Approximate Bayesian Computations to fit and compare insurance loss models

Pierre-Olivier Goffard, Patrick Laub

▶ **To cite this version:**

Pierre-Olivier Goffard, Patrick Laub. Approximate Bayesian Computations to fit and compare insurance loss models. Insurance: Mathematics and Economics, 2021, 100, pp.350-371. 10.1016/j.insmatheco.2021.06.002 . hal-02891046v2

HAL Id: hal-02891046
https://hal.science/hal-02891046v2

Submitted on 29 Apr 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Approximate Bayesian Computations to fit and compare insurance loss models

Pierre-Olivier Goffard and Patrick J. Laub

Univ Lyon, Université Lyon 1, LSAF EA2429, and University of Melbourne, Australia

April 29, 2021

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. 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 two could be associated in an equivalent way with a policyholder, a group of policyholders or even an entire nonlife insurance portfolio. 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 $\boldsymbol{u}_s := (u_{s,1}, u_{s,2}, \dots, u_{s,n_s})$ are distributed as

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

Fitting these distributions is key for claim management purposes. For instance, it allows one to estimate the expected cost of claims and set the premium rate accordingly. The mixed nature of claim data, with a discrete and a continuous component, has lead to two different claim modelling strategies. The first strategy is to handle the claim frequency and the claim severity separately, see for instance Frees [12]. The second approach gathers the two constituents in a compound model for which data in aggregated form suffices. We take the later approach as we assume that the claim count and amounts $\{(n_1, \boldsymbol{u}_1), \dots, (n_t, \boldsymbol{u}_t)\}$ are unobservable. Instead, we only have access to some real-valued *summaries* of the claim data at each time, denoted by

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

Standard actuarial practice uses the aggregated claim sizes, defined as $\Psi(n, \boldsymbol{u}) = \sum_{i=1}^{n} u_i$, and assumes that the claim frequency is Poisson distributed while the severities are governed by a gamma distribution, we refer

to the works of Jørgensen and Souza [20]. This model is named after Tweedie [38] and is commonly used by practitioners for ratemaking, see the paper by Smyth and Jørgensen [35], as well as for claim reserving purposes, see the work of Wüthrich [40]. We want to mention that the Tweedie model is also popular to model the quantity of precipitation, see the work of Dunn [10]. This problem is of interest to insurers due to the impact of heavy rainfall episodes on insurance business, we refer to Lyubchich and Gel [24] for a convincing empirical study. The Tweedie model is already challenging to calibrate, we want to mention here the work of Zhang [42] for likelihood based approaches, but we wish to go beyond it. Our problem is to take some observations of these summaries $\boldsymbol{x} = (x_1, \dots, x_t)$ (or summaries plus frequency $\{\boldsymbol{n}, \boldsymbol{x}\}$) and find the $\boldsymbol{\theta}$ which best explains them for a given parametric model (this model being Tweedie or not Tweedie, see Xacur and Garrido [41]).

Yet another goal is to consider functions Ψ other than the sum because such incomplete data situations arise in reinsurance practice. Reinsurance treaties allows insurance companies to cede a part of their liability over a given time period to a reinsurance company. The reinsurer then only observes its payout at each time period that can be a proportion of the aggregated claim sizes

$$x_s = \alpha \sum_{i=1}^{n_s} u_{s,i}, \quad s = 1, \dots, t, \tag{2}$$

where $\alpha \in (0,1)$ in a quota-share treaty. In the case of a stop loss agreement, the reinsurer covers the risk that the insurer's total claim amount exceeds a threshold $c > 0$ and therefore only observes

$$x_s = \Big(\sum_{i=1}^{n_s} u_{s,i} - c \Big)_+, \quad s = 1, \dots, t. \tag{3}$$

Being able to gain insights into the claim frequency and the claim severity distributions based on the data (2) or (3) would help the reinsurer to better understand the risk they have underwritten. Additionally, it could be a preliminary analysis before suggesting the insurance company an excess of loss reinsurance treaty (XOL) where the reinsurance company takes on the part of each loss (instead of the overall sum) exceeding some threshold.

New methods of claims analysis must be able to handle an increase in the dimension of the data. Modern casualty and property insurance products usually include more than one type of coverage. If actuaries must provide a separate analysis of the claim data for each type of coverage, they could also consider jointly the data for two types of coverage to account for their inter-relation. A car accident can result in bodily injury and material damages thus triggering two indemnifications under each of the guarantees of the automobile insurance contract. Both losses are part of the same claim and are of course linked to the scale of the unfortunate event. The use of data at the aggregated level to fit multivariate Tweedie models has been investigated in the work of Shi et al. [33] for instance. We therefore show how to adapt our procedure to consider the bivariate extension of the data (1) but note that the method can also cope with higher dimensions.

The data considered in (1) may also be seen as the increments of a stochastic process $(Z_t)_{t \geq 0}$ observed at equispaced discrete points in time. If we take the summary to be the sum, then the underlying stochastic process is given by

$$Z_t = \sum_{i=1}^{N_t} U_i, \quad t \geq 0, \tag{4}$$

where $(N_t)_{t \geq 0}$ is a counting process and $(U_i)_{i \geq 1}$ is a sequence of nonnegative random variables. In classical risk theory, the process $(Z_t)_{t \geq 0}$ represents the liability of a nonlife insurance company up to time $t \geq 0$, we refer to

the book of Asmussen and Albrecher [1] for an overview. The number of claims reported at some time $t > 0$ is given by $(N_t)_{t\geq 0}$ and the U_i's are the compensations associated to each claim. The problem of studying the distribution of the jumps based on observations of Z_t was considered, with insurance applications in mind, by Buchmann and Grübel [6]. This problem is also interesting in the field of queueing theory to draw inference on the job size distribution when only having access to the workload. Traditionally, a decompounding (as coined by Buchmann and Grübel [6]) method builds a non-parametric estimate of the claim severity distribution based on the observations of the aggregated sums, see for instance van Es et al. [39], Coca [7] and Gugushvili et al. [18]. The method we propose effectively *decompound* the random sum but assumes that the jump sizes are driven by a parametric model. We then relax the Poisson arrival assumption to consider time dependent data instead of IID.

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

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{5}$$

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. [15].

The posterior distribution (5) 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 IID). If the claim sizes are not IID 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 $\widetilde{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta})$;

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

(iii) if $\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) \leq \epsilon$, where $\epsilon > 0$ is small, then store $\widetilde{\boldsymbol{\theta}}$,

where $\mathcal{D}(\cdot, \cdot)$ denotes a distance measure and ϵ 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 Ψ functions (1) without imposing common simplifying assumptions such as assuming the claim amounts are IID 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.

The basic ABC algorithm outlined above is, arguably, the simplest method of all types of statistical inference in terms of conceptual difficulty. At the same time, this simple method is perhaps the most difficult form

of inference in terms of computational cost. We must use a modified form of this basic regime to minimize (though not eliminate) the gigantic computational costs of ABC.

ABC is a somewhat young field (like machine learning), and the methodology of ABC and the other likelihood-free algorithms are currently the subject of intense research. As such, there are many variations of ABC which are under investigation, and there is no ironclad consensus on which variation of the ABC algorithm is the best. We intend for this work to simplify a reader's first steps into this field of modern computational Bayesian statistics, as we present a restrictive view of ABC instead of an overwhelming exhaustive list of every ABC variation. For a comprehensive overview on ABC, we refer to the monograph of Sisson et al. [34]; in finance and insurance, ABC has been considered in the context of operational risk management by Peters and Sisson [26] and for reserving purposes by Peters et al. [27]. After reading this work we'd encourage interested readers to consider the (subjectively) more conceptually difficult alternatives such as MCMC, ABC-MCMC, ABC-squared, Bayesian synthetic likelihood, variational Bayes, etc.

The rest of the paper is organized as follows. Section 2 provides an introduction to ABC algorithms and presents our specific implementation. Section 3 shows how to use ABC to fit an insurance loss model based on IID univariate, IID bivariate and time dependent data. The performance of our ABC implementation is illustrated on simulated data in Section 5 and on a real world insurance dataset in Section 6.

2 Approximate Bayesian Computation

ABC is a method for approximating the posterior probability $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ without using the likelihood function. It relies on the ability to generate synthetic data from the model being fit. Two ingredients are required for a successful ABC algorithm. First is a distance to measure the dissimilarity between the observed and synthetic data; we will use the Wasserstein distance as suggested in Bernton et al. [3]. Second is an efficient sampling scheme. The acceptance–rejection algorithm laid out in the introduction most often leads to considerable computing time. We instead put together an algorithm based on an adaptive importance sampling strategy called sequential Monte Carlo, see for instance Beaumont et al. [2], Del Moral et al. [8].

Consider some observed data $\boldsymbol{x} = (x_1, \dots x_n)$ and assume that the underlying model does not lead to a tractable likelihood function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Sampling from an approximated version of the posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ can be done in a likelihood-free way through acceptance–rejection. The procedure is summarized in Algorithm 1, where $\mathcal{D}(\cdot, \cdot)$ denotes some dissimilarity measure between the observed and fake data and ϵ corresponds to a tolerance level.

Algorithm 1 ABC acceptance–rejection sampling for continuous data

input observations $\boldsymbol{x} = (x_1, \dots, x_n)$, $\mathcal{D}(\cdot, \cdot)$ distance, $\epsilon > 0$ threshold
for $k = 1 \to K$ **do**
    repeat
        generate $\widetilde{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta})$
        generate $\widetilde{\boldsymbol{x}} \sim p(\boldsymbol{x} \mid \widetilde{\boldsymbol{\theta}})$
    until $\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon$ then **store** $\boldsymbol{\theta}$
    Set $\boldsymbol{\theta}_k = \widetilde{\boldsymbol{\theta}}$
end for
return $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ which are approximately $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ distributed

The procedure depicted in Algorithm 1 allows us to sample from an approximation of the posterior distribution

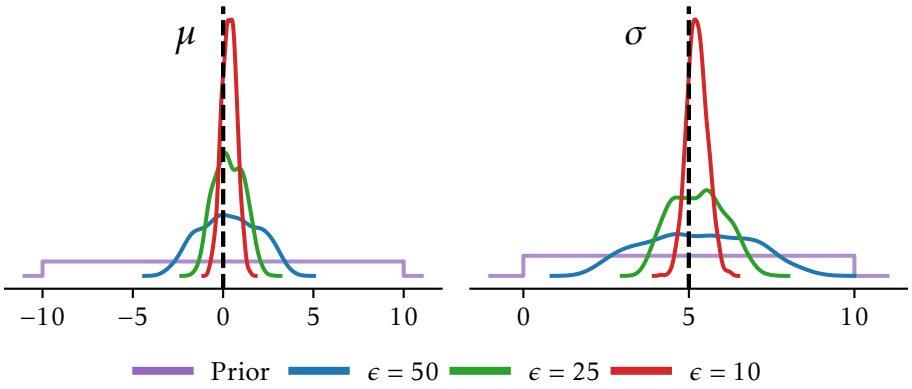

Figure 1: ABC posteriors for 250 simulated Normal(μ, σ) observations. The true parameters $\mu = 0$ and $\sigma = 5$. The fits are generated by Algorithm 1 with $\epsilon = 50$, $\epsilon = 25$, and for $\epsilon = 10$. They become much narrower around the true values as ϵ becomes more restrictive (smaller). All of the posteriors are an improvement over the $\mu \sim \text{Unif}(-10, 10)$, $\sigma \sim \text{Unif}(0, 10)$ **prior** distributions.

given by

$$\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto \pi(\boldsymbol{\theta}) \int_{\mathbb{R}^t} \mathbb{I}_{\{\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon\}}\, p(\widetilde{\boldsymbol{x}} \mid \boldsymbol{\theta})\, \mathrm{d}\widetilde{\boldsymbol{x}}, \tag{6}$$

where

$$\mathbb{I}_{\{\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon\}} = \begin{cases} 1, & \text{if } \mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Distribution (6) is called the ABC *posterior*. If the distance $\mathcal{D}$ is chosen to be

$$\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \mathcal{D}_p(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) \coloneqq \left(\frac{1}{n} \sum_{i=1}^{n} \rho(x_i, \widetilde{x}_i)^p \right)^{1/p} \tag{7}$$

where $\rho(\cdot, \cdot)$ denotes the ground distance in the observation space, for instance ρ is the absolute difference if the data is univariate or the Euclidean norm if the dimension is larger than 1, then the ABC posterior $\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x})$ converges toward the true posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ as ϵ tends to 0, see Rubio and Johansen [30].

Figure 1 shows a simple example of Algorithm 1 in action. It shows the ABC posteriors for some simple normally distributed data when ϵ takes on different values. Notice that as ϵ decreases, the ABC posterior becomes more confident (i.e. narrower) of the true values of $\boldsymbol{\theta} = (\mu, \sigma)$.

The combination of a small ϵ 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. [8], Beaumont et al. [2]. A sequence of ABC posteriors, similar to Figure 1, is constructed by gradually decreasing the tolerance ϵ through a sequence $(\epsilon_g)_{g \geq 1}$ and by leveraging the information about the ϵ_{g-1} approximate posterior to more intelligently create an improved ϵ_g approximate posterior.

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). The parameters of the algorithm are the number of generations G and the number of particles K. 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 $\widetilde{\boldsymbol{\theta}}$ are proposed by sampling repeatedly from $\widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ until the synthetic data $\widetilde{\boldsymbol{x}} \sim p(\boldsymbol{x} \mid \widetilde{\boldsymbol{\theta}})$ satisfies $\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon_{g-1}$. It goes on until K particles $\boldsymbol{\theta}_1^g, \ldots, \boldsymbol{\theta}_K^g$ are selected. We then need to define the next tolerance threshold ϵ_g which is used to calculate the particle weights

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

The tolerance threshold is chosen so as to maintain a specified effective sample size (ESS) of $K/2$ (as in Del Moral et al. [8]). Following Kong et al. [23], the ESS is estimated by $1/\sum_{k=1}^K (w_k^g)^2$. This weighted sampled allows us to update the posterior approximation as

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

where $K_{\boldsymbol{H}}$ is a multivariate KDE with smoothing matrix $\boldsymbol{H}$. A common choice for the KDE is the multivariate Gaussian kernel with a smoothing matrix set to twice the empirical covariance matrix assessed over the population of weighted particles $\{(\boldsymbol{\theta}_k^g, w_k^g)\}_{k=1,\ldots,K}$, see Beaumont et al. [2]. The pseudocode of the algorithm is provided in Algorithm 2.

Algorithm 2 Sequential Monte Carlo Approximate Bayesian Computation

set $\epsilon_0 = \infty$ and $\widehat{\pi}_{\epsilon_0}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \pi(\boldsymbol{\theta})$
for $g = 1 \to G$ **do**
  for $k = 1 \to K$ **do**
    repeat
      generate $\widetilde{\boldsymbol{\theta}} \sim \widehat{\pi}_{\epsilon_{g-1}}(\boldsymbol{\theta} \mid \boldsymbol{x})$
      generate $\widetilde{\boldsymbol{x}} \sim p(\boldsymbol{x} \mid \widetilde{\boldsymbol{\theta}})$
    until $\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon_{g-1}$
    set $\boldsymbol{\theta}_k^g = \widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{x}_k = \widetilde{\boldsymbol{x}}$
  end for
  find $\epsilon_g \leq \epsilon_{g-1}$ so that $\widehat{\text{ESS}} = \left[\sum_{k=1}^K (w_k^g)^2\right]^{-1} \approx K/2$, where

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

  compute $\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{k=1}^K w_k^g K_{\boldsymbol{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_k^g)$
end for

One small variation of Algorithm 2, which we use in the simulations below, is called *particle recycling*. Note that for each generation $g > 1$ we sample K new particles based on the $K^{g-1} := \sum_{k=1}^K \mathbb{I}_{w_k^{g-1} > 0}$ particles from the previous generation. The method above throws away the original K^{g-1} particles in favor of the new generation. But as both sets of particles are equally close to the observed data (both satisfied $\mathcal{D}(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) < \epsilon_{g-1}$), it is less wasteful to combine them into one larger generation, and then proceed with the calculation of ϵ_g using this larger population.

The ABC procedure suffers from the so-called curse of dimensionality [5]. Specifically, if one takes a distance such as defined in (7) 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 $\boldsymbol{x}$, increases. The dimensionality curse can be alleviated by replacing $\boldsymbol{x} \in \mathbb{R}^t$ with summary statistics $S(\boldsymbol{x}) \in \mathbb{R}^d$, where $d < t$. While the choice of the summary statistics $S : \mathbb{R}^t \mapsto \mathbb{R}^d$ is arbitrary, it is desirable to heavily compress the data ($d \ll t$) while limiting the amount of information lost. This is difficult. When the model at hand admits sufficient statistics then these should be taken. In fact, one can show that convergence of $\pi_\epsilon(\boldsymbol{\theta} \mid \boldsymbol{x})$ to $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ as $\epsilon \to 0$ holds when the chosen summary statistics are sufficient [34, Chapter 5], otherwise convergence holds toward $\pi(\boldsymbol{\theta} \mid S(\boldsymbol{x}))$ which may or may not be a sound approximation to $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$. Note that the summary statistics S are not to be confused with the Ψ summaries in Section 1! Rather than resorting to statistical summaries, we follow up on the work of Bernton et al. [4] and measure the dissimilarity between two samples through the Wasserstein distance defined as

$$\mathcal{W}_p(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \left(\inf_{\sigma \in \mathcal{S}_t} \frac{1}{n} \sum_{s=1}^{t} \rho(x_s, \widetilde{x}_{\sigma(s)})^p \right)^{1/p}, \quad p \geq 1, \tag{8}$$

where $\mathcal{S}_t$ denotes the set of all the permutations of $\{1, \ldots, t\}$. In the remainder, we only consider the case where $p = 1$ and further denote $\mathcal{D}(\cdot, \cdot) \coloneqq \mathcal{W}_1(\cdot, \cdot)$. Bernton et al. [3] have shown in their work that the use of the Wasserstein distance uphold the convergence of the ABC posterior toward the true posterior for continuous data. A recent study by Drovandi and Frazier [9] also shows that the Wasserstein distance compares favorably to other measures of dissimilarity between empirical distributions. The problem is that our data is on the border between discrete and continuous. Another obstacle is the practical evaluation of the Wasserstein distance, which can be tricky when dealing with multivariate or time dependent data. We address these points in the next section for each type of claim data considered in this work.

3 ABC for mixed data

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 need to ensure that the convergence result of the ABC posterior distribution toward the exact posterior distribution holds despite the mixed nature of our data. The main task is then to find an efficient way to compute the Wasserstein distance. We handle the case where the data is IID univariate in Section 3.1, IID bivariate in Section 3.2, and we finish with time dependent data in Section 3.3.

3.1 IID univariate data

For each time period, a random number of claims $n \in \mathbb{N}_0$ are filed. The claim frequencies form an IID sample from the probability mass function (PMF) $p_N(n \mid \boldsymbol{\theta}_{\text{freq}})$. Given n, the associated claim sizes $\boldsymbol{u} = (u_1, \ldots, u_n)$ have a joint probability density function (PDF) denoted by $f_{U|N}(\boldsymbol{u} \mid n, \boldsymbol{\theta}_{\text{sev}})$. The available data $x \coloneqq \Psi(n, \boldsymbol{u})$ is univariate, IID (parametrized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{freq}}, \boldsymbol{\theta}_{\text{sev}})$) and mixed because of 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 PDF

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

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

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= p_X(0 \mid \boldsymbol{\theta})^{t_0}[1 - p_X(0 \mid \boldsymbol{\theta})]^{t-t_0} \prod_{s=1}^{t-t_0} f_{X|X>0}(x_s^+ \mid \boldsymbol{\theta}) \\ &= p_X(0 \mid \boldsymbol{\theta})^{t_0}[1 - p_X(0 \mid \boldsymbol{\theta})]^{t-t_0} p(\boldsymbol{x}^+ \mid \boldsymbol{\theta}). \end{aligned} \tag{9}$$

To evaluate the conditional PDF $f_{X|X>0}$ in (9) 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., Ψ is the sum operator. If the claim sizes are* IID *and independent from the claim frequency, which is common in the actuarial science literature, the conditional* PDF *of X taking positive values is*

$$f_{X|X>0}(x \mid \boldsymbol{\theta}) = \frac{1}{1 - p_N(0 \mid \boldsymbol{\theta}_{\text{freq}})} \sum_{n=1}^{\infty} f_U^{(*n)}(x \mid \boldsymbol{\theta}_{\text{sev}}) p_N(n \mid \boldsymbol{\theta}_{\text{freq}}), \tag{10}$$

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 (10) *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* PDF *required for* MCMC *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* (10) *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. [16].*

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 true 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-zero data points as IID continuous data. So, in Algorithm 2 we retain synthetic samples that belong to the set

$$\mathcal{B}_{\epsilon,\boldsymbol{x}} = \left\{ \widetilde{\boldsymbol{x}} \in \mathbb{R}^t ; \boldsymbol{x}^0 = \widetilde{\boldsymbol{x}}^0 \text{ and } \mathcal{D}(\boldsymbol{x}^+, \widetilde{\boldsymbol{x}}^+) < \epsilon \right\}.$$

Algorithm 2 then samples from the approximate posterior distribution

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

where

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

Proposition 1 in Appendix A shows the convergence of π_ϵ toward the true posterior as we let ϵ approach 0. The Wasserstein distance for real-valued, IID observations reduces to

$$\mathcal{W}_p(\boldsymbol{x}^+, \widetilde{\boldsymbol{x}}^+)^p = \frac{1}{t - t_0} \sum_{s=1}^{t-t_0} |x_{(s)}^+ - \widetilde{x}_{(s)}^+|^p,$$

where $x_{(1)} < \ldots < x_{(t-t_0)}$ and $\widetilde{x}_{(1)}^+ < \ldots < \widetilde{x}_{(t-t_0)}^+$ denote the order statistics of the non-zero portions of the observed and synthetic data respectively. Example 2 shows the efficiency of ABC on an example where we can access the true posterior (i.e. the likelihood function is available).

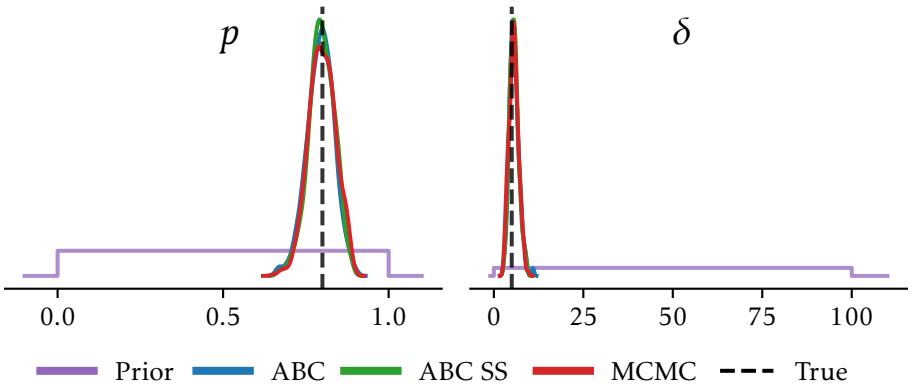

Figure 2: Fitting a Geom(p)–Exp(δ) model to simulated data. The true parameters are $p = 0.8$ and $\delta = 5$. The **ABC posterior**, **ABC summary statistics posterior**, and the **true posterior** (by MCMC) coincide very well (in fact, they overlap), and are considerably narrower than the **prior**.

Example 2. *Let the claim frequency be geometrically distributed*

$$n_1, \dots, n_t \overset{\text{IID}}{\sim} \text{Geom}(p = 0.8),$$

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

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

with PDF *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*

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

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

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

so we can sample from the true 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-zero 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*

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

over the parameters of the Geom(p)–Exp(δ) *parametric model. We set the number of generations to* $G = 10$ *and the number of particles to* $K = 1000$ *for the* ABC *samplers. Figure 2 displays the* KDEs *of the posterior samples produced via* ABC *using the Wasserstein distance,* ABC *with sufficient statistics, and* MCMC. *The* MCMC *posterior sample is generated by the* `PyMC3` *Python library, see Salvatier et al. [31].*

3.2 IID bivariate data

Insurers are typically exposed to more than one type of risk, and it can be beneficial for them to consider the joint risk profile for related products. A joint model for bivariate data like

$$(\{n_s, \boldsymbol{u}_s\}, \{m_s, \boldsymbol{v}_s\}), \quad s = 1, \dots, t$$

could, for example, be used when $\{n_s, \boldsymbol{u}_s\}$ and $\{m_s, \boldsymbol{v}_s\}$ represent the claim data associated to two types of coverage, or two policyholders, or even two nonlife insurance portfolios. The available information is then

$$x_s = \begin{pmatrix} \Psi(n_s, \boldsymbol{u}_s) \\ \Phi(m_s, \boldsymbol{v}_s) \end{pmatrix}, \quad s = 1, \ldots, t.$$

The likelihood function may be written as in the univariate case (9) except that three types of singularity need to be accounted for. Namely, the cases where none of the components, both components and only one component is null. The data is then split four ways as

$$\boldsymbol{x} = (\boldsymbol{x}^{(0,0)}, \boldsymbol{x}^{(+,0)}, \boldsymbol{x}^{(0,+)}, \boldsymbol{x}^{(+,+)}),$$

where $\boldsymbol{x}^{(0,0)} = (x_1^{(0,0)}, \ldots, x_{t_{0,0}}^{(0,0)})$ denotes the portion of the data where both components are null, $\boldsymbol{x}^{(+,0)} = (x_1^{(+,0)}, \ldots, x_{t_{+,0}}^{(+,0)}), \boldsymbol{x}^{(0,+)} = (x_1^{(0,+)}, \ldots, x_{t_{0,+}}^{(0,+)})$ denote the portions of the data where the first or second components are null respectively, and $\boldsymbol{x}^{(+,+)} = (x_1^{(+,+)}, \ldots, x_{t_{+,+}}^{(+,+)})$ corresponds to the portion of the data where both components are nonnegative. The synthetic data simulated within Algorithm 2 is selected only if it belongs to the set

$$\begin{aligned} \mathcal{B}_{\epsilon,\boldsymbol{x}} \quad = \quad & \Big\{ \widetilde{\boldsymbol{x}} \in \mathbb{R}^2 \times \mathbb{R}^t \,;\, t_{0,0} = \widetilde{t}_{0,0},\, t_{0,+} = \widetilde{t}_{0,+},\, t_{+,0} = \widetilde{t}_{+,0} \\ & \mathcal{D}(\boldsymbol{x}^{0,+}, \widetilde{\boldsymbol{x}}^{0,+}) < \epsilon_1,\, \mathcal{D}(\boldsymbol{x}^{+,0}, \widetilde{\boldsymbol{x}}^{+,0}) < \epsilon_2, \text{ and } \mathcal{D}(\boldsymbol{x}^{+,+}, \widetilde{\boldsymbol{x}}^{+,+}) < \epsilon_3 \Big\}. \end{aligned}$$

The tolerance levels ϵ_1, ϵ_2 and ϵ_3 decrease along the sequential Monte Carlo iterations so as to maintain an appropriate effective sample size. The dissimilarity between synthetic and observed data is then measured through the Wasserstein distance. The computation of $\mathcal{D}(\boldsymbol{x}^{+,0}, \widetilde{\boldsymbol{x}}^{+,0})$ and $\mathcal{D}(\boldsymbol{x}^{0,+}, \widetilde{\boldsymbol{x}}^{0,+})$ is similar to that of Section 3.1. Namely, we have

$$\mathcal{D}(\boldsymbol{x}^{+,0}, \widetilde{\boldsymbol{x}}^{+,0}) = \frac{1}{t_{+,0}} \sum_{s=1}^{t_{+,0}} |x_{(s)}^{+,0} - \widetilde{x}_{(s)}^{+,0}|$$

where $x_{(1)}^{(+,0)} < \ldots < x_{(t_{+,0})}^{(+,0)}$ and $\widetilde{x}_{(1)}^{(+,0)} < \ldots < \widetilde{x}_{(t_{+,0})}^{(+,0)}$ and

$$\mathcal{D}(\boldsymbol{x}^{0,+}, \widetilde{\boldsymbol{x}}^{0,+}) = \frac{1}{t_{0,+}} \sum_{s=1}^{t_{0,+}} |x_{(s)}^{0,+} - \widetilde{x}_{(s)}^{0,+}|,$$

where $x_{(1)}^{(0,+)} < \ldots < x_{(t_{0,+})}^{(0,+)}$ and $\widetilde{x}_{(1)}^{(0,+)} < \ldots < \widetilde{x}_{(t_{0,+})}^{(0,+)}$. To compute the Wasserstein $\mathcal{D}(\boldsymbol{x}^{+,+}, \widetilde{\boldsymbol{x}}^{+,+})$ we first set the ground distance ρ to be the Euclidean norm. Finding the optimal permutation in a multivariate setting can be achieved using the Hungarian algorithm at a computational cost of magnitude $\mathcal{O}(t^3)$ (recall that t is the number of observations). Of course, this is significantly higher than the cost required to sort a univariate sample, namely $\mathcal{O}(t \log(t))$. To alleviate the computational burden, we resort to an approximation based on a Hilbert curve. This technique builds a one to one mapping $\phi : \{0, \ldots, 2^k - 1\}^d \mapsto \{0, \ldots, 2^{k \cdot d} - 1\}$ that connects a one-dimensional space to a d-dimensional one, where k is referred to as the Hilbert curve order. Up to rescaling and rounding up our data (we denote by η this data transformation), we can locate it in the space $\{0, \ldots, 2^k - 1\}^d$ by choosing k appropriately and then apply ϕ to it. Consider the transformed data

$$\boldsymbol{y} = (\phi \circ \eta(x_1^{+,+}), \ldots, \phi \circ \eta(x_{t_{+,+}}^{+,+})), \text{ and } \widetilde{\boldsymbol{y}} = (\phi \circ \eta(\widetilde{x}_1^{+,+}), \ldots, \phi \circ \eta(\widetilde{x}_{t_{+,+}}^{+,+})).$$

Denote by $\sigma_{\boldsymbol{y}}$ and $\sigma_{\widetilde{\boldsymbol{y}}}$ the permutations of $\{1, \ldots, t_{+,+}\}$ obtained by sorting $\boldsymbol{y}$ and $\widetilde{\boldsymbol{y}}$ in increasing order. The distance $\mathcal{D}(\boldsymbol{x}^{+,+}, \widetilde{\boldsymbol{x}}^{+,+})$ is then approximated by

$$\mathcal{D}(\boldsymbol{x}^{+,+}, \widetilde{\boldsymbol{x}}^{+,+}) \approx \frac{1}{t_{+,+}} \sum_{s=1}^{t_{+,+}} \rho(x_{\sigma_{\boldsymbol{y}}(s)}^{+,+}, \widetilde{x}_{\sigma_{\widetilde{\boldsymbol{y}}}(s)}^{+,+}). \tag{11}$$

Hilbert curves define a total ordering in a vector space while preserving spatial locality. The approximation (11) performs quite well for two-dimensional data, and the computational cost is the same magnitude as sorting univariate samples.

3.3 Time dependent data

The arrival of claims in insurance is traditionally modelled by a counting process $(N_t)_{t\geq 0}$. The number of claims n_s filed during a given time period s then corresponds to the increments of $(N_t)_{t\geq 0}$. If we take this approach, then our summaries

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

are time dependent instead of IID (unless N_t is a homogeneous Poisson, then the increments are IID and Section 3.1 would apply). To assess the dissimilarity between the observed trajectory $\boldsymbol{x} = (x_1, \dots, x_t)$ and a fake trajectory $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \dots, \widetilde{x}_t)$ we adopt a curve matching strategy as introduced by Bernton et al. [3]. This strategy starts by making the time index part of the data by defining

$$\boldsymbol{y} = \{(x_1, 1), \dots, (x_t, t)\}, \text{ and } \widetilde{\boldsymbol{y}} = \{(\widetilde{x}_1, 1), \dots, (\widetilde{x}_t, t)\}.$$

The ground distance, to be inserted in the Wasserstein distance expression (8), between two data points $y_i = (x_i, i)$ and $\widetilde{y}_j = (\widetilde{x}_j, j)$ is given by

$$\rho_\gamma(y_i, \widetilde{y}_j) = \sqrt{(x_i - \widetilde{x}_j)^2 + \gamma^2 (i-j)^2},$$

where $\gamma \geq 0$ weights the importance of the vertical distance relative to the horizontal distance.

Intuitively, a large value of γ amounts to pairing each point of the observed trajectory with the corresponding time index points in the simulated trajectory. If $\gamma = 0$ then the computation of the Wasserstein distance does not account for the time dependency which brings us back to the case studied in Section 3.1. For an intermediate value of γ, the computation of the Wasserstein distance proceeds in the same way as in the bivariate case studied in Section 3.2.

The effect of any particular γ value will depend on the range of values obtained in the x_s time series. To make γ dimensionless, we first note that the γ variable effectively scales the time axis of the data from s to γs. So we set γ so that the trace plot of the rescaled time series $\{(\gamma s, x_s)\}_{s=1,\dots,t}$ has some desired aspect ratio $H : V$. This is achieved by

$$\gamma = \frac{\max\limits_{s=1,\dots,t} x_s - \min\limits_{s=1,\dots,t} x_s}{t-1} \cdot \frac{H}{V}.$$

as the original $\{(s, x_s)\}_{s=1,\dots,t}$ time series spanned 1 to t and $\min\limits_{s=1,\dots,t} x_s$ to $\max\limits_{s=1,\dots,t} x_s$ on each axis. Each of these γ options are tested in Section 5.4 of the simulation study.

4 Model selection

When it comes to selecting a parametric model for 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. [22, Chapters V & VI]. A decision must be made to find the most suitable models among a set of candidates $\{1, \dots, M\}$. The Bayesian approach to model selection and hypothesis testing uses a categorical random variable m with state space $\{1, \dots, M\}$ and a prior distribution $\pi(m)$. The posterior model evidence is then given by

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

One often compares two models, say 1 and 2, by computing the Bayes factors $B_{12} \coloneqq \pi(2 \mid \boldsymbol{x})/\pi(1 \mid \boldsymbol{x})$. We refer the reader to Kass and Raftery [21] for an overview on Bayesian model selection and Bayes factors. The marginal likelihood of the data according to given model $m \in \{1,\dots,M\}$ is defined by

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

where Θ_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. [17] reduces to adding another step to Algorithm 1 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 3.

Algorithm 3 Acceptance–rejection to compute the model evidence

1: **for** $k = 1 \to K$ **do**
2: **repeat**
3: **generate** $m_k \sim \pi(m)$
4: **generate** $\boldsymbol{\theta}_k \sim \pi(\boldsymbol{\theta} \mid m)$
5: **generate** $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid m_k, \boldsymbol{\theta}_k)$
6: **until** $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon,\boldsymbol{x}}$ then **store** $(m_k, \boldsymbol{\theta}_k)$
7: **end for**

Algorithm 3 is, in essence, the Monte Carlo approach to the computation of models' marginal likelihoods, see for instance McCulloch and Rossi [25]. 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 $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_K \sim \pi(\boldsymbol{\theta} \mid m)$. 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 2 has been adapted in the works of Toni and Stumpf [37] and Prangle et al. [28] 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 $\boldsymbol{x}_k \sim p(\boldsymbol{x} \mid m_k^1, \boldsymbol{\theta}_k^1)$ 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 KDE $K_{\boldsymbol{H}}^m$ with bandwidth $\boldsymbol{H}^{(m)}$ is then fitted to the parameter values associated to each model with

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

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 $\boldsymbol{x}_k \in \mathcal{B}_{\epsilon_{g-1},\boldsymbol{x}}$.[1] Sampling is performed repeatedly until K particles are selected. The acceptance threshold ϵ_g is updated so that the sum of the ESSS for each model is $K/2$. Each particle's weight is 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

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

The algorithm is summarized in Algorithm 4 in the appendix.

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

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

with associated PDF

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

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

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

and a Weibull model $\mathsf{Weib}(r,m)$ with PDF

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

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,

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

[1] It may seem odd that we always sample from $\pi(m)$ instead of $\widehat{\pi}_{\epsilon_{g-1}}(m \mid \boldsymbol{x})$, though we found that by sticking to the model prior we remove the possibility that a good model dies out from the population during the early SMC iterations.

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 functions of the data $\boldsymbol{u} = (u_1, \ldots, u_n)$ may be computed for these loss models and the model probabilities can be estimated using 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 $G = 20$. The model evidences of all three models are reported in Table 1 for sample sizes ranging from 25 to 200.

	PyMC3			ABC		
sample size	Gamma	LogNorm	Weib	Gamma	LogNorm	Weib
25.0	0.44	0.20	0.37	0.46	0.17	0.37
50.0	0.24	0.65	0.11	0.33	0.50	0.17
75.0	0.04	0.95	0.01	0.11	0.83	0.06
100.0	0.01	0.99	0.00	0.04	0.95	0.01
150.0	0.00	1.00	0.00	0.01	0.99	0.00
200.0	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 ABC model selection examples are given in Section 5 and Section 6 for aggregated data.

5 Simulation Study

This section aims at studying the finite sample behavior of our ABC implementation on case studies based on simulated data. In Section 5.1, we assume that the claim sizes are independent from the claim frequency and that the insurer has access to the truncated aggregated sum. In Section 5.2, we consider a model in which the average of the claim sizes depends on the number of claims and the insurer has access to the total claim sizes for each time period. In Section 5.3, we consider a bivariate aggregated claim distribution with dependent claim frequencies. Lastly Section 5.4 considers a time dependent claim arrival process.

Our goal is to check whether our ABC sampling algorithm manages 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 sizes distribution.

In the cases treated in Sections 5.1 through 5.3, the number of generations for ABC is set to $G = 7$ and goes up to $G = 15$ for the time dependent example of Section 5.4. Each generation consists of $K = 1000$, the computing times are reported and discussed in Section 5.5.

5.1 Negative-Binomial Weibull model with truncation

Let the claim frequency be negative binomial distributed

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

with PMF

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

while the claim sizes are Weibull distributed

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

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

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

It corresponds to the data available to a reinsurance company with 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

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

Figure 3 displays the ABC posterior samples when only using the aggregated data (13). The p and k posteriors are quite informative, whereas the scale parameters α and β are slightly skewed in opposite directions so that they compensate for each other.

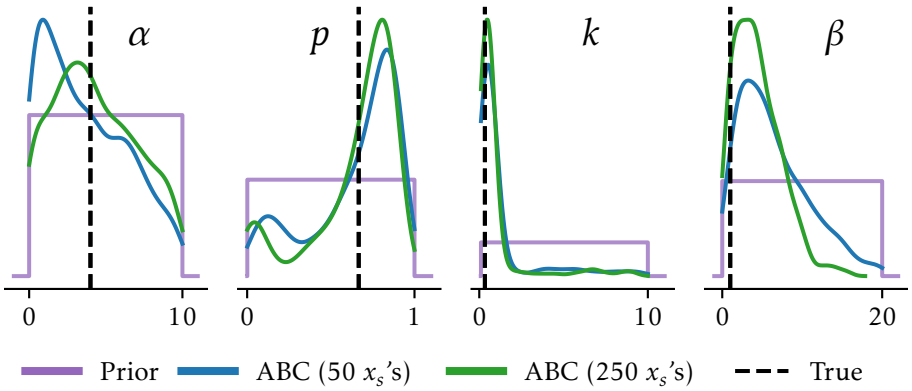

Figure 3: ABC posterior samples of a NegBin(α, p)–Weib(k, β) model fitted to simulated NegBin$(\alpha = 4, p = \frac{2}{3})$–Weib$(k = \frac{1}{3}, \beta = 1)$ data. The posteriors are based on **50 observations** and **250 observations** of the x_s summaries as in (13).

If we observe the claim frequencies n_s as well as the x_s summaries, then we'd expect the ABC algorithm to generate posterior samples even closer to the true values. Figure 4 shows the ABC posteriors for the claim sizes model in this scenario. The ABC posteriors are indeed very strongly concentrated around the true values $k = \frac{1}{3}$ and $\beta = 1$ compared to Figure 3 (the Figure 3 posteriors are drawn with a lower opacity in Figure 4 for ease of reference).

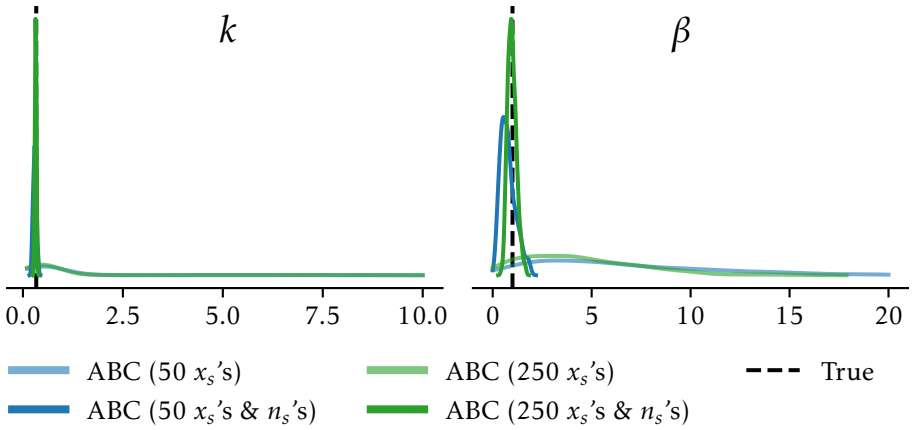

Figure 4: ABC posterior samples of a Weib(k,β) model fitted to data simulated by a NegBin$(\alpha = 4, p = \frac{2}{3})$–Weib$(k = \frac{1}{3}, \beta = 1)$. The data includes each summary x_s as in (13) and each frequency n_s. The posterior with **250 observations** is a slight improvement over the one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

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

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

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

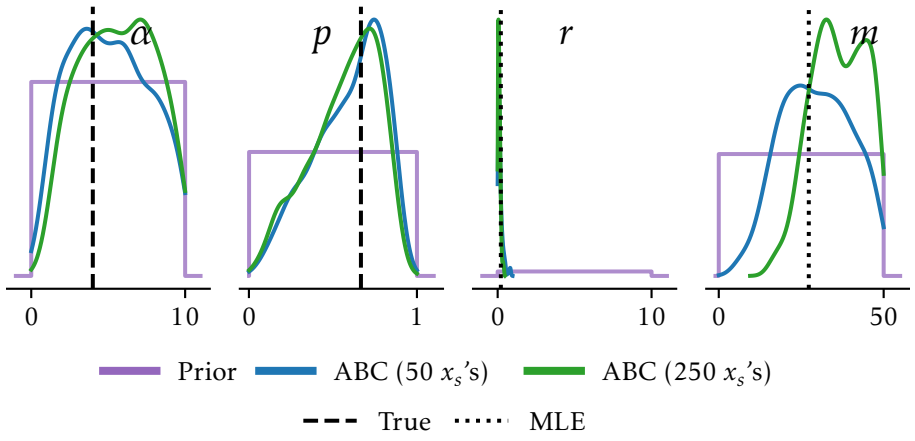

Figure 5: ABC posterior samples of a NegBin(α, p)–Gamma(r, m) model fitted to data simulated by a NegBin$(\alpha = 4, p = \frac{2}{3})$–Weib$(k = \frac{1}{3}, \beta = 1)$ model. The data only includes the summaries x_s as in (13). The target values are the **true values** for α and p and the **MLE estimates** for k and β (based on the individual claim sizes, which are hidden from ABC).

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

Figure 6 displays the ABC posterior samples for the parameters of the gamma distribution when the claim frequency data is available in addition to the summaries (13).

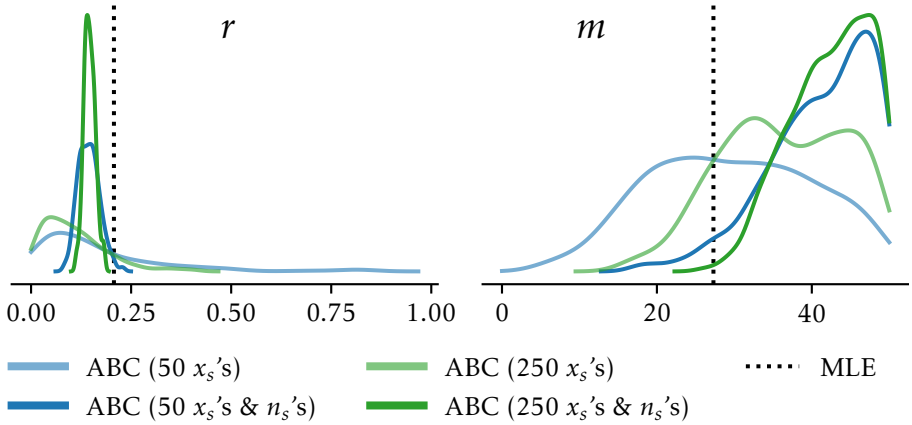

Figure 6: ABC posterior samples of a Gamma(r, m) model fitted to data simulated by a NegBin$(\alpha = 4, p = \frac{2}{3})$–Weib$(k = \frac{1}{3}, \beta = 1)$ model. The data includes each summary x_s as in (13) and each frequency n_s.

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.

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.

Sample Sizes	Frequency Model	
	Negative Binomial	Observed Frequencies
50	0.57	1.00
250	0.59	1.00

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

When only the summaries x_s are available and the claim frequency is modeled by a negative binomial distribution then ABC slightly favours the (true) Weibull over the gamma distributions. When the claim counts n_s are also available then ABC firmly concludes that the Weibull model is the correct model for the claim sizes.

5.2 Dependence between the claim frequency and severity

Let the claim frequency be Poisson distributed

$$n_1, \ldots, n_t \overset{\text{IID}}{\sim} \text{Poisson}(\lambda = 4),$$

with PMF

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

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

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

We denote this $\boldsymbol{u}_s \sim \text{DepExp}(n_s; \beta, \delta)$. The resulting conditional PDF is

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

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 [29]. Empirical studies, like the one conducted in Frees et al. [13], 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. [32]. Equivalently we can scale the expectation of the severity distribution by a factor of $e^{\delta n_s}$. Our case study is inspired by Garrido et al. [14, 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, and selected the prior distributions

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

Figure 7 displays the posterior samples of λ the parameter of the Poisson distribution, β the scale parameter of the exponential parameter and δ the frequency/severity correlation parameter.

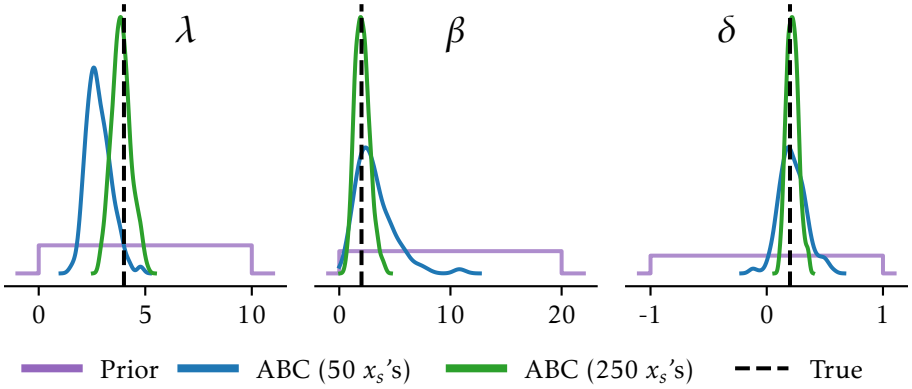

Figure 7: ABC posterior samples of a Poisson(λ)–DepExp($n;\beta,\delta$) model fitted to data simulated by a Poisson($\lambda = 4$)–DepExp($n;\beta = 2,\delta = 0.2$). The data only includes the summaries x_s as in (16).

The algorithm does a tremendous job on this example even without including the claim frequencies at each time period. Figure 8 displays the ABC posterior samples associated to the claim sizes distribution DepExp($n;\beta,\delta$) when including the frequency information in addition to the summaries (16). As observed earlier, the inclusion of the claim frequency information greatly improves the quality of the ABC posterior samples.

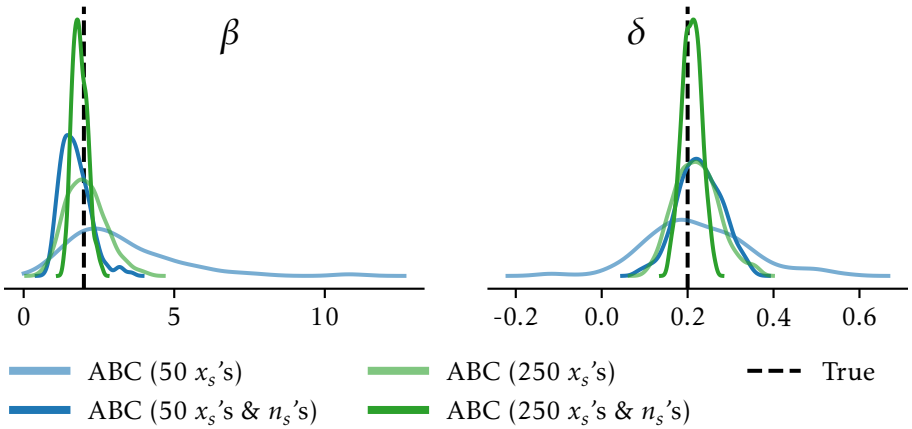

Figure 8: ABC posterior samples of a DepExp$(n;\beta,\delta)$ model fitted to data simulated by a Poisson$(\lambda = 4)$–DepExp$(n;\beta = 2,\delta = 0.2)$. The data includes each summary x_s as in (16) and each frequency n_s.

5.3 Bivariate aggregated claim distribution

This section considers a joint model for the frequency of claims reported for two nonlife insurance portfolios. The claim counts are Poisson distributed with respective intensity Λw_1 and Λw_2 where Λ is some non-negative random variable. The frequency data $(n_1, m_1), \ldots, (n_t, m_t)$ is IID according to a bivariate counting distribution with joint PMF given by

$$p_{N,M}(n,m) = \int \frac{\mathrm{e}^{-\lambda w_1}(\lambda w_1)^n}{n!} \frac{\mathrm{e}^{-\lambda w_2}(\lambda w_2)^m}{m!} \,\mathrm{d}\mathbb{P}_\Lambda(\lambda), \quad n, m \in \mathbb{N}_0.$$

This setting aligns with that of model C in the work of Hesselager [19], and we refer to this model as the BPoisson(Λ, w_1, w_2) bivariate Poisson model. The severities associated to a given time period $s = 1, \ldots, t$ form two mutually independent, IID sequences of exponentially distributed random variables,

$$u_{s,1}, \ldots, u_{s,n_s} \overset{\text{IID}}{\sim} \text{Exp}(m_1 = 10) \quad \text{and} \quad v_{s,1}, \ldots, v_{s,m_s} \overset{\text{IID}}{\sim} \text{Exp}(m_2 = 40).$$

The model encapsulate the link between the frequencies of two insurance portfolios while accommodating for the well known overdispersed nature of the claim count data. Following up on the work of Streftaris and Worton [36], we let Λ be a lognormal random variable LogNorm$(\sigma = 0.2)$ (the mean log parameter is set to 0) as it is consistent with the use of a generalized linear model equipped with a log link function to estimate the Poisson intensity given a set of covariates. The marginal components of the claim frequency distribution are set to $w_1 = 15$ and $w_2 = 5$. The available data is the aggregated claim sizes for each portfolio

$$x_s = \left(\sum_{k=1}^{n_s} u_{s,k}, \sum_{k=1}^{m_s} v_{s,k} \right), \quad s = 1, \ldots, t. \tag{17}$$

We consider data histories of length $t = 50$ and 250. Uniform prior distributions are set over the claim frequency parameters

$$\sigma \sim \text{Unif}(0,2), \quad w_1 \sim \text{Unif}(0,50), \quad w_2 \sim \text{Unif}(0,50),$$

and the claim sizes parameters

$$m_1 \sim \text{Unif}(0,100), \quad m_2 \sim \text{Unif}(0,100).$$

The discrepancy between observed and fake data follows from the approximation of the Wasserstein distance via the projection onto the Hilbert filling curve space detailed in Section 3.2. Figure 9 shows the resulting posterior distribution based on data histories of length 50 and 250.

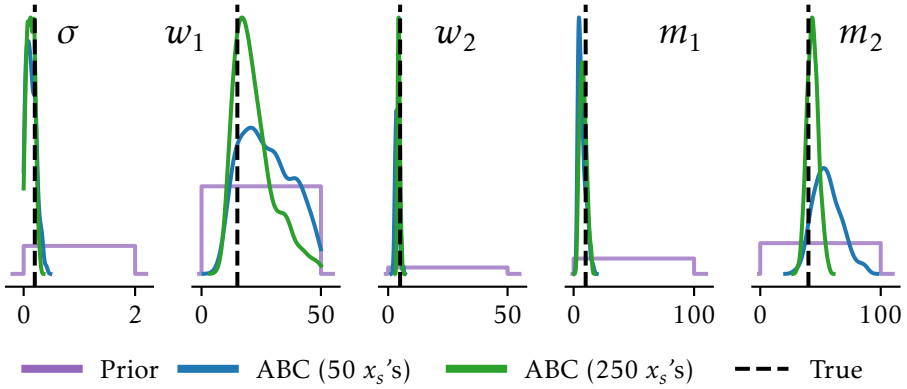

Figure 9: ABC posterior samples of a BPoisson(Λ, w_1, w_2)–Exp(m_1, m_2) fitted to simulated data from a model BPoisson$(\Lambda \sim \text{LogNorm}(\sigma = 0.2), w_1 = 15, w_2 = 5)$–Exp$(m_1 = 10, m_2 = 40)$. The data includes **50 observations** and **250 observations** of the summaries x_s as in (17).

ABC manages to identify the parameters linked to the marginal distributions and the dependence structure.

5.4 Compound sums with nonhomogenous Poisson claim arrival

We can generalize the discrete time model to continuous time by modelling the arrival of claims with a counting process $(N_t)_{t\geq 0}$. The liability of the insurer, taking into account the randomly sized compensation associated to each claim, takes the form of a pure jump process

$$Z_t = \sum_{i=1}^{N_t} U_i, \quad t \geq 0,$$

as in (4) above. Our goal in this section is to see whether our ABC routine enables us to estimate the model parameters from the knowledge of a trajectory of such a stochastic process. We move away from the standard Poisson assumption by assuming that the claim arrival is governed by a nonhomogenous Poisson process with instantaneous arrival rate $\lambda(t)$. We observe the increments of the process $(Z_t)_{t\geq 0}$ defined by

$$X_s := Z_s - Z_{s-1} = \sum_{i=1}^{N_s - N_{s-1}} U_i, \quad s = 1, \ldots, t,$$

where the increments of the counting processes are independent and Poisson distributed with parameter $\mu(s) = \int_s^{s+1} \lambda(s)\,\mathrm{d}s$. We consider a cyclical claim arrival rate by setting

$$\lambda(t) = a + b[1 + \sin(2\pi c t)], \quad t \geq 0.$$

We refer to this model as the cyclical Poisson model CPoisson(a, b, c). We wish to see whether our ABC algorithm allows us to draw inference on the parameters a, b, and c on the intensity function. For this example, we set $a = 1$, $b = 5$, and $c = 1/50$. The claim amounts follow a lognormal distribution LogNorm$(\mu = 0, \sigma = 0.5)$ and we

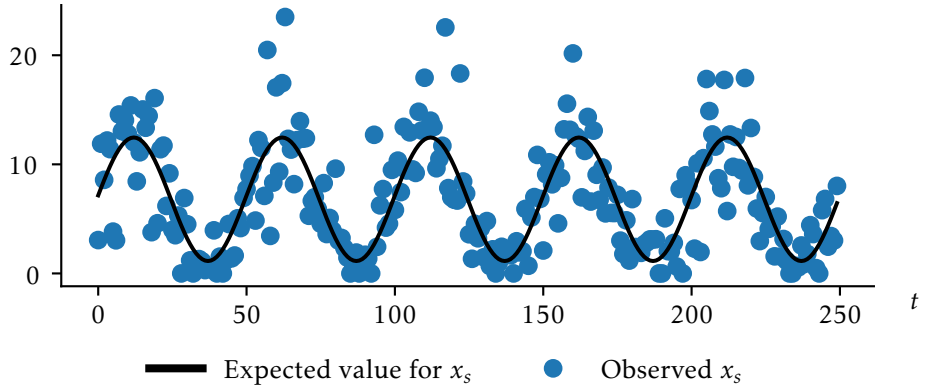

Figure 10: Observations and theoretical mean of the increments of a cyclical compound Poisson process CPoisson($a = 1, b = 5, c = 1/50$)–LogNorm($\mu = 0, \sigma = 0.5$).

consider two time horizons $t \in \{50, 250\}$. Figure 10 displays the observed increments of the nonhomogenous compound Poisson process together with their the theoretical means as a function of time over 250 time periods.

Uniform prior distributions are set over the claim frequency parameters

$$a \sim \text{Unif}(0, 50),\ b \sim \text{Unif}(0, 50),\ \text{and}\ c \sim \text{Unif}(1/1000, 1/10),$$

and the claim sizes parameter

$$\mu \sim \text{Unif}(-10, 10)\ \text{and}\ \sigma \sim \text{Unif}(0, 3).$$

Our aim is to compare the ABC posterior samples resulting from different choices of γ which parametrize the ground distance

$$\rho_\gamma(y_i, \widetilde{y}_j) = \sqrt{(x_i - \widetilde{x}_j)^2 + \gamma^2 (i-j)^2},$$

which is used in the Wasserstein distance (8). We consider the extremal cases where $\gamma = 0$ and $\gamma = \infty$. Recall that $\gamma = \infty$ forces the pairs of data points to have the same time index, whereas $\gamma = 0$ ignores the time dependency altogether. Stated another way, the $\gamma = \infty$ case calculates the L^1 distance between the observed and fake data, and $\gamma = 0$ calculates the L^1 distance between the sorted versions of each data vector. We also look for a tradeoff by setting γ to be

$$\gamma^* = 2 \cdot \frac{\max\limits_{s=1,\ldots,t} x_s - \min\limits_{s=1,\ldots,t} x_s}{t-1}.$$

This transforms the data so the x_s's trace plot has an aspect ratio of $2:1$ and leads to an acceptable compromise between $\gamma = 0$ and $\gamma = \infty$. Figure 11 shows the ABC posterior distributions based on these three different values for γ.

There is a clear distinction between the $\boldsymbol{\theta}$ parameters which affect the quantiles of the x_s summaries (i.e., a, b, μ, and σ) and the parameters which do not (i.e., c). The L^1 distance on the sorted data ($\gamma = 0$) is effectively fitting $\boldsymbol{\theta}$ based on the quantiles of the x_s distribution, so it is incapable of fitting the c parameter. The L^1 distance on the unsorted data ($\gamma = \infty$) fits the c parameter best, though without the sorting it struggles to fit the remaining parameters particularly well. The $\gamma = \gamma^*$ tradeoff choice does fit all parameters moderately well. Comparing time series in ABC is an ongoing and difficult research problem. A promising direction could be to search for an optimal value of γ.

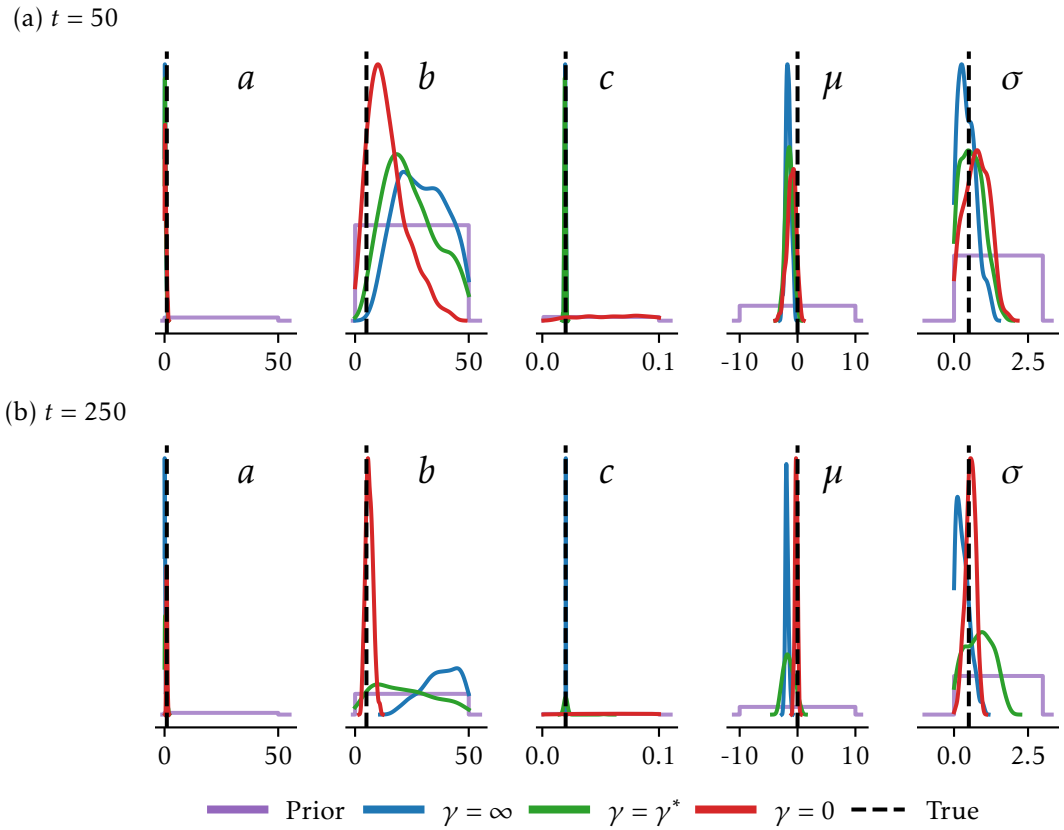

Figure 11: ABC posterior samples of a CPoisson(a, b, c)–LogNorm(μ, σ) fitted to simulated data from a model CPoisson$(a = 1, b = 5, c = 1/50)$–LogNorm$(\mu = 0, \sigma = 0.5)$. The ABC posterior obtained using the **L^1 distance** ($\gamma = \infty$), **the curve matching distance** ($\gamma = \gamma^*$) and the **L^1 distance between sorted data** ($\gamma = 0$) are compared. Sample sizes of length (a) 50 and (b) 250 are considered.

5.5 Computational runtimes and practical considerations

ABC transforms a difficult statistical problem into a difficult computational problem. Luckily, ABC is 'embarrassingly parallelizable', and our Python implementation uses multiprocessing to leverage central processing units (CPUs) which have a large number of physical cores.[2] Using a CPU with large core count allows us to run ABC in parallel and achieve a near-linear speedup when compared to a single-process implementation.

We ran the previous experiments on a virtual machine rented from Amazon Web Services. In particular, we used a 'c6g.16xlarge' instance which has 64 physical cores in its ARM CPU. Amazon currently charges $2.8416 USD/hour for this instance type in their Sydney data center. The runtimes and corresponding costs for each simulation are presented in Table 3 and are based on these rates. We also measured the same runtimes when running on a Mac Mini (Late 2020 model) which has 8 physical cores (4 high performance cores, and 4 high efficiency cores) with the results given in Table 4. These devices are currently the cheapest Macintosh computers on sale. If the CPUs in the Amazon instance and the Mac Mini were equivalent then we'd expect the Mac to be $\approx 8\times$ slower at ABC, but the Apple's M1 chip is only $\approx 3\times$ slower.

[2]Most x86 CPUs also include 'hyperthreaded' virtual cores, which can speed up certain data-intensive workloads. However ABC does not benefit from these virtual cores, and in fact using them is detrimental to overall speed due to the increased context switching costs and cache invalidation.

Sample size	Figure Number								Total
	$G = 7$						$G = 15$		
	3	4	5	6	7	8	9	11	
50	29 s (2.3 ¢)	22 s (1.7 ¢)	22 s (1.7 ¢)	42 s (3.3 ¢)	23 s (1.8 ¢)	19 s (1.5 ¢)	98 s (7.7 ¢)	173 s (13.7 ¢)	24 m
250	55 s (4.3 ¢)	35 s (2.8 ¢)	18 s (1.4 ¢)	94 s (7.4 ¢)	22 s (1.7 ¢)	19 s (1.5 ¢)	171 s (13.5 ¢)	607 s (47.9 ¢)	($1.14)

Table 3: Runtimes (in seconds or minutes) and the associated server rental costs (in USD or cents) for the ABC fits showcased in the figures in this section on a 'c6g.16xlarge' (64 ARM Neoverse cores) instance. Each entry corresponds to one ABC fit, except for the Figure 11 times which are the total of three ABC fits.

Sample size	Figure Number								Total
	$G = 7$						$G = 15$		
	3	4	5	6	7	8	9	11	
50	49 s	27 s	19 s	96 s	27 s	5 s	318 s	424 s	74 m
250	192 s	81 s	56 s	311 s	59 s	10 s	656 s	2149 s	

Table 4: Runtimes (in seconds or minutes) for the ABC fits showcased in the figures in this section on a Late 2020 Mac Mini (8 ARM Apple Silicon cores). Each entry corresponds to one ABC fit, except for the Figure 11 times which are the total of three ABC fits.

Table 3 clearly shows that ABC is very computationally demanding. Even when fully utilizing the 64 cores of the CPU it takes some minutes to complete these fits. This is somewhat comparable to fitting a moderate-sized artificial neural network model. One should definitely not use ABC in scenarios when a likelihood is available! On the other hand, the overall rental cost for these fits ($1.14) is quite small. As ARM processors have a high performance-per-watt, Amazon can rent us these ARM machines for about half the price of the equivalent x86 machines. Porting ABC to a GPU would further reduce costs.

Another conclusion from Table 3 is that the runtime of ABC does not have a linear relationship to the sample size of observed data. In some cases, ABC takes longer to fit the 50 observations than it does to fit 250 observations. This can happen when ABC-SMC quickly finds a 'good' fit for the 50 observations so it aggressively decreases the ϵ_g targets and then it spends a long time trying to find a 'great' fit in the final iterations. In general we observe an exponential increase in the runtime of each ABC-SMC iteration. This is why we set the number of iterations G by trial-and-error, as a small increase in G can increase the ABC-SMC runtime from minutes to days. The Python code written for this paper may be downloaded from GitHub https://github.com/LaGauffre/ABCFitLoMo.

6 Application to a real-world insurance dataset

We consider an open source insurance dataset named `ausautoBI8999` consisting of 22,036 settled personal injury insurance claims in Australia, the first five observations are displayed in Table 5.

The data is accessible from the `R` package `CASDatasets`, see Dutang and Charpentier [11]. 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 6.

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

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 5: An extract of the ausautoBI8999 personal injury claim data.

Month	Claim Frequency	Total Claim Severity
49	149	1,550,000
50	188	3,210,000
51	196	4,810,000
52	203	4,220,000
53	226	5,270,000

Table 6: An extract of the monthly aggregated data.

Table 7.

Statistics	Claim Severity	Claim Frequency	Total Claim Severity
Count	22,000	69	69
Mean	38,400	319	12,300,000
Std	91,000	109	5,220,000
Min	9.96	94	1,550,000
25%	6,300	231	8,210,000
50%	13,900	312	12,000,000
75%	35,100	381	15,500,000
Max	4,490,000	606	26,300,000

Table 7: Descriptive statistics of the claim data.

We are going to use ABC to fit and compare loss models using only the monthly aggregated data in Table 6. We would like to know whether the results differ from fitting the same loss models but using the individual claim sizes data in Table 5.

We start by studying the individual loss distribution. We fit a gamma, a lognormal and a Weibull model to the data shown in Table 5 using maximum likelihood estimation. The estimates of the parameters are given in Table 8 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 12.

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 13 to 15 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 13 and 14, while the parameters of the lognormal distribution are more stable, see Figure 15. The model evidences, displayed in Figure 16, are computed using the Schwarz criterion that approximates the Bayes factor using the maximum

Severity model	Parameters	MLE	BIC
Gamma	r m	4.09 5,350	64,600
Weibull	k β	0.708 28,600	50,300
Lognormal	σ μ	9.56 1.46	50,000

Table 8: Maximum likelihood estimates of a gamma, Weibull and lognormal distribution based on the individual claim sizes data.

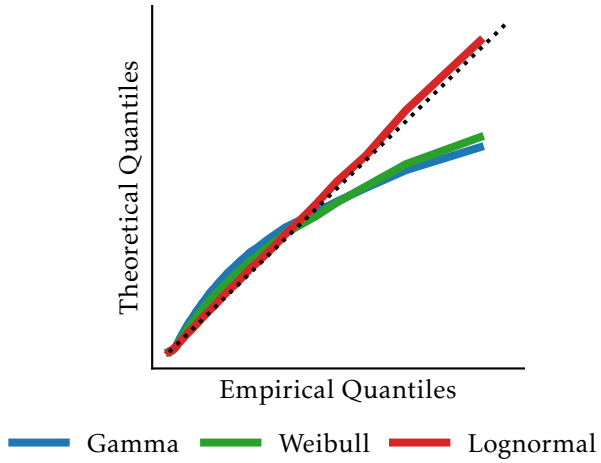

Figure 12: Quantile-quantile plots associated to the gamma, Weibull and lognormal models fitted to the individual claim sizes data.

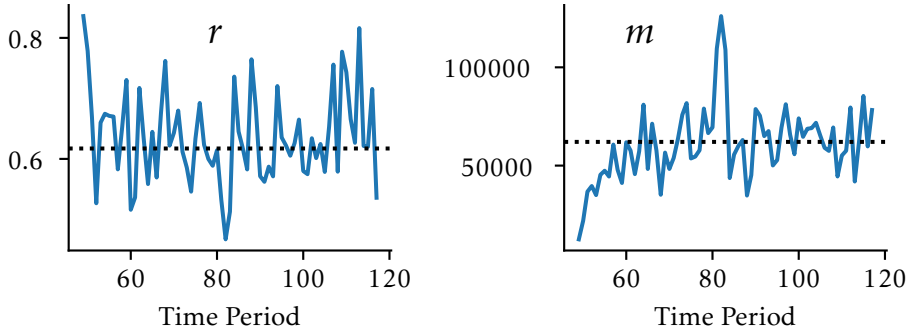

Figure 13: Parameters of the gamma model.

likelihood estimators and the BIC.

The model probabilities mostly favor the lognormal model.

We use ABC to fit a NegBin(α, p)–LogNorm(μ, σ) model to the total claim severities data in Table 6 which consists

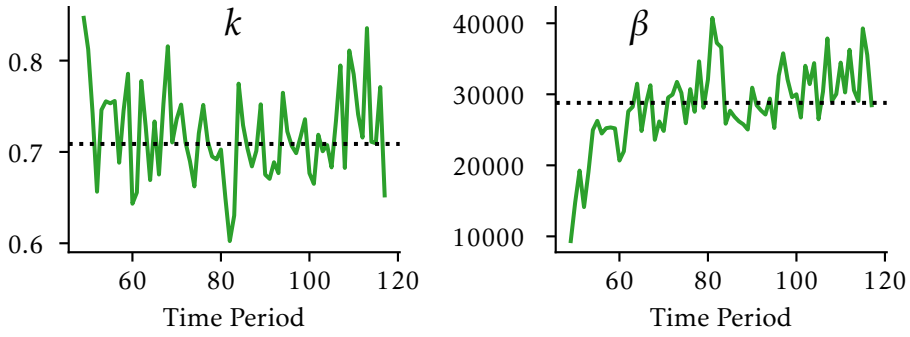

Figure 14: Parameters of the Weibull model.

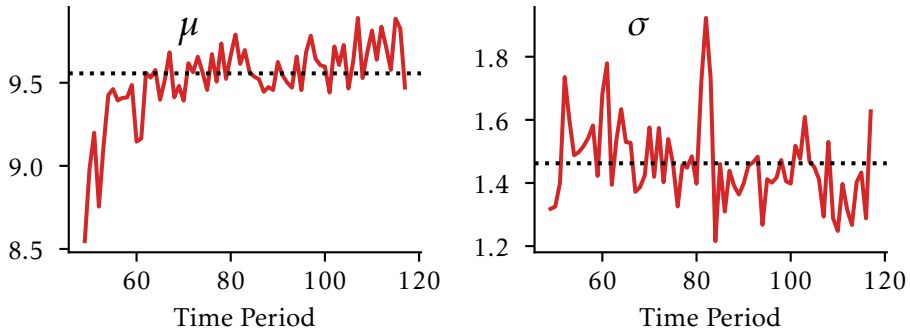

Figure 15: Parameters of the lognormal model.

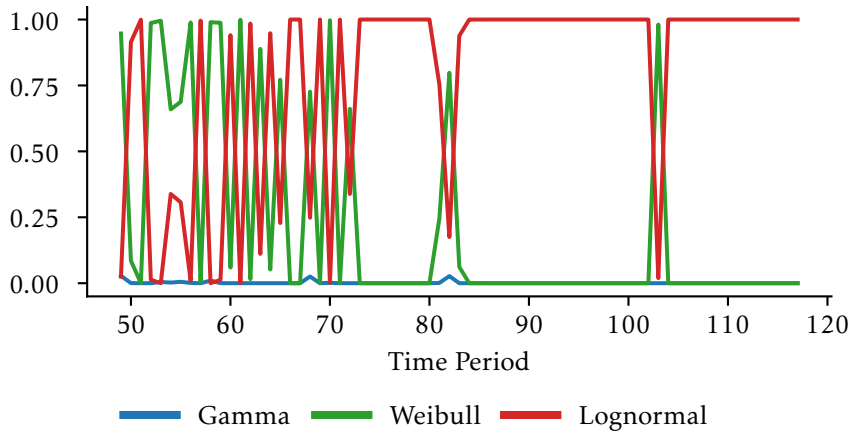

Figure 16: Model evidence for the gamma, lognormal and Weibull models.

of $t = 69$ summaries of the form

$$x_s = \sum_{k=1}^{n_s} u_{s,k}, \quad s = 1, \ldots, t. \tag{18}$$

We consider two sets of prior assumptions over the parameters:

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

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

Prior settings 1 and 2 only differ in the boundaries of the uniform distribution of μ. We opt for a more intensive ABC calibration compared to that of Section 5. 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 NegBin(α, p)–LogNorm(μ, σ) model using only the summaries x_s in (18) are shown in Figure 17.

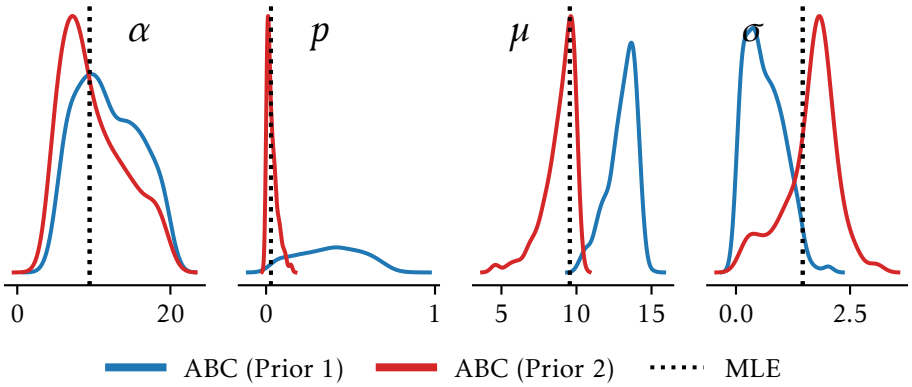

Figure 17: ABC posterior samples of a NegBin(α, p)–LogNorm(μ, σ) model fitted to a real world insurance dataset. The data includes the total claim severities (18) data in Table 6. The posterior samples are closer to the **MLE estimates** with **prior 2** than with **prior 1**.

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, μ and σ.

The ABC posterior samples when including the claim frequency information are shown in Figure 18. We keep the same prior assumptions over μ and σ. These posteriors which use the claim frequency data are less affected by the differing prior settings.

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

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

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

- Weib(k, β) with prior distributions

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

- Gamma(r, m) with prior distributions

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

- LogNorm(μ, σ) with prior distributions

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

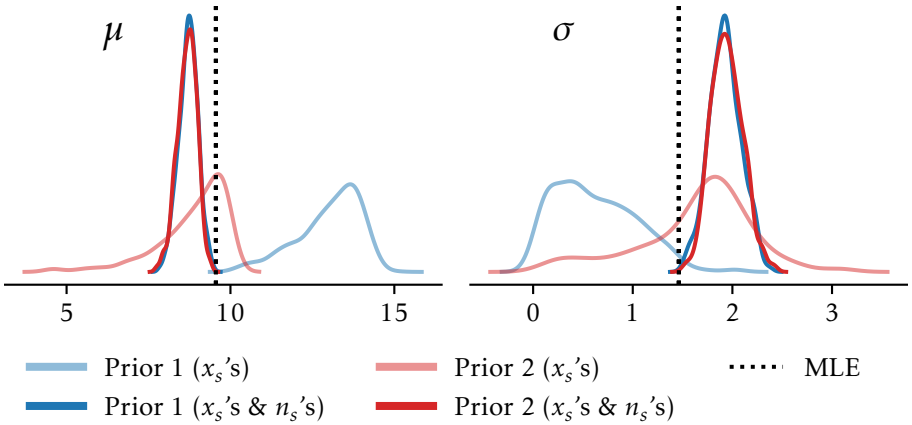

Figure 18: ABC posterior samples of a LogNorm(μ, σ) model fitted to a real world insurance dataset. The data includes the total claim severities and the claim frequencies in Table 6. When the x_s's and n_s's are both observed, the posterior samples with **Prior 1** and **Prior 2** almost totally overlap and are reasonably close to the **MLE estimates**.

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

Table 9: ABC model evidence with the claim frequency and the aggregated claim sizes data.

The bounds of the uniform distributions are set to reflect the variability of the parameters in Figures 13 to 15. The model evidences are reported in Table 9.

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.

7 Conclusion

This paper is a case study of an ABC applications in insurance. We showed how to use this method to calibrate insurance loss models with limited information (one data point per time period). As ABC is not restricted to models which have a known likelihood, we can explore more realist models and discard the classical assumptions of independence in and between the claim frequencies and 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. Our Python code which reproduces the results in this work, as well as a Python package to apply ABC-SMC more generally, are available on Github.

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

The authors are thankful for the relevant comments of the two anonymous referees that help in greatly improve our original manuscript. Patrick J. Laub conducted part of this research while in 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] Søren Asmussen and Hansjörg Albrecher. *Ruin Probabilities*, volume 14 of *Advanced Series on Statistical Science and Applied Probability*. World Scientific, 2nd edition, 2010.

[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. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 10 2019. ISSN 2049-8772. doi: 10.1093/imaiai/iaz003. URL https://doi.org/10.1093/imaiai/iaz003.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] Christopher Drovandi and David T Frazier. A comparison of likelihood-free methods with and without summary statistics, 2021.

[10] Peter K. Dunn. Occurrence and quantity of precipitation can be modelled simultaneously. *International Journal of Climatology*, 24(10):1231–1239, jul 2004. doi: 10.1002/joc.1063.

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

[12] Edward W. Frees. Frequency and severity models. In Edward W. Frees, Richard A. Derrig, and Glenn Meyers, editors, *Predictive Modeling Applications In Actuarial Science*, pages 138–164. Cambridge University Press. doi: 10.1017/cbo9781139342674.006.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] Ole Hesselager. Recursions for certain bivariate counting distributions and their compound distributions. *ASTIN Bulletin*, 26(1):35–52, 1996. doi: 10.2143/AST.26.1.563232.

[20] Bent Jørgensen and Marta C. Paes De Souza. Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93, jan 1994. doi: 10.1080/03461238.1994.10413930.

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

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

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

[24] Vyacheslav Lyubchich and Y. R. Gel. Can we weather proof our insurance? *Environmetrics*, 28(2):e2433, dec 2016. doi: 10.1002/env.2433.

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

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

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

[33] Peng Shi, Xiaoping Feng, and Jean-Philippe Boucher. Multilevel modeling of insurance claims using copulas. *The Annals of Applied Statistics*, 10(2):834–863, jun 2016. doi: 10.1214/16-aoas914.

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

[35] Gordon K. Smyth and Bent Jørgensen. Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, 32(1):143–157, 2002. doi: 10.2143/AST.32.1.1020.

[36] George Streftaris and Bruce J. Worton. Efficient and accurate approximate Bayesian inference with an application to insurance data. *Computational Statistics & Data Analysis*, 52(5):2604–2622, jan 2008. doi: 10.1016/j.csda.2007.09.006.

[37] 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.

[38] Maurice CK Tweedie. An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, volume 579, pages 579–604, 1984.

[39] 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.

[40] Mario V. Wüthrich. Claims reserving using Tweedie's compound Poisson model. *ASTIN Bulletin*, 33(2):331–346, nov 2003. doi: 10.1017/s0515036100013490.

[41] Oscar Alberto Quijano Xacur and José Garrido. Generalised linear models for aggregate claims: to Tweedie or not? *European Actuarial Journal*, 5(1):181–202, jun 2015. doi: 10.1007/s13385-015-0108-5.

[42] Yanwei Zhang. Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Statistics and Computing*, 23(6):743–757, aug 2012. doi: 10.1007/s11222-012-9343-7.

A Convergence of the ABC posterior to the true posterior with mixed data

The following result shows the convergence of π_ϵ toward the true posterior as we let ϵ approach 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

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

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

$$\pi_\epsilon(\boldsymbol{\theta}\mid\boldsymbol{x}) = \frac{\pi(\boldsymbol{\theta})\int_{\mathbb{R}^t}\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\,p(\widetilde{\boldsymbol{x}}\mid\boldsymbol{\theta})\,\mathrm{d}\widetilde{\boldsymbol{x}}}{\int_{\boldsymbol{\Theta}}\pi(\boldsymbol{\theta})\int_{\mathbb{R}^t}\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\,p(\widetilde{\boldsymbol{x}}\mid\boldsymbol{\theta})\,\mathrm{d}\widetilde{\boldsymbol{x}}\,\mathrm{d}\boldsymbol{\theta}} = \frac{\pi(\boldsymbol{\theta})p_\epsilon(\boldsymbol{x}\mid\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}}\pi(\boldsymbol{\theta})p_\epsilon(\boldsymbol{x}\mid\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}, \tag{19}$$

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

$$p_\epsilon(\boldsymbol{x}\mid\boldsymbol{\theta}) = \frac{\int_{\mathbb{R}^t}\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\,p(\widetilde{\boldsymbol{x}}\mid\boldsymbol{\theta})\,\mathrm{d}\widetilde{\boldsymbol{x}}}{\int_{\mathbb{R}^t}\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})\,\mathrm{d}\widetilde{\boldsymbol{x}}}. \tag{20}$$

Since the data is IID, 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 (20) as the product

$$\mathbb{I}_{\mathcal{B}_{\epsilon,\boldsymbol{x}}}(\widetilde{\boldsymbol{x}}) = \mathbb{I}_{\{\boldsymbol{x}^0=\widetilde{\boldsymbol{x}}^0\}}\cdot\mathbb{I}_{\{\mathcal{D}(\boldsymbol{x}^+,\widetilde{\boldsymbol{x}}^+)\le\epsilon\}}. \tag{21}$$

Inserting (21) into the quasi-likelihood (20) leads to

$$\begin{aligned} p_\epsilon(\boldsymbol{x}\mid\boldsymbol{\theta}) &= p_X(0\mid\boldsymbol{\theta})^{t_0}[1-p_X(0\mid\boldsymbol{\theta})]^{t-t_0}\,\frac{\int_{\mathbb{R}^{t-t_0}}\mathbb{I}_{\{\mathcal{D}(\boldsymbol{x}^+,\widetilde{\boldsymbol{x}}^+\le\epsilon\}}p(\widetilde{\boldsymbol{x}}^+\mid\boldsymbol{\theta})\,\mathrm{d}\widetilde{\boldsymbol{x}}}{\int_{\mathbb{R}^{t-t_0}}\mathbb{I}_{\{\mathcal{D}(\boldsymbol{x}^+,\widetilde{\boldsymbol{x}}^+\le\epsilon\}}\,\mathrm{d}\widetilde{\boldsymbol{x}}} \\ &\underset{\epsilon\to 0}{\longrightarrow} p_X(0\mid\boldsymbol{\theta})^{t_0}[1-p_X(0\mid\boldsymbol{\theta})]^{t-t_0}p(\boldsymbol{x}^+\mid\boldsymbol{\theta}) = p(\boldsymbol{x}\mid\boldsymbol{\theta}), \end{aligned} \tag{22}$$

where the limit in (22) follows from applying Proposition 1 of Rubio and Johansen [30], see also Bernton et al. [4, Proposition 2]. Taking the limit as ϵ tends to 0 in (19) yields the announced result. □

B Model selection algorithm

Algorithm 4 ABC-SMC for model selection

for $k = 1 \rightarrow K$ **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 $m = 1, \ldots, M$ **do**
  compute $\widehat{\pi}_{\epsilon_1}(m \mid \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}_{\{m_k^1 = m\}}$
  compute $\widehat{\pi}_{\epsilon_1}(\boldsymbol{\theta} \mid m, \boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\widehat{\pi}_{\epsilon_1}(m|\boldsymbol{x})} K_{\boldsymbol{H}}^m(\boldsymbol{\theta} - \boldsymbol{\theta}_k^1) \mathbb{I}_{\{m_k^1 = m\}}$
end for
for $g = 2 \rightarrow I$ **do**
  for $k = 1 \rightarrow K$ **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 ϵ_g so the sum of the ESSS is $K/2$
  for $k = 1 \rightarrow K$ **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 $m = 1, \ldots, M$ **do**
    compute $\widehat{\pi}_{\epsilon_g}(m \mid \boldsymbol{x}) = \sum_{k=1}^{K} w_k^g \mathbb{I}_{\{m_k^g = m\}}$
    compute $\widehat{\pi}_{\epsilon_g}(\boldsymbol{\theta} \mid m, \boldsymbol{x}) = \sum_{k=1}^{K} \frac{w_k^g}{\widehat{\pi}_{\epsilon_g}(m|\boldsymbol{x})} K_{\boldsymbol{H}}^m(\boldsymbol{\theta} - \boldsymbol{\theta}_k^g) \mathbb{I}_{\{m_k^g = m\}}$
  end for
end for