Sequential Monte Carlo samplers to fit and compare insurance loss models

Pierre-Olivier Goffard

Sequential Monte Carlo samplers to fit and compare insurance loss models

Pierre-O. Goffard[1]

[1] Univ Lyon, Université Lyon 1, LSAF EA2429

June 17, 2021

Abstract

Insurance loss distributions are characterized by a high frequency of small amounts and a lower, but not insignificant, occurrence of large claim amounts. Composite models, which link two probability distributions, one for the "belly" and the other for the "tail" of the loss distribution, have emerged in the actuarial literature to take this specificity into account. The parameters of these models summarize the distribution of the losses. One of them corresponds to the breaking point between small and large claim amounts. The composite models are usually fitted using maximum likelihood estimation. A Bayesian approach is considered in this work. Sequential Monte Carlo samplers are used to sample from the posterior distribution and compute the posterior model evidences to both fit and compare the competing models. The method is validated via a simulation study and illustrated on insurance loss datasets.

MSC 2010: 60G55, 60G40, 12E10.

Keywords: Composite model, Bayesian statistics, Sequential Monte Carlo sampler.

1 Introduction

The distribution of losses in property and casualty insurance is characterized by a high frequency of small claim amounts and a lower, but not insignificant, frequency of considerably larger claim amounts. Composite models, that combine two models one for the body and the other for the tail of the loss distribution have emerged in the actuarial science litterature as a response to this specificity. The occurrence of extreme values

makes the assessment of the average claim cost unreliable and leads to an increase in insurance premia. In practice, actuaries tackle this problem by determining a threshold that tells apart small from large claim amounts. The average cost of the small claim amounts is then estimated using statistical learning techniques such as generalized linear models. The risk associated to the losses above the threshold is partly transferred to the reinsurer while the remainder is reflected in the insurance premium as a safety loading.

The occurrence of extreme claims concerns all branches of non-life insurance and must be studied in order to correctly allocate the solvency capital necessary to compensate for the mismatch between the coolected premia and the extreme claim amounts. The determination of the breaking point between small and large claim sizes is a key factor when analysing risks. Threshold selection methods borrow tools from extreme value theory, see for instance the textbook of Beirlant et al. [3]. It includes famous graphical visualizations such as the mean-excess plot, the Hill plot [22] or the Gerstengarbe plot [19]. In the present work, a different approach is considered. All the parameters, including the threshold parameter, of the composite model are estimated simultaneously. This technique was originally developed by Cooray and Ananda [11]. Maximum likelihood estimation is used to fit several combinations of models for the belly and the tail of the claim amounts distribution. The adequacy of each model is then measured using standard information criteria. Extensive studies have been carried out by Abu Bakar et al. [1] and Grün and Miljkovic [21]. For a recent survey on the use of composite models and threshold selection methods on insurance data I refer the reader to the work of Wang et al. [40].

The present work proposes to fit and compare composite models in a Bayesian way. Bayesian statistics take the model parameters to be random variables. Inference is drawn from the posterior distribution of the parameters obtained by updating the *a priori* assumptions via the likelihood function of the data, for an overview see the book by Geldman et al. [17]. Bayesian inference accounts for the uncertainty around the estimated parameters and compensate the lack of data by the possibility of encapsulating expert field knowledge through the prior distribution. The posterior distribution is often unavailable and must be approximated by an empirical distribution. Markov chain Monte Carlo (MCMC) simulation schemes, such as the well known Metropolis-Hasting and Gibbs samplers, have become the go to techniques to sample from the posterior distribution. Probabilistic programming softwares like WINBUGS [28], JAGS [32], STAN [8] and PYMC [34], have been designed over the years so that practitioners do not have to worry about the fine tuning of these sophisticated algorithms. Bayesian inference of composite models have been considered before, the specific case of the lognormal-Pareto model did attract a lot of attention. Pigeon and Denuit [31] started by randomizing the threshold parameter while Cooray and Chang [12] derived later the posterior distribution for both conjugate and Jeffrey's priors. The

lognormal-Pareto composite model of Scollnick [36] is actually an example of the WINBUGS documentation, see [28, Examples Vol. III].

Instead of the standard MCMC sampler, a Sequential Monte Carlo Sampler (SMC) is put together. This algorithm builds a sequence of empirical distributions, made of weighted particles, that targets the posterior distribution during the final iteration. Generic SMC samplers are described in the seminal paper of Del Moral et al. [29]. An approximation of the posterior distribution normalizing constant, referred to as the marginal likelihood, follows from the weights of the successive particle clouds. MCMC algorithms bypass the evaluation of this constant which is nevertheless necessary for the evaluation of Bayes factors to select the right model, see Kass and Raftery [26]. In addition to providing an approximation of the marginal likelihood, SMC samplers can sample from complicated multimodal posterior distributions, save the trouble of tuning some hyperparameters and are easy to paralellize which is a key feature in the era of multi-core processor computers.

The remainder of the paper is organized as follows. Section 2 provides a brief overview on insurance loss models and Bayesian statistics. Section 3 presents the algorithmic details of the sequential Monte Carlo samplers used to fit the composite models. A simulation experiment is conducted in Section 4 to assess the consistency and finite-sample performance of the estimation and model selection procedures. Section 5 illustrates the application of the SMC algorithm on a real life insurance dataset.

2 Preliminaries

Losses in insurance are usually modelled by nonnegative random variables with probability density function (PDF) denoted by $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}_+$ is the parameter space. The goal is then to find the parameter value $\widehat{\theta}$ that best explain the data $\boldsymbol{x} = (x_1, \ldots, x_n)$. Maximum likelihood estimation takes the parameter $\theta \in \Theta$ that maximises the likelihood function $L(\boldsymbol{x}|\theta)$ as

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\boldsymbol{x}|\theta).$$

In the case of independent and identically distributed (IID) data (which is the case considered here), the likelihood function is given by

$$L(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

The gamma distribution, denoted by $\mathsf{Gamma}(r, m)$, with PDF given by

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

is often used in practice for ratemaking purposes within generalized linear models. The tail of the gamma distribution is said light because of the exponential decrease of its survival function. Namely, it holds that

$$\mathbb{P}(X > x) \sim e^{-x/m}, \text{ when } x \to +\infty,$$

which is problematic to model large claim amounts. Probability distributions with "sub-exponential" tails have been used to circumvent this problem like the Weibull distribution $\mathsf{Weib}(k, \beta)$ with PDF given by

$$f(x; k, \beta) = \frac{k}{\beta} \left(\frac{x}{\beta} \right)^{k-1} e^{-(x/\beta)^k}, \, x > 0.$$

and the lognormal distribution $\mathsf{LogNorm}(\mu, \sigma)$ with PDF given by

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

If the tail of the Weibull and lognormal model are not heavy enough, one has to turn to extreme value probability distributions of which the Pareto distribution $\mathsf{Par}(\alpha, \gamma)$ with PDF

$$f(x; \gamma, \alpha) = \begin{cases} 0, & x \leq \gamma, \\ \frac{\alpha \gamma^\alpha}{x^{\alpha+1}}, & x > \gamma. \end{cases}$$

is a prominent member.

The danish fire insurance claim dataset is a famous example of heavy tailed loss data. Figure 1 provides the histogram and boxplot of the distribution of the danish fire loss data retrieved from the companion R package `SMPracticals` of the book of Davison [13]. The empirical distribution in Figure 1 shows the high frequency of small claim amounts and the lower occurrence of much larger claims that stretches the loss distribution to the right. Figure 2 shows the quantile-quantil plots associated to the gamma, Weibull, lognormal, Pareto models fitted to the danish fire insurance loss data using maximum likelihood estimation. The gamma, Weibull and lognormal models tend to underestimate the higher order quantiles, see Figures 2a, 2b and 2c, while the Pareto model tend to overestimate them, see Figure 2d. The lack of fit of these simple models lead to consider more flexible models, referred to as composite models, in Section 2.1.

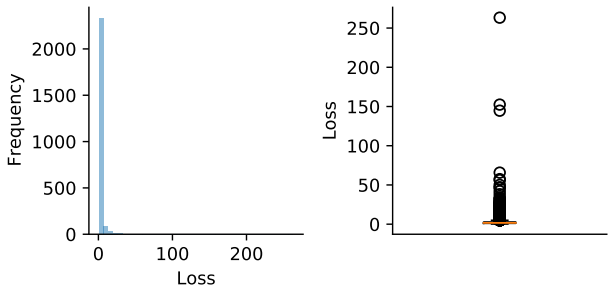
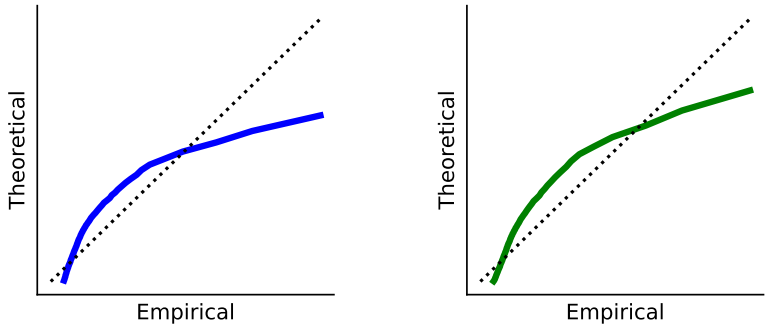
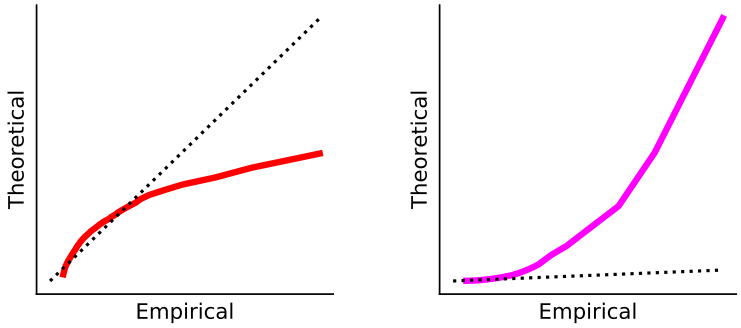


Figure 1: Empirical disribution of the danish fire insurance losses.



(a) Gamma($\widehat{r} = 1.30, \widehat{m} = 2.61$)

(b) Weib($\widehat{k} = 0.96, \widehat{\beta} = 3.29$)



(c) LogNorm($\widehat{\mu} = 0.79, \widehat{\sigma} = 0.72$)

(d) Par($\widehat{\alpha} = 0.54, \widehat{\gamma} = 0.31$)

Figure 2: Quantile-quantile plots associated to the gamma, Weibull, lognormal and Pareto models fitted to the danish fire insurance loss data using maximum likelihood estimation.

2.1 Composite models

Composite models result from the combination of two models, one for the "belly" and the other for the tail of the distribution. The PDF of a composite model is defined as

$$f(x) = \begin{cases} p\frac{f_1(x)}{F_1(\gamma)}, \text{ si } x \leq \gamma, \\ (1-p)\frac{f_2(x)}{1-F_2(\gamma)}, \text{ si } x > \gamma, \end{cases} \tag{1}$$

where f_1, F_1, f_2, and F_2 are the PDF and cumulative distribution function (CDF) of the models for the belly and the tail of the loss distribution respectively. The parameter $p \in (0,1)$ is referred to as the mixing parameter. The threshold parameter $\gamma > 0$ is the breaking point that distinguishes small claim amounts from large ones. Regularity conditions are usually impose over the PDF of the composite model at $x = \gamma$ with

$$f(\gamma^-) = f(\gamma^+), \text{ et } f'(\gamma^-) = f'(\gamma^+). \tag{2}$$

We consider in this work the gamma, Weibull and lognormal model for the belly of the distribution and the Pareto model for the tail of the distribution. The regularity conditions lead to fix some parameters of the models. The settings of the $\mathsf{Gamma}(r,m) - \mathsf{Par}(\alpha,\gamma)$, $\mathsf{Weib}(k,\beta) - \mathsf{Par}(\alpha,\gamma)$ and $\mathsf{LogNorm}(\mu,\sigma) - \mathsf{Par}(\alpha,\gamma)$ composite models are specified in the Examples 1, 2, and 3 below.

Example 1. *If f_1 is the PDF of the gamma distribution $\mathsf{Gamma}(r,m)$ and f_2 is the PDF of the Pareto distribution $\mathsf{Par}(\alpha,\gamma)$, then conditions (2) lead to express m and p in terms of the other parameters as*

$$m = \frac{\gamma}{k+\alpha}, \; p = \frac{\alpha\Gamma(k)F_1(\gamma;r,m)e^{k+\alpha}(k+\alpha)^{-k}}{1+\alpha\Gamma(k)F_1(\gamma;r,m)e^{k+\alpha}(k+\alpha)^{-k}},$$

where $F_1(\gamma;r,m)$ is the gamma distribution CDF.

Example 2. *If f_1 is the PDF of the Weibull distribution $\mathsf{Weib}(k,\beta)$ and f_2 is the PDF of the Pareto distribution $\mathsf{Par}(\alpha,\gamma)$, then conditions (2) lead to express β and p in terms of the other parameters as*

$$\beta = \left(\frac{k}{k+\alpha}\right)^{1/k}\gamma, \; p = \frac{\frac{\alpha}{\gamma}\left[1 - e^{\frac{k+\alpha}{k}}\right]}{\frac{\alpha}{\gamma} + \frac{k}{\gamma}e^{-\frac{k+\alpha}{k}}}.$$

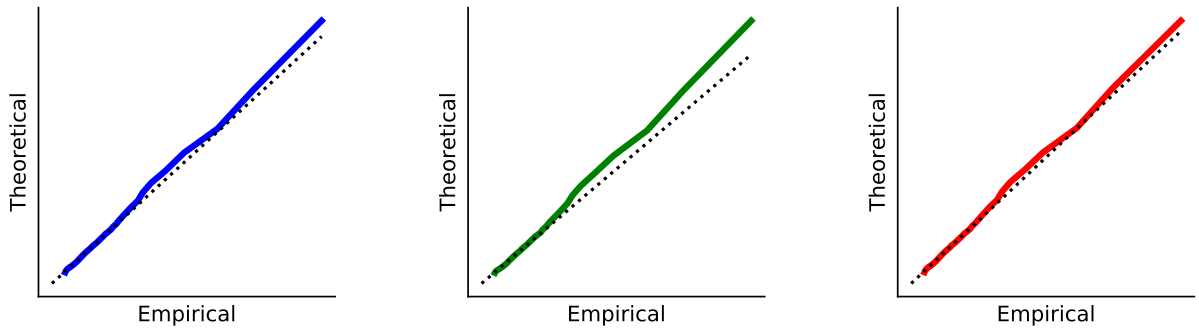
The $\mathsf{Weib}(k,\beta) - \mathsf{Par}(\alpha,\gamma)$ model has been studied in the work of Scollnik and Sun [37].

Example 3. *If f_1 is the PDF of the lognormal distribution $\mathsf{LogNorm}(\mu,\sigma)$ and f_2 is the PDF of the Pareto distribution $\mathsf{Par}(\alpha,\gamma)$, then conditions (2) lead to express β and p in terms of the other parameters as*

$$\mu = \ln(\gamma) - \alpha\sigma^2, \; p = \frac{\alpha\sigma\sqrt{2\pi}\Phi(\alpha\sigma)}{\alpha\sigma\sqrt{2\pi}\Phi(\alpha\sigma) + e^{-\alpha^2\sigma^2/2}},$$

where Φ denotes the CDF of the standard normal distribution. The $\mathsf{LogNorm}(\mu,\sigma) - \mathsf{Par}(\alpha,\gamma)$ model has been studied in the works of Cooray and Ananda [11] and Scollnik [36].

Figure 3 shows the quantile-quantile plots of the composite model fitted to the the danish fire losses data using maximum likelihood estimation. The fit of the composite models looks better than that of the simple



(a) Gamma($\widehat{r} = 35.68$) – Par($\widehat{\alpha} = 1.31, \widehat{\gamma} = 1.16$)

(b) Weib($k = 14.03$) – Par($\widehat{\alpha} = 1.26, \widehat{\gamma} = 1.00$)

(c) LogNorm($\sigma = 0.19$) – Par($\widehat{\alpha} = 1.32, \widehat{\gamma} = 1.21$)

Figure 3: Quantile-quantile plots of the composite models fitted to the danish fire loss data using maximum likelihood estimation.

loss models. Among the composite models the lognormal-Pareto one seems to provide the best adequacy with the empirical quantiles. In addition to graphical tools, adequacy can be measured through information criteria. The Akaike Information Criterion (AIC) is defined by

$$\text{AIC} = 2d - 2l(\mathbf{x}|\widehat{\theta}),$$

where d denotes the number of parameters, see Akaike [2]. The Bayesian information criterion (BIC) is given by

$$\text{BIC} = d\ln(n) - 2l(\mathbf{x}|\widehat{\theta}),$$

where n is the sample size and d is the number of parameters of the model, see Schwarz [35]. The best models minimize the deviance, defined by $-2l(\mathbf{x}|\theta)$, penalized by the number of parameters. Table 1 reports the AIC and BIC of the loss models introduced so far and fitted to the danish fire insurance loss data using using maximum likelihood estimation. The Weibull-Pareto model is associated to the lowest AIC and BIC despite the slight overestimation of the higher order quantile observed on Figure 3b. This work aims at looking into the Bayesian inference of the composite models instead of relying on the maximum likelihood estimators. A brief overview on Bayesian inference and model selection is provided in Section 2.2.

Loss models			AIC	BIC
Gamma	$\widehat{r} =$	1.26	10,490.05	10,501.69
	$\widehat{m} =$	2.43		
Weib	$\widehat{k} =$	0.947	10,544.94	10,556.58
	$\widehat{\delta} =$	2.95		
LogNorm	$\widehat{\mu} =$	0.67	8,931.19	8,942.83
	$\widehat{\sigma} =$	0.73		
Par	$\widehat{\alpha} =$	0.54	11,354.19	11,365.83
	$\widehat{\theta} =$	0.31		
Gamma – Par	$\widehat{r} =$	35.68	7,723.68	7,741.14
	$\widehat{\alpha} =$	1.31		
	$\widehat{\theta} =$	1.15		
Weib – Par	$\widehat{k} =$	14.03	7,686.75	7,704.21
	$\widehat{\alpha} =$	1.26		
	$\widehat{\gamma} =$	1.00		
LogNorm – Par	$\widehat{\sigma} =$	0.19	7,737.73	7,755.19
	$\widehat{\alpha} =$	1.32		
	$\widehat{\gamma} =$	1.20		

Table 1: AIC and BIC of the loss models introduced in Section 2.1 and fitted to the danish fire insurance dataset.

2.2 Bayesian inference

Bayesian statistics defines the posterior distribution of the model parameters θ given the data $\boldsymbol{x} = (x_1, \ldots, x_n)$ as

$$\pi(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\theta)\pi(\theta)}{Z(\boldsymbol{x})}. \tag{3}$$

The posterior distribution (3) follows from applying Bayes' rule to update the prior distribution $\pi(\theta)$ using the likelihood function $L(\boldsymbol{x}|\theta)$. Credible sets as well as point estimates can then be derived from $\pi(\theta|\boldsymbol{x})$ to draw inference on θ. The only issue is the denominator in (3) which is a normalizing constant given by

$$Z(\boldsymbol{x}) = \int_{\Theta} L(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta. \tag{4}$$

The above integral rarely admits a closed-form expression except when the model for the data has a conjugate prior distribution, as in Example 4.

Example 4. *Assume that the claim sizes data is exponentially distributed* $\boldsymbol{x} \overset{\textit{i.i.d.}}{\sim} \mathsf{Gamma}(1, 1/\delta)$ *and take gamma prior distribution over the model parameter* $\delta \sim \mathsf{Gamma}(a, 1/b)$*, then the posterior distribution is also gamma with*

$$\delta|\boldsymbol{x} \sim \mathsf{Gamma}\left(n + a, \frac{1}{b + \sum_{i=1}^{n} x_i}\right),$$

and the normalizing constant is given by

$$Z(\boldsymbol{x}) = \frac{b^a}{\left(\sum_{i=1}^{n} x_i + b\right)^{a+n}} \prod_{i=1}^{n} (a + i). \tag{5}$$

Unfortunately conjugate priors almost only arise in exponential families of probability distributions, see Diaconis and Ylvisaker [14]. In practice, one samples from the posterior distribution via Markov Chain Monte Carlo (MCMC) schemes. The Metropolis-Hasting random walk builds a sequence $(\theta^i)_{i\geq 0}$ by applying a Markov kernel $K_H(\cdot|\theta^i)$ to the current parameter value θ^i, $i \geq 0$. The parameter H corresponds to the magnitude of the perturbation. A new parameter value $\theta^* \sim K_H(\cdot|\theta^i)$ is accepted with probability

$$\alpha(\theta^i, \theta^*) = \max\left[1, \frac{L(\boldsymbol{x}|\theta^*)\pi(\theta^*)K_H(\theta^*|\theta^i)}{L(\boldsymbol{x}|\theta^1)\pi(\theta^i)K_H(\theta^i|\theta^*)}\right], \tag{6}$$

in which case $\theta^{i+1} = \theta^*$, otherwise $\theta^{i+1} = \theta^i$. The resulting sequence $(\theta^i)_{i\geq 0}$ forms a Markov chain trajectory having the posterior distribution as limiting distribution. A standard choice for the Markov kernel is the multivariate normal distribution

$$K_H(\cdot|\theta) \sim \mathsf{Norm}(\mu = \theta, \Sigma = H), \tag{7}$$

where H is a matrix that matches the dimension of θ. The Metropolis-Hasting random walk efficiency, understood as the speed of convergence of the Markov chain toward its asymptotic distribution, decreases with the dimension of the parameters. The workarround consists in turning to another well known MCMC technique that generate a sequence $(\theta^i)_{i\geq 0}$ called Gibbs sampling. To sample from a multivariate posterior distribution $\pi(\theta|\boldsymbol{x})$, a Gibbs sampler samples from the univariate conditional distributions defined as

$$\pi(\theta_j|\boldsymbol{x}, \theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d), \text{ for } j = 1, \ldots, d,$$

where $\theta = (\theta_1, \ldots, \theta_d)$. The current parameter value θ^i is updated component per component starting with the first one

$$\theta_1^i \sim \pi(\cdot|\boldsymbol{x}, \theta_2^i, \ldots, \theta_d^i), \tag{8}$$

before moving to the second one

$$\theta_2^i \sim \pi(\cdot|\boldsymbol{x}, \theta_1^i, \theta_3^i, \ldots, \theta_d^i), \tag{9}$$

and so on. The sequence $(\theta^i)_{i\geq 0}$ forms a Markov chain trajectory whose limiting distribution is the posterior distribution $\pi(\theta|\boldsymbol{x})$. The marginal distributions in (8), (9), etc., are usually unknown and one actually uses the Metropolis-Hasting scheme within the Gibbs iterations to sample from them. The *Metropolis-Hasting within Gibbs* type algorithms admit some drawbacks. First the algorithm must be initialized. In practice, several chains are launched from different starting points θ^0 to verify if they all converge toward the same distribution. Second, the H parameter of the multivariate normal kernel in (7) must be tuned to ensure good sampling properties. It should reflect the variance of the posterior distribution which is unknown. The practical solution consists in putting together an adaptive procedure to adjust H on the fly to reach an acceptance rate of 23.4% which is deemed optimal, see the work of Roberts et al. [33]. Third, the trajectory generation cannot be paralellized. Lastly, MCMC algorithms allows one to sample the posterior distribution of any models as long as the likelihood function has a tractable expression by avoiding the evaluation of the normalizing constant in (3). Indeed, the latter does not appear in the acceptance probability expression in (6). The normalizing constant is nevertheless important for Bayesian model selection as explained below.

Consider a set of competing models $\mathcal{M} = \{m_1, \ldots, m_J\}$ and define a random variable M having a Probability Mass Function (PMF) concentrated on $\mathcal{M}$. A prior distribution such that $\mathbb{P}(M = m_j) = \pi(m_j) \geq 0$, for $j = 1, \ldots, J$, and $\sum_{j=1}^{J} \pi(m_j) = 1$ can then be specified and updated given the data to yield the posterior model evidence as

$$\pi(m_j|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|m_j)\pi(m_j)}{\sum_{i=1}^{J} L(\boldsymbol{x}|m_i)\pi(m_i)}, \; j = 1, \ldots, J. \tag{10}$$

The likelihood $L(\boldsymbol{x}|m)$ of model $m \in \mathcal{M}$ follows from integrating over the possible values of the parameter θ as

$$L(\boldsymbol{x}|m) = \int_{\Theta} L(\boldsymbol{x}|m, \theta)\pi(\theta|m)\mathrm{d}\theta,$$

which corresponds exactly to the normalizing constant in (4). The best model achieves the highest model evidence (10).

The next section describes a sequential Monte Carlo algorithm which allows one to sample from any posterior distributions while providing an approximation of the normalization constant. The implementation is effortless to parallelize and its hyperparameters are straigtforward to tune.

3 Sequential Monte Carlo samplers

Section 3.1 provides a quick reminder of the importance sampling principle required to understand the SMC algorithm detailed in Section 3.2. The SMC algorithms are then applied to fit composite models to danish fire insurance loss data in Section 3.3.

3.1 Importance sampling

Bayesian inference reduces to evaluating quantities such as

$$\mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\varphi) = \int_{\Theta} \varphi(\theta)\pi(\theta|x)\mathrm{d}\theta, \tag{11}$$

where $\mathbb{E}_{\pi(\theta|\boldsymbol{x})}$ is the expectation operator with respect to the posterior distribution and φ is some measurable application. The posterior mean, often used as point estimate, corresponds to the case $\varphi(\theta) = \theta$. The expectation (11) is evaluated through its Monte Carlo approximation

$$\mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\varphi) \approx \frac{1}{N}\sum_{i=1}^{N} \varphi(\theta_i), \tag{12}$$

where $\theta_1,\ldots,\theta_N$ is an IID sample distributed as $\pi(\theta|\boldsymbol{x})$. Importance sampling consists in sampling from a distribution g on Θ either because it is more convenient than sampling from $\pi(\theta|\boldsymbol{x})$ or because it reduces the variance associated to the Monte Carlo estimator (12). The approximation of the normalizing constant relies on the following identity

$$\begin{aligned}
\mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\varphi) &= \int_{\Theta} \varphi(\theta)\pi(\theta|x)\mathrm{d}\theta \\
&= \int_{\Theta} \varphi(\theta)\frac{L(\boldsymbol{x}|\theta)\pi(\theta)}{Z(\boldsymbol{x})}\mathrm{d}\theta \\
&= Z(\boldsymbol{x})^{-1}\int_{\Theta} \varphi(\theta)\frac{L(\boldsymbol{x}|\theta)\pi(\theta)}{g(\theta)}g(\theta)\mathrm{d}\theta \\
&= Z(\boldsymbol{x})^{-1}\int_{\Theta} \varphi(\theta)w(\theta)g(\theta)\mathrm{d}\theta \\
&= Z(\boldsymbol{x})^{-1}\mathbb{E}_g(\varphi\cdot w),
\end{aligned}$$

where $w(\theta) = L(\boldsymbol{x}|\theta)\pi(\theta)/g(\theta)$ is an unnormalized weight function. Taking $\varphi(\theta) = 1$ yields the following expression of the normalizing constant

$$Z(\boldsymbol{x}) = \mathbb{E}_g(w),$$

which may be approximated by

$$Z(\boldsymbol{x}) \approx \frac{1}{N}\sum_{i=1}^{N} w(\widetilde{\theta}_i),$$

where $\widetilde{\theta}_1,\ldots,\widetilde{\theta}_N$ is an IID sample generated from the proposal g. Importance sampling ultimately yields a cloud of weighted particles $\{W_i,\widetilde{\theta}_i\}$, where

$$W_i = \frac{w(\widetilde{\theta}_i)}{\sum_{j=1}^{N} w(\widetilde{\theta}_j)},\ i = 1,\ldots,N,$$

whose empirical distribution targets the posterior distribution in the sense that

$$\sum_{i=1}^{N} W_i \varphi(\widetilde{\theta}_i) \to \mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\varphi),\ \text{pour } N \to \infty,$$

for any measurable application φ. The main challenge when using importance sampling is to find a suitable importance distribution g. If the purpose of g is to be substitute for $\pi(\cdot|\boldsymbol{x})$ then the Effective Sample Size (ESS) of the particle cloud must be high enough. The ESS is an indicator taking values between 1 and N that measures the degeneracy of the cloud of particles. It corresponds to the size of an IID sample that would match the empirical variance of the cloud of weighted particles $\{(W_i,\widetilde{\theta}_i),\ i = 1,\ldots,N\}$. The ESS is estimated by

$$\text{ESS} \approx \frac{1}{\sum_{i=1}^{N} W_i^2}.$$

as suggested in Kong et al. [27].

The sequential Monte Carlo algorithm presented in the next section bypasses the choice of a proposal distribution by constructing a sequence of intermediary distributions while maintaining an appropriate effective sample size.

3.2 Sequential Monte Carlo algorithmic details

A sequential Monte Carlo algorithm builds a sequence of distribution $\pi_s(\theta|\boldsymbol{x})$, $s = 0,\ldots,t$ starting from the prior distribution $\pi_0(\theta|\boldsymbol{x}) = \pi(\theta)$ and ending on the posterior $\pi_t(\theta|\boldsymbol{x}) = \pi(\theta|\boldsymbol{x})$. Two ways of constructing the sequence $\pi_s(\theta|\boldsymbol{x})$, $s = 0,\ldots,t$ are considered in this work. The first consists in introducing the data by batch, see the work of Chopin [10], as

$$\pi_s(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}_{1:n_s}|\theta)\pi(\theta)}{Z_s},\ s = 0,\ldots,t, \tag{13}$$

where n_s, $s = 0,\ldots,t$ is a sequence of integers such that $0 = n_0 < n_1 < \ldots < n_t = n$, the normalizing constant is given by

$$Z_s = \int_{\Theta} L(\boldsymbol{x}_{1:n_s}|\theta)\pi(\theta)\mathrm{d}\theta,$$

and $\boldsymbol{x}_{1:n_s} = (x_1,\ldots,x_{n_s})$ is a sub-sample of $\boldsymbol{x}$. The second gradually activates the likelihood function as

$$\pi_s(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\theta)^{\tau_s}\pi(\theta)}{Z_s},\ s = 0,\ldots,t, \tag{14}$$

where τ_s, $s = 0,\dots,t$ is a sequence of real numbers such that $0 = \tau_0 < \tau_1 < \dots < \tau_t = 1$, and the normalizing constant is given by

$$Z_s = \int_\Theta L(\boldsymbol{x}|\theta)^{\tau_s}\pi(\theta)\mathrm{d}\theta.$$

This approach is inspired from the simulated annealing technique introduced by Neal [30]. The SMC algorithm initializes a cloud of particles using the prior distribution as

$$\theta_i^{(0)} \overset{\text{i.i.d.}}{\sim} \pi(\theta), \text{ and } W_i^{(0)} = \frac{1}{N}, \text{ for } i = 1,\dots,N.$$

To move from one intermediary distribution π_s to the next π_{s+1}, the SMC algorithm takes the cloud of particles $\{(W_i^s, \theta_i^s),\ i = 1,\dots,N\}$ and apply three operations to get $\{(W_i^{s+1}, \theta_i^{s+1}),\ i = 1,\dots,N\}$.

1. (Reweighting step) This step prepares the current cloud to target the next distribution. A particle θ_s^i is reweighted by

   $$W_i^{s+1} \propto w_i^{s+1} = \frac{\pi_{s+1}(\theta_i^s)}{\pi_s(\theta_i^s)}, \text{ for } i = 1,\dots,N,$$

   where $\propto$ stands for "proportional to" and the w_i^{s+1}'s are unnormalized weights, useful to estimate the normalizing constant as we shall see later. Because the weights $W_1^{s+1},\dots,W_N^{s+1}$ are actually importance weights, the targeted distribution π_{s+1} is chosen so that the weights satisfy

   $$\text{ESS} \approx \frac{1}{\sum_{i=1}^N (W_i^{s+1})^2} \geq \rho N,$$

   where $\rho \in (0,1)$. The selection of the next target reduces to picking a suitable sample size n_{s+1} or temperature τ_{s+1}. This is done via binary search and ρ is set to 1/2 following up on the recommendation of Jasra et al. [25].

2. (Resampling step) Particles $\widetilde{\theta}_1^s,\dots,\widetilde{\theta}_N^s$ are sampled from the particle clouds $\{(W_i^{s+1}, \theta_i^s),\ )i = 1,\dots,N\}$. A simple multinomial resampling is used here, but note that alternative schemes discussed for instance in the work of Gerber et al. [18] are also possible.

3. (Move step) Metropolis-Hasting within Gibbs moves are applied to the particles $\widetilde{\theta}_1^s,\dots,\widetilde{\theta}_N^s$ to yield the new generation of particles $\theta_1^{s+1},\dots,\theta_N^{s+1}$. The matrix H of the Markov Kernel K is given by $\widehat{\Sigma}\cdot 2.38/\sqrt{d}$, where $\widehat{\Sigma}$ is the empirical variance-covariance matrix of particles system $\{(W_i^{s+1}, \theta_i^s),\ i = 1,\dots,N\}$. The number of transitions $k \in \mathbb{N}$ to be applied is set to ensure the diversification of the particle cloud. In practice, the Markov kernel is applied once to each particle. The acceptance rate $\widehat{p_a}$ is estimated after this first round and k is then given by

   $$k = \max\left\{k_{\max}, \min\left[k_{\min}, \frac{\log(1-c)}{\log(1-\widehat{p_a})}\right]\right\},$$

where $k_{\min}$ and $k_{\max}$ denotes the minimum and maximum number of transitions, and $c \in (0,1)$ is the probability that each particle is moved at least once. Note that $k_{\min}, k_{\max}$ and c are the user-defined parameters of the SMC algorithm. The new particles $\theta_1^{s+1}, \ldots, \theta_N^{s+1}$ are sampled from π_{s+1} and are equally weighted with $W_i^{s+1} = 1/N$ for $i = 1, \ldots, N$.

The adaptative choice of the target distribution in step 1 and the calibration of H and k in step 3 are standard SMC algorithmic tricks used for instance in the paper of South et al. [38] and the SMC sampler of the Python package PYMC of Salvatier et al. [34]. The move step is easy to paralellize to optimize the computing time. A summary of the algorithm is given in Algorithm 1. The unnormalized weights $\{w_i^s,\ 1 \le i \le N,\ 1 \le s \le t\}$ yield an approximation of the normalizing constant as

$$Z(\boldsymbol{x}) = Z_t = \prod_{s=1}^{t} \frac{Z_s}{Z_{s-1}} \approx \prod_{s=1}^{t} \left(\frac{1}{N} \sum_{i=1}^{N} {w_i}^s \right).$$

Figure 4 shows the distribution of the SMC approximations of the normalizing constant of the exponential model of Example 4 computed on a IID sample $\boldsymbol{x} \sim \text{Gamma}(1, 1/3)$ of size 50 with prior distribution $\delta \sim \text{Gamma}(0.1, 10)$ and population sizes varying in $N \in \{500, 2000, 5000\}$. The accuracy of the estimator depends on the population

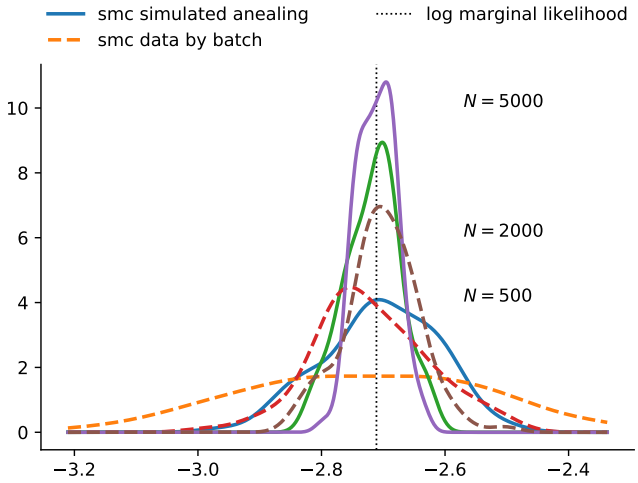


Figure 4: Histogram of the SMC approximations of the posterior distribution normalizing constant for an exponential model $\text{Gamma}(1, 1/\delta)$ with prior assumption $\delta \sim \text{Gamma}(0.1, 10)$ depending on the population size $N \in \{500, 2000, 5000\}$. The data is made of 100 IID samples of 50 observations drawn from an exponential model $\boldsymbol{x} \sim \text{Gamma}(1, 1/3)$.

size (the higher the better) which is chosen by the user according to a computing time budget. The SMC algorithm is applied to the danish fire insurance loss data in the following section.

Algorithm 1 SMC sampler for $\pi(\theta|\boldsymbol{x})$

Set $\rho \in (0,1)$; $k_{\min} \in \mathbb{N}$; $k_{\max} \in \mathbb{N}$; $c \in (0,1)$

Initialize $s \leftarrow 0$; $\pi_0(\theta) \leftarrow \pi(\theta)$;

for $i = 1 \rightarrow N$ **do**

  $\theta_i^0 \sim \pi(\theta)$; $W_i^0 \leftarrow 1/N$

end for

while $\pi_s(\theta) \neq \pi(\theta|\boldsymbol{x})$ **do**

  Search for π_{s+1} such that

$$\frac{1}{\sum_{i=1}^{N}(W_i^{s+1})^2} \geq \rho N, \text{ with } W_i^{s+1} \propto w_i^{s+1} = \pi_{s+1}(\theta_i^s)/\pi_s(\theta_i^s),\ i = 1,\ldots,N$$

  Compute $\widehat{\Sigma} = \mathrm{Cov}\left(\{(W_i^{s+1}, \theta_i^s),\ i = 1,\ldots,N\}\right)$

  for $i = 1 \rightarrow N$ **do**

    Sample $\widetilde{\theta}_i \sim \{\theta_1^{(s)},\ldots,\theta_N^{(s)}\}$ with probabilities W_j^{s+1}, pour $1 \leq j \leq N$

  end for

  for $i = 1 \rightarrow N$ **do**

    $\widetilde{\theta}_i^* \leftarrow K_H(\widetilde{\theta}_i, \cdot)$ où $K_H(\widetilde{\theta}_i, \cdot)$ where $H = \frac{2.38}{\sqrt{d}} \cdot \widehat{\Sigma}$

  end for

  Compute $p_a = N^{-1} \sum_{i=1}^{N} \mathbb{I}_{\widetilde{\theta}_i^* = \widetilde{\theta}_i}$; $k = \max\left\{k_{\max}, \min\left[k_{\min}, \frac{\log(1-c)}{\log(1-p_a)}\right]\right\}$

  for $i = 1 \rightarrow N$ **do**

    $\theta_i^{s+1} \leftarrow K_H^{*(k-1)}(\widetilde{\theta}_i^*, \cdot)$ where $K_H^{*(k-1)}(\widetilde{\theta}_i^*, \cdot)$ corresponds to $k-1$ Metropolis-Hasting-Gibbs moves

    $W_i^{s+1} \leftarrow 1/N$

  end for

end while

Return $(W_1^t, \theta_1^t),\ldots,(W_N^t, \theta_N^t)$

3.3 Application to composite models

Posterior model evidences have been criticized in the literature because the marginal likelihood is too sensitive to the prior distribution and measures the adequacy of the model to the data that used for the fit. Two information criteria are computed to compliment the analysis of the posterior model evidences. Let $\theta_1,\ldots,\theta_N$ be an IID sample from the posterior distribution $\pi(\theta|\boldsymbol{x})$. The Deviance Information Criterion (DIC), introduced

in the work of Spiegelhalter et al. [39], is defined by

$$\text{DIC} = -2[l(\boldsymbol{x}|\widetilde{\theta}) - p_{DIC}],$$

where $\widetilde{\theta} = \mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\theta)$ is the mean of the posterior distribution and

$$p_{DIC} = \mathbb{E}_{\pi(\theta|\boldsymbol{x})}[l(\boldsymbol{x}|\theta)] - l(\boldsymbol{x}|\widetilde{\theta}) \approx \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{x}|\theta_i) - l(\boldsymbol{x}|\widetilde{\theta}),$$

is a correction term that tends toward the number of parameters of the model. The DIC is less sensitive to the pior assumptions and is similar in nature to the AIC and BIC as it relies on the deviance augmented by the number of parameters. To assess the predictive capacity of the model, it would be better to build an information criterion based on the log pointwise predictive density

$$\text{lppd} = \sum_{j=1}^{n} \log \mathbb{E}_{\pi(\theta|\boldsymbol{x})}\Big[L(x_j^*|\theta)\Big] \approx \sum_{j=1}^{n} \log\left[\frac{1}{N}\sum_{i=1}^{N} L(x_j^*|\theta_i)\right],$$

where $\boldsymbol{x}^*$ is a left-out sample of data. One way to achieve this consists in resorting to a leave-one-out cross validation procedure. The computing time associated to fitting the models several times is often prohibitive. One workarround is to compute the log pointwise density on the observed data and add a correction term to it. This is the idea of the Widely Applicable Information Criterion (WAIC), defined as

$$WAIC = -2\left\{\sum_{j=1}^{n} \log \mathbb{E}_{\pi(\theta|\boldsymbol{x})}\Big[L(x_j|\theta)\Big] - p_{WAIC}\right\}, \tag{15}$$

where

$$p_{WAIC} = \mathbb{V}_{\pi(\theta|\boldsymbol{x})}[l(\boldsymbol{x}|\theta)].$$

Watanabe [41] has shown that applying this correction makes the log pointwise predictive density in (15) asymptotically equivalent to computing the log pointwise predictive density using a leave-one-out cross validation procedure. For a comprehensive discussion about the information criteria used in Bayesian statistics, I refer the reader to the work of Gelman et al. [16].

The SMC algorithm is applied to the danish fire loss data using the simulated annealing and data by batch approaches. The algorithm hyperparameters are set as follows

$$N = 1000, \rho = \frac{1}{2}, k_{\min} = 2, k_{\max} = 25, \text{ and } c = 0.99.$$

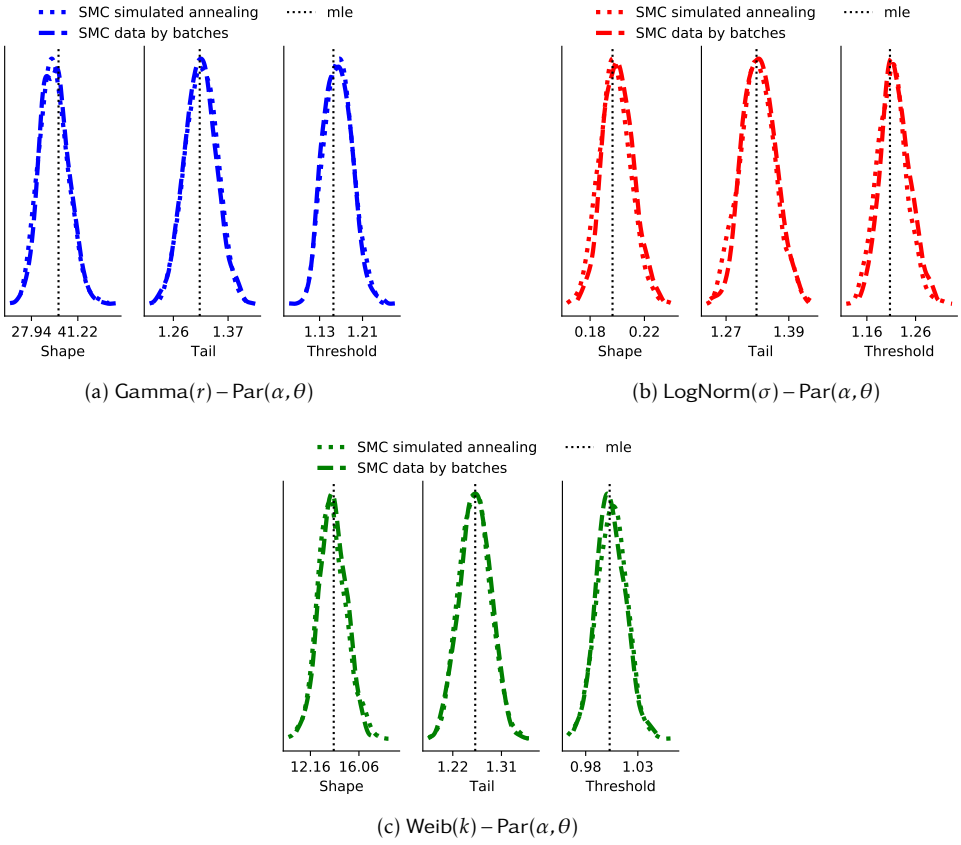
The prior distribution over the parameters are independent gamma distributions with

$$r \sim \text{Gamma}([0.1, 10]),\ \sigma \sim \text{Gamma}([0.1, 10]),\ k \sim \text{Gamma}([0.1, 10]),$$

for the small claim sizes

$$\alpha \sim \text{Gamma}([0.1, 10]),\ \theta \sim \text{Gamma}([0.1, 10])),$$

for the larger claim sizes portion of the data. The posterior distribution of the composite models parameters are given on Figure 5. The posterior distributions are similar for the two SMC algorithms and are concentrate



(a) Gamma(r) – Par(α, θ)

(b) LogNorm(σ) – Par(α, θ)

(c) Weib(k) – Par(α, θ)

Figure 5: Posterior distributions of the composite models parameters fitted to the danish fire loss data using the SMC samplers based on the simulated annealing (dotted) and data by batch (dashed) approaches.

arround the maximum likelihood estimators. This result was expected since the composite models satisfy the regularity conditions for the convergence of Bayes estimators toward point estimates that minimize the Kullback-Leibler divergence (e.g. maximum likelihood estimators) to hold, see the work of Bunke and Milhaud [6] and the references therein. The SMC algorithm returns an estimation of the log marginal likelihood that enables the evaluation of the posterior evidences of the competing models as

$$\pi(m_j|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|m_j)\pi(m_j)}{\sum_{i=1}^{J} L(\boldsymbol{x}|m_i)\pi(m_i)},\ j = 1, \ldots, J.$$

A discrete uniform prior $\pi(m_j) = 1/3,\ j = 1, 2, 3$, is set over the three composite models. The posterior model evidences and the information criteria for the composite models are reported in Table 2. The posterior model

Methods	Models	$\log Z(\boldsymbol{x})$	$\pi(m\|\boldsymbol{x})$	DIC	WAIC	Time
Simulated	LogNorm – Par	−3882.53	0	7725.72	7743.71	–
annealing	Weib – Par	−3858.50	1	7674.48	7689.55	
	Gamma – Par	−3878.20	0	7711.70	7730.08	
Data	LogNorm – Par	−3897.10	0	7725.99	7744.63	4×
by	Weib – Par	−3857.55	1	7674.45	7689.83	
batch	Gamma – Par	−3877.91	0	7711.60	7729.29	

Table 2: Posterior model evidences and information criteria of the composite models fitted to the danish fire insurance loss data.

evidence, the DIC and the WAIC all favor the Weibull-Pareto model which is consistent with the results of Section 2.1. We further note that the computing time associated to the SMC sampler for which the data is introduced gradually is four times higher. This is why only the simulated annealing approach is considered in the simulation study of Section 4 and the real data analysis in Section 5.

4 Simulation study

The SMC sampler is applied to fit the lognormal-Pareto, gamma-Pareto and Weibull-Pareto models on data generated by a $\mathsf{LogNorm}(\sigma = 1/2) - \mathsf{Par}(\alpha = 1, \gamma = 5)$ model. The hyperparameters of the SMC sampler are given by

$$N = 1000, \rho = \frac{1}{2}, k_{\min} = 2, k_{\max} = 25, \text{ and } c = 0.99.$$

In Section 4.1, the models are fitted on samples of sizes $50, 100, 250$. Bayesian posterior consistency holds as composite models satisfy the required regularity conditions, see for instance the survey of Hong and Martin [24]. The goal of Section 4.1 is to appreciate the speed at which the posterior distributions concentrate around the true value of the model parameters in the case of the lognormal-Pareto model and around the pseudotrue value of the parameter in the case of the Weibull-Pareto and gamma-Pareto models. The pseudotrue value is given by the maximum likelihood estimator computed on a sample of size $100,000$. Section 4.2 repeats the experiment of Section 4.1 $1,000$ times and also consider samples of size 500. The goal is to see how often the posterior model evidence and the information criteria point to the model that generated the data.

As part of the study of the distribution of the amounts of insurance claim, particular attention is paid to the estimation of higher order quantiles that characterize the risk associated to some insurance coverage. The posterior distribution over the model parameters leads to the definition of a posterior distribution of any quantity of interest that may be estimated through the considered model. The posterior distributions of the 95% and 99% quantiles are studied in Section 4.1.

Let Δ be a quantity of interest (e.g. the 95% quantile). It is possible to combine the estimation $\widehat{\Delta}_j$ of each competing model m_j for $j = 1,\dots,J$ through their posterior model evidence as

$$\widehat{\Delta} := \mathbb{E}(\Delta|\boldsymbol{x}) \approx \sum_{j=1}^{J} \widehat{\Delta}_j \pi(m_j|\boldsymbol{x}), j = 1,\dots,J. \tag{16}$$

This ensemble estimation procedure, known as Bayesian Model Averaging (BMA), is detailed in the work of Hoeting et al. [23] and used to estimate the 95% and 99% quantiles in Section 4.2.

4.1 Finite sample estimator consistency

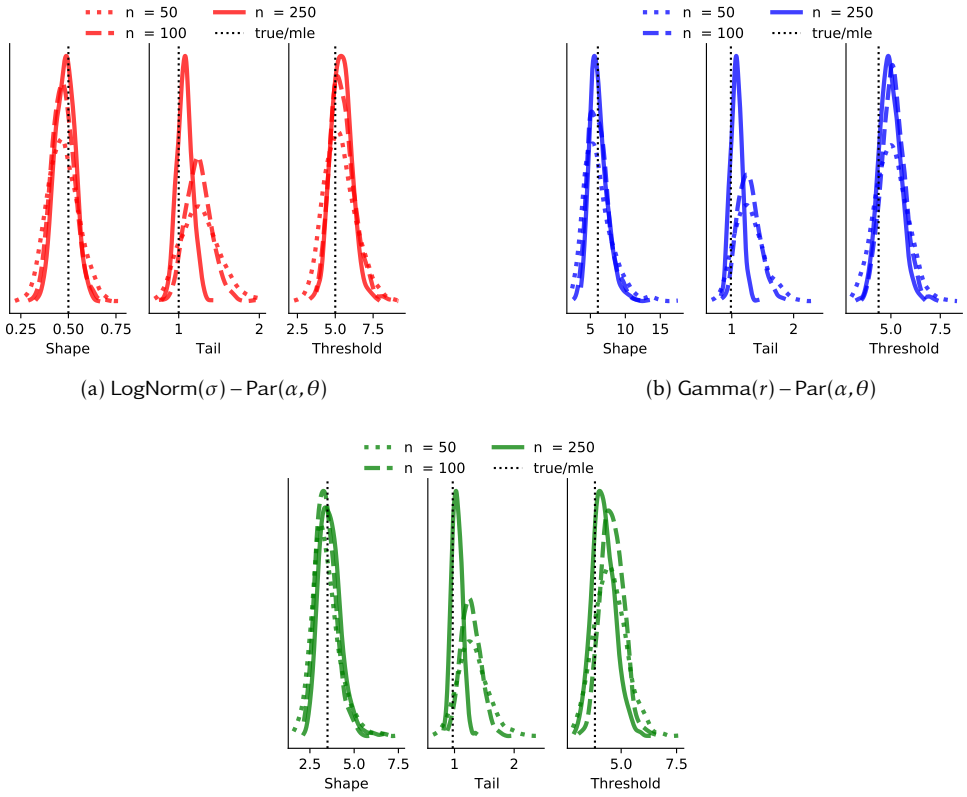
The $\mathsf{Gamma}(r)-\mathsf{Par}(\alpha,\theta)$, $\mathsf{LogNorm}(\sigma)-\mathsf{Par}(\alpha,\theta)$, and $\mathsf{Weib}(r)-\mathsf{Par}(\alpha,\theta)$ models are fitted on samples of sizes $50, 100$, and 250 generated by a $\mathsf{LogNorm}(\sigma = 1/2)-\mathsf{Par}(\alpha = 1, \gamma = 5)$ model. The prior assumptions on the attritional part of the severity distribution are given by

$$r \sim \mathsf{Gamma}(0.1, 10),\ \sigma \sim \mathsf{Gamma}(0.1, 10),\ k \sim \mathsf{Gamma}(0.1, 10),$$

The prior asumptions on the extreme part of the claim sizes distributions are given by

$$\alpha \sim \mathsf{Gamma}(0.1, 10),\ \theta \sim \mathsf{Gamma}(0.1, 10).$$

The resulting posterior distributions are shown on Figure 6. The posterior distribution concentrates arround the true and pseudotrue values of the parameters as the sample size increases. The posterior distribution of the 95% and 99% quantiles estimated through the composite models are shown on Figure 7. The true value of the quantiles fall inside the credible sets of all the models and for all the sample size considered. Taking a gamma-Pareto or Weibull-Pareto model instead of a lognormal-Pareto model only slightly deteriorates the precision on the right tail estimate. We note that the gap with the true value can be quite significant, especially for the 99% quantile when having only 50 observations.



(a) LogNorm(σ) – Par(α, θ)

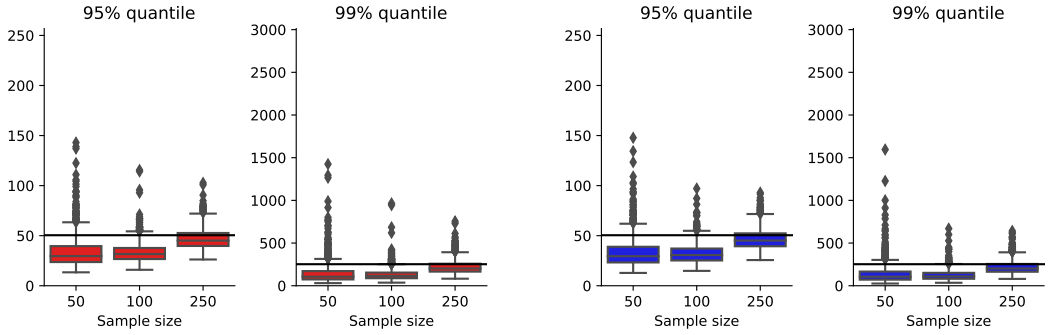
(b) Gamma(r) – Par(α, θ)

(c) Weib(k) – Par(α, θ)

Figure 6: Posterior distributions of the composite models parameters fitted to data simulated from a LogNorm($\sigma = 1/2$) – Par($\alpha = 1, \gamma = 5$) model. The samples contain 50 (dotted), 100 (dashed) et 250 (solid) observations.

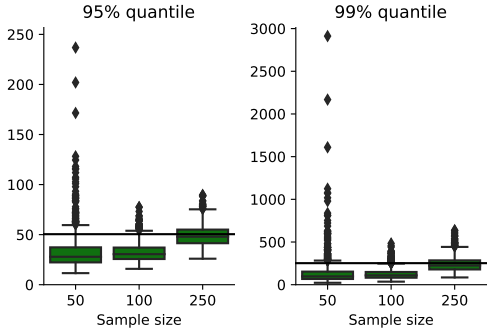
4.2 Finite sample model choice consistency

The experiment conducted in the previous section is repeated 1,000 times. The first objective is to study the behavior of the model evidences and information criteria as a function of the sample size. If only one model must be kept, it has to be the one associated to the highest the model evidence or the lowest DIC or WAIC. Figure 8 shows how often each model got selected over the 1,000 simulation runs by each criteria for sample of sizes 50, 100, 250, and 500. Note that LMD stands for Log Marginal Deviance and corresponds to the highest posterior model evidence (equivalently the highest log marginal likelihood). The DIC consistently picks the lognormal-Pareto model but the number of times improves in a slower fashion when increasing the sample size. The LMD and WAIC hesitate between the gamma-Pareto and the lognormal-Pareto models for sample of sizes $n \in \{50, 100\}$ before recommending the lognormal-Pareto model for larger sample sizes $n \in \{250, 500\}$. All in all



(a) LogNorm(σ) – Par(α, θ)

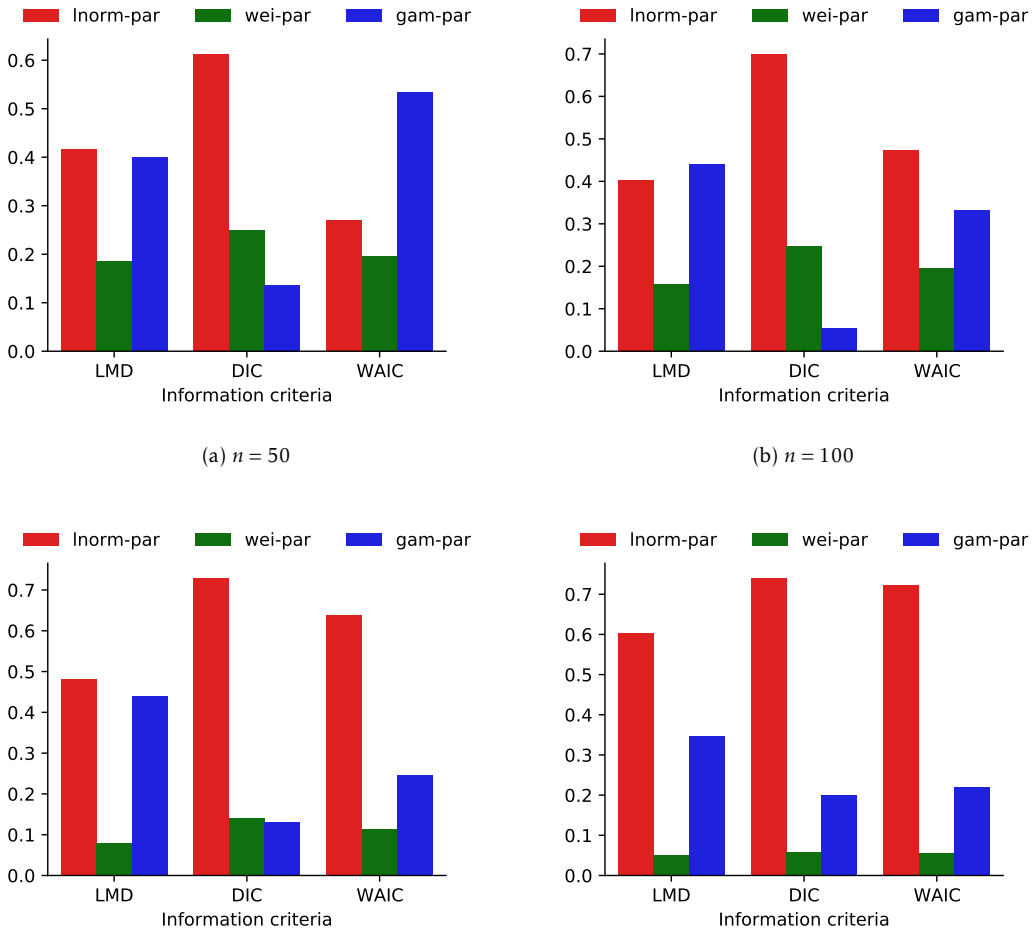
(b) Gamma(r) – Par(α, θ)



(c) Weib(k) – Par(α, θ)

Figure 7: Posterior distribution of the 95% and 99% quantiles of the composite models fitted to data drawn from a LogNorm – Par($\sigma = 1/2, \alpha = 1, \gamma = 5$) model. Samples of size 50, 100, and 250 are considered, and the solid line indicates the true 95% and 99% quantiles of the LogNorm($\sigma = 1/2$) – Par($\alpha = 1, \gamma = 5$) model.

the composite models considered are difficult to tell apart which is not surprising as they have a Pareto tail in common. The accuracy of the estimate of the 95% and 99% returned by the composite models is compared to the nonparametric estimates. The combination of the composite models estimates of the quantiles through Bayesian Model averaging is also considered. Table 3 report the mean absolute error over the 1,000 simulation runs. The composite model do a better a job at estimating the quantile than the non parametric method, especially for the 99% quantiles. The gamma-Pareto model achieves the best accuracy when the sample size is small $n \in \{50, 100\}$, the lognormal-Pareto takes over for larger sample sizes $n \in \{250, 500\}$. The BMA approach does not bring much improvement, it could benefit from the addition of more composite models. Now that the algorithm have been backtested successfully on artificial data, we are ready to apply it on a real insurance loss dataset.



(a) $n = 50$

(b) $n = 100$

(c) $n = 250$

(d) $n = 500$

Figure 8: How often each model is selected for data drawn from a LogNorm – Par($\sigma = 1/2, \alpha = 1, \gamma = 5$) model with sample size varying in $n \in \{50, 100, 250, 500\}$.

5 Application to real insurance data

The `ausautoBI8999` dataset includes 22,0365 closed auto insurance bodily injury claims in Australia. The data is retrieved from the R package `CASDatasets` maintained by Dutang and Charpentier [15] that accompanies the textbook of Charpentier [9]. An extract of the dataset is provided in Table 4. The variable `AggClaim` indicates the claim amount, the variable `FinDate` indicates the settlement date and `FinMth` is an index for the month of settlement. Descriptive statistics for the claim severities are provided in Table 5. The loss distribution is highly dispersed, note that the maximum exceed 4 millions and that the standard deviation is greater than

	$n = 50$		$n = 100$		$n = 250$		$n = 500$	
modèle	95%	99%	95%	99%	95%	99%	95%	99%
Empirical	23.48	531.88	17.31	205.45	11.33	126.41	7.72	85.76
LogNorm – Par	24.27	240.95	**15.50**	136.56	**9.30**	**77.80**	**6.16**	**50.69**
Weib – Par	23.30	221.10	16.49	148.49	10.09	88.59	6.90	60.46
Gamma – Par	**21.68**	**195.71**	**15.50**	**134.61**	9.59	80.87	6.35	53.18
BMA	23.57	228.57	15.75	139.02	9.54	80.69	6.28	52.28

Table 3: Mean absolute error when estimating the 95% and 99% quantiles empirically and with the composite models for $1,000$ datasets drawn from a $\mathsf{LogNorm} - \mathsf{Par}(\sigma = 1/2, \alpha = 1, \gamma = 5)$ model and sample of sizes $50, 100, 250$, and 500.

FinDate	FinMth	AggClaim
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 4: Extract of the `ausautoBI8999` dataset that contains the losses associated to closed bodily injury motor insurance claims.

the mean. The loss distribution is summarized through an histogram and a boxplot on Figure 9. As usual, the empirical loss distribution exhibits a high frequency of small claim amounts and a few significantly larger claim amounts. The $\mathsf{Gamma}(r) - \mathsf{Par}(\alpha, \theta)$, $\mathsf{LogNorm}(\sigma) - \mathsf{Par}(\alpha, \theta)$, and $\mathsf{Weib}(r) - \mathsf{Par}(\alpha, \theta)$ composite models are fitted to the data using the SMC algorithm. The hyperparameters of the SMC sampler are given by

$$N = 1000, \rho = \frac{1}{2}, k_{\min} = 2, k_{\max} = 25, \text{ and } c = 0.99.$$

The prior asumptions over the attritional part of the loss distributions are as follows

$$r \sim \mathsf{Gamma}(0.1, 10),\ \sigma \sim \mathsf{Gamma}(0.1, 10),\ k \sim \mathsf{Gamma}(0.1, 10).$$

The prior assumptions over the tail of the loss distribution are given by

$$\alpha \sim \mathsf{Gamma}(0.1, 10),\ \theta \sim \mathsf{Gamma}(0.1, 10).$$

	AggClaim
Number of observations	22,036.00
Mean	38,367.22
Standard Deviation	90,981.11
Minimum	9.96
25% Quantile	6,296.97
50% Quantile	13,853.87
75% Quantile	35,123.42
Maximum	4,485,797.21

Table 5: Descriptive statistics of the variable `AggClaim` part of the `ausautoBI8999` dataset.

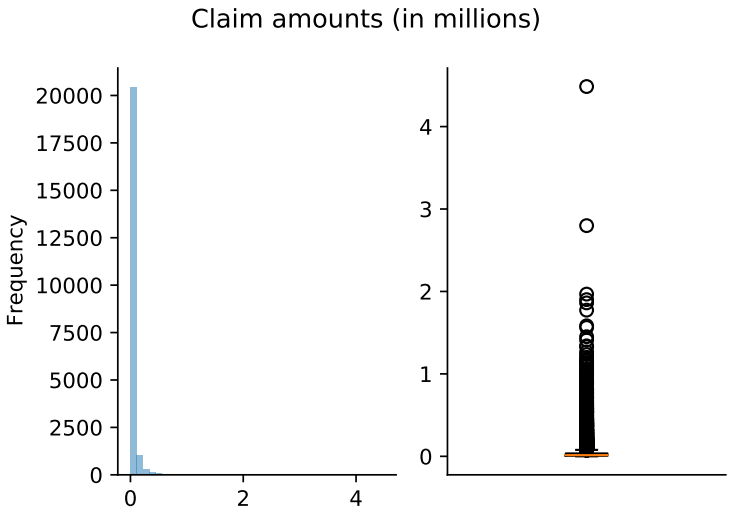
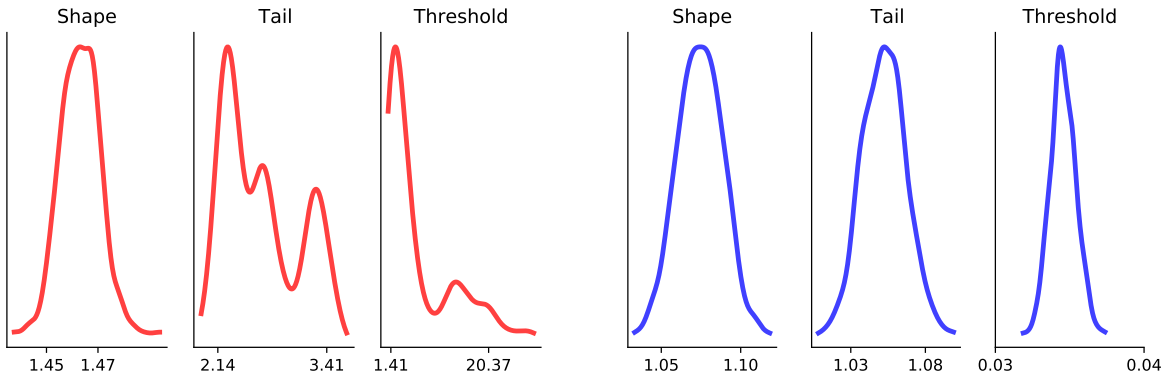


Figure 9: Loss distribution of the `ausautoBI8999` dataset.

Section 5.1 considers the whole dataset while Section 5.2 subsets the data on a monthly basis.

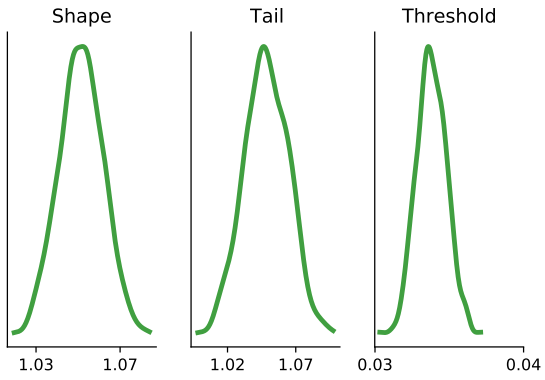
5.1 Overall analysis

The posterior distributions of the composite models parameters are given on Figure 10. The posterior distributions of the tail and threshold parameters of the lognormal-Pareto model are not as expected. The algorithm sets the threshold parameter to very high levels leaving very few observations to infer the tail parameter whose posterior distribution is quite wide. This means that the method tries to model the losses only



(a) LogNorm(σ) – Par(α, θ)

(b) Gamma(r) – Par(α, θ)



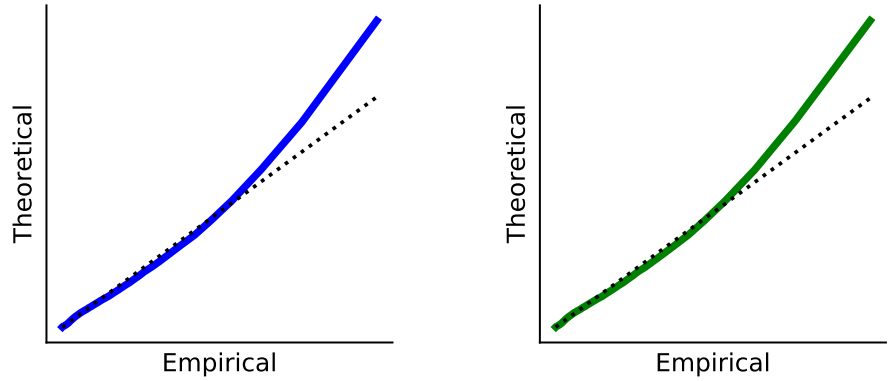
(c) Weib(k) – Par(α, θ)

Figure 10: Posterior distributions of the composite models parameters fitted to the australian motor insurance loss data.

using the lognormal distribution. This could be explained by the good adequacy of the lognormal distribution to the data. The lognormal model LogNorm(μ, σ) is fitted to the data with asociated prior assumptions

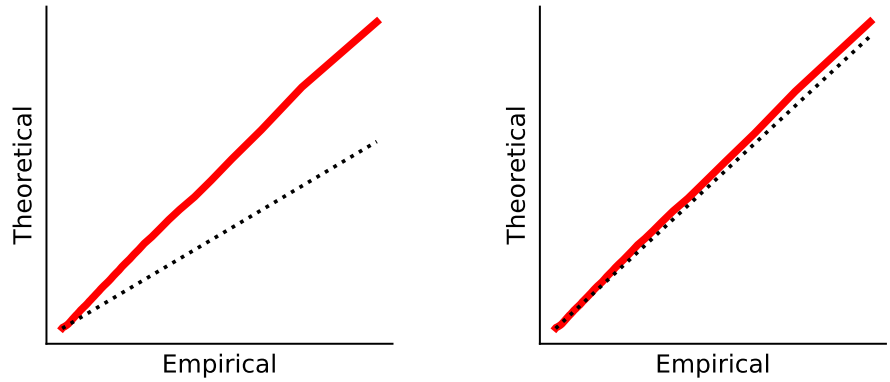
$$\mu \sim \text{Norm}(0, 10), \ \sigma \sim \text{Gamma}(0.1, 10).$$

Figure 11 shows the quantile-quantile plots of the composite and lognormal models fitted to the data. The parameters values are given by the posterior mean. The fit of the lognormal model is indeed superb, see Figure 11d. The right tail of the Weibull-Pareto and gamma-Pareto models looks slightly too heavy, see Figures 11a and 11b. Both the posterior model evidence and the information criteria favor the gamma-Pareto model which contradicts the graphical hint provided by Figure 11.



(a) Gamma($\widehat{r} = 1.07$) – Par($\widehat{\alpha} = 1.05, \widehat{\gamma} = 0.03$)

(b) Weib($\widehat{k} = 1.05$) – Par($\widehat{\alpha} = 1.04, \widehat{\gamma} = 0.03$)



(c) LogNorm($\widehat{\sigma} = 1.46$) – Par($\widehat{\alpha} = 2.65, \widehat{\gamma} = 3.68$)

(d) LogNorm($\widehat{\mu} = -4.25, \widehat{\sigma} = 1.46$)

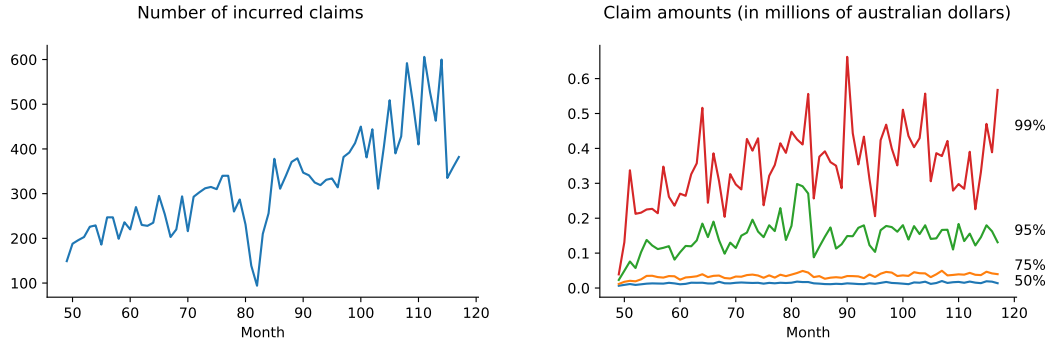
Figure 11: Quantile-quantile plots of the composite and lognormal models fitted to the australian motor insurance loss data.

5.2 Monthly analysis

The number of datapoints per month lies between 100 and 600. Figure 12 gives the number of observations and the upper empirical quantiles for each month. The 95% and 99% quantiles exhibits a high variance compared to the 50% and 75% quantiles. The posterior distributions of the parameters are summarized by the mean surrounded by the 5% and 95% quantiles for each month on Figure 13. The posterior distributions of the tail and threshold parameters of the lognormal-Pareto are too wide to be reliable. The posterior distribution for the Weibull-Pareto and gamma-Pareto model are of acceptable quality except for month 82. The number of

Models	log marginal likelihood	Model evidence	DIC	WAIC
LogNorm – Par	54,193.14	0.00	−107,797.02	−108,415.71
Weib – Par	54,558.40	0.02	−109,163.09	−109,149.82
Gamma – Par	54,562.08	0.98	−109,170.88	−109,157.51
LogNorm	54,196.78	0.00	−108,423.77	−108,420.43

Table 6: Posterior model evidence and information criteria of the composite and lognormal models fitted to the australian motor insurance loss data.

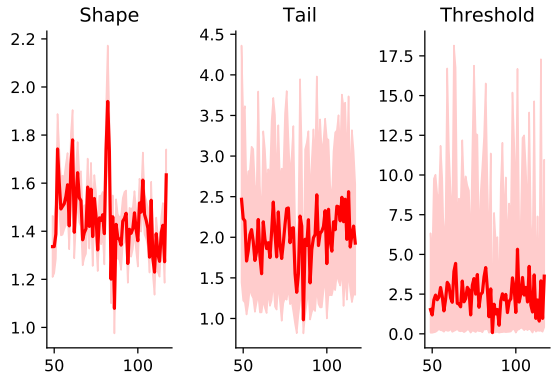


(a) Number of observations per month

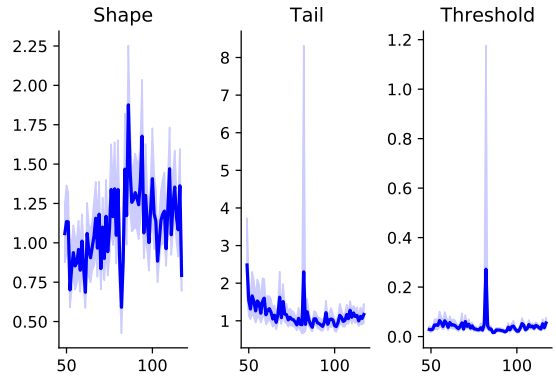
(b) 50%, 75%, 95% and 99% quantile of the loss distribution for each month

Figure 12: Number of observations and empirical quantiles fo each month.

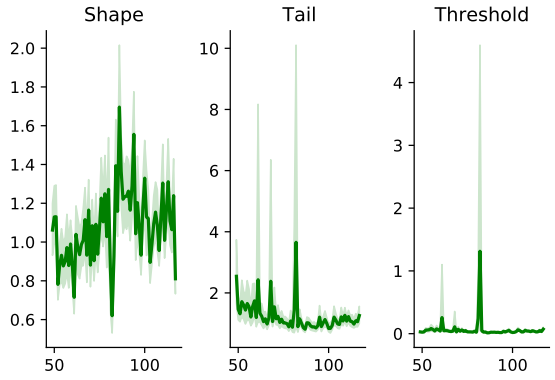
observations for this particular monthly drops to a low 94 which explains the poor quality of the posterior distributions. The model evidences of the composite models for each month are shown on Figure 14. The model selection methods favor (quite equally) the Weibull-Pareto and gamma-Pareto models for most of the month. It is a bit disapointing that the lognormal-Pareto model gets picked from time to time in spite of the poor quality of the posterior distribution over its parameters. This is reflected on the estimation of the 95% and 99% quantiles of the loss distribution through the composite models and their combination resulting from the Bayesian model averaging approach shown on Figure 15. The estimations given by the Weibull-Pareto and the gamma-Pareto models are quite close to the empirical estimation for the 95% quantiles, it is significantly higher for the 99% quantile.



(a) LogNorm(σ) – Par(α, θ)



(b) Gamma(r) – Par(α, θ)



(c) Weib(k) – Par(α, θ)

Figure 13: mean, 5% and 95% quantiles of the posterior distribution of the composite models parameters for each month.

The lognormal-Pareto model is replaced by a simple lognormal model LogNorm(μ, σ) with posterior assumptions given by

$$\mu \sim \text{Norm}(0, 10), \ \sigma \sim \text{Gamma}(0.1, 10).$$

The posterior distributions are summarized by the mean surrounded by the 5% and 95% quantiles of the posterior distribution for each month on Figure 16. The posterior distribution of the LogNorm(μ, σ) model is stable over the months. The updated model evidences are shown on Figure 17 and the estimations of the 95% and 99% quantiles of the loss distribution on Figure 18. The lognormal model is favored some months which improve the estimation of the quantiles of the loss distribution when using the Bayesian model averaging approach.

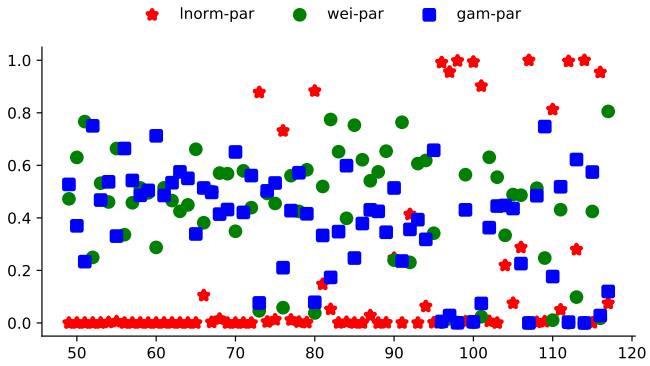


Figure 14: Posterior evidences of the Gamma(r) – Par(α, θ), LogNorm(σ) – Par(α, θ), and Weib(r) – Par(α, θ) models for each month.

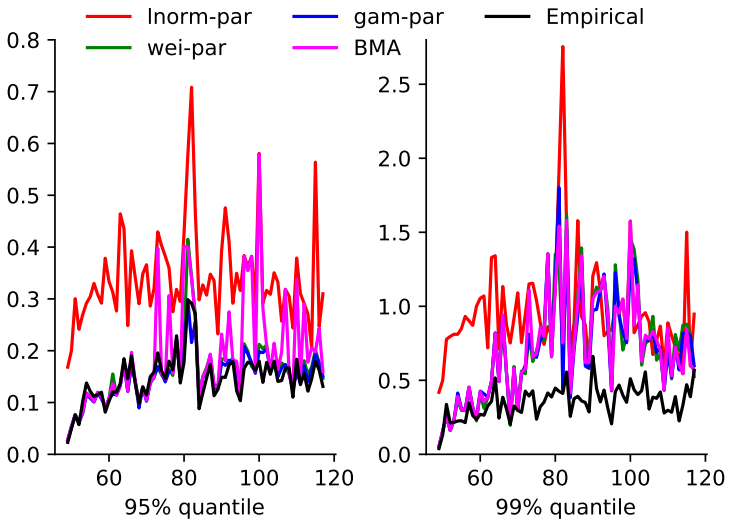


Figure 15: 95% and 99% quantiles estimated empirically and using the composite models for each month.

5.3 Discussion of the results

The analysis carried out in this section shows that a Pareto tail is too heavy for the data at hand. One workaround would be to consider alternative models for the tail, like the stoppa or the Burr distributions, see for instance the works of Calderín-Ojeda and Kwok [7], Abu Bakar et al. [1] and Grün and Miljkovic [21].

The inference method of composite models used here and referred to as the "simultaneous" approach in the survey of Wang et al. [40], is failing when one of the constituent of the composite model fits well the data.

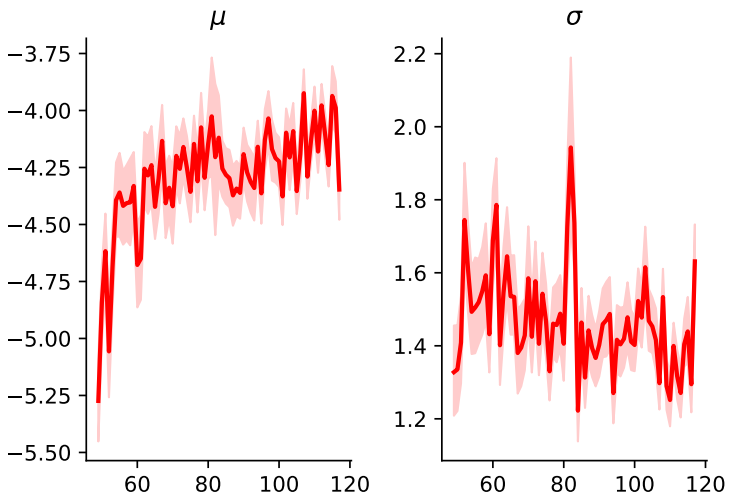


Figure 16: mean, 5% and 95% quantile of the posterior distribution of parameters of the LogNorm(μ, σ) model for each month.

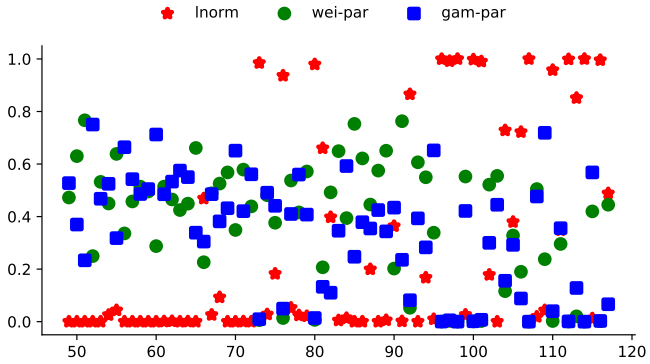


Figure 17: Posterior probabilities of the LogNorm(μ, σ), Gamma(r) – Par(α, θ), and Weib(r) – Par(α, θ) models for each month.

The problem we have encountered is easy to reproduce by simply fitting the lognormal-Pareto model to data generated by a lognormal model for instance. To the best of my knowledge, this limitation has never been pointed out before in the litterature. This is fine as long as the model selection procedure rejects the problematic model. In view of the results of Section 5.2, this is not always the case, especially when the number of observations is not sufficient.

Despite the lack of fit of the composite models to the right tail of the data, the likelihood-based criteria still

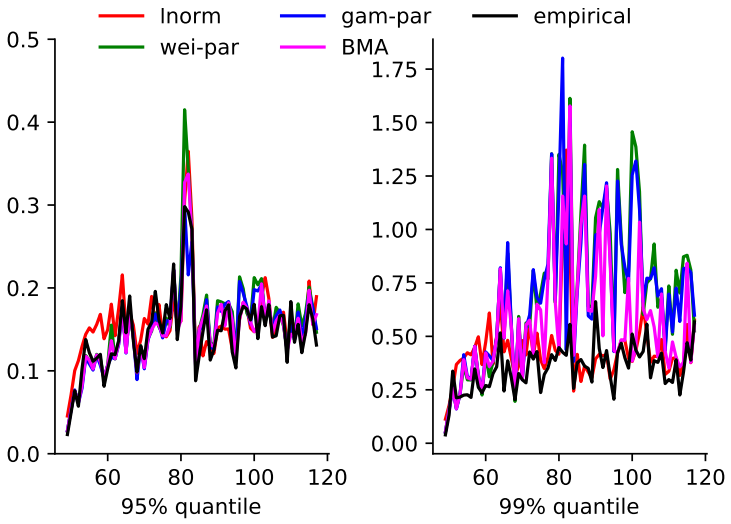


Figure 18: 95% and 99% quantiles estimated empirically and through the LogNorm(μ,σ), Gamma$(r)-$Par(α,θ), and Weib$(r)-$Par(α,θ) models for each month.

favor the composite models. The likelihood measures the overall fit of a model, the high frequency of small claims then gives too much weight to the belly of the distribution. If the ultimate goal is to estimate the high order quantiles then maybe inference method that do not rely on the likelihood function should be preferred. Minimum distance estimators that minimizes a discrepancy measure between the quantiles of the model and the empirical ones, would be better suited. The work of Bernton et al. [4] focuses on parameter estimates that minimize the Wasserstein distance which reduces in our case (IID and univariate data) to a distance between quantiles. Posterior distributions may be obtained by applying an Approximate Bayesian Computation (ABC) algorithm. ABC combined to the Wasserstein distance have been considered in the work of Bernton et al. [5] and applied to aggregated insurance data in the work of Goffard and Laub [20].

6 Conclusions and perspectives

This paper presents an implementation of a SMC sampler to fit and compare composite models in a Bayesian framework. The `python` code can be freely downloaded from the following `github` repository `https://github.com/LaGauffre/SMCCompoMo`. Likelihood functions of other composite models can be added to better the odds of finding the perfect fit. The Bayesian approach, compared to the frequentist approach, takes into account the uncertainty around the parameter estimates and enables to encapsulate expert knowledge in the prior distribution. SMC samplers have three advantages over the standard MCMC algorithm: (1) It avoids the fine

tuning of some hyperparameters, (2) it provides an appoximation of the normalizing constant as a byproduct, and (3) it is very easy to paralellize to take advantage of the multi-core processors that equip modern computers.

The simulation experiment showed the capacity of the algorithm to identify the model that generated the data. The analysis of real insurance data revealed a weak spot when one of the components of the composite model fits the data too well. The selection of a model using likelihood based criteria may not be optimal if the goal is to accurately estimate the higher order quantile. Inference and model selection procedures that rely on the minimization of a distance to the empirical quantiles will be investigated in a future research endeavor.

Acknowledgmements

The author's work is partially funded by the DIALog – Digital Insurance And Long-term risks – Chair under the aegis of the Fondation du Risque, a joint initiative by UCBL and CNP Assurances.

References

[1] S. A. Abu Bakar, N. A. Hamzah, M. Maghsoudi, and S. Nadarajah. Modeling loss data using composite models. *Insurance: Mathematics and Economics*, 61:146 – 154, 2015. ISSN 0167-6687. doi: https://doi.org/10.1016/j.insmatheco.2014.08.008. URL http://www.sciencedirect.com/science/article/pii/S0167668714001024.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Springer Series in Statistics*, pages 199–213. Springer New York, 1998. doi: 10.1007/978-1-4612-1694-0_15.

[3] G. Beirlant, Segers J. L., and Teugels D. *Statistics of Extremes*. John Wiley & Sons, 2004. ISBN 0471976474. URL https://www.ebook.de/de/product/3611778/beirlant_goegebeur_segers_jozef_l_teugels_daniel_de_waal_statistics_of_extremes.html.

[4] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, oct 2019. doi: 10.1093/imaiai/iaz003.

[5] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, feb 2019. doi: 10.1111/rssb.12312.

[6] O. Bunke and X. Milhaud. Asymptotic behavior of bayes estimates under possibly incorrect models. *The Annals of Statistics*, 26(2), apr 1998. doi: 10.1214/aos/1028144851.

[7] E. Calderín-Ojeda and C. F. Kwok. Modeling claims data with composite stoppa models. *Scandinavian Actuarial Journal*, 2016(9):817–836, apr 2015. doi: 10.1080/03461238.2015.1034763.

[8] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. doi: 10.18637/jss.v076.i01.

[9] A. Charpentier. *Computational Actuarial Science with R*. Taylor & Francis Ltd., August 2014. URL https://www.ebook.de/de/product/25483298/computational_actuarial_science_with_r.html.

[10] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, aug 2002. doi: 10.1093/biomet/89.3.539.

[11] K. Cooray and M. Ananda. Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian Actuarial Journal*, 2005(5):321–334, sep 2005. doi: 10.1080/03461230510009763.

[12] K. Cooray and C.-I. Cheng. Bayesian estimators of the lognormal–pareto composite distribution. *Scandinavian Actuarial Journal*, 2015(6):500–515, dec 2013. doi: 10.1080/03461238.2013.853368.

[13] A. C. Davison. *Statistical Models*. Cambridge University Press, October 2011. ISBN 0521773393. URL https://www.ebook.de/de/product/4229672/a_c_davison_statistical_models.html.

[14] P. Diaconis and D.S Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, mar 1979. doi: 10.1214/aos/1176344611.

[15] C. Dutang and A. Charpentier. Package 'casdatasets'. 2020.

[16] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, aug 2013. doi: 10.1007/s11222-013-9416-2.

[17] A. J. B. Gelman, H. S. Carlin, S. D. B. Dunson, and A. Vehtari. *Bayesian Data Analysis*. Taylor & Francis Ltd, 2013. ISBN 1439840954. URL http://www.stat.columbia.edu/~gelman/book/.

[18] Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4), aug 2019. doi: 10.1214/18-aos1746.

[19] F W Gerstengarbe and P C Werner. A method for the statistical definition of extreme-value regions and their application to meteorological time series. Zeitschrift fuer Meteorologie; (German Democratic Republic), January 1989.

[20] P.-O. Goffard and P. Laub. Approximate Bayesian Computations to fit and compare insurance loss models. working paper or preprint, April 2021. URL https://hal.archives-ouvertes.fr/hal-02891046.

[21] B. Grün and T. Miljkovic. Extending composite loss models using a general framework of advanced computational tools. *Scandinavian Actuarial Journal*, 2019(8):642–660, apr 2019. doi: 10.1080/03461238.2019.1596151.

[22] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5): 1163–1174, 1975. ISSN 00905364. URL http://www.jstor.org/stable/2958370.

[23] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999. ISSN 08834237. URL http://www.jstor.org/stable/2676803.

[24] L. Hong and R. Martin. A review of bayesian asymptotics in general insurance applications. *European Actuarial Journal*, 7(1):231–255, apr 2017. doi: 10.1007/s13385-017-0151-5.

[25] A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris. Inference for Lévy-driven stochastic volatility models via adaptive sequential monte carlo. *Scandinavian Journal of Statistics*, 38(1):1–22, dec 2010. doi: 10.1111/j.1467-9469.2010.00723.x.

[26] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, jun 1995. doi: 10.1080/01621459.1995.10476572.

[27] A. Kong, J. S. Liu, and W. H. 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.

[28] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[29] P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, jun 2006. doi: 10.1111/j.1467-9868.2006.00553.x.

[30] R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. doi: 10.1023/a:1008923215028.

[31] M. Pigeon and M. Denuit. Composite lognormal–pareto model with random threshold. *Scandinavian Actuarial Journal*, 2011(3):177–192, sep 2011. doi: 10.1080/03461231003690754.

[32] M. Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

[33] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, feb 1997. doi: 10.1214/aoap/1034625254.

[34] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016. doi: 10.7717/peerj-cs.55.

[35] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), mar 1978. doi: 10.1214/aos/1176344136.

[36] D. P. M. Scollnik. On composite lognormal-pareto models. *Scandinavian Actuarial Journal*, 2007(1):20–33, mar 2007. doi: 10.1080/03461230601110447.

[37] D. P. M. Scollnik and C. Chenchen Sun. Modeling with weibull-pareto models. *North American Actuarial Journal*, 16 (2):260–272, apr 2012. doi: 10.1080/10920277.2012.10590640.

[38] L. F. South, A. N. Pettitt, and C. C. Drovandi. Sequential monte carlo samplers with independent markov chain monte carlo proposals. *Bayesian Analysis*, 14(3):753–776, sep 2019. doi: 10.1214/18-ba1129.

[39] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, oct 2002. doi: 10.1111/1467-9868.00353.

[40] Y. Wang, I. Hobæk Haff, and A. Huseby. Modelling extreme claims via composite models and threshold selection methods. *Insurance: Mathematics and Economics*, 91:257–268, mar 2020. doi: 10.1016/j.insmatheco.2020.02.009.

[41] S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116):3571–3594, 2010. URL `http://jmlr.org/papers/v11/watanabe10a.html`.