# Approximate Bayesian Computations to fit and compare insurance loss models

Pierre-Olivier Goffard and Patrick J. Laub

Univ Lyon, Université Lyon 1, LSAF EA2429

July 6, 2020

**Abstract**

Approximate Bayesian Computation (ABC) is a statistical learning technique to calibrate and select models by comparing observed data to simulated data. This technique bypasses the use of the likelihood and requires only the ability to generate synthetic data from the models of interest. We apply ABC to fit and compare insurance loss models using aggregated data. We present along the way how to use ABC for the more common claim counts and claim sizes data. A state-of-the-art ABC implementation in Python is proposed. It uses sequential Monte Carlo to sample from the posterior distribution and the Wasserstein distance to compare the observed and synthetic data.

*MSC 2010*: 60G55, 60G40, 12E10.
*Keywords*: Bayesian statistics, approximate Bayesian computation, likelihood-free inference, risk management.

## 1 Introduction

Over a fixed time period, an insurance company experiences a random number of claims called the *claim frequency*, and each claim requires the payment of a randomly sized compensation called the *claim severity*. The claim frequency is a counting random variable while the claim sizes are non-negative continuous random variables. Let us say that the claim frequency and the claim severity distributions are specified by the parameters $\boldsymbol{\theta}_{\text{freq}}$ and $\boldsymbol{\theta}_{\text{sev}}$ respectively, with $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{freq}}; \boldsymbol{\theta}_{\text{sev}})$. For each time $s = 1, \dots, t$ the number of claims $n_s$ and the claim sizes $\boldsymbol{u}_s \coloneqq (u_{s,1}, u_{s,2}, \dots, u_{s,n_s})$ are distributed as

$$n_s \sim p_N(n\,;\boldsymbol{\theta}_{\text{freq}}) \quad \text{and} \quad (\boldsymbol{u}_s \mid n_s) \sim f_U(\boldsymbol{u}\,; n, \boldsymbol{\theta}_{\text{sev}}).$$

We wish to fit these distributions, however, we assume that these independent and identically distributed (i.i.d.) values $\{(n_1, \boldsymbol{u}_1), \dots, (n_t, \boldsymbol{u}_t)\}$ are unobservable. Instead, we only have access to some real-valued *summaries* of the claim data at each time, denoted by

$$x_s = \Psi(n_s, \boldsymbol{u}_s) \quad \text{for } s = 1, \dots, t. \tag{1}$$

The summaries could be the aggregated claims if $\Psi(n, \boldsymbol{u}) = \sum_{i=1}^{n} u_i$ or the maximum claims if $\Psi(n, \boldsymbol{u}) = \max_{1 \leq i \leq n} u_i$. Our problem is to take some observations of these summaries $\boldsymbol{x} = (x_1, \ldots, x_t)$ and find the $\boldsymbol{\theta}$ which best explains them for a given parametric model.

Such incomplete data situations arise in reinsurance, see the monograph of Albrecher et al. [1, Chapter I, Section 3]. For instance, within a global non-proportional reinsurance agreement, the reinsurance company covers the risk that the insurer's total claim amount is in excess of a threshold $c > 0$. The reinsurer is only observing its payout at each time period $x_s = (\sum_{i=1}^{n_s} u_{s,i} - c)_+$. Being able to infer the parameters of the claim frequency and the claim severity distributions would help the reinsurer to better understand the risk they have underwritten.

**Remark 1.1.** *When the summary is the aggregated loss $\Psi(n, \boldsymbol{u}) = \sum_{i=1}^{n} u_i$, we effectively* decompound *the random sum. Traditionally, a decompounding method builds a non-parametric estimate of the claim severity distribution based on the observations of the aggregated sums, see Buchmann and Grübel [7] or Bøgsted and Pitts [6]. A popular application is the study of discretely observed compound Poisson processes, see for instance van Es et al. [32], Coca [8] and Gugushvili et al. [17] where a Bayesian non-parametric approach is used.*

A Bayesian approach to estimating $\boldsymbol{\theta}$ would be to treat $\boldsymbol{\theta}$ as a random variable and find (or approximate) the *posterior distribution* $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Bayes' theorem tells us that

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{2}$$

where $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ is the *likelihood* and $\pi(\boldsymbol{\theta})$ is the *prior distribution.* The prior represents our beliefs about $\boldsymbol{\theta}$ before seeing any of the observations and is informed by our domain-specific expertise. The posterior distribution is a very valuable piece of information that gathers our knowledge over the parameters. A point estimate $\widehat{\boldsymbol{\theta}}$ may be derived by taking the mean or mode of the posterior. For an overview on Bayesian statistics, we refer to the book of Gelman et al. [14].

The posterior distribution (2) rarely admits a closed-form expression, so it is approximated by an empirical distribution of samples from $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Posterior samples are typically obtained using Markov Chain Monte Carlo (MCMC), yet a requirement for MCMC sampling is the ability to evaluate (at least up to a constant) the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. When considering the definition of $\boldsymbol{x}$ in (1), we can see that there is little hope of finding an expression for the likelihood function even in simple cases (e.g. when the claim sizes are i.i.d.). If the claim sizes are not i.i.d. or if the number of claims influences their amount, then the chance that a tractable likelihood for $\boldsymbol{x}$ exists is extremely low. Even when a simple expression for the likelihood exists, it can be prohibitively difficult to compute (such as in a big data regime), and so a likelihood-free approach can be beneficial.

We advertise here a likelihood-free estimation method known as *approximate Bayesian computation* (ABC). This technique has attracted a lot of attention recently due to its wide range of applicability and its intuitive underlying principle. One resorts to ABC when the model at hand is too complicated to write the likelihood function but still simple enough to generate artificial data. Given some observations $\boldsymbol{x}$, the basic principle consists in iterating the following steps:

(i) generate a potential parameter from the prior distribution $\boldsymbol{\theta}^* \sim \pi(\boldsymbol{\theta})$;

(ii) simulate 'fake data' $\boldsymbol{x}^*$ from the likelihood $(\boldsymbol{x}^* \mid \boldsymbol{\theta}^*) \sim p(\boldsymbol{x} \mid \boldsymbol{\theta})$;

(iii) if $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \epsilon$, where $\epsilon > 0$ is small, then store $\boldsymbol{\theta}^*$,

where $\| \cdot \|$ denotes a distance measure and $\epsilon$ is an acceptance threshold. The algorithm provides us with a sample of $\boldsymbol{\theta}$'s whose distribution is close to the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$.

The ABC algorithm presented in this work allows us to consider a wide variety of $\Psi$ functions (1) without imposing common simplifying assumptions such as assuming the claim amounts are i.i.d. and independent from the claim frequency. In addition to parameter estimation, ABC allows us to perform model selection in a Bayesian manner. This direction is also investigated. For a comprehensive overview on ABC, we refer to the monograph of Sisson et al. [29]. In finance and insurance, ABC has been considered in the context of operational risk management [21] and for reserving purposes [22].

The rest of the paper is organized as follows. Section 2 provides a gentle introduction to ABC algorithms. We start by presenting the ABC routines used on count data and continuous data, then show how to use ABC to fit an insurance loss model based on aggregated data. Section 3 explains how to adapt the ABC algorithm to compare models by computing the a posteriori model probability of each competing model. The performance of our ABC implementation are illustrated on simulated data in Section 4 and on a real world insurance data set in Section 5.

# 2 Model calibration

ABC is a method for approximating the posterior probability $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ without using the likelihood function. The implementation of ABC is tied to the nature of the data at hand. In our problem, the frequency data is discrete, the individual claim sizes are continuous and the aggregated data is a mixture of discrete and continuous (due to the atom at 0). We take advantage of this fact to introduce ABC algorithms for discrete data in Section 2.1, continuous data in Section 2.2, and mixed data in Section 2.3. The acceptance–rejection algorithm laid out in the introduction most often leads to considerable computing time, so Section 2.4 explains how to speed up ABC using sequential Monte Carlo (SMC). Section 2.5 shows the validity of our ABC implementation on an illustrative example.

## 2.1 ABC for count data

Consider some count data $n_1, \ldots, n_t \in \mathbb{N}_0$ which are i.i.d. with the probability mass function (p.m.f.) $p_N(n \mid \boldsymbol{\theta})$; for example, the $n_s$'s could be claim frequencies. The likelihood of such data is $p(\boldsymbol{n} \mid \boldsymbol{\theta}) = \prod_{s=1}^{t} p_N(n_s \mid \boldsymbol{\theta})$, where $\boldsymbol{n} = (n_1, \ldots, n_t)$. For common discrete distributions, such as the Poisson or negative binomial, the likelihood function is tractable and may be plugged into an MCMC sampling algorithm to produce samples from the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{n}) \propto p(\boldsymbol{n} \mid$

$\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Alternatively, we can sample from the posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{n})$ in a likelihood-free way by acceptance–rejection, which is detailed in Algorithm 1.

**Algorithm 1** Acceptance–rejection sampling the posterior of count data.

```
input observations $\boldsymbol{n} = (n_1, \dots, n_t)$
for $k = 1 \to K$ do
    repeat
        generate $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta})$
        generate $\boldsymbol{n}_k \sim p(\boldsymbol{n} \mid \boldsymbol{\theta}_k)$
    until $\boldsymbol{n}_k = \boldsymbol{n}$ then store $\boldsymbol{\theta}_k$
end for
return $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ which are i.i.d. samples from $\pi(\boldsymbol{\theta} \mid \boldsymbol{n})$
```

Algorithm 1 gives samples from $\pi_0(\boldsymbol{\theta} \mid \boldsymbol{n})$ which is exactly the desired posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{n})$:

$$\pi_0(\boldsymbol{\theta} \mid \boldsymbol{n}) \propto \pi(\boldsymbol{\theta}) \int_{\mathbb{N}_0^t} \mathbb{I}_{\{\boldsymbol{n}=\widetilde{\boldsymbol{n}}\}}\, p(\widetilde{\boldsymbol{n}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{n}} = \pi(\boldsymbol{\theta}) p(\boldsymbol{n} \mid \boldsymbol{\theta}),$$

where

$$\mathbb{I}_{\{\boldsymbol{n}=\widetilde{\boldsymbol{n}}\}} = \begin{cases} 1, & \text{if } \boldsymbol{n} = \widetilde{\boldsymbol{n}}, \\ 0, & \text{otherwise.} \end{cases}$$

As we collect more data, the probability of seeing an exact match $\{\boldsymbol{n} = \widetilde{\boldsymbol{n}}\}$ decreases exponentially. This, combined with a diffuse prior distribution, will result in a cumbersome waiting time before getting a posterior sample. A natural refinement is to require an exact correspondence between the samples sorted in ascending order. The acceptance rate may still be too low to be practical, and in this case an approximate match between the observed and fake data must be considered. We discuss this matter within the continuous data case in the following section.

## 2.2 ABC for continuous data

Let $u_1, \dots, u_n \in \mathbb{R}$ be an i.i.d. sample of continuous data with a probability density function (p.d.f.) denoted $f_U(u \mid \boldsymbol{\theta})$. An example of such data would be the claim sizes. With the notation $\boldsymbol{u} = (u_1, \dots, u_n)$, we can write the likelihood as $p(\boldsymbol{u} \mid \boldsymbol{\theta}) = \prod_{i=1}^n f_U(u_i \mid \boldsymbol{\theta})$. If the data is fitted to a standard probability model, say gamma or normal, then we can sample from the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{u}) \propto \pi(\boldsymbol{\theta}) p(\boldsymbol{u} \mid \boldsymbol{\theta})$, with an MCMC scheme. If the likelihood is unavailable, then we can adapt Algorithm 1 to the case of continuous data for which exact correspondence between observed and fake data is not possible. Synthetic samples are then accepted whenever they fall sufficiently close to the observed data. That is, if the dissimilarity between two samples, assessed by a norm $\|\cdot\|$ on $\mathbb{R}^n$, is smaller than some tolerance threshold $\epsilon > 0$, see Algorithm 2.

The procedure depicted in Algorithm 2 allows us to sample from an approximation of the posterior distribution given by

$$\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{u}) \propto \pi(\boldsymbol{\theta}) \int_{\mathbb{R}^t} \mathbb{I}_{\{\|\boldsymbol{u}-\widetilde{\boldsymbol{u}}\| < \epsilon\}}\, p(\widetilde{\boldsymbol{u}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{u}}, \tag{3}$$

**Algorithm 2** ABC acceptance–rejection sampling for continuous data.

```
1: input observations $\boldsymbol{u} = (u_1, \dots, u_n)$, $\epsilon > 0$ threshold, $\|\cdot\|$ norm
2: for $k = 1 \to K$ do
3:     repeat
4:         generate $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta})$
5:         generate $\boldsymbol{u}_k \sim p(\boldsymbol{u} \mid \boldsymbol{\theta}_k)$
6:     until $\|\boldsymbol{u} - \boldsymbol{u}_k\| < \epsilon$ then store $\boldsymbol{\theta}_k$
7: end for
8: return $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ which are approximately $\pi(\boldsymbol{\theta} \mid \boldsymbol{u})$ distributed
```

where

$$\mathbb{I}_{\{\|\boldsymbol{u}-\widetilde{\boldsymbol{u}}\|<\epsilon\}} = \begin{cases} 1, & \text{if } \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\| < \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Distribution (3) is called the *ABC posterior* and it has the desirable theoretical property of converging toward the standard posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{u})$ as $\epsilon$ tends to 0, see Rubio and Johansen [26], Prangle et al. [24] or Bernton et al. [3].

The ABC procedure suffers from the so-called curse of dimensionality [4]. Specifically, if one takes the Euclidean distance or some variation of it to measure the dissimilarity between observed and fake data then the odds of getting an acceptable match will plummet as the number of observations, i.e. the dimension of $\boldsymbol{u}$, increases. The dimensionality curse can be alleviated by replacing $\boldsymbol{u} \in \mathbb{R}^n$ with summary statistics $S(\boldsymbol{u}) \in \mathbb{R}^d$, where $d < n$, in Algorithm 2 (specifically, in line 5 the norm becomes $\|S(\boldsymbol{u}) - S(\boldsymbol{u}_k)\|$). While the choice of the summary statistics $S : \mathbb{R}^n \mapsto \mathbb{R}^d$ is arbitrary, it is desirable to have $d \ll n$ while limiting the information loss. This is difficult. When the model at hand admits sufficient statistics then these should be taken. In fact, the only statistics which uphold the convergence of $\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{u})$ to $\pi(\boldsymbol{\theta} \mid \boldsymbol{u})$ as $\epsilon \to 0$ are sufficient statistics. Note that the summary statistics $S$ are not to be confused with the $\Psi$ summaries in Section 1!

**Remark 2.1.** *When dealing with frequency data (see Section 2.1), it is possible to define a map $S : \mathbb{N}_0^t \mapsto \mathbb{R}^d$ which allows us to reduce the dimension and adopt the ABC procedure for continuous data. Consider for instance, the case where the claim frequency are Poisson distributed and the map $S$ corresponds to the empirical mean.*

Most often, we will not be able to find sufficient statistics. Many research papers have been dedicated to designing ad hoc summary statistics in the ABC literature, we refer to the survey of Blum et al. [5]. The problem is that it always implies a loss of information along with a convergence toward the posterior distribution conditionally to the summary statistics instead of the true posterior. We illustrate the use of summary statistics in Section 2.5, but do not use this technique in the other examples.

Bernton et al. [3] recommend the Wasserstein distance to measure the dissimilarity between two samples. The Wasserstein distance is deemed difficult to compute but for real-valued i.i.d. observations it reduces to

$$\mathcal{W}_p(\boldsymbol{u},\widetilde{\boldsymbol{u}}) = \frac{1}{n}\sum_{k=1}^{n}\left|u_{(k)} - \widetilde{u}_{(k)}\right|^p, \text{ for } p \geq 1,$$

where $u_{(1)} < \ldots < u_{(n)}$ and $\widetilde{u}_{(1)} < \ldots < \widetilde{u}_{(n)}$ denote the order statistics of the observed and synthetic data respectively. The use of the order statistics as summary statistics is not new, it was investigated for instance in the work of Sousa et al. [30] and Fearnhead and Prangle [11]. Now that we have reviewed the use of ABC in the case of discrete and continuous data, we turn to the case of mixed data which is of primary interest for the actuarial application at the center of this work.

### 2.3 ABC for mixed data

We return to the problem of fitting a model to aggregated insurance data. Recall that, for each time period, a random number $n \in \mathbb{N}_0$ of claims are filed. The claim frequencies form an i.i.d. sample from the p.m.f. $p_N(n \mid \boldsymbol{\theta}_{\text{freq}})$. Given $n$, the associated claim sizes $\boldsymbol{u} = (u_1, \ldots, u_n)$ have a joint p.d.f. denoted by $f_{U|N}(\boldsymbol{u} \mid n, \boldsymbol{\theta}_{\text{sev}})$.

The distribution of the available information $x := \Psi(n, \boldsymbol{u})$ is parametrized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{freq}}, \boldsymbol{\theta}_{\text{sev}})$ and admits a point mass $p_X(0 \mid \boldsymbol{\theta})$ at 0. Zeros can occur if no claims are filed ($n = 0$) which occurs with probability $p_N(0 \mid \boldsymbol{\theta}_{\text{freq}})$, or because of censoring effects like in the non-proportional reinsurance treaty case, see Section 1. The continuous part of $x$'s distribution is characterized by the conditional p.d.f.

$$[1 - p_X(0 \mid \boldsymbol{\theta})]\, f_{X|X>0}(x \mid \boldsymbol{\theta}) \quad \text{for } x > 0.$$

For a data history $\boldsymbol{x} = (x_1, \ldots, x_t)$ of $t$ time periods, we separate the zeros from the non-negative data points, so

$$\boldsymbol{x} = (\boldsymbol{x}^0, \boldsymbol{x}^+) = (\underbrace{0, \ldots, 0}_{t_0 \text{ zeros}}, \underbrace{x_1^+, \ldots, x_{t-t_0}^+}_{t-t_0 \text{ non-zeros}}).$$

The likelihood function may be written as

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= p_X(0 \mid \boldsymbol{\theta})^{t_0}[1 - p_X(0 \mid \boldsymbol{\theta})]^{t-t_0} \prod_{s=1}^{t-t_0} f_{X|X>0}(x_s^+ \mid \boldsymbol{\theta}) \qquad (4) \\ &= p_X(0 \mid \boldsymbol{\theta})^{t_0}[1 - p_X(0 \mid \boldsymbol{\theta})]^{t-t_0} p(\boldsymbol{x}^+ \mid \boldsymbol{\theta}). \end{aligned}$$

To evaluate the conditional p.d.f. $f_{X|X>0}$ in (4) we must consider all possible values of $n$ which often leads to an infinite series without closed-form expression, as illustrated in Example 1.

**Example 1.** *Consider the case where we only observe the aggregate claim sizes $x_s = \sum_{i=1}^{n_s} u_{s,i}$ for $s = 1, \ldots, t$, i.e., $\Psi$ is the sum operator. If the claim sizes are i.i.d. and independent from the claim frequency, which is common in the actuarial science literature, the conditional p.d.f. of $X$ taking positive values is*

$$f_{X|X>0}(x \mid \boldsymbol{\theta}) = \frac{1}{1 - p_N(0 \mid \boldsymbol{\theta}_{\text{freq}})} \sum_{n=1}^{\infty} f_U^{(*n)}(x \mid \boldsymbol{\theta}_{\text{sev}}) p_N(n \mid \boldsymbol{\theta}_{\text{freq}}), \qquad (5)$$

*where* $f_U^{(*n)}(x \mid \boldsymbol{\theta}_{\text{sev}})$ *denotes the* $n$*-fold convolution product of* $f_U(x \mid \boldsymbol{\theta}_{\text{sev}})$ *with itself. A closed-form expression of* (5) *is available only in a few cases. For the remaining cases, quite some energy has been dedicated by actuarial scientists to finding convenient numerical approximations. Note that none of the aforementioned numerical routines would be suited to the multiple evaluations of the conditional p.d.f. required for Bayesian inference or maximum likelihood inference via some optimization algorithm. We begin our numerical illustration of the ABC method on some cases where a closed-form expression of* (5) *is available, as we will be able to sample from the true posterior via an MCMC simulation scheme. Point estimates may also be compared to frequentist estimators such as the maximum likelihood or the method of moment estimators. The latter has been used in a similar situation in the work of Goffard et al. [15].*

The lack of analytical expression for the likelihood function justifies the use of a likelihood-free inference method such as ABC. The distribution of $x$ is of mixed type which means we cannot directly apply Algorithm 2 as we would lose the convergence toward the standard posterior distribution. To address this issue, we ask that the number of zeros in the synthetic samples $\widetilde{t}_0$ matches the number of zeros in the observed data $t_0$ and we treat the non-negative data points as i.i.d. continuous data. So, in Algorithm 3 we retain synthetic samples that belong to the set

$$\mathcal{B}_{\epsilon,\boldsymbol{x}} = \left\{\widetilde{\boldsymbol{x}} \in \mathbb{R}^t\,;\, \boldsymbol{x}^0 = \widetilde{\boldsymbol{x}}^0 \text{ and } \|\boldsymbol{x}^+ - \widetilde{\boldsymbol{x}}^+\| < \epsilon\right\}. \tag{6}$$

**Algorithm 3** ABC acceptance–rejection sampling for mixed data.

```
input observations $\boldsymbol{x} = (x_1, \dots, x_t)$, $\epsilon > 0$ threshold, $\|\cdot\|$ norm
for $k = 1 \to K$ do
    repeat
        generate $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta})$
        generate $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k)$
    until $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon,\boldsymbol{x}}$, then store $\boldsymbol{\theta}_k$        ▷ $\mathcal{B}_{\epsilon,\boldsymbol{x}}$ defined by (6)
end for
return $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ which are approximately $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ distributed
```

Algorithm 3 samples from the approximate posterior distribution

$$\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto \pi(\boldsymbol{\theta}) \int_{\mathbb{R}^t} \mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\, p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{x}}, \tag{7}$$

where

$$\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}}) = \begin{cases} 1, & \text{if } \boldsymbol{x}^0 = \widetilde{\boldsymbol{x}}^0 \text{ and } \|\boldsymbol{x}^+ - \widetilde{\boldsymbol{x}}^+\| < \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

The following result shows the convergence of $\pi_\epsilon$ toward the standard posterior as we let $\epsilon$ approaching 0.

**Proposition 1.** *Suppose that*

$$\sup_{(\widetilde{\boldsymbol{x}},\boldsymbol{\theta}) \in \mathcal{B}_{\epsilon,\boldsymbol{x}} \times \boldsymbol{\Theta}} p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{\theta}) < \infty,$$

*for some* $\epsilon > 0$*. Then, for each* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, we have*

$$\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x}) \longrightarrow \pi(\boldsymbol{\theta} \mid \boldsymbol{x}), \text{ as } \epsilon \to 0.$$

*Proof.* The modified prior $\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x})$ is defined as

$$\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta}) \int_{\mathbb{R}^t} \mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\, p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{x}}}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) \int_{\mathbb{R}^t} \mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\, p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{x}}\, \mathrm{d}\boldsymbol{\theta}} = \frac{\pi(\boldsymbol{\theta}) p_\epsilon(\boldsymbol{x} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) p_\epsilon(\boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}, \tag{8}$$

where $p_\epsilon(\boldsymbol{x} \mid \boldsymbol{\theta})$ is an approximation of the likelihood

$$p_\epsilon(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{\int_{\mathbb{R}^t} \mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\, p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{x}}}{\int_{\mathbb{R}^t} \mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\, \mathrm{d}\widetilde{\boldsymbol{x}}}. \tag{9}$$

Since the data is i.i.d., we rearrange the vectors $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$ to set aside the zeros in the data, so $\boldsymbol{x} = (\boldsymbol{x}^0, \boldsymbol{x}^+)$ and $\widetilde{\boldsymbol{x}} = (\widetilde{\boldsymbol{x}}^0, \widetilde{\boldsymbol{x}}^+)$, respectively. It allows us to write the indicator function in (9) as the product

$$\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}}) = \mathbb{I}_{\{\boldsymbol{x}^0 = \widetilde{\boldsymbol{x}}^0\}} \cdot \mathbb{I}_{\{\|\boldsymbol{x}^+ - \widetilde{\boldsymbol{x}}^+\| \leq \epsilon\}}. \tag{10}$$

Inserting (10) into the quasi-likelihood (9) leads to

$$\begin{aligned} p_\epsilon(\boldsymbol{x} \mid \boldsymbol{\theta}) &= p_X(0 \mid \boldsymbol{\theta})^{t_0} [1 - p_X(0 \mid \boldsymbol{\theta})]^{t-t_0} \frac{\int_{\mathbb{R}^{t-t_0}} \mathbb{I}_{\{\|\boldsymbol{x}^+ - \widetilde{\boldsymbol{x}}^+\| \leq \epsilon\}} p(\widetilde{\boldsymbol{x}}^+ \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{x}}}{\int_{\mathbb{R}^{t-t_0}} \mathbb{I}_{\{\|\boldsymbol{x}^+ - \widetilde{\boldsymbol{x}}^+\| \leq \epsilon\}}\, \mathrm{d}\widetilde{\boldsymbol{x}}} \\ &\underset{\epsilon \to 0}{\longrightarrow} p_X(0 \mid \boldsymbol{\theta})^{t_0} [1 - p_X(0 \mid \boldsymbol{\theta})]^{t-t_0} p(\boldsymbol{x}^+ \mid \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \boldsymbol{\theta}), \end{aligned} \tag{11}$$

where the limit in (11) follows from applying Proposition 1 of Rubio and Johansen [26], see also Bernton et al. [3, Proposition 2]. Taking the limit as $\epsilon$ tends to 0 in (8) yields the announced result. □

Following up on the discussion in Section 2.2, we take the Wasserstein distance to evaluate the dissimilarities between the non-negative portions of the fake and observed data. When comparing the non-negative data points, a small $\epsilon$ leads to an accurate but potentially slow ABC algorithm. The combination of a small $\epsilon$ and a prior more diffuse than the posterior distribution makes ABC rejection sampling inefficient as acceptance almost never occurs. We therefore move from the acceptance–rejection simulation scheme to a Sequential Monte Carlo (SMC) scheme inspired by the work of Del Moral et al. [9].

## 2.4 ABC using SMC

Sequential Monte Carlo (ABC-SMC) is an ABC approach where a sequence of distributions is constructed by gradually decreasing tolerance $\epsilon$ through a sequence $(\epsilon_g)_{g\geq 1}$. The ABC-SMC algorithm starts by sampling a finite number of parameter sets (particles) from the prior distribution and each intermediate distribution (called a generation) is obtained as a weighted sample approximated via a multivariate Kernel Density Estimator (KDE).

We start by setting the number of generation $G$ and the number of particles $K$. For the first generation ($g = 1$), the tolerance level is set to $\epsilon_1 = \infty$. Particles are proposed from the prior distribution $\boldsymbol{\theta}_k^1 \sim \pi(\boldsymbol{\theta})$ and retained if the synthetic data $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid \theta_k^1)$ satisfies $\boldsymbol{x}_k \in \mathcal{B}_{\infty,\boldsymbol{x}}$. It goes on until $K$ particles are selected. Note that the condition $\boldsymbol{x}_k \in \mathcal{B}_{\infty,\boldsymbol{x}}$ simply means that the number of zeros in the fake data matches the number of zeros in the observed data. A first approximation of the posterior distribution follows from fitting a multivariate Kernel Density Estimator (KDE) $K_h$ to the first generation of particles

$$\widehat{\pi}_{\epsilon_1}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} K_h\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^1\|\big),$$

where $h$ denotes the bandwidth. For a given generation $g > 1$, we hold an approximation $\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ of the posterior distribution based on the $(g-1)^{\text{th}}$ generation of particles. New particles $\boldsymbol{\theta}_k^g$ are proposed by sampling repeatedly from $\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ until the synthetic data $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k^g)$ satisfies $\boldsymbol{x}_k \in \mathcal{B}_{\infty,\boldsymbol{x}}$. It goes on until $K$ particles are selected, the synthetic data is also kept. An acceptance threshold $\epsilon_g$ is defined as the empirical quantile of order $\alpha \in (0,1)$ of the distances $||\boldsymbol{x}^+ - \boldsymbol{x}_k^+||$, $k = 1,\dots,K$. Each particle is assigned a weight

$$w_k^g \propto \frac{\pi(\boldsymbol{\theta}_k^g)}{\widehat{\pi}_{g-1}(\boldsymbol{\theta}_k^g \mid \boldsymbol{x})}\mathbb{I}_{\mathcal{B}_{\epsilon_g,\boldsymbol{x}}}(\boldsymbol{x}_k), \quad k = 1,\dots,K,$$

which is used to update the posterior approximation to

$$\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{k=1}^{K} w_k^g K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^g\|).$$

The pseudocode of the algorithm is provided in Appendix A, see Algorithm 5.

A common choice for the kernel is the multivariate Gaussian kernel with covariance matrix set to twice the empirical covariance matrix assessed over the cloud of weighted particles $\{(\boldsymbol{\theta}_k^g, w_k^g)\}_{k=1,\dots,K}$, see Beaumont et al. [2].

The behavior of the algorithm can be investigated by calculating the Effective Sample Size (ESS), defined in Del Moral et al. [9] as

$$\text{ESS}^g = \Big[\sum_{k=1}^{K}(w_k^g)^2\Big]^{-1},\ g = 1,\dots,G.$$

The effective sample size ranges from 1 to $N$ and indicates whether the algorithm is efficient in sampling from the targeted distribution. An ESS falling below a certain threshold, typically $N/2$ see Del Moral et al. [9], should trigger a resampling step. We close this section by illustrating the performance of our ABC implementation on an example where both the likelihood and sufficient summary statistics are available.

## 2.5 Illustrations on total claim amounts data

Let the claim frequency be geometrically distributed

$$n_1,\dots,n_t \overset{\text{i.i.d.}}{\sim} \mathsf{Geom}(p = 0.8),$$

with p.m.f. given by $p_N(n\,;\,p) = (1-p)p^n$, $n \in \mathbb{N}_0$. Assume that the claim amounts are exponentially distributed

$$u_{s,1},\dots,u_{s,n_s} \overset{\text{i.i.d.}}{\sim} \mathsf{Exp}(\delta = 5), \quad s = 1,\dots,t.$$

with p.d.f. defined as $f(x\,;\,\delta) = (1/\delta)\mathrm{e}^{-x/\delta}$, $x > 0$, irrespective of the claim frequency. The available data is the aggregated claim sizes

$$x_s = \sum_{k=1}^{n_s} u_{s,k}, \quad s = 1, \ldots, t,$$

and we assume that $t = 100$ data points are available to conduct the inference. The likelihood function of the data is given by

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = (1-p)^t \Big(\frac{p}{\delta}\Big)^{t-t_0} \exp\Big[-\frac{1-p}{\delta} \sum_{s=1}^{t-t_0} x_s^+\Big],$$

and allows us to sample from the standard posterior distribution via an MCMC scheme. This compound geometric-exponential model admits $t_0$ (the number of zeros in the data) and $\sum_{s=1}^{(t-t_0)} x_s^+$ (sum of the non-negative data points) as sufficient statistics which in turn allows us to sample from an ABC posterior based on sufficient summary statistics. We set uniform priors

$$p \sim \mathsf{Unif}(0,1), \quad \delta \sim \mathsf{Unif}(0,100)$$

over the parameters of the $\mathsf{Geom}(p)$–$\mathsf{Exp}(\delta)$ we want to fit. We set the number of generation to $G = 10$, the number of particles to $K = 1000$ and the order of the quantile to $\alpha = 0.5$ for the ABC sampler. Figure 1 displays the histograms of the posterior samples produced via MCMC, ABC with sufficient statistics and ABC using the Wasserstein distance.

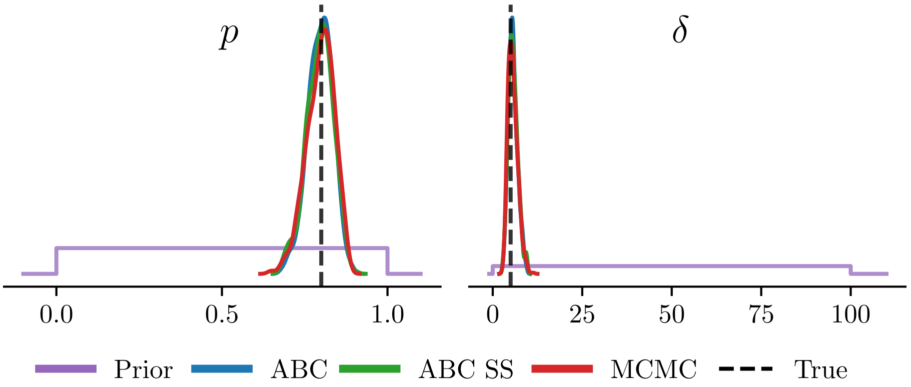

Figure 1: Fitting a $\mathsf{Geom}(p)$–$\mathsf{Exp}(\delta)$ model to simulated data. The true parameters are $p = 0.8$ and $\delta = 5$. The **ABC posterior**, **ABC summary statistics posterior**, and the **true posterior** (by MCMC) coincide very well, and are considerably narrower than the **prior**.

The MCMC posterior sample has been obtained by using the dedicated function in the `PyMC3` Python library, see Salvatier et al. [27].

# 3 Model selection

When it comes to modeling claim data, one has plenty of options for both the claim frequency and the claim sizes, see for instance the book of Klugman et al.

[19, Chapters V & VI]. A decision must be made to find the most suitable models among a set of candidates $\{1, \ldots, M\}$. The Bayesian approach to model selection and hypothesis testing consists in defining a categorical random variable $m$ with state space $\{1, \ldots, M\}$ and a priori distribution $\pi(m)$. The a posteriori model evidence is then given by

$$\pi(m \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid m)\pi(m)}{\sum_{\widetilde{m}=1}^{M} p(\boldsymbol{x} \mid \widetilde{m})\pi(\widetilde{m})}, \quad m \in \{1, \ldots, M\}.$$

One often compares two models, say 1 and 2, by computing the Bayes factors $B_{12} = \pi(2 \mid \boldsymbol{x})/\pi(1 \mid \boldsymbol{x})$. For an overview on Bayesian model selection and Bayes factor, we refer the reader to Kass and Raftery [18]. The marginal likelihood of the data according to given model $m \in \{1, \ldots, M\}$ is defined by

$$p(\boldsymbol{x} \mid m) = \int_{\Theta_m} p(\boldsymbol{x} \mid m, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid m)\,\mathrm{d}\boldsymbol{\theta}, \quad \text{for } m \in \{1, \ldots, M\}, \tag{12}$$

where $\Theta_m$ denotes the parameter space of model $m$. The evaluation of (12) is challenging from a computational point of view, even when the likelihood is available. The acceptance–rejection implementation of ABC proposed in Grelaud et al. [16] reduces to add a layer to the standard Algorithm 3 by first drawing a model from $\pi(m)$. The posterior probability of a model is then proportional to the number of times this model was selected, see Algorithm 4.

**Algorithm 4** Acceptance–rejection to compute the model evidence.

```
for $k = 1 \to K$ do
    repeat
        generate $m_k \sim \pi(m)$
        generate $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta} \mid m)$
        generate $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid m_k, \boldsymbol{\theta}_k)$
    until $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon,\boldsymbol{x}}$ then store $(m_k, \boldsymbol{\theta}_k)$
end for
```

The spirit of Algorithm 4 relates to the Monte Carlo approach to the computation of models' marginal likelihood, see for instance McCulloch and Rossi [20]. Namely, the model evidence is evaluated by

$$p(\boldsymbol{x} \mid m) \approx \frac{1}{K}\sum_{k=1}^{K} p(\boldsymbol{x} \mid m, \boldsymbol{\theta}_k),$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K \sim \pi(\boldsymbol{\theta} \mid m)$. This procedure might be inefficient as most of the $\boldsymbol{\theta}_i$ have small likelihoods when the posterior is more concentrated than the prior distribution. Importance sampling strategies have been proposed to address this issue. The sequential Monte Carlo idea used in Algorithm 5 have been adapted in the works of Toni and Stumpf [31] and Prangle et al. [23] to improve the sampling efficiency. Our implementation is described hereafter.

We fix the number of generations $G$ and the number of particles $K$. When several models are competing, a particle is a combination of a model and its parameters.

For the first generation ($g = 1$), for each particle $k = 1, \dots, K$, a model $m_k^1$ is drawn from $\pi(m)$ with parameter $\boldsymbol{\theta}_k^1$ sampled from the prior distribution $\pi(\boldsymbol{\theta} \mid m_k^1)$ until the synthetic data $\boldsymbol{x}_k \sim p(\boldsymbol{x}|m_k^1, \boldsymbol{\theta}_k^1)$ satisfies $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon_1, \boldsymbol{x}}$, where $\epsilon_1 = \infty$. A first approximation of the posterior model probability is given by

$$\widehat{\pi}_{\epsilon_1}(m \mid \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{\{m_k^1 = m\}}.$$

A multivariate Kernel Density Estimator (KDE) $K_h$ with bandwidth $h$ is then fitted to the parameter values associated to each model with

$$\widehat{\pi}_{\epsilon_1}(\boldsymbol{\theta} \mid m, \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\widehat{\pi}_{\epsilon_1}(m \mid \boldsymbol{x})} K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^1\|) \mathbb{I}_{\{m_k^1 = m\}}, \ m \in \{1, \dots, M\}.$$

At a given generation $g \in \{1, \dots, G\}$ and for each model $m \in \{1, \dots, M\}$, we hold an approximation of the posterior model evidence $\widehat{\pi}_{\epsilon_{g-1}}(m \mid \boldsymbol{x})$ and the posterior distribution of the parameters $\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid m, \boldsymbol{x})$. New particles $(m_k^g, \boldsymbol{\theta}_k^g)$ are proposed by sampling from $\pi(m)$ and $\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid m_k^g, \boldsymbol{x})$ until the synthetic data $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid m_k^g, \boldsymbol{\theta}_k^g)$ satisfies $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon_{g-1}, \boldsymbol{x}}$. Sampling is performed repeatedly until $K$ particles are selected. The acceptance threshold $\epsilon_g$ becomes the empirical quantile of order $\alpha \in (0, 1)$ of the distances $\|\boldsymbol{x}^+ - \boldsymbol{x}_k^+\|$, $k = 1, \dots K$. To each particle is assigned a weight given by

$$w_k^g \propto \frac{\pi(\boldsymbol{\theta}_k^g \mid m_k^g)}{\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta}_k^g \mid m_k^g, \, \boldsymbol{x})} \mathbb{I}_{\mathcal{B}_{\epsilon_g, \boldsymbol{x}}}(\boldsymbol{x}_k), \quad k = 1, \dots, K.$$

The model probability is then updated

$$\widehat{\pi}_{\epsilon_g}(m \mid \boldsymbol{x}) = \sum_{k=1}^{K} w_k^i \mathbb{I}_{\{m_k^g = m\}},$$

along with the posterior distribution of the parameters associated to each model

$$\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta} \mid m, \boldsymbol{x}) = \sum_{k=1}^{K} \frac{w_k^g}{\widehat{\pi}_{\epsilon_g}(m \mid \boldsymbol{x})} K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^g\|) \, \mathbb{I}_{\{m_k^g = m\}}, \ m = 1, \dots, M.$$

The algorithm is summarized in Algorithm 6 of Appendix A.

Our ABC implementation when evaluating posterior model probabilities is tested on a simple example where we aim at fitting individual claim sizes generated from a lognormal distribution

$$u_1, \dots, u_n \overset{\text{i.i.d.}}{\sim} \mathsf{LogNorm}(\mu = 0, \sigma = 1),$$

with associated p.d.f.

$$f(x\,;\,\mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\Big[-\frac{(\ln x - \mu)}{2\sigma^2}\Big], \quad x > 0.$$

The lognormal model is compared to a gamma model $\mathsf{Gamma}(r, m)$ with p.d.f.

$$f(x\,;\,r, m) = \frac{\mathrm{e}^{-x/m} x^{r-1}}{m^r \Gamma(r)}, \quad x > 0,$$

and a Weibull model $\mathsf{Weib}(r, m)$ with p.d.f.

$$f(x\,;\,k,\beta) = \frac{k}{\beta}\big(\frac{x}{\beta}\big)^{k-1}\exp\big[-\big(\frac{x}{\beta}\big)^{k}\big], \quad x > 0.$$

Uniform priors are set over the parameters of all the model:

$$\mu \sim \mathsf{Unif}(-20, 20), \text{ and } \sigma \sim \mathsf{Unif}(0, 5),$$

for the lognormal model,

$$r \sim \mathsf{Unif}(0, 5), \text{ and } m \sim \mathsf{Unif}(0, 100),$$

for the gamma model, and

$$k \sim \mathsf{Unif}(\tfrac{1}{10}, 5), \text{ and } \beta \sim \mathsf{Unif}(0, 100),$$

for the Weibull model. The likelihood function of the data $\boldsymbol{u} = u_1, \dots, u_n$ may be computed for these loss models and the model probability can be estimated through the Sequential Monte Carlo sampler of the `PyMC3` library. The computation of model probabilities via ABC is more demanding than simply estimating parameters. Namely, the number of iterations must be larger to lead to an accurate model probability estimation. We therefore set the number of iterations to $G = 25$. The model evidences of all three models are reported in Table 1 for samples of size ranging from 25 to 200.

| | PyMC3 | | | ABC | | |
|---|---|---|---|---|---|---|
| | Gamma | LogNorm | Weib | Gamma | LogNorm | Weib |
| sample size | | | | | | |
| 25 | 0.42 | 0.18 | 0.40 | 0.51 | 0.15 | 0.34 |
| 50 | 0.25 | 0.64 | 0.11 | 0.31 | 0.51 | 0.18 |
| 75 | 0.04 | 0.95 | 0.01 | 0.15 | 0.79 | 0.07 |
| 100 | 0.01 | 0.99 | 0.00 | 0.07 | 0.91 | 0.02 |
| 150 | 0.00 | 1.00 | 0.00 | 0.01 | 0.99 | 0.00 |
| 200 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |

Table 1: Model evidence for individual claim sizes data simulated by a $\mathsf{LogNorm}(\mu = 0, \sigma = 1)$ model. The model evidences computed via ABC fare well compared to the model evidences computed by relying on the likelihood function.

Further approximate Bayesian model evidence computations are proposed in Section 4 and Section 5 when the data at hand is aggregated.

# 4 Simulation Study

This section aims at studying the finite sample behavior of our ABC implementation on two case studies based on simulated data. In Section 4.1, we assume that the claim sizes are independent from the claim frequency and that the insurer have access to the right truncated aggregated sum. In Section 4.2, we consider a model in which the average of the claim sizes depends on the number of claims and the insurer have access to the total claim sizes for each time period.

Our goal is to check whether our ABC sampling algorithm manage to return a posterior sample that concentrates around the true value when the model is well specified. Another question is how does the ABC posterior behave when the model is misspecified? The ABC posterior samples are compared, in that case, to the maximum likelihood estimates of the parameters.

Finally, we assume that the claim frequency data is available in addition to the aggregated data. The number of claims is then input directly in our ABC implementation to specify how many claim sizes should be generated for each time period. It reduces the computing time, and allow us to drop the parametric assumption over the claim frequency distribution and direct our focus on the claim amounts distribution.

In both examples, the number of generations for ABC is set to $G = 7$ and each consists of $K = 1000$ particles when only one model is considered and when the claim frequency is not available. Knowing the number of claims leads to a reduction in calculation time, which in turn allows us to bring the number of iterations to $G = 10$.

## 4.1 Negative-Binomial Weibull model with truncation

Let the claim frequency be negative binomial distributed

$$n_1, \dots, n_t \overset{\text{i.i.d.}}{\sim} \mathsf{NegBin}(\alpha = 4,\, p = \tfrac{2}{3}),$$

with p.m.f.

$$p_N(n\,;\,\alpha, p) = \binom{\alpha + n - 1}{n} p^{\alpha}(1-p)^n, \quad n \geq 0,$$

while the claim sizes are Weibull distributed

$$u_{s,1}, \dots, u_{s,n_s} \overset{\text{i.i.d.}}{\sim} \mathsf{Weib}(k = \tfrac{1}{3},\, \beta = 1), \quad s = 1, \dots, t.$$

The available data is the aggregated claim size in excess of a threshold $c$, given by

$$x_s = \Big(\sum_{i=1}^{n_s} u_{s,i} - c\Big)_+, \quad s = 1, \dots, t. \tag{13}$$

It corresponds to the data available to a reinsurance company within the frame of a global non-proportional treaty over a non-life insurance portfolio. The cases $t = 50$ and $t = 250$ are considered. The prior distributions over the four parameters are

$$\alpha \sim \mathsf{Unif}(0, 10),\ p \sim \mathsf{Unif}(\tfrac{1}{1000}, 1),\ k \sim \mathsf{Unif}(\tfrac{1}{10}, 10),\ \text{and } \beta \sim \mathsf{Unif}(0, 20). \tag{14}$$

Figure 2 displays the ABC posterior samples when only using the aggregated data (13).

The $p$ and $k$ posteriors are quite informative, whereas the scale parameters $\alpha$ and $\beta$ are skewed in opposite directions and seem to compensate for each other.

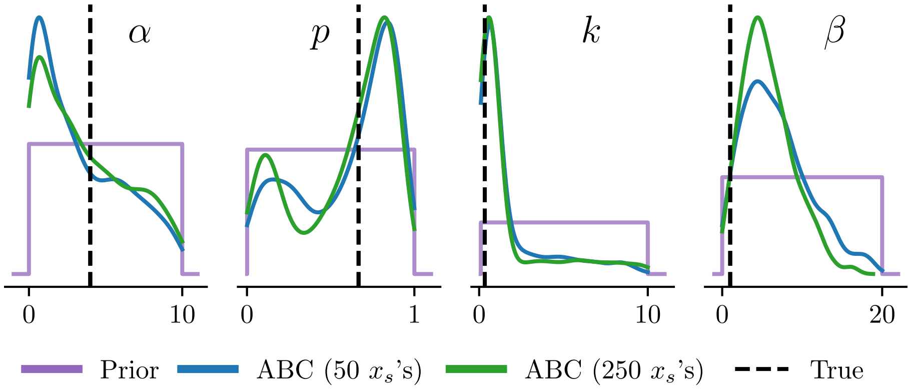

Figure 2: ABC posterior samples of a $\mathsf{NegBin}(\alpha, p)$–$\mathsf{Weib}(k, \beta)$ model fitted to data simulated by a $\mathsf{NegBin}(\alpha = 4, p = \frac{2}{3})$–$\mathsf{Weib}(k = \frac{1}{3}, \beta = 1)$. The posteriors are based on **50 observations** and **250 observations** of the $x_s$ summaries as in (13).

We then include the claim frequencies $n_s$ in the input data of our ABC algorithm to see if this helps in getting posterior samples closer to the target. Figure 3 displays the ABC posterior samples of the claim sizes model when the claim frequency data is available in addition to the summaries (13).

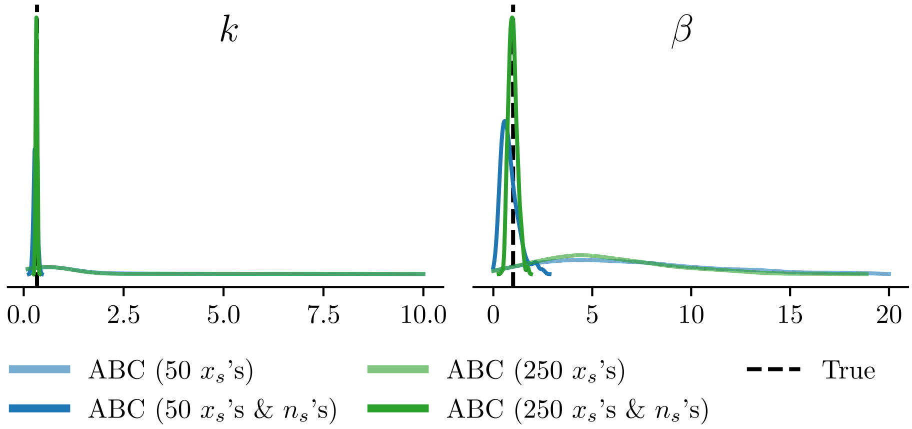

Figure 3: ABC posterior samples of a $\mathsf{Weib}(k, \beta)$ model fitted to data simulated by a $\mathsf{NegBin}(\alpha = 4, p = \frac{2}{3})$–$\mathsf{Weib}(k = \frac{1}{3}, \beta = 1)$. The data includes each summary $x_s$ as in (13) and each frequency $n_s$. The posterior with **250 observations** is a slight improvement over the one with **50 observations**.

The ABC posteriors are very strongly concentrated around the true values $k = \frac{1}{3}$ and $\beta = 1$ compared to that of Figure 2.

We now turn to the case where the model is misspecified. The same data simulated from a $\mathsf{NegBin}(\alpha = 4, p = \frac{2}{3})$–$\mathsf{Weib}(k = \frac{1}{3}, \beta = 1)$ model is used to fit a $\mathsf{NegBin}(\alpha, p)$–$\mathsf{Gamma}(r, m)$ model. The prior distributions over the four parameters are uniform with

$$\alpha \sim \mathsf{Unif}(0,20), \quad p \sim \mathsf{Unif}(\tfrac{1}{1000},1), \quad r \sim \mathsf{Unif}(0,10), \text{ and } m \sim \mathsf{Unif}(0,20). \tag{15}$$

The true values for the gamma distribution parameters are replaced by the maximum likelihood estimators based on a large sample of Weibull distributed individual losses. Figure 4 displays the ABC posterior samples when only using the aggregated data (13).

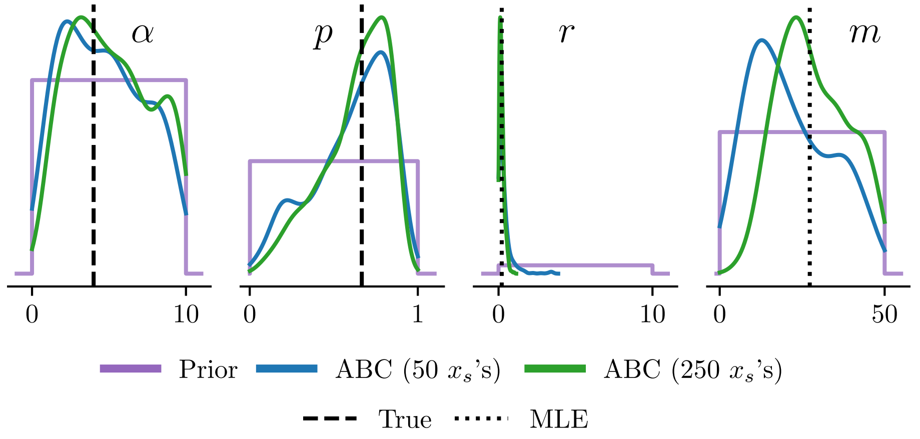

Figure 4: ABC posterior samples of a $\mathsf{NegBin}(\alpha, p)$–$\mathsf{Gamma}(r, m)$ model fitted to data simulated by a $\mathsf{NegBin}(\alpha = 4, p = \frac{2}{3})$–$\mathsf{Weib}(k = \frac{1}{3}, \beta = 1)$ model. The data only includes the summaries $x_s$ as in (13). The target values are the **true values** for $\alpha$ and $p$ and the **MLE estimates** for $k$ and $\beta$ given the claim sizes.

The ABC posterior distributions are informative regarding $p$, $r$ and $m$, however the algorithm does not improve significantly the prior assumption over $\alpha$.

Figure 5 displays the ABC posterior samples for the parameter of the gamma distribution when the claim frequency data is available in addition to the summaries (13).

The posterior sample for $m$ does not seem to center around the maximum likelihood estimator. Note that the situation improves greatly when considering a larger sample, of size 500 say. Also note that by fitting a gamma model on the individual losses, the mean *a posteriori* for $m$ is around 40, which may explain why our ABC posterior somewhat miss the target.

To perform model selection, we specify to our ABC algorithm the Weibull and the gamma distribution as competing models for the claim sizes and we set uniform priors as in (14) and (15) over the parameters. The model evidences computed via ABC are reported in Table 2.

When only the summaries $x_s$ are available and the claim frequency is modeled by a negative binomial distribution then ABC cannot decide between the Weibull and the gamma distributions. When the claim counts $n_s$ are also available then ABC favors greatly the Weibull model for the claim sizes.

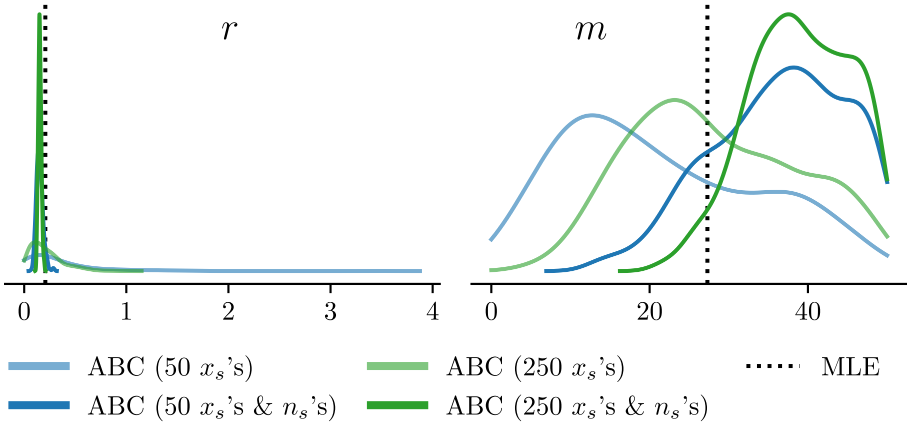

Figure 5: ABC posterior samples of a $\mathsf{Gamma}(r, m)$ model fitted to data simulated by a $\mathsf{NegBin}(\alpha = 4, p = \frac{2}{3})$–$\mathsf{Weib}(k = \frac{1}{3}, \beta = 1)$ model. The data includes each summary $x_s$ as in (13) and each frequency $n_s$.

| Sample Sizes | Frequency Model | |
|---|---|---|
| | Negative Binomial | Observed Frequencies |
| 50 | 0.51 | 0.88 |
| 250 | 0.44 | 1.00 |

Table 2: Model evidence in favor of a $\mathsf{Weib}(k, \beta)$ model when compared against a $\mathsf{Gamma}(r, m)$ model for data simulated by a $\mathsf{NegBin}(\alpha = 4, p = \frac{2}{3})$–$\mathsf{Weib}(k = \frac{1}{3}, \beta = 1)$ model. The values should increase to 1 as the sample size increases.

## 4.2 Dependence between the claim frequency and severity

Let the claim frequency be Poisson distributed

$$n_1, \dots, n_t \overset{\text{i.i.d.}}{\sim} \mathsf{Poisson}(\lambda = 4),$$

with p.m.f.

$$p_N(k\,;\,\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \quad k \geq 0.$$

The claim sizes are assumed to be exponentially distributed with a scale parameter depending on the observed claim frequency with

$$u_{s,1}, \dots, u_{s,n_s} \mid n_s \overset{\text{i.i.d.}}{\sim} \mathsf{Exp}(\mu = \beta\,\mathrm{e}^{\delta n_s}), \text{ for } s = 1, \dots, t.$$

We denote this $\boldsymbol{u}_s \sim \mathsf{DepExp}(n_s\,;\,\beta, \delta)$, and take $\beta = 2$ and $\delta = 0.2$. The resulting conditional p.d.f. is

$$f_U(x \mid n\,;\,\beta, \delta) = \frac{1}{\beta\mathrm{e}^{\delta n}} \exp\bigl(-\frac{x}{\beta\mathrm{e}^{\delta n}}\bigr), \quad x > 0.$$

This dependence structure relates to the insurance ratemaking practice where premiums are computed using the average claim frequency and severity predicted by a generalized linear models (GLM). In the classical setting, the claim frequency is assumed to be Poisson distributed and the claim sizes are gamma distributed.

The GLM are then fitted independently for the claim frequency and the claim severity, we refer to Renshaw [25]. Empirical studies, like the one conducted in Frees et al. [12], have shown how the claim sizes may vary with the claim frequency. A standard practice is then to include the predicted claim frequency as a covariate within the claim sizes model, see for instance Shi et al. [28]. It then reduces to bump the expectation of the severity by a factor $\mathrm{e}^{\delta n_s}$. Our case study is inspired by Garrido et al. [13, Example 3.1]. The available data is the aggregated claim sizes

$$x_s = \sum_{k=1}^{n_s} u_{s,k}, \quad s = 1, \dots, t. \tag{16}$$

We consider data histories of length $t = 50$ and 250.

Uniform prior distributions are set over the model parameters as

$$\lambda \sim \mathsf{Unif}(0, 10), \ \beta \sim \mathsf{Unif}(0, 20), \ \text{and } \delta \sim \mathsf{Unif}(-1, 1).$$

Figure 6 displays the posterior samples of $\lambda$ the parameter of the Poisson distribution, $\beta$ the scale parameter of the exponential parameter and $\delta$ the frequency/severity correlation parameter.

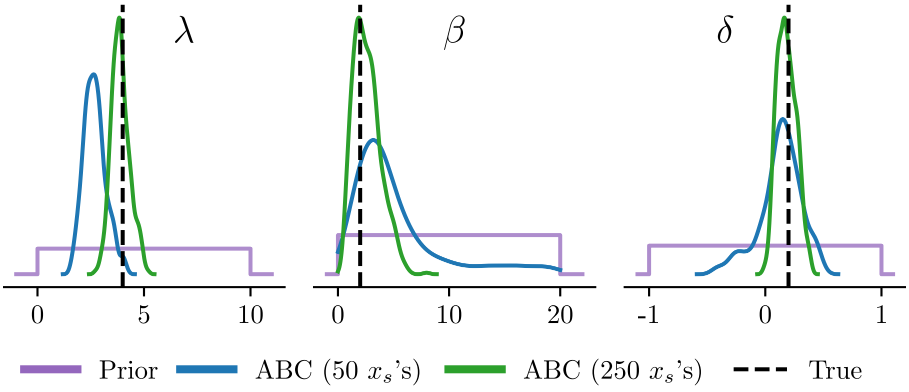

Figure 6: ABC posterior samples of a $\mathsf{Poisson}(\lambda)$–$\mathsf{DepExp}(n; \beta, \delta)$ model fitted to data simulated by a $\mathsf{Poisson}(\lambda = 4)$–$\mathsf{DepExp}(n; \beta = 2, \delta = 0.2)$. The data only includes the summaries $x_s$ as in (16).

The algorithm does a tremendous job on this example even without including the claim count information of each time period.

Figure 7 displays the ABC posterior samples associated to the claim sizes distribution $\mathsf{DepExp}(n; \beta, \delta)$ when including the frequency information in addition to the summaries (16).

As already noted, the inclusion of the claim frequency information improves the ABC posterior samples.

## 5 Application to a real-world insurance data set

We consider an open source insurance data set named `ausautoBI8999` consisting of 22, 036 settled personal injury insurance claims in Australia, the first five

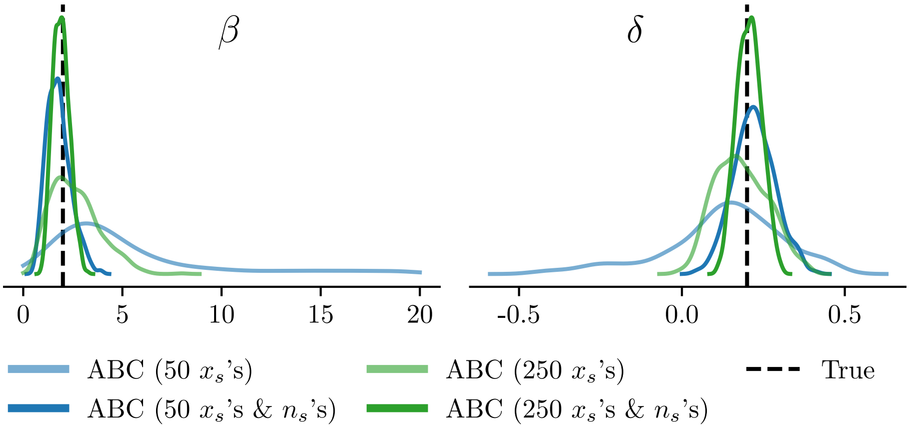

Figure 7: ABC posterior samples of a $\mathsf{DepExp}(n; \beta, \delta)$ model fitted to data simulated by a $\mathsf{Poisson}(\lambda = 4)$–$\mathsf{DepExp}(n; \beta = 2, \delta = 0.2)$. The data includes each summary $x_s$ as in (16) and each frequency $n_s$.

observations are displayed in Table 3.

| Date | Month | Claim Severity |
|---|---:|---:|
| 1993-10-01 | 52 | 87.75 |
| 1994-02-01 | 56 | 353.62 |
| 1994-02-01 | 56 | 688.83 |
| 1994-05-01 | 59 | 172.80 |
| 1994-09-01 | 63 | 43.29 |

Table 3: `ausautoBI8999` personal injury claim data.

The data is accessible from the `R` package `CASDatasets`, see Dutang and Charpentier [10]. The data is then aggregated monthly by reporting the number of claims along with the sum of all the compensations associated to each month, see Table 4.

| Month | Claim Frequency | Total Claim Severity |
|---|---:|---:|
| 49 | 149 | 1.55e+06 |
| 50 | 188 | 3.21e+06 |
| 51 | 196 | 4.81e+06 |
| 52 | 203 | 4.22e+06 |
| 53 | 226 | 5.27e+06 |

Table 4: Monthly aggregated data.

Descriptive statistics for the claim sizes, claim frequencies and the aggregated claims sizes are reported in Table 5.

We are going to use ABC to fit and compare loss models using only the monthly aggregated data in Table 4. We would like to know whether the results differ from fitting the same loss models but using the individual claim sizes data in Table 3.

| Statistics | Claim Severity | Claim Frequency | Total Claim Severity |
|---|---|---|---|
| Count | 2.20e+04 | 6.90e+01 | 6.90e+01 |
| Mean | 3.84e+04 | 3.19e+02 | 1.23e+07 |
| Std | 9.10e+04 | 1.09e+02 | 5.22e+06 |
| Min | 9.96e+00 | 9.40e+01 | 1.55e+06 |
| 25% | 6.30e+03 | 2.31e+02 | 8.21e+06 |
| 50% | 1.39e+04 | 3.12e+02 | 1.20e+07 |
| 75% | 3.51e+04 | 3.81e+02 | 1.55e+07 |
| Max | 4.49e+06 | 6.06e+02 | 2.63e+07 |

Table 5: Descriptive statistics of the claim data.

| Severity model | Parameters | MLE | BIC |
|---|---|---|---|
| Gamma | $r$ | 4.09e+0 | 6.46e+5 |
| | $m$ | 5.35e+3 | |
| Weibull | $k$ | 7.08e-1 | 5.03e+5 |
| | $\beta$ | 2.86e+4 | |
| Lognormal | $\sigma$ | 9.56e+0 | 5.00e+5 |
| | $\mu$ | 1.46e+0 | |

Table 6: Maximum likelihood estimates of a gamma, Weibull and lognormal distribution based on the individual claim sizes data.

We start by studying the individual loss distribution. We fit a gamma, a lognormal and a Weibull model to the data shown in Table 3 using maximum likelihood estimation. The estimates of the parameters are given in Table 6 and will serve as benchmark for our ABC posterior samples.

The lognormal distribution seems to provide the best fit when looking at the values of the Bayesian Information Criteria (BIC). This result is visually confirmed by the quantile-quantile plots displayed in Figure 8.

We then investigate the stationarity of the individual loss distribution by fitting the three loss models to the data associated to each time period separately. Figures 9 to 11 display the parameters of the gamma, Weibull and lognormal distribution respectively depending on the time period considered.

The parameters of the Weibull and gamma distributions exhibit a high variability, see Figures 9 and 10, while the parameters of the lognormal distribution are more stable, see Figure 11. The model evidences, displayed in Figure 12, are computed using the Schwarz criterion that approximates the Bayes factor using the maximum likelihood estimators and the BIC.

The model probabilities mostly favor the lognormal model.

We use ABC to fit a $\mathsf{NegBin}(\alpha, p)$–$\mathsf{LogNorm}(\mu, \sigma)$ model to the total claim severities data in Table 4 which consists of $t = 69$ summaries of the form

$$x_s = \sum_{k=1}^{n_s} u_{s,k}, \quad s = 1, \dots, t. \tag{17}$$

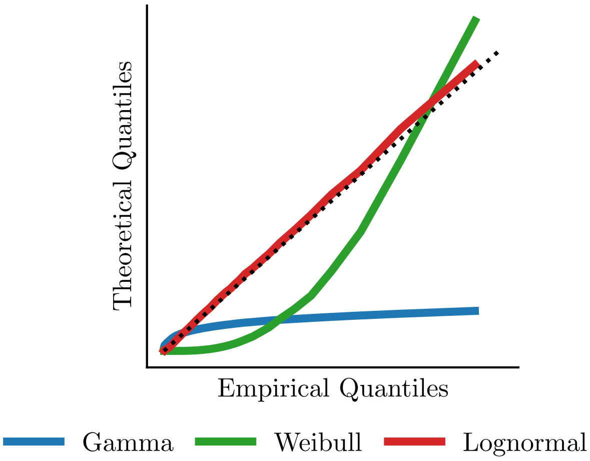

Figure 8: Quantile-quantile plots associated to the gamma, Weibull and lognormal models fitted to the individual claim sizes data.

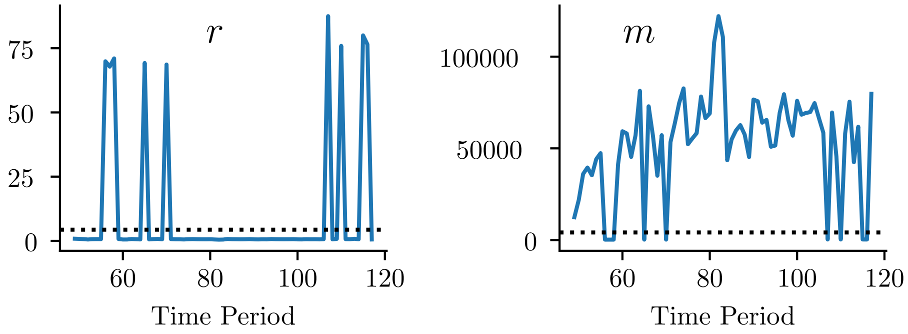

Figure 9: Parameters of the gamma model.

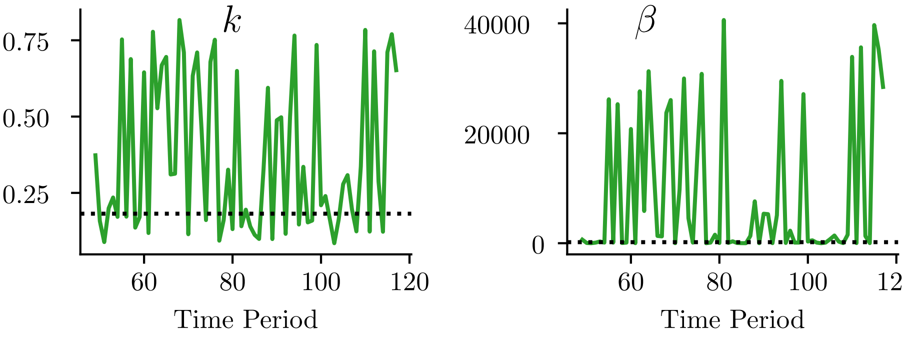

Figure 10: Parameters of the Weibull model.

We consider two sets of prior assumptions over the parameters:

1. $\alpha \sim \mathsf{Unif}(0, 20)$, $p \sim \mathsf{Unif}(\frac{1}{1000}, 1)$, $\mu \sim \mathsf{Unif}(-10, 10)$, and $\sigma \sim \mathsf{Unif}(0, 10)$,
2. $\alpha \sim \mathsf{Unif}(0, 20)$, $p \sim \mathsf{Unif}(\frac{1}{1000}, 1)$, $\mu \sim \mathsf{Unif}(0, 20)$, and $\sigma \sim \mathsf{Unif}(0, 10)$.

Prior settings 1 and 2 only differ in the boundaries of the uniform distribution

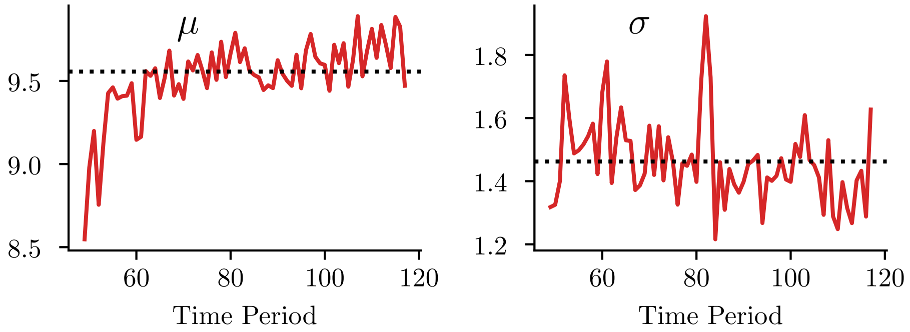

Figure 11: Parameters of the lognormal model.

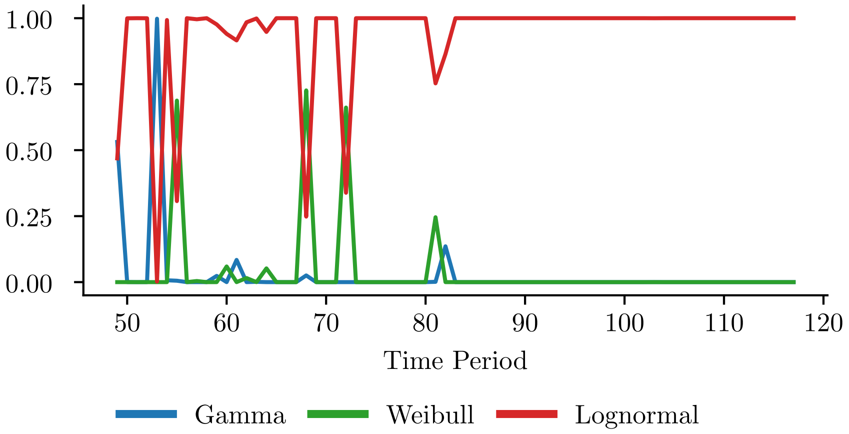

Figure 12: Model evidence for the gamma, lognormal and Weibull models.

of $\mu$. We opt for a more intensive ABC calibration compared to that of section 4. The number of iterations is fixed at $G = 20$ when the claim frequencies are known and $G = 15$ when they are not. The ABC posterior samples of the $\mathsf{NegBin}(\alpha, p)$–$\mathsf{LogNorm}(\mu, \sigma)$ model using only the summaries $x_s$ in (17) are shown in Figure 13.

The results with prior settings 1 and 2 are noticeably different. More specifically, the ABC posterior are tighter and more centered around the MLE estimates with prior 2 at least when it comes to estimating the parameters $p$, $\mu$ and $\sigma$.

The ABC posterior samples when including the claim frequency information are shown in Figure 14. We keep the same prior assumptions over $\mu$ and $\sigma$.

Including the claim frequency data helps in making the results consistent from one prior setting to the other.

We now turn to the problem of selecting a model for the claim sizes, so we specify a negative binomial distribution $\mathsf{NegBin}(\alpha, p)$ with uniform prior distributions

$$\alpha \sim \mathsf{Unif}(0, 20), \quad p \sim \mathsf{Unif}(0, 1)$$

to model the claim frequency and let our ABC algorithm pick a claim amounts models among the following:

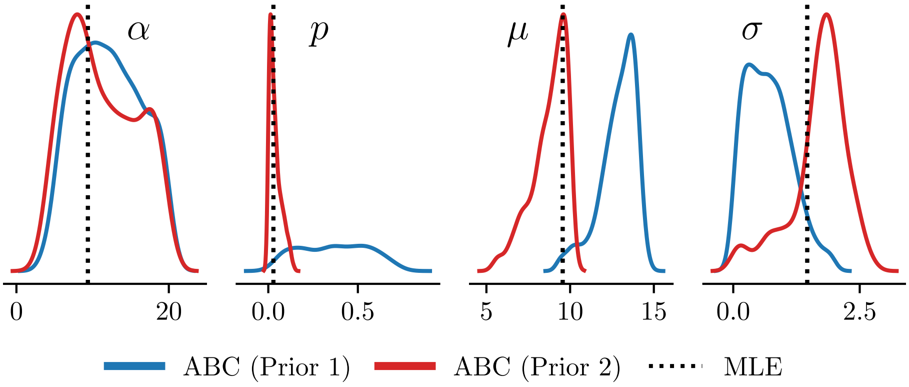

Figure 13: ABC posterior samples of a $\mathsf{NegBin}(\alpha, p)$–$\mathsf{LogNorm}(\mu, \sigma)$ model fitted to a real world insurance data set. The data includes the total claim severities (17) data in Table 4. The posterior samples are closer to the **MLE estimates** with **prior 2** than with **prior 1**.

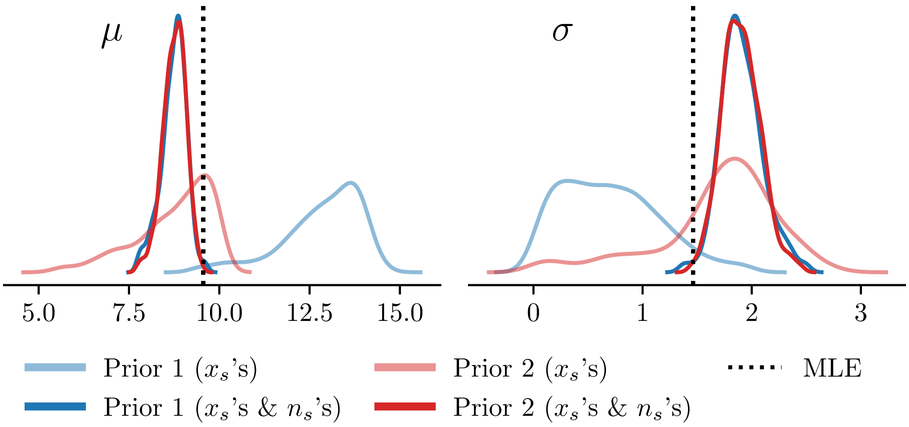

Figure 14: ABC posterior samples of a $\mathsf{LogNorm}(\mu, \sigma)$ model fitted to a real world insurance data set. The data includes the total claim severities and the claim frequencies in Table 4. When the $x_s$'s and $n_s$'s are both observed, the posterior samples with **Prior 1** and **Prior 2** almost totally overlap and are reasonably close to the **MLE estimates**.

- $\mathsf{Weib}(k, \beta)$ with prior distributions

$$k \sim \mathsf{Unif}(\tfrac{1}{1000}, 1), \quad \beta \sim \mathsf{Unif}(0, 4 \times 10^4),$$

- $\mathsf{Gamma}(r, m)$ with prior distributions

$$r \sim \mathsf{Unif}(0, 100), \quad \beta \sim \mathsf{Unif}(0, 1.5 \times 10^5),$$

- $\mathsf{LogNorm}(\mu, \sigma)$ with prior distributions

$$\mu \sim \mathsf{Unif}(5, 10), \quad \sigma \sim \mathsf{Unif}(0, 3).$$

The bounds of the uniform distributions are set to reflect the variability of the parameters in Figures 9 to 11. The model evidences are reported in Table 7.

| Frequency Model | Severity Model | | |
|---|---|---|---|
| | Gamma | Lognormal | Weibull |
| Negative Binomial | 0.92 | 0.01 | 0.07 |
| Observed Frequencies | 0.00 | 0.49 | 0.51 |

Table 7: ABC model evidence with the claim frequency and the aggregated claim sizes data.

We see that ABC strongly favors the gamma model when the claim frequency is assumed to have a negative binomial distribution. When including the claim count, ABC discards the gamma model but is unable to decide between the Weibull or the lognormal model. This result is of course a little disappointing but probably means that ABC would need more than 69 observations to pick the right model.

# 6 Conclusion

This paper is a case study of an ABC application in insurance. We showed how to use this method to calibrate insurance loss models with limited information (one data point per time period). The fact that the method does not require the knowledge of the likelihood function permits to go beyond the classical setting where independence is assumed between the claim frequency and the claim sizes.

An ABC routines essentially relies on two things: (i) an efficient sampling strategy and (ii) a reliable measure of dissimilarity between samples of data. We put together an ABC routine that implements a parallel sequential Monte Carlo sampler and uses the Wasserstein distance to compare the synthetic data to the observed one. The python code may be downloaded from the following GitHub repository https://github.com/LaGauffre/ABCFitLoMo.

ABC has become over the years a common practice in a variety of fields ranging from ecology to genetics. We believe that ABC could be also applied to a wide range of sophisticated models that arise in finance and insurance.

# Acknowledgments

Patrick J. Laub conducted this research within the DAMI – Data Analytics and Models for Insurance – Chair under the aegis of the Fondation du Risque, a joint initiative by UCBL and BNP Paribas Cardif.

# References

[1] Hansjörg Albrecher, Jan Beirlant, and Jozef L Teugels. *Reinsurance: Actuarial and Statistical Aspects*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2017.

[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] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2): 235–269, 2019.

[4] Michael GB Blum. Approximate Bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association*, 105(491): 1178–1187, 2010.

[5] Michael GB Blum, Maria Antonieta Nunes, Dennis Prangle, and Scott A Sisson. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

[6] Martin Bøgsted and Susan M Pitts. Decompounding random sums: a nonparametric approach. *Annals of the Institute of Statistical Mathematics*, 62(5):855–872, 2010.

[7] Boris Buchmann and Rudolf Grübel. Decompounding: an estimation problem for Poisson random sums. *Ann. Statist.*, 31(4):1054–1074, 08 2003. doi: 10.1214/aos/1059655905. URL `https://doi.org/10.1214/aos/1059655905`.

[8] Alberto J Coca. Efficient nonparametric inference for discretely observed compound Poisson processes. *Probability Theory and Related Fields*, 170 (1-2):475–523, 2018.

[9] 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.

[10] C Dutang and A Charpentier. CASdatasets: Insurance datasets (official website). *http://cas.uqam.ca/*, 2016.

[11] Paul Fearnhead and Dennis Prangle. 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, 2012.

[12] Edward W. Frees, Jie Gao, and Marjorie A. Rosenberg. Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 15(3):377–392, 2011. doi: 10.1080/10920277.2011.10597626. URL `https://doi.org/10.1080/10920277.2011.10597626`.

[13] J. Garrido, C. Genest, and J. Schulz. Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 70:205 – 215, 2016. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2016.06.006. URL `http://www.sciencedirect.com/science/article/pii/S0167668715303358`.

[14] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[15] Pierre-Olivier Goffard, S Rao Jammalamadaka, and Simos G. Meintanis. Goodness-of-fit tests for compound distributions with applications in insurance. 2019.

[16] Aude Grelaud, Christian P Robert, Jean-Michel Marin, Francois Rodolphe, and Jean-François Taly. ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, 4(2):317–335, 2009.

[17] Shota Gugushvili, Frank van der Meulen, and Peter Spreij. A non-parametric Bayesian approach to decompounding from high frequency data. *Statistical Inference for Stochastic Processes*, 21(1):53–79, 2018.

[18] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[19] Stuart A Klugman, Harry H Panjer, and Gordon E Willmot. *Loss Models: From Data to Decisions*, volume 715. John Wiley & Sons, 2012.

[20] Robert McCulloch and Peter E Rossi. A Bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics*, 49(1-2):141–168, 1991.

[21] Gareth Peters and Scott Sisson. Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk*, 1(3), 2006.

[22] 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.

[23] Dennis Prangle, Paul Fearnhead, Murray P Cox, Patrick J Biggs, and Nigel P French. Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, 13(1):67–82, 2014.

[24] Dennis Prangle, Richard G Everitt, and Theodore Kypraios. A rare event approach to high-dimensional approximate Bayesian computation. *Statistics and Computing*, 28(4):819–834, 2018.

[25] 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.

[26] FJ Rubio and Adam M Johansen. A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654, 2013.

[27] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

[28] Peng Shi, Xiaoping Feng, and Anastasia Ivantsova. Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 64:417 – 428, 2015. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2015.07.006. URL `http://www.sciencedirect.com/science/article/pii/S0167668715001183`.

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

[30] Vitor C Sousa, Marielle Fritz, Mark A Beaumont, and Lounès Chikhi. Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics*, 181(4):1507–1519, 2009.

[31] Tina Toni and Michael P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26 (1):104–110, 10 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp619. URL `https://doi.org/10.1093/bioinformatics/btp619`.

[32] Bert van Es, Shota Gugushvili, and Peter Spreij. A kernel type nonparametric density estimator for decompounding. *Bernoulli*, 13(3):672–694, 08 2007. doi: 10.3150/07-BEJ6091. URL `https://doi.org/10.3150/07-BEJ6091`.

# A Algorithmic details

**Algorithm 5** Sequential Monte Carlo Approximate Bayesian Computation Algorithm.

```
for $k = 1 \to K$ do
    repeat
        generate $\boldsymbol{\theta}_k^1 \sim \pi(\boldsymbol{\theta})$
        generate $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k^1)$
    until $\boldsymbol{x}_k \in \mathcal{B}_{\infty,\boldsymbol{x}}$
end for
compute $\widehat{\pi}_{\epsilon_1}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^1\|)$
for $g = 2 \to G$ do
    for $k = 1 \to K$ do
        repeat
            generate $\boldsymbol{\theta}_k^g \sim \widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid \boldsymbol{x})$
            generate $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid \boldsymbol{\theta}_k^g)$
        until $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon_{g-1},\boldsymbol{x}}$
    end for
    set $\epsilon_g = \text{Quantile}\big(\|\boldsymbol{x}^+ - \boldsymbol{x}_1^+\|, \ldots, \|\boldsymbol{x}^+ - \boldsymbol{x}_K^+\|\,;\, \alpha\big)$
    for $k = 1 \to K$ do
        set $w_k^g \propto \frac{\pi(\boldsymbol{\theta}_k^g)}{\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta}_k^g|\boldsymbol{x})} \mathbb{I}_{\mathcal{B}_{\epsilon_g,\boldsymbol{x}}}(\boldsymbol{x}_k)$
    end for
    compute $\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{k=1}^{K} w_k^g K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^g\|)$
end for
```

**Algorithm 6** ABC-SMC for model selection.

```
for $k = 1 \to K$ do
    repeat
        generate $m_k^1 \sim \pi(m)$
        generate $\boldsymbol{\theta}_k^1 \sim \pi(\boldsymbol{\theta} \mid m_k^1)$
        generate $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid m_k^1, \boldsymbol{\theta}_k^1)$
    until $\boldsymbol{x}_k \in \mathcal{B}_{\infty,\boldsymbol{x}}$
end for
for $m = 1, \dots, M$ do
    compute $\widehat{\pi}_{\epsilon_1}(m \mid \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}_{\{m_k^1=m\}}$
    compute $\widehat{\pi}_{\epsilon_1}(\boldsymbol{\theta} \mid m, \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{\widehat{\pi}_{\epsilon_1}(m|\boldsymbol{x})} K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^1\|)\mathbb{I}_{\{m_k^1=m\}}$
end for
for $g = 2 \to I$ do
    for $k = 1 \to K$ do
        repeat
            generate $m_k^g \sim \pi(m)$
            generate $\boldsymbol{\theta}_k^g \sim \widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid m_k^g, \boldsymbol{x})$
            generate $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid m_k^g, \boldsymbol{\theta}_k^g)$
        until $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon_{g-1},\boldsymbol{x}}$
    end for
    set $\epsilon_g = \mathrm{Quantile}(\|\boldsymbol{x}^+ - \boldsymbol{x}_1^+\|, \dots, \|\boldsymbol{x}^+ - \boldsymbol{x}_K^+\|\,;\, \alpha)$
    for $k = 1 \to K$ do
        set $w_k^g \propto \frac{\pi(\boldsymbol{\theta}_k^g|m_k^g)}{\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta}_k^g|m_k^g, \boldsymbol{x})}\mathbb{I}_{\mathcal{B}_{\epsilon_g,\boldsymbol{x}}}(\boldsymbol{x}_k)$
    end for
    for $m = 1, \dots, M$ do
        compute $\widehat{\pi}_{\epsilon_g}(m \mid \boldsymbol{x}) = \sum_{k=1}^{K} w_k^g \mathbb{I}_{\{m_k^g=m\}}$
        compute $\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta} \mid m, \boldsymbol{x}) = \sum_{k=1}^{K} \frac{w_k^g}{\widehat{\pi}_{\epsilon_g}(m|\boldsymbol{x})} K_h(\|\boldsymbol{\theta} - \boldsymbol{\theta}_k^g\|)\mathbb{I}_{\{m_k^g=m\}}$
    end for
end for
```