Mathematics and Computers in Simulation 230 (2025) 193–206

Contents lists available at ScienceDirect

Mathematics and Computers in Simulation

journal homepage: www.elsevier.com/locate/matcom

Original articles

A parsimonious dynamic mixture for heavy-tailed distributions[☆]

Marco Bee

Department of Economics and Management, University of Trento, Italy

ARTICLE INFO

Keywords:
Dynamic mixtures
Exponential distribution
Noisy cross-entropy
Weight function

ABSTRACT

Dynamic mixture distributions are convenient models for highly skewed and heavy-tailed data. However, estimation has proved to be challenging and computationally expensive. To address this issue, we develop a more parsimonious model, based on a one-parameter weight function given by the exponential cumulative distribution function. Parameter estimation is carried out via maximum likelihood, approximate maximum likelihood and noisy cross-entropy. Simulation experiments and real-data analyses suggest that approximate maximum likelihood is the best method in terms of RMSE, albeit at a high computational cost. With respect to the version of the dynamic mixture with weight equal to the two-parameter Cauchy cumulative distribution function, the reduced flexibility of the present model is more than compensated by better statistical and computational properties.

1. Introduction

Mixture distributions with dynamic weights, introduced by [1], are very flexible models for skewed and heavy-tailed data. The density function is defined as follows:

$$f(x;\boldsymbol{\theta}) = \frac{(1-p(x;\boldsymbol{\gamma}_0))f_1(x;\boldsymbol{\gamma}_1) + p(x;\boldsymbol{\gamma}_0)f_2(x;\boldsymbol{\gamma}_2)}{Z}, \quad x \in \mathbb{R}^+, \tag{1}$$

where $p(x;\boldsymbol{\gamma}_0) : \mathbb{R} \to [0,1]$ is a non-decreasing function, $f_1(x;\boldsymbol{\gamma}_1)$ and $f_2(x;\boldsymbol{\gamma}_2)$ are densities of absolutely continuous random variables and Z is a normalizing constant. The vectors $\boldsymbol{\gamma}_i$ $(i=0,1,2)$ contain the parameters of the weight function $p(x;\boldsymbol{\gamma}_0)$, of the body density $f_1(x;\boldsymbol{\gamma}_1)$ and of the tail density $f_2(x;\boldsymbol{\gamma}_2)$, respectively; $\boldsymbol{\theta} = (\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2)' \in \boldsymbol{\Theta} \subset \mathbb{R}^d$ is the vector of all parameters. A natural choice for $p(x;\boldsymbol{\gamma}_0)$ is the cumulative distribution function (cdf) of some absolutely continuous random variable. Whereas (1) apparently resembles the density of "static" finite mixture distributions (e.g., 2), the difference between the two families is non-negligible, and has relevant consequences on estimation.

Models for skewed and heavy-tailed data have long been investigated in actuarial mathematics and statistical finance: see [3,4] for overviews related to insurance analytics and operational risk measurement. More specifically, spliced distributions obtained by joining a lognormal and a Pareto model have been proposed by [5,6]; in operational risk, quantile-based distributions, such as the g-and-h, have been advocated for datasets with large skewness and/or kurtosis (7; 8; 9).

With respect to spliced distributions, dynamic mixtures (1) have two major advantages. First, no continuity and differentiability constraints are necessary, since the density (1) is continuous and differentiable, as long as the component densities are. Second, there is no threshold separating the two distributions. Since setting or estimating such a cut-off point is a non-trivial issue (see, e.g. 10, 11 and 12 for some methods aimed at identifying the threshold in a lognormal-Pareto setup), often handled in a somewhat subjective manner, (1) can be the building block of a more rigorous, fully unsupervised, approach. From this point of view, our method is similar in spirit to the technique devised by [13,14], which has recently been used by [15] for modeling cyber risk.

[☆] The FitDynMix R package, available on CRAN, contains codes for simulation and estimation of the lognormal-GPD dynamic mixture.

E-mail address: marco.bee@unitn.it.

https://doi.org/10.1016/j.matcom.2024.11.011

Received 17 January 2024; Received in revised form 26 September 2024; Accepted 12 November 2024

Available online 20 November 2024

Despite their attractive modeling properties, the usefulness of dynamic mixtures in practical applications has been limited by generally challenging parameter estimation since, regardless of the specification of the weight function and of the component distributions, the normalizing constant Z cannot be evaluated in closed form.

As far as we know, all versions of (1) studied in the literature are based on the Cauchy cdf for the weight function $p(x;\boldsymbol{\gamma}_0)$, on some two-parameter density $f_1(x;\boldsymbol{\gamma}_1)$ for the body and on the zero-location Generalized Pareto distribution (GPD) for the tail density $f_2(x;\boldsymbol{\gamma}_2)$. In this setup, two approaches to estimation have been proposed. [1] approximate the normalizing constant via quadrature methods and maximize numerically the resulting likelihood function. [16] uses approximate maximum likelihood (AMLE), a computer-intensive method which only requires the ability to simulate the data-generating process. Simulation-based comparisons suggest that the latter approach yields better results. However, while avoiding the evaluation of the likelihood, it is computationally demanding.

Since estimation is complicated, it is worth considering whether the model can be simplified, possibly by reducing the number of parameters to be estimated: the rationale is that it may be preferable to devise a more parsimonious model, even at the price of giving up some flexibility, with the goal of obtaining better parameter estimates.

According to this idea, we need to assess how the complexity of the model can be reduced. As said above, (1) is typically based on a size distribution for the body, such as the Weibull or the lognormal, and a GPD for the tail. The latter is motivated by theoretical justifications grounded in Extreme Value Theory (EVT): hence, it is not sensible to replace it with a simpler distribution. As of the former, any size distribution could in principle be used, but all such models are (at least) two-parameter distributions. Hence, the most reasonable way of setting up a model with fewer parameters should be focused on the weight function.

There are two additional reasons supporting this approach. First, [1,16] find that estimating the weight function is by far more complicated than estimating the two mixture distributions. Second, even though the estimates of the weight parameters are poor, usually the overall fit of the distribution remains acceptable; both [1,16] note that there seem to be "compensation effects" between the parameters, so that the reduced flexibility of $p(x;\boldsymbol{\gamma}_0)$ is likely to be balanced by adjustments of the densities parameters.

The novelty of this paper is twofold. First, building on the preceding remarks, we propose a parsimonious model aimed at improving parameter estimation without causing any major goodness-of-fit reduction. Specifically, we develop a five-parameter version of the lognormal-GPD dynamic mixture, where the weight function is the exponential cdf. Second, we introduce a new estimation approach, based on the noisy cross-entropy method, and compare it to numerical maximum likelihood and AMLE, in terms of both statistical and computational efficiency. The analysis is carried out using simulation experiments and real data.

In addition to the lognormal-GPD, we employ the Weibull-GPD mixture. The Weibull is chosen as a possible alternative to the lognormal because it has a different tail behavior (17, tables 3.4.2 to 3.4.4). Another reasonable replacement of the lognormal would be the Gamma, but the Gamma and the lognormal are both in the Maximum Domain of Attraction of the Gumbel distribution, hence we conjecture that the results for the Gamma-GPD would be quite similar to the lognormal-GPD.

The rest of the paper is organized as follows. In Section 2 we detail the dynamic mixture model and the estimation methods. In Section 3 we describe the simulation experiments and their outcomes. In Section 4 we present two real-data examples related to operational risk and non-life insurance. Finally, in Section 5 we discuss the main findings. Appendix reports simulation results for the Weibull-GPD mixture.

2. Model definition and estimation methods

In (1), the weight function $p(x;\boldsymbol{\gamma}_0)$ is the cdf of an absolutely continuous random variable, so that larger values of x are sampled with larger probabilities by the second component distribution. $p(x;\boldsymbol{\gamma}_0)$ can be defined either on $\mathbb{R}$ or on $\mathbb{R}^+$, and the corresponding difference is observed mainly near the origin: in the former case, even when $x \to 0$, the weight of $f_2(x;\boldsymbol{\gamma}_1)$ cannot be exactly 0. On the other hand, the right-tail behavior is determined by the tail distribution in both cases, since $\lim_{x\to\infty} p(x;\boldsymbol{\gamma}_0) = 1$, so that $\lim_{x\to\infty} f(x;\boldsymbol{\theta}) = \lim_{x\to\infty} f_2(x;\boldsymbol{\gamma}_2)$.

With parsimony in mind, we aim at reducing the number of parameters to be estimated. As regards f_1 and f_2, implementing this plan is quite difficult: since the former is supposed to fit well the body of the data, some two-parameter size distribution (see, e.g., 18) should be used. As of the latter, it should be an appropriate model for the tail; thus, well-known EVT theoretical results [17] strongly suggest to use the GPD.

In existing implementations, the weight function is defined on the whole real line. In particular, both [1,16] employ the cdf of the Cauchy distribution, which is a location-scale family with parameters μ_c and τ:

$$p(x;\mu_c,\tau) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu_c}{\tau}\right), \quad x \in \mathbb{R},\ \mu_c \in \mathbb{R},\ \tau \in \mathbb{R}^+. \tag{2}$$

Unfortunately, both papers find that estimating the Cauchy parameters is difficult, especially as concerns τ. On the other hand, simulation experiments suggest that even poor estimates of μ_c and τ do not have a major impact on the goodness of fit of the distribution. Hence, we conjecture that the decrease in flexibility caused by the use of a one-parameter weight function will be more than compensated, in terms of goodness-of-fit, by better estimates of the remaining parameters.

Thus, we propose a one-parameter weight function, given by the exponential cdf:

$$p(x;\lambda) = (1 - e^{-x/\lambda})\mathbb{1}_{x\geq 0}, \quad \lambda \in \mathbb{R}^+. \tag{3}$$

Accordingly, (1) can be rewritten as

$$f(x;\boldsymbol{\theta}) = \frac{(1-p(x;\lambda))f_1(x;\mu,\sigma^2) + p(x;\lambda)f_2(x;\xi,\beta)}{Z},\ x \in \mathbb{R}^+, \tag{4}$$

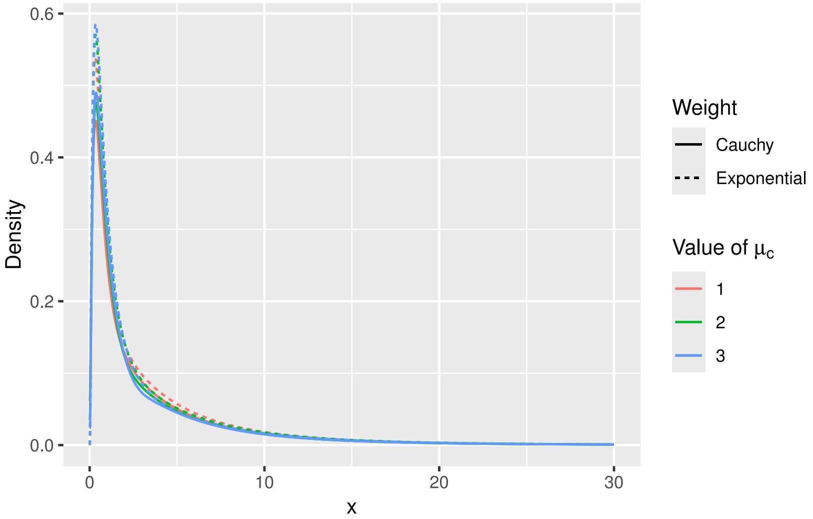

Fig. 1. Cauchy and exponential weight functions.

where $f_1(x;\mu,\sigma^2)$ and $f_2(x;\xi,\beta)$ are the pdfs of $X_1 \sim \text{Logn}(\mu,\sigma^2)$ and $X_2 \sim \text{GPD}(\xi,0,\beta)$, respectively. As of the normalizing constant, it is easy to show that is given by

$$Z = \int_0^\infty \left[\frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2} - \frac{1}{\beta}\left(1+\frac{\xi x}{\beta}\right)^{-1/\xi-1} \right] e^{-\lambda x} dx. \tag{5}$$

The main difference between a one- and a two-parameter weight function based on a location-scale family is the behavior near the origin, as the latter yields dynamic mixtures where the second component has a non-negligible weight, even when $x \to 0$. However, the exponential-based weight function can approximate such a setup, since, for large values of λ, the cdf is very steep and quickly reaches 1. To illustrate this, Fig. 1 shows the weight functions (2) and (3). Denoting with $X_c^{\mu_c,\tau}$ and X_e^{λ} the Cauchy (μ_c,τ) and the exponential (λ) distributions, respectively, and recalling that $P(X_c^{\mu_c,\tau} \le \mu_c) = 0.5$, to determine an appropriate setup for the comparison we use a root-finding procedure to find $\lambda^*(\mu_c) : P(X_e^{\lambda^*(\mu_c)} \le \mu_c) = 0.5$, so that a weight equal to 0.5 is associated to $x = \mu_c$ in both cases. Equivalently, with both weight functions observations smaller than μ_c are more likely to be lognormal, whereas observations larger than μ_c are more likely to be GPD. The three values of μ_c used in Fig. 1 are $\mu_c \in \{1,2,3\}$, with $\tau = 1$; the corresponding numerical values of λ in the exponential weight function are $\lambda^*(\mu_c) \in \{0.693, 0.347, 0.231\}$.

In these specific parameter combinations, the functions are comparable, even for small x. Of course, the Cauchy weight function depends also on τ: when it is larger, the weight function is flatter and converges to 1 more slowly.

Fig. 2 displays the corresponding densities, in a setup where the remaining parameters are $\mu = 0$, $\sigma = 1$, $\xi = 0.25$, $\beta = 3.5$. Again, the difference is negligible: both the shape of the density and the height of the mode are essentially the same. As expected, the densities become indistinguishable as $x \to \infty$.

The dynamic mixture (1) is based on three parametric assumptions, concerning the distribution of the weight, the body and the tail of the distribution. An alternative choice for the body is the Weibull distribution originally proposed by [1]. Since we will employ it in the next sections, we recall here its density and normalizing constant:

$$f(x;\theta) = \frac{(1-p(x;\lambda))f_1(x;\alpha,\sigma_w) + p(x;\lambda)f_2(x;\xi,\beta)}{Z}, \; x \in \mathbb{R}^+, \tag{6}$$

$$Z = 1 + \frac{1}{\pi}\int_0^\infty \left[\frac{1}{\beta}\left(1+\frac{\xi x}{\beta}\right)^{-1/\xi-1} - \alpha\sigma_w^\alpha x^{\alpha-1} e^{-(\sigma_w x)^\alpha} \right] e^{-\lambda x} dx,$$

where $f_1(x;\alpha,\sigma_w)$ and $f_2(x;\xi,\beta)$ are the pdfs of $X_1 \sim \text{Weib}(\alpha,\sigma_w)$ and $X_2 \sim \text{GPD}(\xi,0,\beta)$, respectively; α and σ_w are the shape and scale parameter of the Weibull, respectively.

2.1. Estimation

Broadly speaking, the methods developed in the literature for estimation of a dynamic mixture can be grouped into two categories: maximization of a likelihood computed via numerical evaluation of the integral (5) and simulation-based approaches.

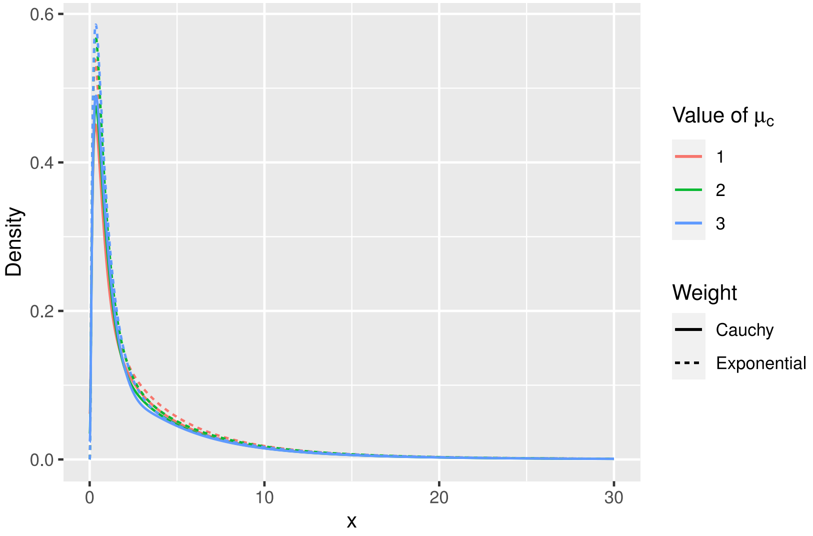

Fig. 2. Densities based on Cauchy and exponential weights.

2.1.1. *Maximum likelihood*

The exact likelihood function is not available, but an approximation can be computed numerically, by first evaluating the integral in the normalizing constant via quadrature methods and then plugging it into the likelihood function derived from (4). This can be finally maximized using standard optimization routines.

The main difficulty with this approach is the first step. We follow the advice of [1], who rewrite the integral on $[0,\infty)$ as a sum of $\tilde{n}$ integrals on disjoint intervals $[n-1,n]$, $n \in \mathbb{N}$, so that $\hat{I}_{\tilde{n}} = \sum_{i=1}^{\tilde{n}} I_{[i-1,i]}$. Since the integrand is a monotonically decreasing function, a natural stopping rule is $\tilde{n} = \min n : I_{[n-1,n]} < \epsilon_I$. However, even for "small" values of ϵ_I, $\hat{I}_{\tilde{n}}$ has a negative bias, which impacts estimation precision: see [1,16] for details.

2.1.2. *Approximate maximum likelihood*

Approximate Maximum Likelihood Estimation (AMLE) is a simulation-based approach that only requires the ability of simulating the distribution of interest. Hence, neither the log-likelihood function nor the normalizing constant have to be evaluated.

From the algorithmic point of view, AMLE [19] is analogous to Approximate Bayesian Computation (ABC), with a simple restriction that makes it a frequentist approach. A step by step description is as follows.

Algorithm 1 (*AMLE*).

1. Obtain a sample $\boldsymbol{\theta}_\epsilon^* = (\boldsymbol{\theta}_{\epsilon,1}^*, \dots, \boldsymbol{\theta}_{\epsilon,\ell}^*)'$ from the approximate posterior $\hat{\pi}_\epsilon(\boldsymbol{\theta}|\boldsymbol{x})$; ℓ is commonly called ABC sample size;
2. Use this sample to construct a non-parametric estimator $\hat{\phi}$ of $\hat{\pi}_\epsilon(\boldsymbol{\theta}|\boldsymbol{x})$;
3. Compute the maximum of $\hat{\phi}$, $\tilde{\boldsymbol{\theta}}_{\ell,\epsilon}$. This is an approximation of the MLE $\hat{\boldsymbol{\theta}}$.

To simulate the ABC sample mentioned at Step 1 of Algorithm 1, one exploits the ABC rejection algorithm below [20].

Algorithm 2 (*ABC Rejection Algorithm*).

1. Simulate $\boldsymbol{\theta}^*$ from the uniform prior $\pi(\cdot)$;
2. Generate $\boldsymbol{z} = (z_1, \dots, z_n)'$ from $f(\cdot|\boldsymbol{\theta}^*)$;
3. Accept $\boldsymbol{\theta}^*$ with probability proportional to a normalized Markov kernel $K_\epsilon(\boldsymbol{x}|\boldsymbol{z})$; otherwise, return to Step 1.

In Bayesian ABC implementations, a non-uniform prior is used at Step 1 of Algorithm 2; the uniform prior employed here is the reason why AMLE is a frequentist approach. Under regularity conditions, the AMLE approximation converges pointwise to the posterior distribution. Furthermore, if $\hat{\pi}_\epsilon(\cdot|\boldsymbol{x})$ is equicontinuous, the mode of $\hat{\pi}_\epsilon(\cdot|\boldsymbol{x})$ converges to the mode of the likelihood $\pi(\cdot|\boldsymbol{x})$; see [19] for details.

Analogously to [16], who used AMLE for the estimation of a dynamic mixture with Cauchy-based weights, the Markov kernel is defined on the space of the true and simulated data:

$$K_\epsilon(\boldsymbol{x}|\boldsymbol{z}) \propto \begin{cases} 1 & \text{if } \rho(\mathcal{P}_0^n, \mathcal{P}_\theta^n) < \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where ρ is the Cramér-von Mises distance, and $\mathcal{P}_0^n$ and $\mathcal{P}_{\theta}^n$ are the distributions of $\boldsymbol{x}$ and of $\boldsymbol{z}|\boldsymbol{\theta}$, respectively. Since they are unknown, in practice they are replaced by the corresponding empirical counterparts.

The simulation results in [16] suggest that AMLE is clearly preferable to MLE in terms of statistical efficiency. However, the price to pay is an heavy computational burden: with a sample size $n = 500$, AMLE needs about 55 min, whereas MLE only takes 32 s. Considering that the standard errors of the estimators must be obtained via bootstrap techniques, AMLE's computing times are a major issue.

2.1.3. Noisy cross-entropy

Historically, the cross-entropy (CE) method was introduced by Rubinstein [21] in an importance sampling setup. Here we focus on the version of CE for continuous optimization; see [22].

The approach is based on two steps [23]. First, the problem is randomized, i.e. the parameters are treated as random variables; we denote the family of instrumental pdfs of the parameter vector $\boldsymbol{\theta}$ by $\{f_{\boldsymbol{\theta}}(\cdot;\boldsymbol{u}); \boldsymbol{u} \in \mathcal{V} \subset \mathbb{R}^k\}$. In a second step, the actual optimization task is linked to the Associated Stochastic Program (ASP)

$$s(\gamma) = P_{\boldsymbol{u}}(\ell(\boldsymbol{\theta}) \geq \gamma) = \mathrm{E}_{\boldsymbol{u}}(\mathbf{1}_{\{\ell(\boldsymbol{\theta})\geq\gamma\}}), \tag{7}$$

where γ is an unknown parameter and ℓ is the target function. The probability $s(\gamma)$ is now estimated for some γ (γ^*, say) close to the maximum of ℓ: if γ^* is large enough, $\{\ell(\boldsymbol{\theta}) \geq \gamma^*\}$ is a rare event. This connects the CE method for optimization and for estimation of rare-events probabilities.

On the theoretical side, if ℓ is a real-valued function on a finite set, it can be shown that $f_{\boldsymbol{\theta}}(\cdot)$ converges to the Dirac delta density centered at $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ is the maximizer of ℓ [22, p. 132]. Equivalently, the sequence $\hat{\boldsymbol{\theta}}^{(t)}$ converges to $\boldsymbol{\theta}^*$ with probability 1, provided the initial variance of the distribution of $\boldsymbol{\theta}$ is "sufficiently large", so that the whole parameter space is explored.

A generalization called *noisy CE* [22, Chap. 6] deals with cases where the objective function is not in closed form. In particular, if an unbiased estimate of $\ell(\boldsymbol{\theta})$ is obtained via simulation, the noise is Monte Carlo sampling variability.

In the present setup, an unbiased estimate of the normalizing constant can be computed by estimating the integral I as follows:

1 Simulate $y_1, \dots, y_B$ from an instrumental random variable Y, where B is the number of replications. The tail of the instrumental distribution cannot be lighter than the function to be integrated, hence we take the GPD: $Y \sim GPD(\beta_I, \xi_I)$.

2 For all y_i, $i = 1, \dots, B$, simulated at Step 1, evaluate the integrand in (5):

$$\hat{h}_i = \left[\frac{1}{\sqrt{2\pi}\sigma y_i} e^{-\frac{1}{2}\left(\frac{\log y_i - \mu}{\sigma}\right)^2} - \frac{1}{\beta}\left(1 + \frac{\xi y_i}{\beta}\right)^{-1/\xi-1}\right] e^{-\lambda y_i}.$$

3 Compute $\hat{I}^{MC} = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{h}_i / f_{GPD}(x_i)\right)$, where f_{GPD} is the density of Y.

Even though basic properties of the Monte Carlo method (e.g., 24) imply that $\mathrm{E}(\hat{I}^{MC}) = I$, the log-likelihood is a non-linear function of I; hence, it cannot be estimated in an unbiased manner by simply replacing I with $\hat{I}^{MC}$. Nevertheless, under regularity conditions, the estimator obtained by maximizing the simulated likelihood is asymptotically equivalent to the MLE. In particular, consistency and asymptotic efficiency have been proved by [25].

In light of the previous remarks, the following pseudo-code describes the continuous noisy CE optimization procedure.

Algorithm 3 (*Cross-Entropy for Continuous Optimization*).

```
1 Set a starting value $\boldsymbol{v}^{(0)}$ for the parameters of the instrumental distributions, choose numerical values for the smoothing parameter $\alpha \in [0,1]$, the quantile level $1-\rho$ ($\rho \in [0,1]$) and the number of simulated parameter values, $M$.
2 Set $t = 1$ and repeat steps (a)–(e) below.
    (a) Simulate a sample $\boldsymbol{\theta}_1^{(t)}, \dots, \boldsymbol{\theta}_M^{(t)}$ from the instrumental density $f_{\boldsymbol{v}}(\cdot;\boldsymbol{v}^{(t-1)})$. Compute $\ell(\boldsymbol{\theta}_1^{(t)}), \dots, \ell(\boldsymbol{\theta}_M^{(t)})$, i.e. evaluate the log-likelihood.
    (b) Compute the sample $(1-\rho)$-quantile $\hat{\gamma}_t$ of $\ell(\boldsymbol{\theta}_1^{(t)}), \dots, \ell(\boldsymbol{\theta}_M^{(t)})$.
    (c) Use the same sample to solve the ASP (7). Call the solution $\boldsymbol{v}^{(t)}$.
    (d) Smooth out $\boldsymbol{v}^{(t)}$: $\boldsymbol{v}^{(t)} = \alpha\boldsymbol{v}^{(t)} + (1-\alpha)\boldsymbol{v}^{(t-1)}$.
    (e) If a stopping criterion is satisfied, stop. Else, set $t = t+1$ and go back to Step (a).
```

As can be seen, the implementation depends on some inputs. The smoothing parameter α at Step 2(d) governs the speed of convergence: as it gets close to 1, the algorithm converges faster, at the cost of increasing the risk of identifying a local optimum [22, p. 189]. The quantile level $1-\rho$ is related to the rarity of the event $\{\ell(\boldsymbol{\theta}) \geq \gamma^*\}$ (see (7)): a smaller ρ, i.e., a higher quantile level, increases the rarity.

To set the numerical values of these parameters, a guideline is usually given by previous applications in the literature, possibly integrated by small pilot simulations. A natural choice for $\theta^{(0)}$ is the vector of MLEs. As for the smoothing parameter α and the quantile level $1-\rho$, [22] suggest $\alpha = 0.5$ and $\rho \in [0.01, 0.1]$. Some experiments based on these two values lead us to use $\alpha = 0.5$ and

Table 1
Case $\xi = 0.25$. Bias and RMSE of the estimators. True parameter values: $\lambda = 1$, $\mu = 0$, $\sigma = 0.5$, $\xi = 0.25$, $\beta = 3.5$. RMSEs in bold identify the minimum-RMSE estimator of each parameter across the three methods.

n			λ	μ	σ	ξ	β
100	Bias	AMLE	0.325	0.029	0.004	0.033	−0.249
		CE	−0.163	−0.042	−0.036	−0.052	0.396
		MLE	0.060	−0.027	−0.018	−0.064	0.437
	RMSE	AMLE	0.675	**0.163**	**0.106**	**0.093**	**0.598**
		CE	**0.411**	0.213	0.119	0.179	1.065
		MLE	0.677	0.229	0.170	0.164	1.000
500	Bias	AMLE	0.182	0.014	−0.009	0.053	−0.303
		CE	−0.179	−0.047	−0.029	−0.022	0.190
		MLE	0.022	−0.015	−0.011	−0.025	0.155
	RMSE	AMLE	**0.091**	**0.082**	0.072	0.091	0.456
		CE	0.266	0.086	**0.065**	0.083	0.489
		MLE	0.330	0.095	0.085	**0.072**	**0.449**

$\rho = 0.1$, since small perturbations did not have any major impact on the estimates. M is set to 500, since larger values did not seem to improve accuracy.

Furthermore, a decision about the parametric model corresponding to the randomization step, i.e. the density f_v at Step 2(a), needs to be made. The chosen distribution is normal for μ, which is a real number, and lognormal for the remaining parameters, which are non-negative[1], since both distributions yield an explicit solution of the ASP [22, p. 82].

As shown by Rubinstein and Kroese [22, p. 188], the solutions of the ASP are analogous to the MLEs of the instrumental distributions. Hence, if the instrumental distribution of the ith parameter is normal, the updating formulas are the mean and variance of the ith elements of the samples of $\boldsymbol{\theta}$ such that $\ell(\boldsymbol{\theta})$ exceeds $\hat{\gamma}_t$:

$$\hat{\mu}_i = \frac{\sum_{j:\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t} \theta_{ij}}{\sum_{j=1}^{M} \mathbb{1}_{\{\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t\}}}; \quad \hat{\sigma}_i^2 = \frac{\sum_{j:\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t} \theta_{ij}^2}{\sum_{j=1}^{M} \mathbb{1}_{\{\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t\}}} - \hat{\mu}_i^2.$$

Analogously, in the lognormal case, the solutions are the mean and variance of the logarithm of the ith elements of the samples of $\boldsymbol{\theta}$ such that $\ell(\boldsymbol{\theta})$ exceeds $\hat{\gamma}_t$:

$$\tilde{\mu}_i = \frac{\sum_{j:\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t} \log \theta_{ij}}{\sum_{j=1}^{M} \mathbb{1}_{\{\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t\}}}; \quad \tilde{\sigma}_i^2 = \frac{\sum_{j:\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t} (\log \theta_{ij})^2}{\sum_{j=1}^{M} \mathbb{1}_{\{\ell(\boldsymbol{\theta}_j)>\hat{\gamma}_t\}}} - \tilde{\mu}_i^2.$$

The closed-form solution of the ASP is an appealing feature of the CE method, also because it considerably eases the computational burden [23, Sect. 2.3].

Various stopping rules have been developed in the noisy optimization case. We use the first criterion proposed by Rubinstein and Kroese [22, p. 207]:

1. At iteration t, compute the moving average of order O of the sample $(1-\rho)$ quantiles $\hat{\gamma}_t$.
2. For each t, compute the minimum B_t^- and the maximum B_t^+ of the aforementioned moving averages over iterations $t, t+1, \ldots, t+r$.
3. If $(B_t^+ - B_t^-)/B_t^-$ is smaller than a predefined tolerance ϵ_O, stop; else, set $t = t+1$ and go to Step 1.

In the rest of the paper we employ $O = 10$, $r = 5$ and $\epsilon_O = 10^{-4}$; if convergence has not been reached in 200 iterations, the algorithm stops. No theoretical optimality result is available for this stopping rule, but some simulation-based analyses suggest that no modification of these values significantly improves the properties of the estimators.

3. Simulation experiments

To compare the estimators obtained with the three methods in a finite-sample setup, we perform some simulation experiments. Specifically, we sample observations from the dynamic mixture (4) with parameters $\lambda = 1$, $\mu = 0$, $\sigma = 0.5$, $\xi \in \{0.25, 0.5\}$, $\beta = 3.5$, where the two values of ξ correspond to different degrees of tail heaviness. The sample size is $n \in \{100, 500\}$, and the number of replications is $B = 200$. To simulate the distribution, we employ the algorithm suggested by [1] and implemented in the `FitDynMix` R package. The same experiments have been carried out for the Weibull-GPD mixture; see Appendix for details.

Tables 1 and 2 show the bias and RMSE of the estimators when $\xi = 0.25$ and $\xi = 0.5$, respectively.

When $\xi = 0.25$ (Table 1), AMLEs have the smallest RMSE for most parameters; the gain is more significant when $n = 100$ and for the estimators of the GPD parameters. In the $\xi = 0.5$ case (Table 2), the advantage of AMLE is clearer, and again especially noticeable when $n = 100$.

[1] In principle, the shape parameter ξ of the GPD can be negative. However, a negative ξ implies a finite right endpoint, which is not reasonable in a lognormal-GPD mixture.

Table 2
Case $\xi = 0.5$. Bias and RMSE of the estimators. True parameter values: $\lambda = 1$, $\mu = 0$, $\sigma = 0.5$, $\xi = 0.5$, $\beta = 3.5$. RMSEs in bold identify the minimum-RMSE estimator of each parameter across the three methods.

n			λ	μ	σ	ξ	β
100	Bias	AMLE	0.177	0.039	0.013	0.000	−0.119
		CE	−0.044	−0.065	−0.063	0.218	0.335
		MLE	0.011	−0.020	−0.008	0.160	0.706
	RMSE	AMLE	0.579	**0.168**	**0.103**	**0.116**	**0.744**
		CE	**0.433**	0.186	0.131	0.320	1.387
		MLE	0.644	0.210	0.169	0.266	1.491
500	Bias	AMLE	0.047	−0.002	−0.000	−0.012	0.031
		CE	−0.037	−0.067	−0.058	0.272	−0.067
		MLE	0.012	−0.0142	−0.010	0.222	0.197
	RMSE	AMLE	0.250	**0.062**	**0.057**	**0.058**	**0.353**
		CE	**0.227**	0.098	0.082	0.288	0.501
		MLE	0.334	0.092	0.086	0.237	0.535

A comparison with the Cauchy-based setup is possible by considering the outcomes in [16], where the same simulation experiments are carried out with the Cauchy weight function. In terms of the RMSE of the AMLE estimators of μ, σ, ξ and β, when $\xi = 0.25$ the results are similar, whereas when $\xi = 0.5$ the exponential case is slightly better. As of the MLEs, they are mostly similar in both parameter configurations. Even though a direct comparison between the estimates of the Cauchy parameters in the first case and of the exponential parameter in the second is not entirely justified, it is worth noting that the very large bias and RMSE of the MLEs in the Cauchy case with $n = 100$ and $\xi = 0.25$ are not observed in the exponential case. Also for AMLE, the estimate of λ in the exponential setup has bias and RMSE smaller than the corresponding measures for the estimates of the Cauchy parameters.

Fig. 3 shows the simulated distributions of the AMLEs with $\xi = 0.25$ and $n = 100$: despite the rather small sample size, the distributions are smooth and bell-shaped.

As of the computational burden of the three estimation methods, on a Windows machine with an i7-6700 CPU 3.40 GHz, the computing times in seconds, when $n = 100$ and $\xi = 0.25$, are 6s for MLE, 99 s for CE and 613 s for AMLE. Also from this point of view, the present model is preferable with respect to the Cauchy-based one, whose computing times are reported in [16].

3.1. A misspecified model

The parametric assumptions underlying (1) are not easy to verify empirically. Whereas the analysis of a mixture based on a different weight function (Cauchy instead of exponential) has already been performed (16,26) and the GPD tail is justified by strong theoretical reasons, the lognormal distribution of the body is difficult to assess, and not always justified by theoretical reasons. Given this premise, it is of interest to study the fit of the lognormal-GPD model when the true f_1 is actually different. According to this remark, in this section we estimate a lognormal-GPD mixture with data simulated from the Weibull-GPD mixture (6).

The numerical values of the parameters of the weight function and of the GPD are the same used in the previous experiments: $\lambda = 1$, $\beta = 3.5$, $\xi \in \{0.25, 0.5\}$; α and σ_w are such that the expectation and variance of the Weibull are identical to the corresponding moments of the lognormal. This amounts to solving for α and σ_w the system

$$\begin{cases} \sigma_w \Gamma\left(1 + \frac{1}{\alpha}\right) = e^{\mu + \frac{\sigma^2}{2}}, \\ \sigma_w^2 \left[\left(\Gamma(1 + \frac{2}{\alpha}\right) - \left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^2\right] = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}, \end{cases}$$

with $\mu = 0$ and $\sigma = 0.5$. The solution obtained by means of the `nleqslv` R command is $(\alpha, \sigma_w) = (1.957, 1.278)$, which are therefore the numerical values of the parameters employed in the simulation experiments.

Fig. 4 displays the histogram of 500 simulated values from the Weibull-GPD mixture with parameters $\lambda = 1$, $\alpha = 1.957$, $\sigma_w = 1.278$, $\xi = 0.25$ and $\beta = 3.5$, along with the true Weibull-GPD density (dashed) and the lognormal-GPD density estimated via CE (continuous)[2]. The two densities appear quite close to each other, and the estimated density seems to provide a good fit. To assess more formally the goodness-of-fit of the misspecified distribution, we employ the Kolmogorov–Smirnov (KS) and Anderson–Darling (AD) test. The p-values equal to 0.844 and 0.891, respectively, allow us to conclude that the approach seems quite robust with respect to possible misspecification of f_1.

[2] The results obtained via AMLE are very similar and therefore omitted.

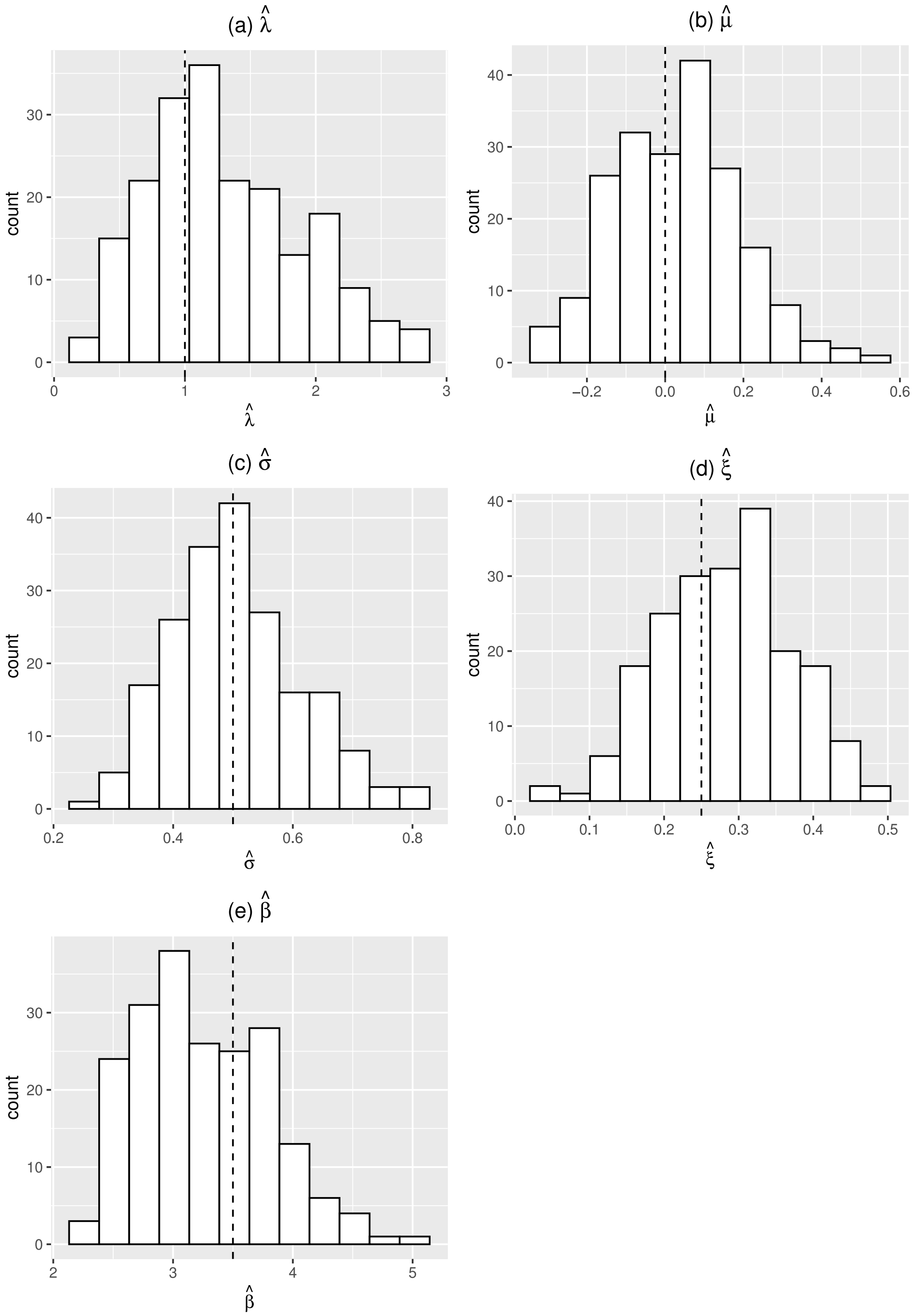

Fig. 3. Simulated distributions of AMLEs in the setup with $\xi = 0.25$. The dashed vertical lines denote the true parameter values.

4. Empirical analysis

4.1. Operational risk

Operational risk losses are a typical example of skewed and heavy-tailed distributions. Here we analyze losses collected in the *Business Disruption and System Failure* (BDSM) business line at the Italian bank Unicredit from 2005 to 2014. The 152 observations are displayed in Fig. 5. The distribution is skewed, with an heavy right tail: sample skewness and kurtosis are equal to 5.42 and 33.17, respectively.

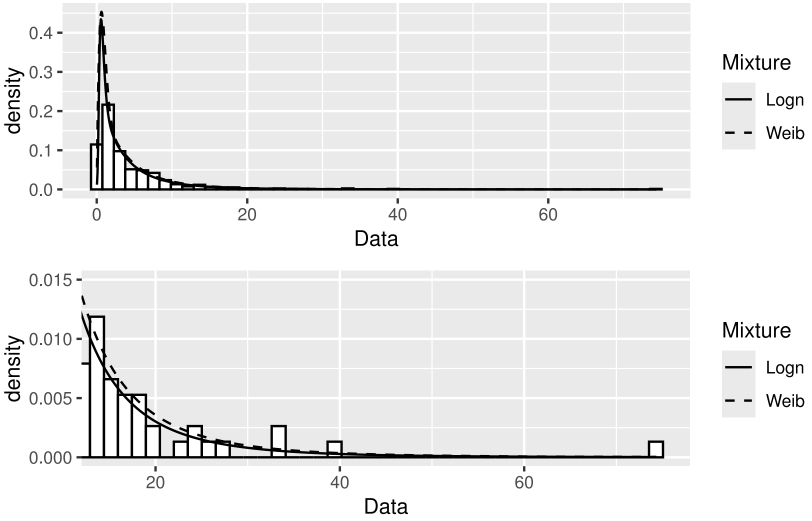

Fig. 4. Upper panel: histogram of 500 simulated observations from a Weibull-GPD mixture with parameters $\lambda = 1$, $\alpha = 1.957$, $\sigma = 1.278$, $\xi = 0.25$ and $\beta = 3.5$, with superimposed the true (Weibull-GPD) and estimated (lognormal-GPD) densities. Lower panel: zoom on the distribution tail.

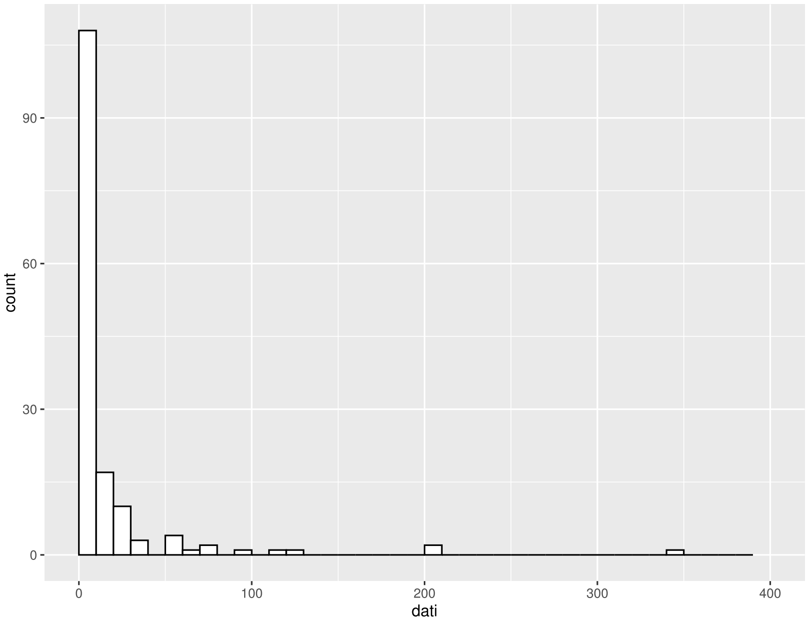

Fig. 5. Operational risk data.

Tables 3 and 4 display parameter estimates, standard errors and p-values of the KS and AD test for the lognormal-GPD and the Weibull-GPD, respectively.

The first notable outcome is that, according to the KS and AD p-values, the dynamic mixture is always an appropriate model for the data, regardless of the estimation method and of the parametric model of the body. For comparison purposes, we have also reported the MLEs and the p-values of the tests obtained when modeling all the data either with the lognormal or the zero-location GPD: there is strong evidence that neither distribution is supported by the KS and AD tests.

As of AMLE and CE, the standard errors are similar for λ, AMLE is better at estimating the GPD parameters, and CE is preferable for the lognormal parameters. MLE is mostly the worst method in terms of variability.

Computing times are negligible for MLE with $\epsilon_I = 10^{-4}$. On a Windows machine with an i7-6700 CPU @3.40 GHz and 16Gb RAM, AMLE takes approximately 31 min and CE 167 s.

Table 3
Lognormal-GPD mixture: parameter estimates, standard errors and p-value of the KS and AD tests in the operational risk example.

	λ	μ	σ	ξ	β	KS	AD
AMLE	0.320 (0.110)	1.190 (0.18)	0.259 (0.093)	0.610 (0.082)	6.642 (0.665)	0.312	0.239
CE	0.463 (0.149)	1.046 (0.056)	0.145 (0.034)	0.729 (0.116)	6.705 (0.907)	0.173	0.110
MLE	0.273 (0.370)	1.110 (0.321)	0.203 (0.117)	0.695 (0.132)	7.519 (1.226)	0.490	0.275
Lognormal	– –	1.976 (0.093)	1.144 (0.097)	– –	– –	0.004	< 0.001
GPD	– –	– –	– –	0.601 (0.119)	7.457 (0.943)	< 0.001	< 0.001

Table 4
Weibull-GPD mixture: parameter estimates, standard errors and p-value of the KS and AD tests in the operational risk example.

	λ	α	σ	ξ	β	KS	AD
AMLE	0.378 (0.122)	3.355 (0.24)	5.760 (0.471)	0.678 (0.097)	6.474 (0.649)	0.401	0.379
CE	0.396 (0.149)	4.101 (0.056)	6.287 (0.034)	0.649 (0.116)	8.431 (0.907)	0.253	0.295

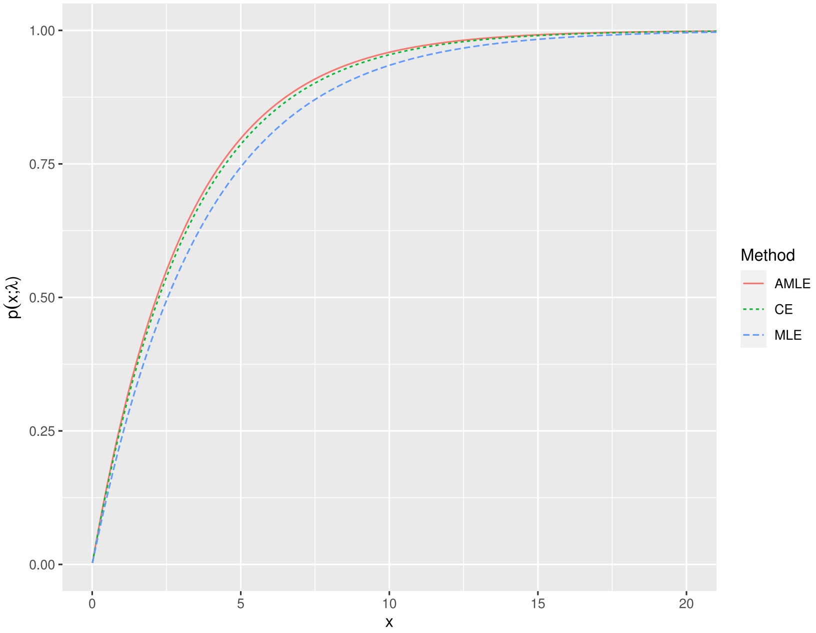

Fig. 6. Estimated weight functions with the three methods in the operational risk example.

For the lognormal-GPD, the estimated weight functions are shown in Fig. 6. They are quite similar to each other. In all cases, observations larger than 10 are GPD with high probability. Since $\#\{x : x < 10\} = 108$ and $\#\{x : x \geq 10\} = 44$, this suggests that the body of the distribution is lognormal, but there is a non-negligible GPD tail.

Value-at-Risk estimates at levels $1 - \alpha \in \{0.5, 0.9, 0.95, 0.99, 0.995\}$ are displayed in Table 5. For comparison purposes, the VaR computed by means of the Peaks-over-Threshold (POT) method has been reported as well. Standard errors are estimated via non-parametric bootstrap, but are omitted for MLE, since numerical maximization of the likelihood crashed on some bootstrap samples.

There are a few differences between the estimated VaRs obtained with the three methods, but consider that the standard errors are quite large, partly as a consequence of the rather limited sample size and the scarcity of tail data; in particular, four observations are larger than 200, and only two are larger than 300. Hence, it is hard to conclude that the differences are significant. In terms of variability, the CE-based VaR seems to be the best one, especially at the highest levels.

Table 5
Estimated VaR and bootstrap standard errors in the operational risk example.

	50%	90%	95%	99%	99.5%
Empirical	4.99	37.18	73.57	274.32	355.25
	(0.64)	(12.44)	(30.08)	(88.29)	(83.94)
AMLE	4.87	32.06	53.84	163.40	251.38
	(0.79)	(9.78)	(23.69)	(155.24)	(339.52)
CE	4.15	35.17	64.37	226.85	377.46
	(0.87)	(6.85)	(15.89)	(93.57)	(192.06)
MLE	4.44	38.07	68.67	235.86	405.68
POT	–	39.94	84.153	257.87	380.21
	–	(12.00)	(34.62)	(91.33)	(319.04)

Table 6
Weibull-GPD. Estimated VaR and bootstrap standard errors in the operational risk example.

	50%	90%	95%	99%	99.5%
AMLE	5.26	34.06	52.17	151.40	234.83
	(0.90)	(10.56)	(23.88)	(144.46)	(322.81)
CE	5.62	36.66	60.12	219.43	356.93
	(1.13)	(8.94)	(22.06)	(152.78)	(244.99)

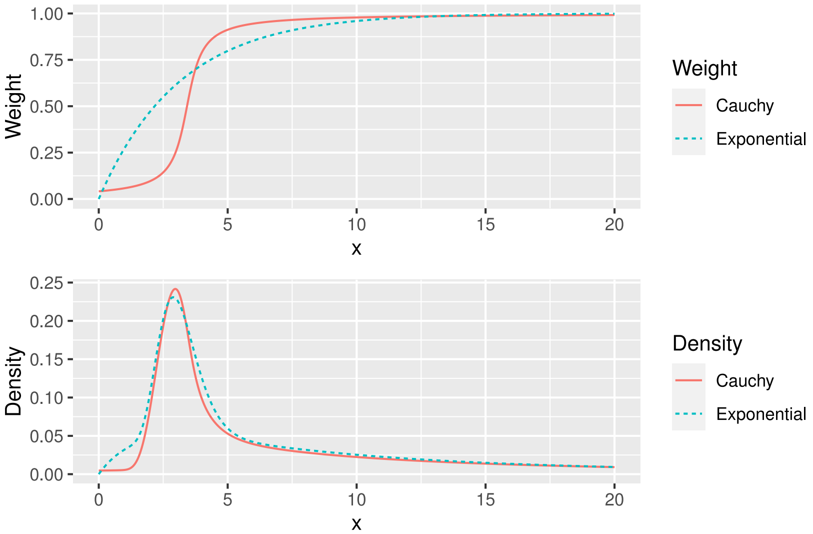

Fig. 7. Estimated Cauchy-based and exponential-based weight functions and corresponding densities in the operational risk example.

Table 6 shows the same risk measures obtained by fitting the Weibull-GPD mixture. The outcomes are analogous to the lognormal-GPD in Table 5.

Finally, for the lognormal-GPD, Fig. 7 shows the estimated weight functions (upper panel) and densities (lower panel) when using the Cauchy and exponential weight function, respectively. For the former case, the estimated parameters are taken from Bee [16, Table 4]. Some difference can be noticed in the estimated weight functions for small values of x (less than 10, say), but overall the densities are very similar, except for the smallest values of x.

4.2. AON Re Belgium

The dataset analyzed in this section, available in the `CASdatasets` R package, contains 1823 fire losses collected by the reinsurance broker AON Re Belgium. The histogram of the data is shown in Fig. 8 where, to improve readability, we have omitted the 10 observations larger than 8000. The distribution is very skewed, with a maximum loss equal to 190 541, and sample skewness and kurtosis equal to 33.93 and 1290, respectively.

Point estimates, bootstrap standard errors and p-values of the KS and AD tests are shown in Table 7.

Similarly to the previous application, the pure lognormal and pure GPD are clearly rejected. However, here also the CE-based distribution is not accepted. The numerical values of the estimated parameters suggest that this may be due to an inaccurate estimate of λ, whose CE-based point estimate $\hat{\lambda}^{CE}$ is much larger (and has a larger standard deviation) than both $\hat{\lambda}^{AMLE}$ and $\hat{\lambda}^{CE}$. When looking at the standard errors of the estimators, AMLE is the best approach for most parameters.

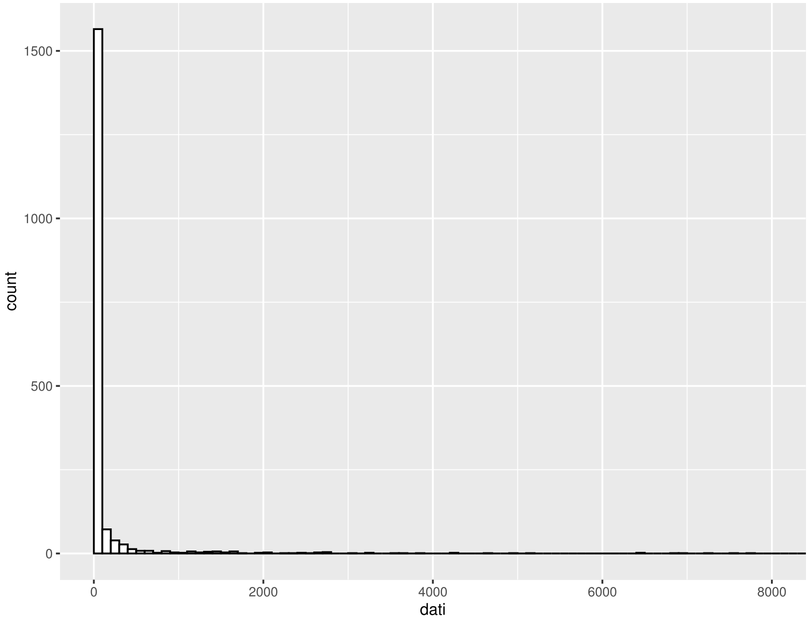

Fig. 8. Belgium reinsurance data.

Table 7
Parameter estimates, standard errors and p-value of the KS and AD tests in the AON Re Belgium reinsurance example.

	λ	μ	σ	ξ	β	KS	AD
AMLE	0.811 (0.110)	1.886 (0.180)	0.783 (0.093)	1.711 (0.082)	5.375 (0.665)	0.360	0.287
CE	5.234 (1.114)	2.126 (0.511)	0.749 (0.097)	1.479 (0.052)	8.389 (0.442)	< 0.001	< 0.001
MLE	0.354 (0.261)	1.430 (0.383)	0.757 (0.347)	1.598 (0.087)	8.613 (1.976)	0.643	0.814
Lognormal	– –	2.606 (0.046)	1.975 (0.041)	– –	– –	< 0.001	< 0.001
GPD	– –	– –	– –	1.472 (0.055)	8.955 (0.434)	< 0.001	< 0.001

The outcomes of the CE method deserve some further comment. The large estimated value of λ results in a weight function that reaches 1 very quickly (see Fig. 9). Equivalently, this implies that almost all observations are GPD with high probability. Two additional consequences are that such a mixture is not accepted by the KS and AD tests, and the point estimates and standard errors of β and ξ are very close to the corresponding quantities in the pure GPD case. In other words, with the CE method we estimate a GPD with almost all observations. Since CE maximizes the likelihood, it makes sense that estimates and standard errors are almost identical to the MLEs of the pure GPD. Hence, the likely reason why the standard errors suggest that the CE method is good at estimating β and ξ is that $\hat{\beta}^{CE}$ and $\hat{\xi}^{CE}$ use more observations than the corresponding AMLE and MLE estimates. Here it is crucial to take into account the outcomes of the goodness-of-fit tests, which are the main tool for concluding that CE is not working well in this case.

The impact of these differences is made clear by Fig. 9. The weight function estimated via CE reaches 1 very quickly, whereas the other two functions are less steep. In all cases, the majority of the observations are GPD, since 1157 observations (i.e., 63.5%) are larger than 5.

5. Conclusion

In this paper we have developed a parsimonious dynamic mixture model for skewed and heavy-tailed data. In particular, the simplification lies in the use of the one-parameter exponential weight function, instead of the two-parameter Cauchy cdf used by [1,16]. According to the outcomes of the simulation experiments, this way of proceeding results in more precise estimates of all parameters. Two empirical applications related to operational risk and actuarial losses suggest that the data-generating process is flexible enough to model skewed and heavy-tailed data. Similarly to the original analysis based on the Cauchy cdf, for which the AMLE approach was the best one, also in the current setup AMLE outperforms CE and MLE, especially in the empirical applications, but is computationally more demanding.

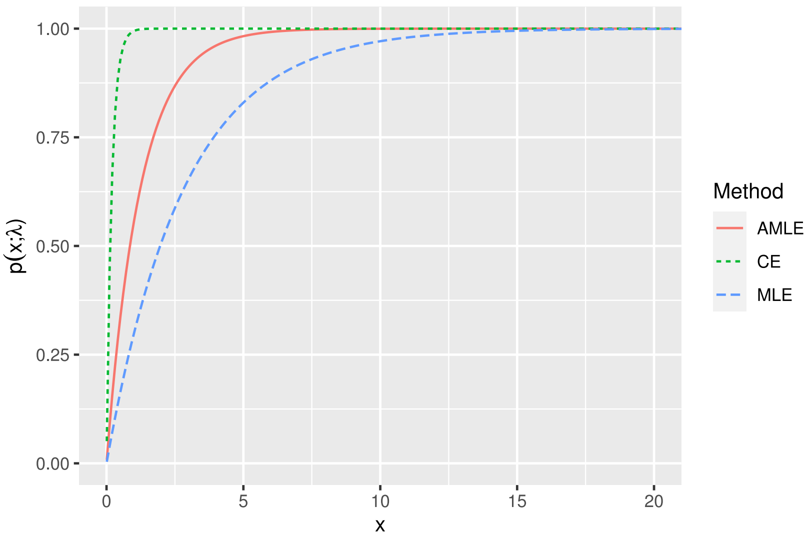

Fig. 9. Estimated weight functions with the three methods in the Belgium reinsurance dataset.

Table A1
Case $\xi = 0.25$. Bias and RMSE of the Weibull-GPD estimators. True parameter values: $\lambda = 1$, $\alpha = 1.957$, $\sigma = 1.278$, $\xi = 0.25$, $\beta = 3.5$. RMSEs in bold identify the minimum-RMSE estimator of each parameter across the two methods.

n			λ	α	σ	ξ	β
100	Bias	AMLE	0.294	−0.017	0.010	0.037	−0.188
		CE	−0.184	−0.026	−0.014	−0.065	0.295
	RMSE	AMLE	0.614	**0.150**	0.115	**0.098**	**0.558**
		CE	**0.402**	0.231	**0.114**	0.166	1.037
500	Bias	AMLE	0.201	0.021	−0.020	0.042	−0.252
		CE	−0.209	−0.045	−0.039	−0.012	0.234
	RMSE	AMLE	**0.102**	0.097	0.085	0.081	**0.402**
		CE	0.304	**0.076**	**0.077**	**0.059**	0.550

CE performs rather well in the simulation analysis and in the first application, but considerably less well in the reinsurance example: this may suggest that it is less robust with respect to violations of the parametric assumptions. This issue requires further investigation.

Finally, to the best of our knowledge, identifiability of dynamic mixtures has never been proved. Hence, on the theoretical side, this is still an open problem.

Acknowledgments

The present work is funded by the European Union - Next Generation EU, within the scope of the PRIN 2022 project "Prediction and causal inference on the tail index for policy decisions", funded by the European Union, Mission 4, Component 2, CUP E53D23006380006. An earlier version has been presented at the 16th International Conference of the ERCIM WG on Computational and Methodological Statistics, Berlin, December 16–18, 2023. I thank the participants for useful comments and suggestions. I also thank the associate editor and an anonymous reviewer for valuable comments that considerably improved an earlier version of the paper.

Appendix. Simulation results for the Weibull-GPD case

The simulation experiments of Section 3 have been performed also for the Weibull-GPD distribution, which is the model originally considered by [1]. The true values of the parameters are $\lambda = 1$, $\alpha = 1.957$, $\sigma = 1.278$, $\xi \in \{0.25, 0.5\}$, $\beta = 3.5$, $n \in \{100, 500\}$. The number of replications is $B = 200$. The results for AMLE and CE are displayed in Tables A1 and A2.

As can be seen from the tables, the outcomes are quite similar to the corresponding lognormal-GPD results in Tables 1 and 2: when $\xi = 0.5$, AMLE is preferable for most parameters, especially for $n = 100$, whereas when $\xi = 0.25$ the difference is smaller and there is no clear winner.

Table A2
Case $\xi = 0.5$. Bias and RMSE of the Weibull-GPD estimators. True parameter values: $\lambda = 1$, $\alpha = 1.957$, $\sigma = 1.278$, $\xi = 0.5$, $\beta = 3.5$. RMSEs in bold identify the minimum-RMSE estimator of each parameter across the two methods.

n			λ	α	σ	ξ	β
100	Bias	AMLE	0.157	0.031	0.014	0.020	−0.087
		CE	−0.055	−0.071	−0.069	0.184	0.377
	RMSE	AMLE	**0.453**	**0.171**	**0.116**	**0.128**	**0.700**
		CE	0.459	0.202	0.136	0.299	1.443
500	Bias	AMLE	0.047	−0.009	−0.008	−0.008	0.027
		CE	−0.037	−0.067	−0.058	0.272	−0.067
	RMSE	AMLE	0.250	**0.063**	**0.060**	**0.068**	**0.312**
		CE	**0.218**	0.096	0.086	0.274	0.479

References

[1] A. Frigessi, O. Haug, H. Rue, A dynamic mixture model for unsupervised tail estimation without threshold selection, Extremes 3 (5) (2002) 219–235.

[2] D. Titterington, A. Smith, U. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, 1985.

[3] S.A. Klugman, H.H. Panjer, G.E. Willmot, Loss Models: From Data to Decisions, second ed., Wiley, 2004.

[4] Panjer H. H., Operational Risk Modeling Analytics, Wiley, 2006.

[5] D. Scollnik, On composite lognormal-Pareto models, Scand. Actuar. J. 1 (2007) 20–33.

[6] M. Bee, On discriminating between lognormal and Pareto tail: an unsupervised mixture-based approach, Adv. Data Anal. Classif. (2022) http://dx.doi.org/10.1007/s11634-022-00497-4.

[7] M. Degen, P. Embrechts, D.D. Lambrigger, The quantitative modeling of operational risk: between g-and-h and EVT, Astin Bull. 37 (2) (2007) 265–291.

[8] M. Cruz, G. Peters, P. Shevchenko, Fundamental Aspects of Operational Risk and Insurance Analytics: A Handbook of Operational Risk, Wiley, 2015.

[9] M. Bee, J. Hambuckers, L. Trapin, Estimating large losses in insurance analytics and operational risk using the g-and-h distribution, Quant. Finance 21 (7) (2021) 1207–1221.

[10] X. Gabaix, R. Ibragimov, Rank-1/2: A simple way to improve the OLS estimation of tail exponents, J. Bus. Econom. Statist. 29 (1) (2011) 24–39.

[11] Y. Malevergne, V. Pisarenko, D. Sornette, Gibrat's law for cities: uniformly most powerful unbiased test of the pareto against the lognormal, in: Swiss Finance Institute Research Paper Series 09-40, Swiss Finance Institute, 2009.

[12] M. Bee, M. Riccaboni, S. Schiavo, Pareto versus lognormal: A maximum entropy test, Phys. Rev. E 84 (2011) 026104.

[13] N. Debbabi, M. Kratz, A new unsupervised threshold determination for hybrid models, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2014, pp. 3440–3444.

[14] N. Debbabi, M. Kratz, M. Mboup, A self-calibrating method for heavy tailed data modelling. application in neuroscience and finance, 2017.

[15] M. Dacorogna, N. Debbabi, M. Kratz, Building up cyber resilience by better grasping cyber risk via a new algorithm for modelling heavy-tailed data, European J. Oper. Res. 311 (2) (2023) 708–729.

[16] M. Bee, Unsupervised mixture estimation via approximate maximum likelihood based on the Cramér - von Mises distance, Comput. Statist. Data Anal. 185 (2023) 107764.

[17] P. Embrechts, C. Klüppelberg, T. Mikosch, Modelling Extremal Events for Insurance and Finance, Springer, 1997.

[18] C. Kleiber, S. Kotz, Statistical Size Distributions in Economics and Actuarial Sciences, Wiley, 2003.

[19] F.J. Rubio, A.M. Johansen, A simple approach to maximum intractable likelihood estimation, Electron. J. Stat. 7 (2013) 1632–1654.

[20] M.A. Beaumont, W. Zhang, D.J. Balding, Approximate Bayesian computation in population genetics, Genetics 162 (2002) 2025–2035.

[21] R.Y. Rubinstein, Optimization of computer simulation models with rare events, European J. Oper. Res. 99 (1997) 89–112.

[22] R.Y. Rubinstein, D. Kroese, The Cross-Entropy Method, Springer, 2004.

[23] D.P. Kroese, R.Y. Rubinstein, P.W. Glynn, Machine Learning: Theory and Applications, in: Handbook of Statistics, chapter Chapter 2 - The Cross-Entropy Method for Estimation, vol. 31, Elsevier, 2013, pp. 19–34.

[24] N. Chan, H. Wong, Simulation Techniques for Financial Risk Management, second ed., Wiley, 2015.

[25] V.A. Hajivassiliou, P.A. Ruud, Handbook of Econometrics, Volume 4, Chapter 40 Classical Estimation Methods for LDV Models using Simulation, Elsevier, 1994, pp. 2383–2441.

[26] M. Bee, Unsupervised tail modeling via noisy cross-entropy minimization, Appl. Stoch. Models Bus. Ind. 40 (2024) 945–959.