Sequential Monte Carlo samplers to fit and compare insurance loss models

Pierre-O. Goffard[1]

[1] Université de Strasbourg, IRMA (UMR 7501).

November 3, 2022

Abstract

Insurance loss distributions are characterized by a high frequency of small claim amounts and a lower, but not insignificant, occurrence of large claim amounts. Composite models, which link two probability distributions, one for the "body" 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 an insurance loss dataset.

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. A model that accounts for both

extreme and moderate claim amounts is necessary to inform the insurance company decision process regarding premium calculation, risk capital allocation and risk transfer optimization.

The breakpoint between small and large claims is a key parameter in composite models. Two approaches are commonly used in practice to determine it. The first, known as the fixed-threshold approach, consists of choosing a threshold using tools from extreme value theory before adjusting the components of the composite models. It relies on famous graphical visualizations such as the mean-excess plot [21], the Hill plot [28] or the Gerstengarbe plot [20]. Automatic threshold selection methods have been implemented to mitigate the subjectivity inherent in locating a stable area on a curve. The goal is often to minimize the root mean squared error around the tail index estimate, see Caeiro and Gomes [9], or to pass a goodness-of-fit test, see Guillou and Hall [25]. These techniques rely on the hypothesis that the data points in excess of the threshold are distributed according to a Generalized Pareto Distribution (GPD). For an overview on extreme value data analysis, I refer the reader to the textbook of Beirlant et al. [4]. Another approach, less used in practice but widely studied in the actuarial literature, treats the threshold like any other parameter of the composite model and performs a simultaneous estimation. I will refer to this procedure as the free-threshold approach. A survey on the use of composite models and threshold selection methods on insurance data is conducted in the work of Wang et al. [49]. Maximum likelihood estimation is commonly used to fit several combinations of models for the body and the tail of the claim amounts distribution. The adequacy of each model is then measured using standard information criteria, see for instance Grün and Miljkovic [24].

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. [18]. 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 [34], JAGS [39], STAN [10] and PYMC [42], have been designed over the years so that practitioners do not have to worry about the fine tuning of these sophisticated algorithms. The lognormal-Pareto composite model of Scollnik [44] is actually an example of the WINBUGS documentation, see [34, Examples Vol. III]. The application of Bayesian techniques to composite models enables to quantify the uncertainty around the threshold parameter, as noted in the survey of Scarrot and MacDonald [43]. 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. [36]. 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 [31]. 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 multimodality of the posterior distribution around the threshold parameter has been encountered in practice by Cabras and Castellanos [8, Figure 3].

The main contribution of this work is to provide an efficient computational tool to fit and compare a myriad of composite models with a fixed or free threshold. This tool takes the form of a Python package available on `pip`[1] for anyone to use. The posterior distribution of the parameters and the posterior probability of the models make it possible to account for the uncertainty around the values of the parameters and the model to be used. This uncertainty can then be taken into account when estimating quantities of interest to risk managers, such as extreme quantiles or probabilities of ruin at one year, via credible sets. It is also possible to use the posterior probabilities of the model as weights to combine the estimate provided by each model. This procedure, known as Bayesian Model Averaging (BMA), see Kass and Raftery [31], is investigated. This work can be viewed as the Bayesian counterpart of the work of Grün and Miljkovic [24]. Besides the inference method, the main difference lies in the regularity assumptions of the probability density function (PDF) at the threshold parameter. Namely, the authors of [24] impose continuity and differentiability at the threshold whereas the models considered here can be discontinuous or only continuous but not differentiable. Another difference is that the algorithm allows the user to set in advance or not the breaking point between low and high severities to enable the comparison between the fixed and free threshold approach.

The remainder of the paper is organized as follows. Section 2 provides a brief overview on composite loss models and their use for risk management purposes. Section 3 recalls the principles of Bayesian statistics and 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

[1] https://pypi.org/project/bayes-splicing/

the famous danish fire insurance data.

2 Composite models

Losses in insurance are usually modelled by nonnegative random variables with probability density function (PDF) denoted by $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ is the parameter space and $d > 0$ its dimension. The goal is then to find the parameter value $\widehat{\theta}$ that best explains the data $\boldsymbol{x} = (x_1, \ldots, x_n)$. This is usually achieved by maximizing the likelihood function $L(\boldsymbol{x}|\theta)$. 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 occurence of extreme losses makes the vast majority of the simple parametric models (2-3 parameters) inadequate. The lack of fit of these simple models leads to consider more flexible ones.

2.1 Definition and assumptions

Composite models, sometimes called splicing or spliced models, combine two models, one for the "body" and the other for the "tail" of the distribution. The PDF of a composite model is defined as

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

where f_1, F_1, f_2, and F_2 are the PDF and cumulative distribution function (CDF) of the body 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. Such models appeared in the statistical literature as early as the work of Mendes and Lopez [14] who used maximum likelihood estimation to infer the parameters. A Bayesian framework was proposed by Berhens et al. [3] to enable the quantification of the uncertainty around the threshold parameter. A commonly used simplification of model (1) sets

$$p = F_1(\gamma), \tag{2}$$

making γ the p-quantile of the F_1 distribution. It has been pointed out several times that the potential discontinuity at $x = \gamma$ of the density (1) could pose problems in some practical situations. Continuity at the threshold $f(\gamma^-) = f(\gamma^+)$ is achieved by setting

$$p = \frac{f_1(\gamma)}{F_1(\gamma)} \Big/ \frac{f_2(\gamma)}{1-F_2(\gamma)}. \tag{3}$$

Composite models appeared in the actuarial science litterature with the work of Cooray and Ananda [11]. The authors further imposed differentiability at the threshold $f'(\gamma^-) = f'(\gamma^+)$ which leads to loose another degree of freedom by letting γ be the solution of

$$\frac{f_1'(\gamma)}{f_1(\gamma)} - \frac{f_2'(\gamma)}{f_2(\gamma)} = 0, \tag{4}$$

see the work of Grün and Miljkovic [24]. Wang et al [49] have reported that imposing such a condition tends to make the model not flexible enough leading to the conclusion that practitionners would be better off by choosing a threshold before fitting a general composite model.

In the present paper, the bulk and tail of composite models are selected from the list in Table 1. In contrast to

Name	Parameters		PDF
Exponential	Exp(δ)	$\delta > 0$	$\delta e^{-\delta x},\ x > 0$
Gamma	Gamma(r, m)	$r, m > 0$	$\frac{x^{r-1}e^{-x/m}}{\Gamma(r)m^r},\ x > 0$
Weibull	Weibull(k, β)	$k, \beta > 0$	$\frac{k}{\beta}\left(\frac{x}{\beta}\right)^{k-1} e^{-(x/\beta)^k},\ x > 0$
Lognormal	Lognormal(μ, σ)	$\mu \in \mathbb{R},\ \sigma > 0$	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right],\ x > 0$
Inverse-Gaussian	Inverse-Gaussian(μ, λ)	$\mu, \lambda > 0$	$\sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(\frac{\lambda(x-\mu)^2)}{2\mu^2 x}\right),\ x > 0$
Inverse-Gamma	Inverse-Gamma(r, m)	$r, m > 0$	$\frac{e^{-m/x}m^r}{x^{r+1}\gamma(r)},\ x > 0$
Inverse-Weibull	Inverse-Weibull(k, β)	$k, \beta > 0$	$k\beta^k x^{-k-1} e^{-(\beta/x)^k},\ x > 0$
Lomax	Lomax(α, σ)	$\alpha, \sigma > 0$	$\frac{\alpha\sigma^\alpha}{(\sigma+x)^{\alpha+1}},\ x > 0$
Log-Logistic	Log-Logistic(β, σ)	$\beta, \sigma > 0$	$\frac{\beta\sigma^\beta x^{\beta-1}}{(\sigma^\beta+x^\beta)^2},\ x > 0$
Burr	Burr(α, β, σ)	$\alpha, \beta, \sigma > 0$	$\frac{\alpha\beta\sigma^{\alpha\beta}x^{\beta-1}}{(\sigma^\beta+x^\beta)^{\alpha+1}},\ x > 0$
Pareto	Pareto(α, γ)	$\alpha, \gamma > 0$	$\frac{\gamma^\alpha}{x^{\alpha+1}},\ x > \gamma$
Generalized Pareto	GPD(ξ, σ, γ)	$\xi, \sigma, \gamma > 0$	$\sigma^{-1}\left[1 + \frac{\xi(x-\gamma)}{\sigma}\right]^{-(\xi+1)/\xi},\ x \geq 0$

Table 1: List of distribution for bulk and the tail of the composite models.

many previous works in the statistical and actuarial science litterature, the tail component of the composite model may not belong to the Pareto distribution family. Three composite model settings are considered. The "discontinuous" composite model assumes that the mixing parameter p has a prior distribution (uniform or

beta), the "simple" composite model sets p as in (2), and lastly the "continuous" composite model sets p as in (3).

Remark 2.1. *Imposing differentiability at $x = \gamma$ entails the loss of two degrees of freedom by letting the threshold be the solution of (4) and therefore depend on the other parameters of the composite models. The practical issue is that the uniqueness of the solutions of (4) is far from being granted. A numerical root finding procedure would be necessary which will slow down the inference process. A tedious case by case study of the 120 possible composite models would be necessary. Such a study has been carried out in the litterature for a few particular cases including the lognormal-Pareto case, see Cooray and Ananda [11] and Scollnik [44], and the Weibull-Pareto case, see Scollnik and Sun [45] and Abu Bakar et al. [1]. In addition to general composite model being discontinuous or continuous, The* `bayes-splicing` *package implements three composite models with continuity and differentiability at the threshold including the lognormal-Pareto model, the Weibull-Pareto model and the gamma-Pareto model described in Example 1.*

Example 1. *If f_1 is the PDF of the gamma distribution Gamma(r, m) and f_2 is the PDF of the Pareto distribution Pareto(α, γ), then continuity and diffrentiability of (1) 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.

The `bayes-splicing` package is able to fit a total of 375 loss models when accounting for all the distributions in Table 1 either considered separately or as component of a splicing model (with three possible settings "discontinuous", "simple" or "continuous"). The models are fitted and compared using state of the art Bayesian statistics computational tools which are presented in Section 3. The goal is to provide the right take on the loss distribution to assess the level of capital required for solvency purposes defined in Section 2.2.

2.2 Composite models usage in actuarial science

Over a given time period, a year say, a non-life insurance company handles a random number of claims N, each of which is associated to a randomly sized compensation forming hereby a sequence $U_1, \ldots, U_N$. Usually, the U_i's are taken to be independent and identically distributed random variables, independent from the claim frequency N. The total cost over one year amounts to

$$S = \sum_{i=1}^{N} U_i. \tag{5}$$

The random sum (5) corresponds to the liability of the insurer within what is called a collective model, see for instance the book of Klugman et al. [32]. Actuaries and risk managers typically want to quantify the risk of

large losses by a single comprehensible number, a risk measure. One popular risk measure is the Value-at-Risk which corresponds to a high order level quantile of the distribution of S. It is used by risk managers in banks, insurance companies, and other financial institutions to allocate risk reserves and to determine risk margins. The right tail of U has a strong influence on the higher order quantile of S and must be modelled appropriately, using for instance a composite model. One way to mitigate the risk associated with large claims is to transfer some of it to a reinsurance company. The aggregate claim sizes is then divided into

$$S = D + R,$$

where D is the amounts that stays with the first-line insurer after reinsurance and R is the amount paid by the reinsurer. The first-line insurer receives premiums, that sum up to Π, from the policyholders and part of it is ceded to the reinsuer. We have

$$\Pi = \Pi_R + \Pi_D,$$

where Π_R is the reinsurance premium ceded to the reinsurer while Π_D is the premium retained by the first-line insurer. A common reinsurance agreement in property and casualty insurance is the Excess-of-Loss (XOL) one with

$$R = \sum_{i=1}^{N} \min\{(U_i - P)_+, L\}, \text{ and } D = \sum_{i=1}^{N} \Big[\min\{U_i, P\}\mathbb{I}_{\{U_i \leq P+L\}} + (U_i - L)\mathbb{I}_{\{U_i > P+L\}}\Big], \tag{6}$$

where $(\cdot)_+$ denotes the positive part, P is the priority and L is the limit. The first-line insurer seeks values of P and L to optimize its expected surplus under some solvency constraints. We focus in this work on the left quantile of the distribution of $\Pi_D - D$ to determine a risk reserve. For a comprehensive overview on the statistical and actuarial aspect of reinsurance, the reader is referred to the book of Albrecher et al. [2]. The claim sizes are assumed to be IID from a composite model. The claim frequency is Poisson distributed with mean $\lambda > 0$. The premiums are defined by the expectation principle with a safety loading of 5% and computed using numerical integration[2]. Once the parameters of the composite model have been inferred then the quantiles of the insurer's surplus are calculated via crude Monte Carlo simulations. Through this application, we see the importance of precisely modeling the entire distribution of losses, which is the point of composite models. They are essential for an accurate assessment of insurance and reinsurance premiums but also of the probability of occurrence of serious claims. Let us move on to the Bayesian estimation procedure to fit and compare the composite models.

[2] `scipy.integrate` method from the `scipy` python library https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.quad.html

3 Bayesian inference via a sequential Monte Carlo sampler

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

$$\pi(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\theta)\pi(\theta)}{Z(\boldsymbol{x})}. \tag{7}$$

The posterior distribution (7) 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 (7) which is a normalizing constant given by

$$Z(\boldsymbol{x}) = \int_{\Theta} L(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta. \tag{8}$$

The above integral rarely admits a closed-form expression except when the model for the data has a conjugate prior distribution. Unfortunately, conjugate prior distributions almost only arise in exponential families of probability distributions, see Diaconis and Ylvisaker [15]. 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

$$\alpha(\theta^i, \theta^*) = \max\left[1, \frac{L(\boldsymbol{x}|\theta^*)\pi(\theta^*)K_H(\theta^*|\theta^i)}{L(\boldsymbol{x}|\theta^1)\pi(\theta^i)K_H(\theta^i|\theta^*)}\right], \tag{9}$$

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

$$K_H(\cdot|\theta) \sim \mathsf{Norm}(\mu = \theta, \Sigma = H), \tag{10}$$

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 workaround is to turn to another well-known MCMC technique that generates 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

$$\theta_1^i \sim \pi(\cdot|\boldsymbol{x}, \theta_2^i, \ldots, \theta_d^i), \tag{11}$$

before moving to the second one

$$\theta_2^i \sim \pi(\cdot|\boldsymbol{x}, \theta_1^i, \theta_3^i, \ldots, \theta_d^i), \tag{12}$$

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 (11), (12), 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 (10) must be tuned to ensure good sampling properties. It should reflect the variance of the posterior distribution which is unknown. The practical solution is to implement 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. [41]. Third, the trajectory generation cannot be parallelized. Lastly, MCMC algorithms allow 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 (7). Indeed, the latter does not appear in the acceptance probability expression in (9). 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

$$\pi(m_j|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|m_j)\pi(m_j)}{\sum_{i=1}^J L(\boldsymbol{x}|m_i)\pi(m_i)}, \; j = 1, \ldots, J. \tag{13}$$

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 (8). The best model achieves the highest model evidence (13). Another use of the posterior probabilities is the weighting of the estimation of the different models. Denote by Δ a quantity. It is possible to combine the estimation $\widehat{\Delta}_j$ of each model m_j as

$$\widehat{\Delta} = \sum_{j=1}^J \widehat{\Delta}_j \pi(m_j|\boldsymbol{x}).$$

This ensemble estimation procedure, known as Bayesian Model Averaging, is detailed in the work of Hoeting et al. [29] and tested out in Section 4.

This 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. Section 3.1 provides a quick reminder of the importance sampling principle required to understand the SMC algorithm detailed in Section 3.2.

3.1 Importance sampling

Bayesian inference reduces to evaluating quantities such as

$$\mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\varphi) = \int_\Theta \varphi(\theta)\pi(\theta|x)\mathrm{d}\theta, \tag{14}$$

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 (14) is evaluated through its Monte Carlo approximation

$$\mathbb{E}_{\pi(\theta|\boldsymbol{x})}(\varphi) \approx \frac{1}{K}\sum_{i=1}^{K}\varphi(\theta_i), \tag{15}$$

where $\theta_1,\ldots,\theta_K$ is an IID sample distributed as $\pi(\theta|\boldsymbol{x})$. Importance sampling samples from a distribution g, instead of $\pi(\theta|\boldsymbol{x})$, either because it is more convenient or because that reduces the variance associated to the Monte Carlo estimator (15). 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

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

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

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

whose empirical distribution targets the posterior distribution in the sense that

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

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

$$\text{ESS} \approx \frac{1}{\sum_{i=1}^{K} W_i^2},$$

as suggested in Kong et al. [33].

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 $\pi_t(\theta|\boldsymbol{x}) = \pi(\theta|\boldsymbol{x})$. To build this sequence of distribution sequence $\pi_s(\theta|\boldsymbol{x})$, $s = 0,\ldots,t$, we consider simulated annealing, see Neal [37]. This technique gradually activates the likelihood function as

$$\pi_s(\theta|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\theta)^{\tau_s}\pi(\theta)}{Z_s},\ s = 0,\ldots,t, \tag{16}$$

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

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

The SMC algorithm initializes a cloud of particles using the prior distribution as

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

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

1. (Reweighting step) This step prepares the current cloud to target the next distribution. A particle θ_s^i is reweighted by
$$W_i^{s+1} \propto w_i^{s+1} = \frac{\pi_{s+1}(\theta_i^s)}{\pi_s(\theta_i^s)},\ \text{for } i = 1,\dots,K,$$
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 $W_1^{s+1},\dots,W_K^{s+1}$ are actually importance weights, the targeted distribution π_{s+1} is chosen so that the weights satisfy
$$\text{ESS} \approx \frac{1}{\sum_{i=1}^K (W_i^{s+1})^2} \geq \rho N,$$
where $\rho \in (0,1)$. The selection of the next target reduces to picking a suitable 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. [30].

2. (Resampling step) Particles $\widetilde{\theta}_1^s,\dots,\widetilde{\theta}_K^s$ are sampled from the particle clouds $\{(W_i^{s+1}, \theta_i^s),\ i = 1,\dots,K\}$. A simple multinomial resampling is used here, but note that alternative schemes discussed for instance in the work of Gerber et al. [19] are also possible.

3. (Move step) Metropolis-Hasting within Gibbs moves are applied to the particles $\widetilde{\theta}_1^s,\dots,\widetilde{\theta}_K^s$ to yield the new generation of particles $\theta_1^{s+1},\dots,\theta_K^{s+1}$. 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 the particles system $\{(W_i^{s+1}, \theta_i^s),\ i = 1,\dots,K\}$. 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 pilot run 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 $\theta_1^{s+1},\dots,\theta_K^{s+1}$ are sampled from π_{s+1} and are equally weighted with $W_i^{s+1} = 1/K$ for $i = 1,\dots,K$.

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. [46] and the SMC sampler of the Python package `pymc` of Salvatier et al. [42]. The Metropolis-Hasting move step, also used in Nguyen et al. [38], is easy to paralellize to optimize the computing time.

A summary of the algorithm can be found in Appendix A, see Algorithm 1.

The unnormalized weights $\{w_i^s, 1 \leq i \leq K, 1 \leq s \leq t\}$ yield an approximation of the normalizing constant as

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

The accuracy of the estimator depends on the population size (the higher the better) which is chosen by the user according to a computing time budget. The `bayes-splicing` package implements the sequential Monte Carlo sampler to fit and compare the composite models introduced in Section 2.1. In addition to the posterior probability, it is possible to compute two Bayesian information criteria. The Deviance Information Criterion (DIC), introduced in the work of Spiegelhalter et al. [47] and the Widely applicable Information Criterion (WAIC) of Watanabe [50]. For a comprehensive discussion about the information criteria used in Bayesian statistics, I refer the reader to the work of Gelman et al. [17]. Although this work focuses on the posterior model probability to compare the competing theories, note that the `bayes-splicing` includes python methods to compute the DIC and WAIC.

4 Simulation study

The experiment involves generating claim data from a known composite model and fitting it using multiple composite models. The PDF of the composite models in competition are continuous at the threshold, and combine the Weibull or the Inverse-Weibull distribution for the body to the Inverse-Weibull, Lomax or Log-Logistic distribution for the tail. This makes six models to choose from. Note that the choice of the data-generating model and competing models is arbitrary. We limit ourselves to 6 models so as not to unreasonably increase the calculation times and for the sake of readability of the graphs on which the results are presented. The sample size of the artificial data ranges from 500 to 5,000. For each of the 1,000 simulation runs, the posterior probability of each model is computed. A first objective is to assess the finite sample consistency of the model selection procedure. Once the composite models have been calibrated, an estimation of the quantiles of the first-line insurer surplus distribution, defined in Section 2.2, is produced and compared to the true value. For the latter application, The average number of claims is $\lambda = 250$. The XOL priority and limit are set to be the 90% and 99% quantiles of the splicing distribution that generated the data. The quantile estimates resulting from the best model (according to the posterior probabilities) and from the Bayesian model averaging procedure are also compared to the true value. Two cases are considered. In Section 4.1, the model is well-specified, meaning that the data generating model belongs to the set of concurrent models. In Section 4.2, the model is misspecified, which means that the data generating model does not belongs to the set

of concurrent models.

4.1 When the model is well-specified

The losses are drawn from a Weibull($k = 1/2, \beta = 1$) – Lomax($\alpha = 2.5, \sigma = 1.5$) continuous composite model with a threshold $\gamma = 1.5$. The prior assumptions of the competing composite models are given in Appendix B, see Table 3. The prior assumption over the threshold parameter is given by $\gamma \sim \text{Gamma}(1, 1)$. The number of particles is $N = 8,000$. The posterior model probabilities of the competing composite models computed for each of the $1,000$ simulation runs for samples of sizes $500, 1000, 2000$, and 5000 are shown in Figure 1. The posterior probabilities discriminate the composite models having the Inverse-Weibull distribution as

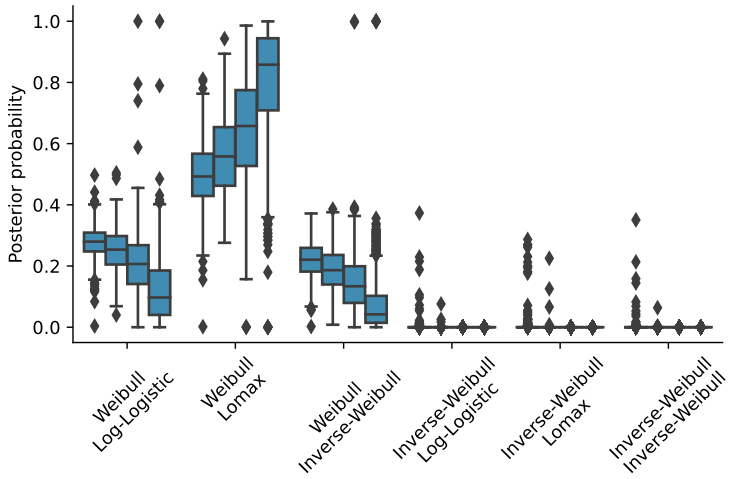
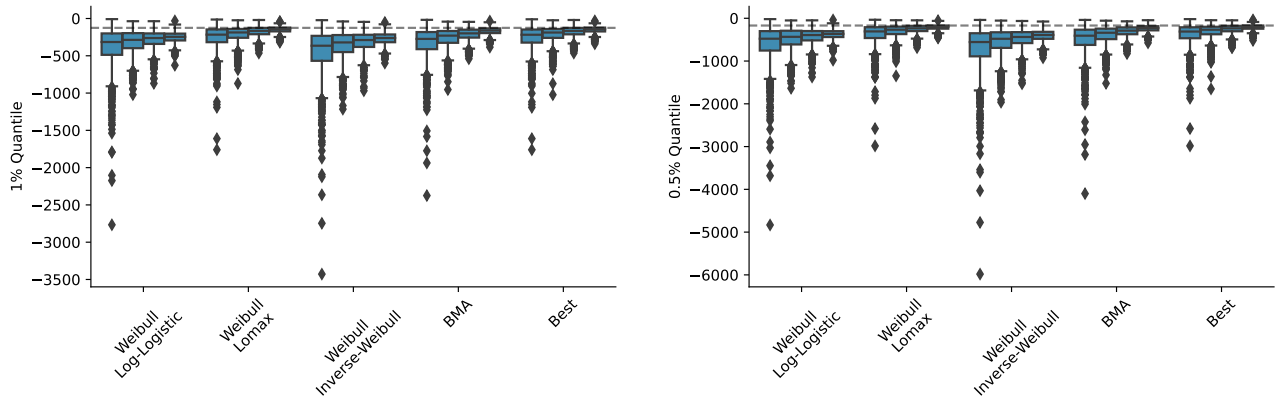


Figure 1: Posterior probabilities assigned to each composite model for sample of sizes $500, 1000, 2000$, and 5000 drawn from a continuous Weibull – Lomax composite model.

bulk distribution. The posterior probability of the Weibull-Lomax model converges toward one as the sample size increases, just as expected. Figure 2 displays the boxplots of the 1% and 0.5% quantiles of the first line insurer surplus $\Pi_D - D$ depending on the composite model used. The "Best" model corresponds, for each simulation run, to the model associated to the highest model probability. The "BMA" model provides the quantile estimate resulting from the weighted average of all the models where the weights are defined by the posterior probabilities. The Weibull-Lomax model provides the most accurate estimation of the quantiles followed closely by the "Best" and "BMA" models which is consistent with the posterior probabilities reported on Figure 1. Note that the composite models with the Inverse-Weibull distribution as the body have been removed because the estimate of the quantiles is too far from the true value to assess the accuracy of the other



(a) 1% quantile

(b) 0.5% quantile

Figure 2: Left quantiles of the profit and loss distribution of the first-line insurer depending on the composite model used. The loss data consists of sample of sizes 500, 1000, 2000, and 5000 drawn from a continuous Weibull – Lomax composite model. The dashed line represents the true value.

models.

4.2 When the model is misspecified

The losses are drawn from a $\text{Exp}(\delta = 1/2) - \text{Burr}(\alpha = 1.8, \beta = 2, \sigma = 3)$ continuous composite model with a threshold $\gamma = 2.5$. The prior assumptions of the competing composite models are given in Table 3. The number of particles is $N = 8,000$. The posterior model probabilities of the competing composite models computed for each of the 1,000 simulation runs for samples of sizes 500, 1000, 2000, and 5000 are shown in Figure 3. The highest posterior probabilities are reported for the composite models having the Weibull distribution as body, and more specifically for the composite models having the Log-logistic and Lomax distributions as tail. This tendency is reinforced when the sample size increases. This result was expected, since the exponential distribution is a special case of the Weibull distribution on the one hand and the Lomax and Log-logistic distributions are special cases of the Burr distribution on the other hand. Figure 4 displays the boxplots of the 1% and 0.5% quantiles of the first line insurer surplus $\Pi_D - D$ depending on the composite model used. When the model is misspecified then the "Best" and "BMA" models provide the most accurate estimation of the quantiles. Note that the BMA approach also seems to decrease the variance of the estimation. This validates the model selection procedure proposed in this work since model misspecification tends to be the rule in real applications.

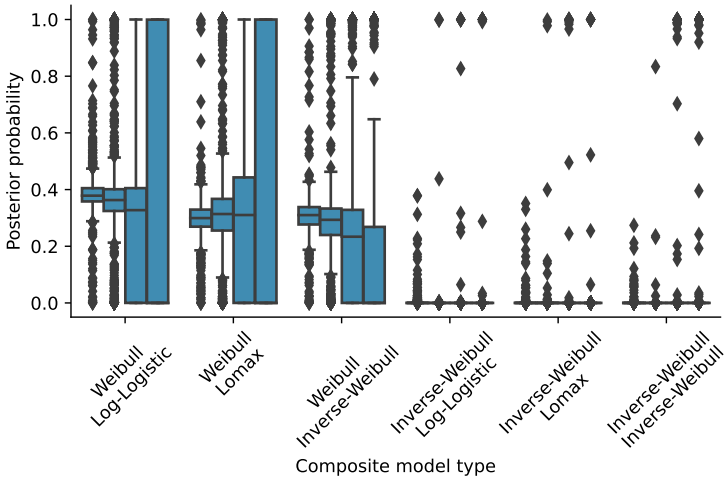
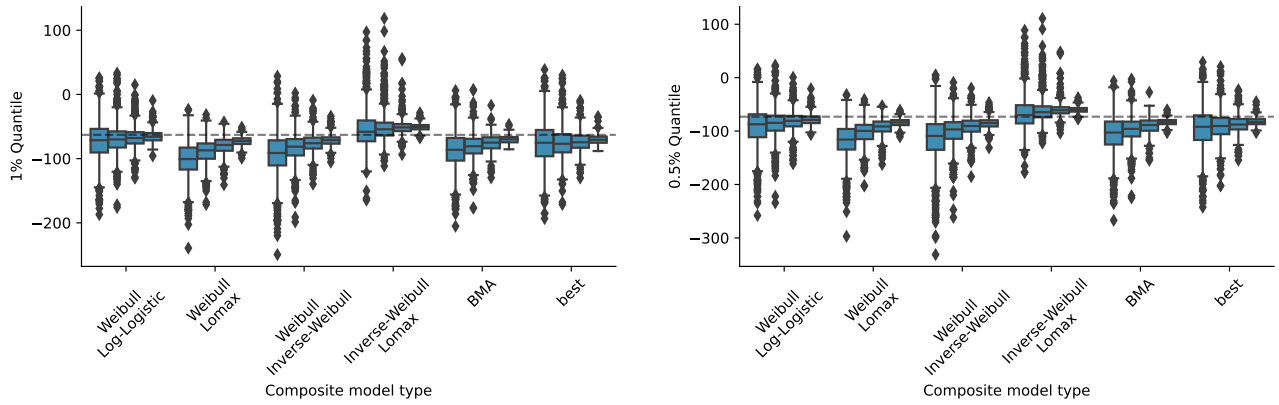


Figure 3: Posterior probabilities assigned to each composite model for sample of sizes 500, 1000, 2000, and 5000 drawn from a continuous Exp – Burr composite model.



(a) 1% quantile

(b) 0.5% quantile

Figure 4: Left quantiles of the profit and loss distribution of the first line insurer depending on the composite model used. The loss data consists of sample of sizes 500, 1000, 2000, and 5000 drawn from a continuous Exp – Burr composite model. The dashed line represents the true value.

5 Application to the Danish fire insurance data

The Danish fire insurance data are highly valued by actuarial researchers for comparing statistical methods dealing with extreme values. The data was retrieved from the R package `SMPracticals` of the book by Davison

[13]. The aim of this study is to find the most suitable composite model for these data among all the possible combinations of distribution for the body and the tail, the different types of composite model including "continuous", "discontinuous" and "simple", and the threshold selection method. The composite models are compared based on log marginal likelihood and a quantile-based distance, which we now present. Denote by $F(x,\theta)$ the CDF of the composite model and let $Q(y,\theta) = \inf\{x \in \mathbb{R} \,;\, F(x,\theta) \geq y\}$ be its generalized inverse function that we refer to as the quantile function. The distance

$$W(F,\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left|Q\left(i/n,\widehat{\theta}\right) - x_{i:n}\right|, \tag{17}$$

aims at comparing the quantiles estimated via the composite model $Q\left(i/n,\widehat{\theta}\right)$, where $\widehat{\theta}$ denotes the mean *a posteriori*, to the empirical quantiles $x_{i:n}$ for $i = 1,\ldots,n$. The notation W comes from the fact that (17) corresponds to the empirical counterpart of a variation of the Wasserstein distance between the empirical measure and the probability measure with density $f(x,\theta)$ with respect to the Lebesgue measure. The distance (17) summarizes in a single number the information contained in a Q-Q plot. Information criteria based on the likelihood function, such as log marginal likelihood, often tend to focus on the part of the distribution that concentrates many data points. A distance based on quantiles, like (17), is more sensitive to extreme values and therefore puts more emphasis on a potential mismatch in the tail of the distribution. In Section 5.1, the threshold parameter is treated like any of the other model parameters. In addition to choosing the most suitable composite model, we want to highlight the impact of the selected composite model on the value of the extreme value threshold within this free-threshold approach. Section 5.2 completes the results of Section 5.1 by considering the case where the value of the extreme value threshold is fixed before estimating the other parameters of the composite model.

5.1 Simultaneous estimation of the extreme value threshold

In this section, the threshold parameter of the composite model is estimated together with the other parameters. The prior assumptions over the threshold and mixing parameters (for the discontinuous composite models) read as follows

$$\gamma \sim \text{Uniform}([\min(\mathbf{x}), \max(\mathbf{x})]), \text{ and } p \sim \text{Uniform}([0,1]).$$

The number of particles is $N = 10,000$. Figure 5 gives the boxplots of the log marginal likelihood and the quantile distance depending on the type of composite model considered. The discontinuity at the threshold improves the overall fit of the model to the data when examining the marginal likelihood values. The difference in terms of matching the empirical quantiles of the data, however, is less obvious. Figure 6 shows the posterior probabilities, denoted by w which stands for weight, of the various composite models grouped by type. The

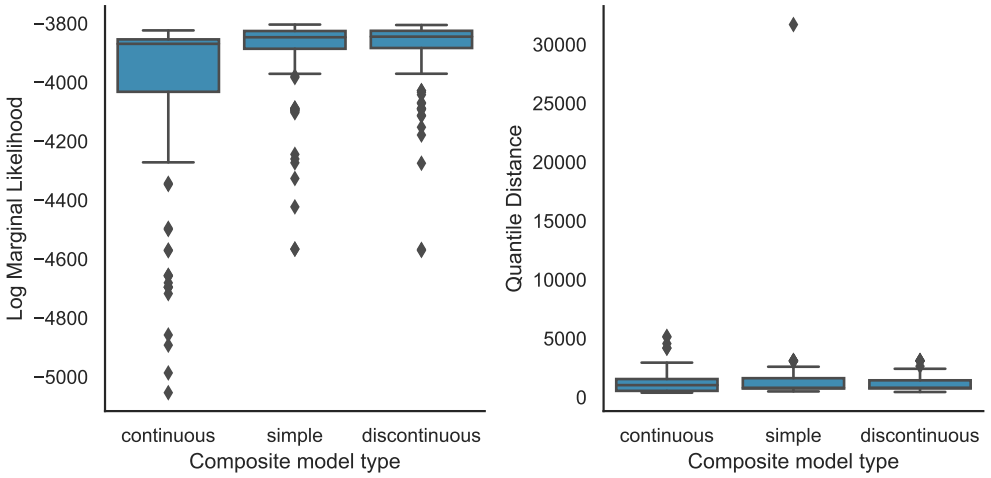


Figure 5: Log marginal likelihood and quantile distance of the fitted composite models depending on their type.

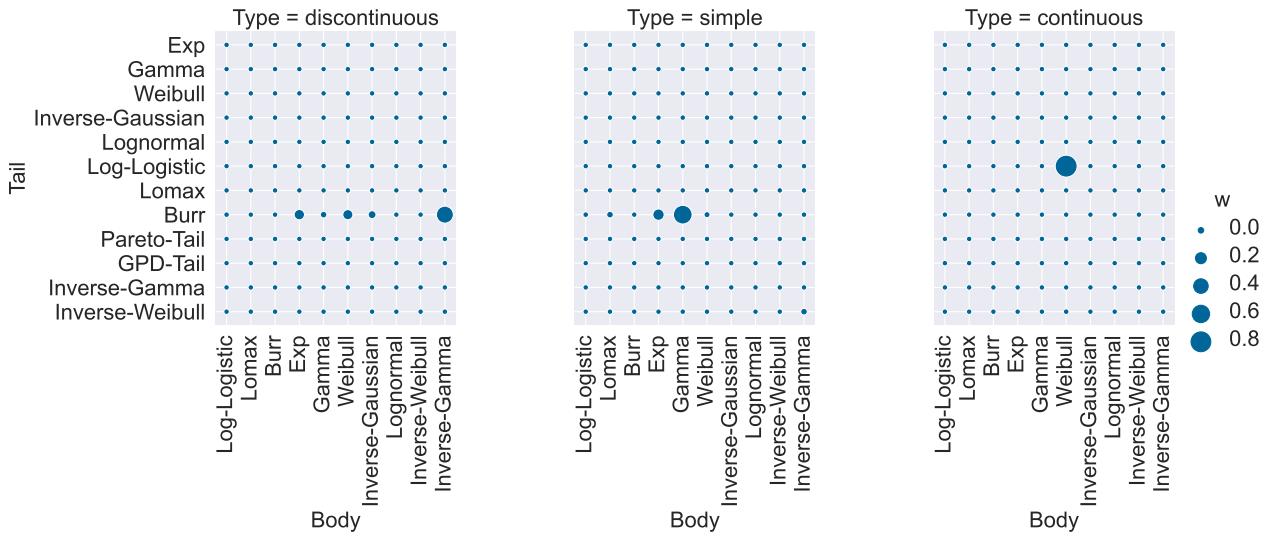
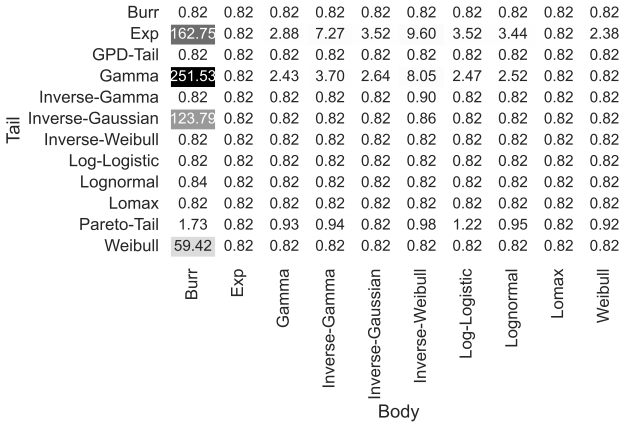
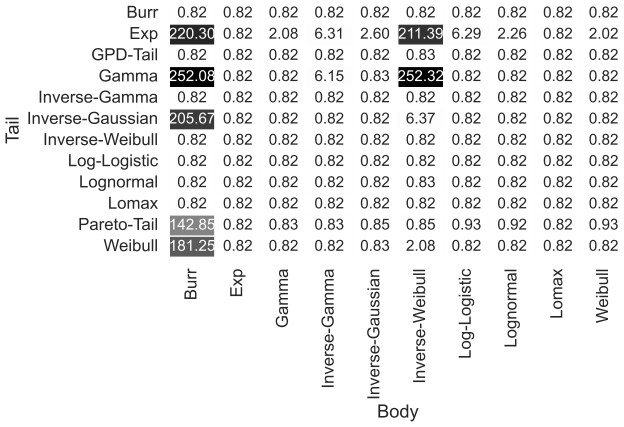


Figure 6: Posterior probabilities of the composite models depending on their type with the free threshold approach.

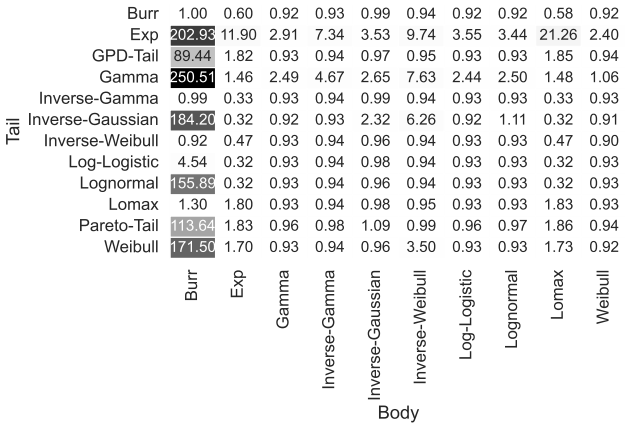
discontinous and simple composite models favor the Burr distribution for the tail combined to either the exponential, gamma, inverse-Gaussian or Weibull distribution for the distribution body. The best continuous composite model is the Weibull-Log logistic one. Let us inspect the extreme value threshold resulting from the simultaneous estimation of the composite model parameters shown in Figure 7. The majority of the



(a) Discontinuous



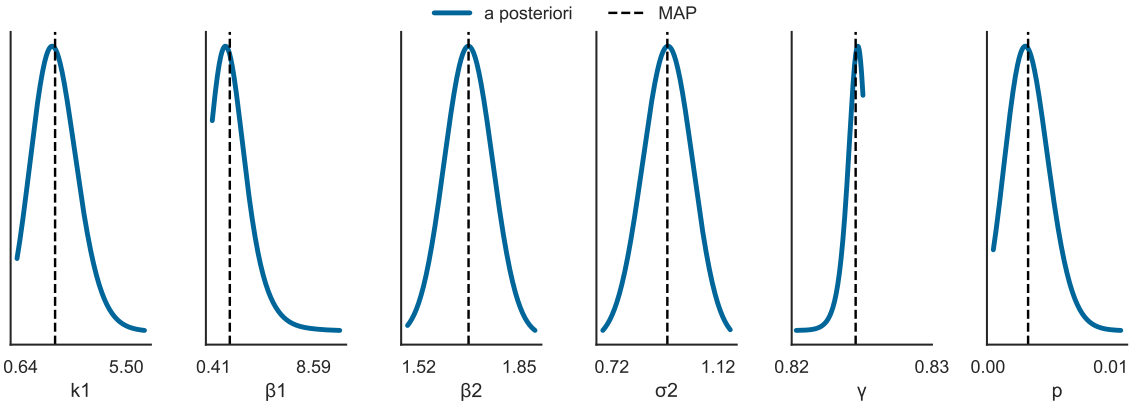
(b) Simple



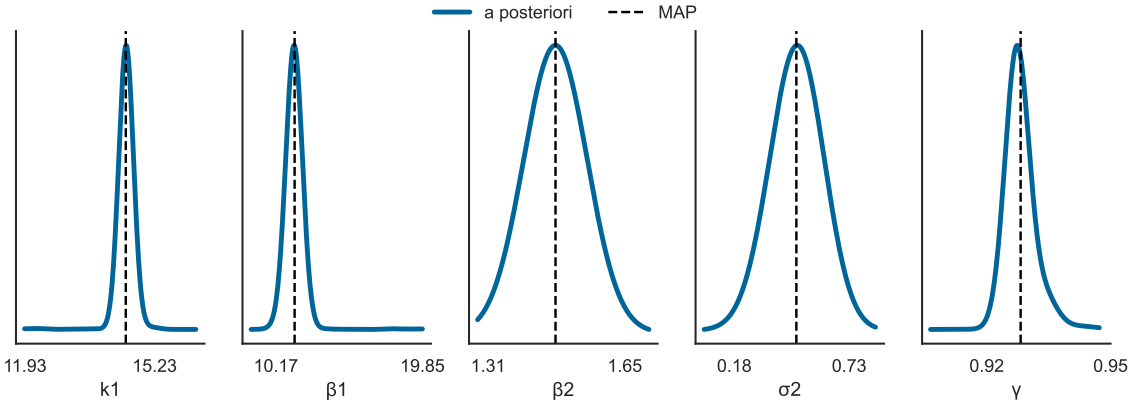
(c) Continuous

Figure 7: Posterior mean of the threshold resulting from the simultaneous estimation of the composite model parameters.

"discontinuous" and "simple" composite models agree on a threshold of 0.82 as if a discontinuity were indeed present in the data. For the "continuous" composite models, the results are more contrasted even if the threshold takes on fairly low values, oscillating between 0.9 and 1. Let us take as an example the Weibull-Log logistic composite model. Figure 8 shows the posterior distributions of the parameters in the "continuous" and "discontinuous" setting. The shape of the posterior distribution of the threshold in the "discontinuous" case indicates a steep downward slope in the likelihood function at the threshold revealing a strong evidence of a discontinuity in the data, see Figure 8a. The shape of the posterior distribution of the γ parameter in the "continuous" case is much more standard, see Figure 8b. Of the single loss models, the Burr distribution is



(a) Discontinuous

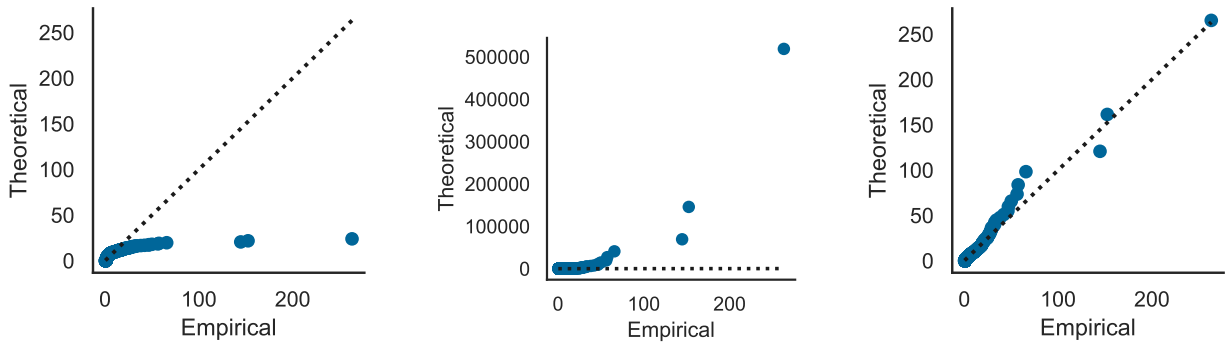


(b) Continuous

Figure 8: Posterior plots of the composite model parameters.

the one that leads to the best fit to the data. If we define the Burr distribution as the body distribution of the composite model, then the threshold takes on high values. The algorithm wants to model the whole data using only the Burr model. When the composite model's body distribution is the exponential distribution, then it's the other way around. The algorithm sets the threshold as low as possible. The conclusion is that it is difficult to base the definition of the extreme value threshold on the simultaneous estimation of the parameters of a composite model. If the composite model is discontinuous, the risk is to capture a discontinuity which could be a simple artefact of the empirical distribution of the data. If the composite model is continuous then the value of the threshold strongly depends on the choice made for the body and the tail of the composite model. Overall, many threshold values reported in Figure 4 are too low because 95% to 99% of the claim amounts are considered extreme. In my humble opinion, one possibility is to define a balanced composite model with a few

parameters, for example, the exponential distribution as the body and the Pareto distribution as the tail. The resulting threshold is 1.83, which is reasonable since 60% of claims are below this threshold. This recalls the procedure of detecting a break in the Q-Q plot of the exponential (see Figure 9a) and Pareto distribution (see Figure 9b). The resulting fit seems fair, especially in the tail, see Figure 9c.



(a) Exp(δ) (b) Pareto(α) (c) Continuous Exp(δ) – Pareto(α)

Figure 9: Q-Q plots of the Exp model, Pareto model, and the Exp-Pareto continuous composite model fitted to the danish fire loss data using the posterior mean.

5.2 Separate estimation of the extreme value threshold

Many methods for selecting the extreme value threshold have been proposed in the statistical literature and they lead to very different threshold values as shown in Table 2. The value reported in Table 2 were obtained using the `tea`[3] R package. The purpose of this work is not to discuss the merits of each of these methods but rather to illustrate the ability of the `bayes-splicing` package to take a fixed threshold value before inferring the rest of the parameters of the composite model. We therefore use an automatic threshold selection method presented in a recent work of Bladt et al. [7]. It consists in minimizing the asymptotic mean squared error of a trimmed version of the Hill estimator of the tail index parameter of the Pareto distribution. The procedure leads to a threshold of 2.78. Since the threshold selection method assumes that the tail is of Pareto type, only the composite models having a Pareto or generalized Pareto tail are considered. The number of particles is $N = 10,000$. Figure 10 shows the posterior probabilities, denoted by w which stands for weight, of the various composite models grouped by type.

The highest posterior probability is obtained by the composite Burr-Pareto model and this for all types

[3] `tea` package documentation

Threshold value	Method	Bibliography entry
2.46	Minimizing the AMSE of the Hill estimator	Caeiro and Gomes [9]
14.39	A Bias-based procedure	Drees and Kaufmann [16]
25.29	Eyeballing and Hill plot	Danielsson et al. [12]
4.61	Eyeballing and Gerstengarbe plot	Gerstengarbe and Werner [20]
12.06	Exponential goodness-of-fit test	Guillou and Hall [25]
1.43	Double bootsrap	Gomes et al. [23]
4.09	Single bootstrap	Hall [26]
1.50	Minimizing the AMSE of the Hill estimator	Hall and Welsh [27]
2.78	Minimizing the AMSE of a trimmed Hill estimator	Bladt et al. [7]

Table 2: Threshold values depending on the extreme value technique used.

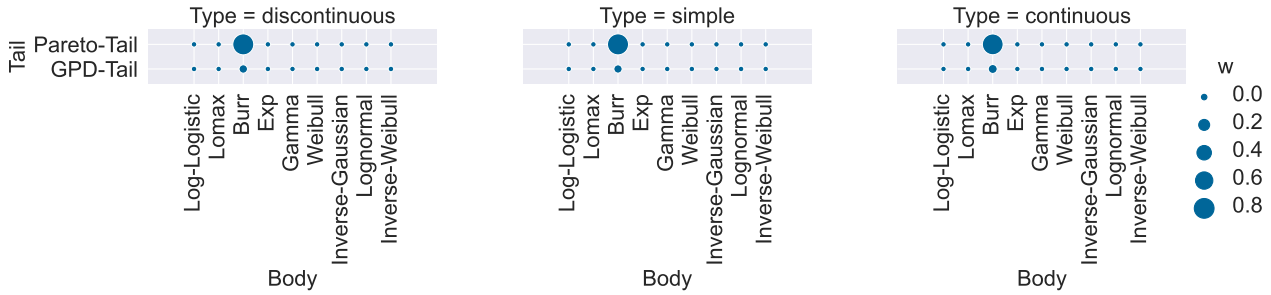


Figure 10: Posterior probabilities of the composite models depending on their type with the fixed threshold approach.

of composite models. This can be explained by the flexibility of the Burr distribution thanks to its three parameters. This result justifies the use of a nonparametric assumption on the distribution of the body of the composite model when the threshold is chosen before the estimation of the other parameters. This makes it possible in particular to overcome the uncertainty surrounding the choice of the distribution of the body of the composite model. We further compare the goodness-of-fit of the composite models with fixed threshold to that of the composite models for which the threshold was set by the algorithm. Figure 11 shows the boxplot of the log marginal likelihood and the quantile distance for the composite models having a Pareto or a generalized Pareto tail. The examination of the marginal likelihood indicates that it is preferable to leave the threshold free while fixing the threshold beforehand improves the distance to the empirical quantiles. This last remark does not seem to hold when the composite model is continuous at the threshold. We conclude that for a "continuous" composite model, performing simultaneous estimation is the way to go because the overall fit is better and the

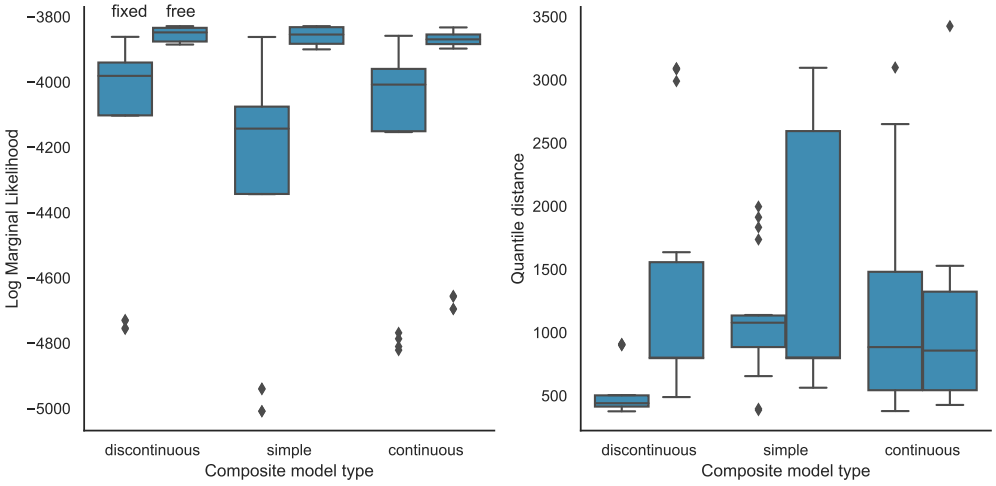


Figure 11: Log marginal likelihood and quantile distance of the composite models according to their type, and if the threshold was estimated before or at the same time as the other parameters.

quantile matching is as good as with a prefixed threshold. If the threshold is fixed beforehand, it is preferable to opt for a "discontinuous" composite model because the loss of the degree of freedom associated with the mixing parameter p seriously affects the fit to the data.

6 Conclusions and perspectives

This paper presents an implementation of a SMC sampler to fit and compare composite models within a Bayesian framework. The python package can be installed from pip, see bayes-splicing and the notebooks to reproduce the results can be freely downloaded from the following `github` repository SMCCompoMo. 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 advantage of using the posterior model probabilities to weight the estimates of the quantile of the insurer's surplus returned by the various models has been demonstrated, in particular in the case where the model generating the data does not belong to the set of concurrent models. The analysis of the Danish fire insurance data highlighted the

advantages and disadvantages of taking a discontinuous or continuous composite model, and also of letting the algorithm select a threshold or fix it beforehand.

There are many avenues for future research. The Monte Carlo sequential sampler could be improved by considering a more sophisticated particle perturbation scheme than the random walk one, similar to what is proposed by South et al. [46]. It is also possible to combine the estimates of the different models using weights based on cross-validation procedures to improve the predictive power of the Bayesian averaging procedure, see the work of Yao et al. [51]. When the threshold is set before the other parameters, then a flexible model for the body of the composite model and a discontinuity at the threshold can be optimal. Along these lines, a nonparametric assumption could be made over the body distribution. Tancredi et al. [48] used a mixture of uniform distributions, MacDonald et al. [35] used a kernel density estimator, Cabras and Castellanos [8] used an orthogonal polynomial expansion, and Reynkens et al. [40] used a mixture of Erlang distributions. The question is which of these methods best suits the Bayesian framework proposed here. In view of the importance given to the tail of the loss distribution, an inference method that does not rely on the likelihood function could be considered. Minimum distance estimators that minimize a measure of the deviation between the empirical quantiles and those of the model could be a valid direction. The work of Bernton et al. [5] 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. [6] and applied to aggregated insurance data in the work of Goffard and Laub [22].

Acknowledgmements

The author thanks the associate editor and the two anonymous reviewers for their insightful comments which greatly improved the manuscript. The author's work is conducted within the Research Chair DIALog under the aegis of the Risk Foundation, an initiative by 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. Albrecher, J. Bierlant, and J. Teugels. *Reinsurance*. John Wiley & Sons Inc, November 2017. ISBN

0470772689. URL https://www.ebook.de/de/product/24938207/hansjoerg_albrecher_jan_bierlant_jozef_teugels_reinsurance.html.

[3] C. N. Behrens, H. F. Lopes, and D. Gamerman. Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244, oct 2004. doi: 10.1191/1471082x04st075oa.

[4] 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.

[5] 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.

[6] 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.

[7] M. Bladt, H. Albrecher, and J. Beirlant. Threshold selection and trimming in extremes. *Extremes*, 23(4):629–665, jul 2020. doi: 10.1007/s10687-020-00385-0.

[8] S. Cabras and M. E. Castellanos. A bayesian approach for estimating extreme quantiles under a semiparametric mixture model. *ASTIN Bulletin*, 41(1):87–106, 2011. doi: 10.2143/AST.41.1.2084387.

[9] F. Caeiro and M. I. Gomes. *Extreme value modeling and risk analysis : methods and applications*, chapter Threshold selection in extreme value analysis, pages 69–82. Chapman-Hall/CRC, Boca Raton, FL, 2015.

[10] 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.

[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] J. Danielsson, L. M. Ergun, L. de Haan, and C. G. de Vries. Tail index estimation: Quantile driven threshold selection. *Available at SSRN 2717478*, 2016.

[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] B. Vaz de Melo Mendes and H. F. Lopes. Data driven estimates for mixtures. *Computational Statistics & Data Analysis*, 47(3):583–598, oct 2004. doi: 10.1016/j.csda.2003.12.006.

[15] 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.

[16] H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172, jul 1998. doi: 10.1016/s0304-4149(98)00017-9.

[17] 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.

[18] 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/.

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

[20] 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.

[21] S. Ghosh and S. Resnick. A discussion on mean excess plots. *Stochastic Processes and their Applications*, 120(8):1492–1517, 2010. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2010.04.002. URL https://www.sciencedirect.com/science/article/pii/S0304414910001079.

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

[23] M. I. Gomes, F. Figueiredo, and M. M. Neves. Adaptive estimation of heavy right tails: resampling-based methods in action. *Extremes*, 15(4):463–489, dec 2011. doi: 10.1007/s10687-011-0146-6.

[24] 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.

[25] A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):293–305, may 2001. doi: 10.1111/1467-9868.00286.

[26] P. Hall. Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, 32(2):177–203, feb 1990. doi: 10.1016/0047-259x(90)90080-2.

[27] P. Hall and A. H. Welsh. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13(1), mar 1985. doi: 10.1214/aos/1176346596.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss Models: From Data to Decisions*. WILEY, January 2019. ISBN 1119523788. URL https://www.ebook.de/de/product/32978767/stuart_a_klugman_harry_h_panjer_gordon_e_willmot_loss_models_from_data_to_decisions.html.

[33] 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.

[34] 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.

[35] A. MacDonald, C.J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell. A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157, jun 2011. doi: 10.1016/j.csda.2011.01.005.

[36] 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.

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

[38] Thi Le Thu Nguyen, François Septier, Gareth W. Peters, and Yves Delignon. Efficient sequential monte-carlo samplers for bayesian inference. *IEEE Transactions on Signal Processing*, 64(5):1305–1319, 2016. doi: 10.1109/TSP.2015.2504342.

[39] 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.

[40] T. Reynkens, R. Verbelen, J. Beirlant, and K. Antonio. Modelling censored losses using splicing: A global fit strategy with mixed erlang and extreme value distributions. *Insurance: Mathematics and Economics*, 77:65–77, nov 2017. doi: 10.1016/j.insmatheco.2017.08.005.

[41] 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.

[42] 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.

[43] C. Scarrott and A. MacDonald. A review of extreme value threshold es-timation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60, 2012.

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

[45] 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.

[46] 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.

[47] 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.

[48] A. Tancredi, C. Anderson, and A. O'Hagan. Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9(2):87–106, aug 2006. doi: 10.1007/s10687-006-0009-8.

[49] 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.

[50] 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.

[51] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3), sep 2018. doi: 10.1214/17-ba1091.

A Summary of the Sequential Monte Carlo algorithm

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 $i = 1 \rightarrow K$ **do**
    $\theta_i^0 \sim \pi(\theta)$; $W_i^0 \leftarrow 1/K$
end for
while $\pi_s(\theta) \neq \pi(\theta|\boldsymbol{x})$ **do**
    Search for π_{s+1} such that

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

    Compute $\widehat{\Sigma} = \mathrm{Cov}\left(\{(W_i^{s+1}, \theta_i^s),\ i = 1,\ldots,K\}\right)$
    for $i = 1 \rightarrow K$ **do**
        Sample $\widetilde{\theta}_i \sim \{\theta_1^{(s)},\ldots,\theta_K^{(s)}\}$ with probabilities W_j^{s+1}, for $1 \leq j \leq K$
    end for
    for $i = 1 \rightarrow K$ **do**
        $\widetilde{\theta}_i^* \leftarrow K_H(\widetilde{\theta}_i, \cdot)$ where $K_H(\widetilde{\theta}_i, \cdot)$ where $H = \frac{2.38}{\sqrt{d}} \cdot \widehat{\Sigma}$
    end for
    Compute $p_a = N^{-1} \sum_{i=1}^{K} \mathbb{I}_{\widetilde{\theta}_i^* = \widetilde{\theta}_i}$; $k = \max\left\{k_{\max}, \min\left[k_{\min}, \frac{\log(1-c)}{\log(1-p_a)}\right]\right\}$
    for $i = 1 \rightarrow K$ **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
        $W_i^{s+1} \leftarrow 1/K$
    end for
end while
Return $(W_1^t, \theta_1^t),\ldots,(W_K^t, \theta_N^t)$

B A priori assumptions

Table 3 provides an overview of the prior assumtions used to do the Bayesian fit of the composite models in Sections 4 and 5. Note that the `bayes-splicing` allows the user to choose different hyper-parametrization than what is specified in Table 3. Note also that beta and uniform distribution can be used as in addition to the

Name	Parameters		prior
Exponential	Exp(λ)	$\lambda > 0$	$\lambda \sim$ Gamma$(1,1)$
Gamma	Gamma(r,m)	$r,m > 0$	$r,m \sim$ Gamma$(1,1)$
Weibull	Weibull(k,β)	$k,\beta > 0$	$k,\beta \sim$ Gamma$(1,1)$
Lognormal	Lognormal(μ,σ)	$\mu \in \mathbb{R},\ \sigma > 0$	$\mu \sim$ Normal$(0,0.5)$ and $\sigma \sim$ Gamma$(1,1)$
Inverse-Gaussian	Inverse-Gaussian(μ,λ)	$\mu,\lambda > 0$	$\mu,\lambda \sim$ Gamma$(1,1)$
Inverse-Gamma	Inverse-Gamma(r,m)	$r,m > 0$	$r,m \sim$ Gamma$(1,1)$
Inverse-Weibull	Inverse-Weibull(k,β)	$k,\beta > 0$	$k,\beta \sim$ Gamma$(1,1)$
Lomax	Lomax(α,σ)	$\alpha,\sigma > 0$	$\alpha,\sigma \sim$ Gamma$(1,1)$
Log-Logistic	Log-Logistic(β,σ)	$\beta,\sigma > 0$	$\beta,\sigma \sim$ Gamma$(1,1)$
Burr	Burr(α,β,σ)	$\alpha,\beta,\sigma > 0$	$\alpha,\beta,\sigma \sim$ Gamma$(1,1)$
Pareto	Pareto(α,γ)	$\alpha,\gamma > 0$	$\alpha \sim$ Gamma$(1,1)$
Generalized Pareto	GPD(ξ,σ,γ)	$\xi,\sigma,\gamma > 0$	$\xi,\sigma \sim$ Gamma$(1,1)$

Table 3: Prior assumptions over the parameters of distribution for the bulk and tail of the composite models.

gamma and normal distributions. The γ parameter in the Pareto and Generalized Pareto cases corresponds to a threshold. This parameter should be set to the minimum of the data points if the models are considered alone. The prior specification of the mixing parameter p and the threshold parameter γ of the composite models are given in the main text.

C Computing time analysis

This section gives an overview of the time required to fit composite models using the home-made SMC algorithm of the `bayes-splicing` package. Calculation times are provides for composite models of varying complexity and with changing regularity assumptions at the threshold. The impact of the size of the particle population is studied as well as whether the move step is parallelized or not. The models were fitted to the danish fire insurance dataset. The experiment has been conducted on a Dell Latitude 7290 equipped with an Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz. When the paralellized version of the algorithm is used, 4 cores are mobilized simultaneously. Table 4 shows the computation times associated with fitting discontinuous composite models using the fixed-threshold approach. It is observed that the computation time increases with the number of parameters and the size of the particle population. The parallelization of the move step improves the calculation times, its interest is all the more clear as the number of particles increases. Table 5 shows the computation times associated with fitting continuous composite models using the fixed-threshold approach.

Model Body	Tail	Not Paralell $K = 5,000$	Not Paralell $K = 10,000$	Not Paralell $K = 20,000$	Paralell $K = 5,000$	Paralell $K = 10,000$	Paralell $K = 20,000$
Exp	Pareto-Tail	23	45	77	22	33	57
	GPD-Tail	33	75	108	27	41	73
Weibull	Pareto-Tail	135	301	468	71	128	248
	GPD-Tail	157	337	632	97	161	321
Burr	Pareto-Tail	219	512	970	132	229	452
	GPD-Tail	264	638	1049	158	263	557

Table 4: Computing time required to fit discontinuous composite models within the fixed threshold setting depending on the composite model, the number of particles and whether the move step is paralellized.

Compared to Table 4, the calculation times are slightly lower since one less parameter is to be estimated. Table 6 and Table 7 shows the computation times associated with fitting discontinuous and continuous composite models using the free-threshold approach. Compared to Table 4 and Table 5, the calculation times are slightly longer since one more parameter has to be estimated. Otherwise, the same remarks apply for the impact of the number of parameters and the size of the particle population.

		Not Paralell			Paralell		
Model		$K = 5,000$	$K = 10,000$	$K = 20,000$	$K = 5,000$	$K = 10,000$	$K = 20,000$
Body	Tail						
Exp	Pareto-Tail	17	33	57	16	23	40
	GPD-Tail	26	55	72	21	31	54
Weibull	Pareto-Tail	86	198	326	57	94	179
	GPD-Tail	146	235	561	72	131	257
Burr	Pareto-Tail	178	418	712	106	186	373
	GPD-Tail	253	468	966	128	220	433

Table 5: Computing time required to fit continuous composite models within the fixed threshold setting depending on the composite model, the number of particles and whether the move step is paralellized.

		Not Paralell			Paralell		
Model		$K = 5,000$	$K = 10,000$	$K = 20,000$	$K = 5,000$	$K = 10,000$	$K = 20,000$
Body	Tail						
Exp	Pareto-Tail	41	81	164	38	58	103
	GPD-Tail	65	116	228	44	73	132
Weibull	Pareto-Tail	112	270	545	75	137	291
	GPD-Tail	125	244	483	72	144	258
Burr	Pareto-Tail	251	427	1085	113	243	472
	GPD-Tail	189	374	790	121	175	379

Table 6: Computing time required to fit discontinuous composite models within the free threshold setting depending on the composite model, the number of particles and whether the move step is paralellized.

		Not Paralell			Paralell		
Model		$K = 5,000$	$K = 10,000$	$K = 20,000$	$K = 5,000$	$K = 10,000$	$K = 20,000$
Body	Tail						
Exp	Pareto-Tail	27	48	95	28	40	66
	GPD-Tail	33	64	125	30	48	85
Weibull	Pareto-Tail	93	175	373	57	105	203
	GPD-Tail	139	280	457	78	140	306
Burr	Pareto-Tail	181	435	842	99	183	446
	GPD-Tail	248	495	839	105	227	394

Table 7: Computing time required to fit continuous composite models within the free threshold setting depending on the composite model, the number of particles and whether the move step is paralellized.