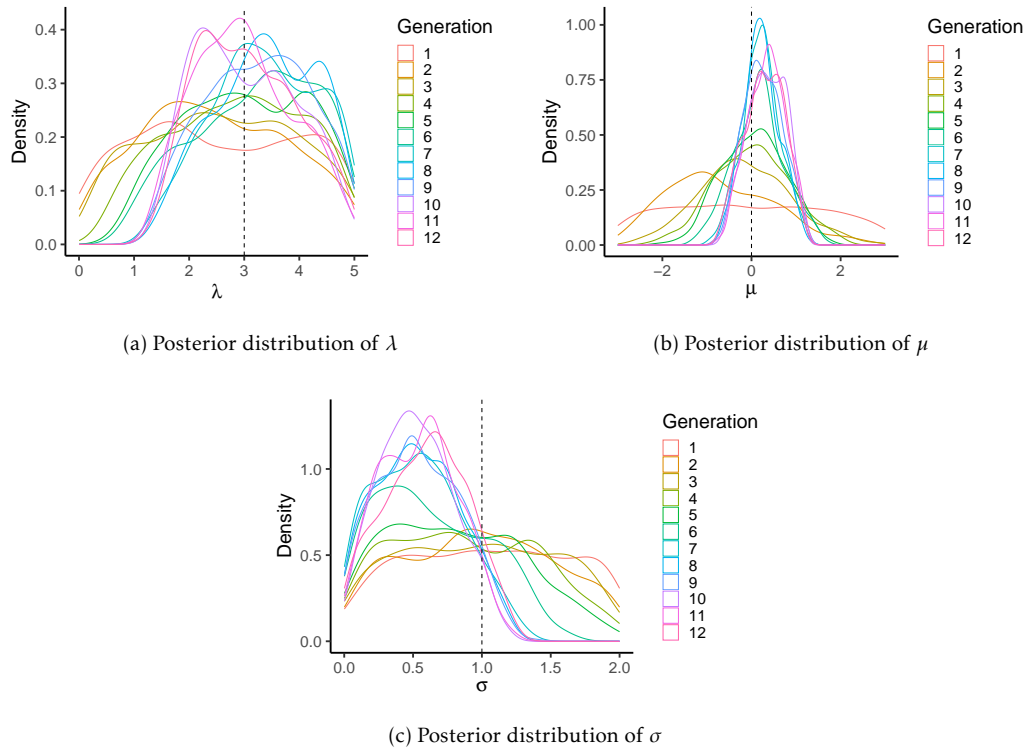
(a) Posterior distribution of $\lambda$

(b) Posterior distribution of $\mu$

(c) Posterior distribution of $\sigma$

Figure 11: Posterior distributions of the parameters of the $\mathrm{Pois}(\lambda) - \mathrm{LogNormal}(\mu, \sigma)$ risk model.