Supporting Materials: Using Artificial Censoring to Improve Extreme Tail Quantile Estimates

Yang Liu

Department of Statistics, University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, BC, Canada V6T 1Z4

E-mail: yang.liu@stat.ubc.ca

Matías Salibián-Barrera

Department of Statistics, University of British Columbia

E-mail: matias@stat.ubc.ca

Ruben H. Zamar

Department of Statistics, University of British Columbia

E-mail: ruben@stat.ubc.ca

James V. Zidek

Department of Statistics, University of British Columbia

E-mail: jim@stat.ubc.ca

1. Model settings in the simulation on lower quantile estimation

Based on the complexity of the models from which we simulated the data, we classify them into three categories, parametric models, mixture models and nonparametric models and introduce them separately.

1.1. Parametric models

Here is a brief description of the four parametric models considered in Section 2

- Weibull distribution is commonly used to describe the distribution of particle size and material strength. We only consider a two–parameter Weibull distribution, whose PDF is,

$$f(x;\kappa,\eta) = \frac{\kappa}{\eta}\left(\frac{x}{\eta}\right)^{\kappa-1}\exp\left(-\left(\frac{x}{\eta}\right)^{\kappa}\right), \qquad x \geq 0.$$

  $\kappa > 0$ is the shape parameter and $\eta > 0$ is the scale parameter.

- Log–normal distribution is the exponential transformation of normal distribution, which means that the logarithm of a log–normal random variable will be normally distribution. Its PDF is,

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right), x \geq 0.$$

Table 1. Parameters for the censored parametric models to imitate the left tail of MOR1 and MOR2

Model	MOR1		MOR2	
Weibull	$\alpha = 6.822$	$\eta = 7.173$	$\alpha = 7.378$	$\eta = 6.738$
Log-normal	$\mu = 2.072$	$\sigma = 0.336$	$\mu = 1.976$	$\sigma = 0.2916$
Gamma	$k = 12.93$	$s = 0.601$	$k = 16.16$	$s = 0.4407$
Minimum Gumbel	$a = 6.620$	$b = 0.650$	$a = 6.315$	$b = 0.5997$

- Gamma distribution is also parameterized by a shape parameter k and a scale parameter s,

$$f(x; k, s) = \frac{x^{k-1}}{\Gamma(k)s^k} \exp\left(-\frac{x}{s}\right), x \geq 0.$$

- Gumbel distribution (maximum or minimum) is the type I extreme value distribution introduced by Frechét (1927) and Fisher and Tippet (1928). Only the Gumbel distribution of the minimum is considered in our simulation. It is necessary to point out that people usually treat the maximum type Gumbel as the default Gumbel distribution. So we will specify the Gumbel distribution we use here as "minimum Gumbel", whose PDF is:

$$f(x; a, b) = \frac{1}{b} \exp\left(\frac{x-a}{b} - \exp\left(\frac{x-a}{b}\right)\right).$$

All these four models can be fitted to the lower tail of the two real data sets with the artificial censorship as in (1). Here their parameters are summarized in Table 1.

1.2. Mixture models

From the histogram of the MOR1 and MOR2 in Figure 1, there are two modes in each data set. Then, it is natural to consider our data as generated by a mixture of two uni–modal distributions with PDF:

$$f(x) = p_1 f_1(x) + (1-p) f_2(x), \tag{1}$$

where the f_1, f_2 are the density functions of the two sub–populations. They can be either from the same distribution families or from different distribution families. Here p indicates the proportion from the first sub–population, which also means that a random variable X from (1) has probability p of being from the first sub–population and probability $1-p$ from the second sub–population. Of course, we can also consider mixture models with more than two sub–populations.

Our simulations suggest that we can consider mixture models with only two sub–populations and the sub–populations are from the same distribution family. The distributions considered here are normal, log–normal and Weibull. The parameters in this mixture model can be estimated by the EM algorithm (Dempster et al., 1977). The estimated parameters of the three mixture models are summarized in Table 2. Here we does not apply censoring in the EM algorithm for the mixture models, as they have more parameters and thus more flexibility to approximate the tail even without censoring.

Table 2. Mixture models to imitate MOR1 and MOR2

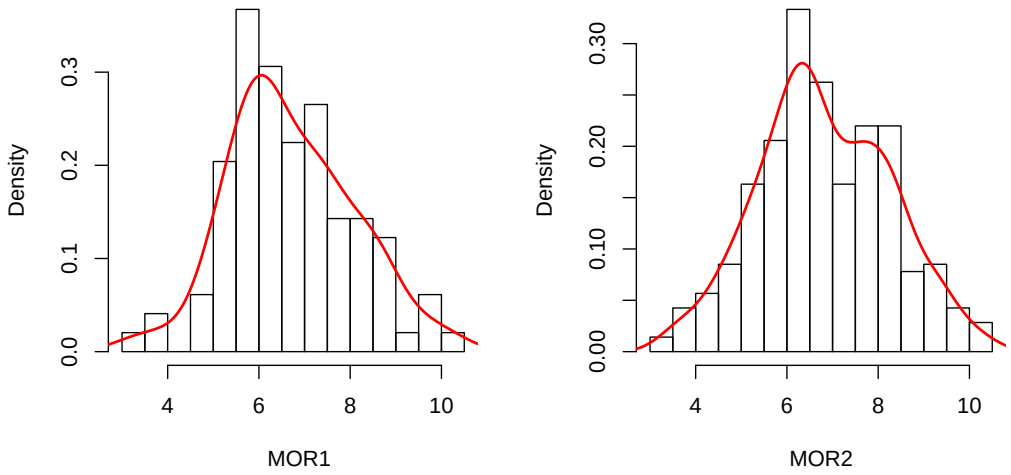
	Model	p	Majority Population		Minority Population	
MOR1	Normal	0.5629	$\mu_1 = 5.953$	$\sigma_1 = 0.970$	$\mu_2 = 7.676$	$\sigma_2 = 1.215$
	Log-normal	0.9758	$\mu_1 = 1.897$	$\sigma_1 = 0.189$	$\mu_2 = 1.245$	$\sigma_2 = 0.102$
	Weibull	0.7448	$\alpha_1 = 5.494$	$\eta_1 = 7.599$	$\alpha_2 = 15.81$	$\eta_2 = 5.983$
MOR2	Normal	0.5406	$\mu_1 = 5.924$	$\sigma_1 = 1.042$	$\mu_2 = 7.859$	$\sigma_2 = 1.095$
	Log-normal	0.6649	$\mu_1 = 1.976$	$\sigma_1 = 0.167$	$\mu_2 = 1.736$	$\sigma_2 = 0.226$
	Weibull	0.7932	$\alpha_1 = 5.427$	$\eta_1 = 7.642$	$\alpha_2 = 12.01$	$\eta_2 = 6.186$

1.3. Non–parametric models

As we have discussed above, we can never be certain about the validity of our model assumptions for the data set. So the model we use to simulate data might not represent the real data set (even if we consider only the tail). On the other hand, we can use the kernel density estimate (Sheather and Jones, 1991) to simulate data sets similar to our real data set, which can be viewed as a kind of smoothed bootstrap. To simulate data from a kernel density estimate with Gaussian kernel and bandwidth b of a real data set, we will first sample with replacement from the original data set. Then $i.i.d.$ random noise from the normal distribution with mean 0 and standard deviation b will be added to each observation in the re–sample of the real data set. In this way, we will obtain a simulated data set re–sampled from the original data set plus some noise.

In this paper, we will use the "Solve-the-Equation" approach (Sheather and Jones, 1991) to select the bandwidth for the kernel density estimate. The bandwidth selected for MOR1 is 0.0419 while the bandwidth for MOR2 is 0.04056. The curves of these two kernel density estimates are shown in Figure 1.

Fig. 1. Kernel density models for MOR1 and MOR2



2. Belief introduction to copula models

For the four one–parameter copula families considered in this paper, we list their copula functions and domains of the parameter δ below. A full treatment of them can be found in Joe (2014). For notational simplicity, define $\tilde{u} = -\log(u)$, $\tilde{v} = -\log(v)$, $\bar{u} = 1 - u$, and $\bar{v} = 1 - v$.

- Galambous, for $\delta \in (0, \infty)$

$$C(u, v; \delta) = uv \exp\left(\tilde{u}^{-\delta} + \tilde{v}^{-\delta}\right)^{-1/\delta}.$$

- Gumbel, for $\delta \in [1, \infty)$

$$C(u, v; \delta) = uv \exp\left(\tilde{u}^{-\delta} + \tilde{v}^{-\delta}\right)^{-1/\delta}.$$

- Joe, for $\delta \in [1, \infty)$

$$C(u, v; \delta) = 1 - \left(\bar{u}^{\delta} + \bar{v}^{\delta} - \bar{u}^{\delta}\bar{v}^{\delta}\right)^{1/\delta}.$$

- MTCJ, for $\delta \in [1, \infty)$

$$C(u, v; \delta) = u + v - 1 + \left(\bar{u}^{-\delta} + \bar{v}^{-\delta} - 1\right)^{-1/\delta}.$$

3. Additional plots for Section 2

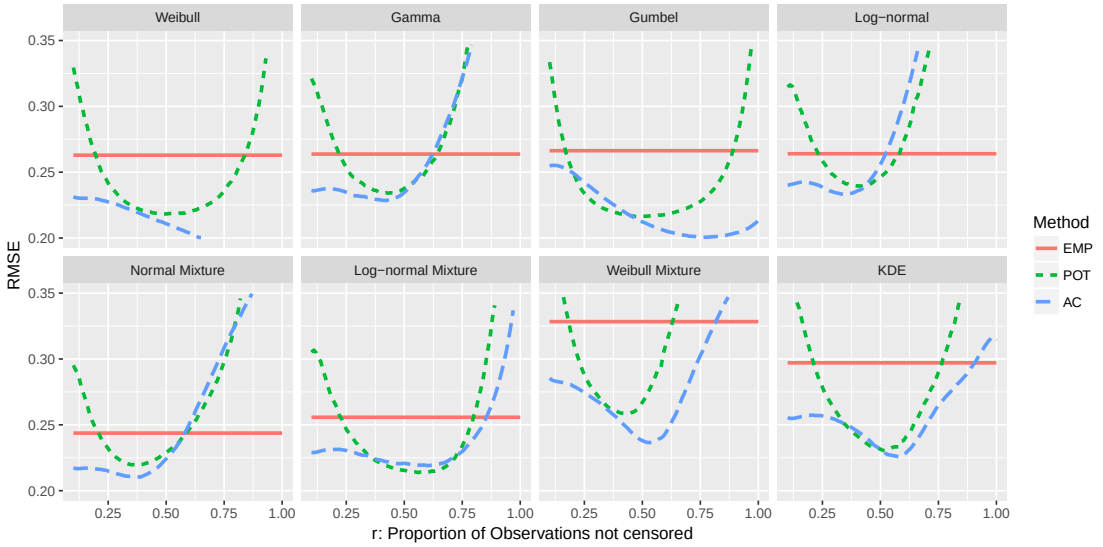


Fig. 2. RMSE of the fifth percentile estimates with sample size 100 from AC approach and POT with different amount of censoring (proportion of observations used in the parametric part).

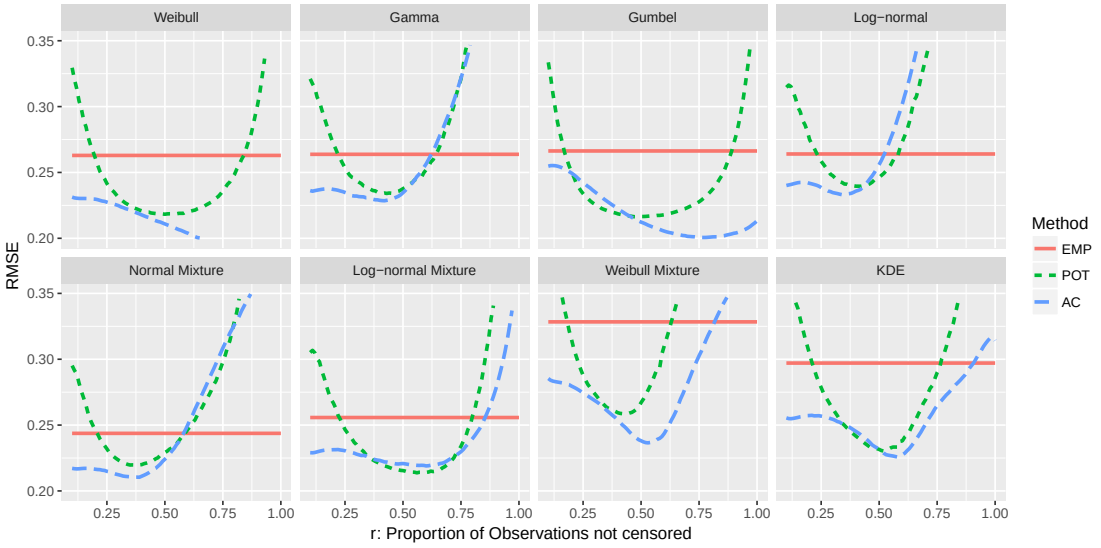


Fig. 3. RMSE of the fifth percentile estimates with sample size 500 from AC approach and POT with different amount of censoring (proportion of observations used in the parametric part).

References

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Joe, H. (2014) *Dependence modeling with copulas*. CRC Press.

Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**, 683–690.

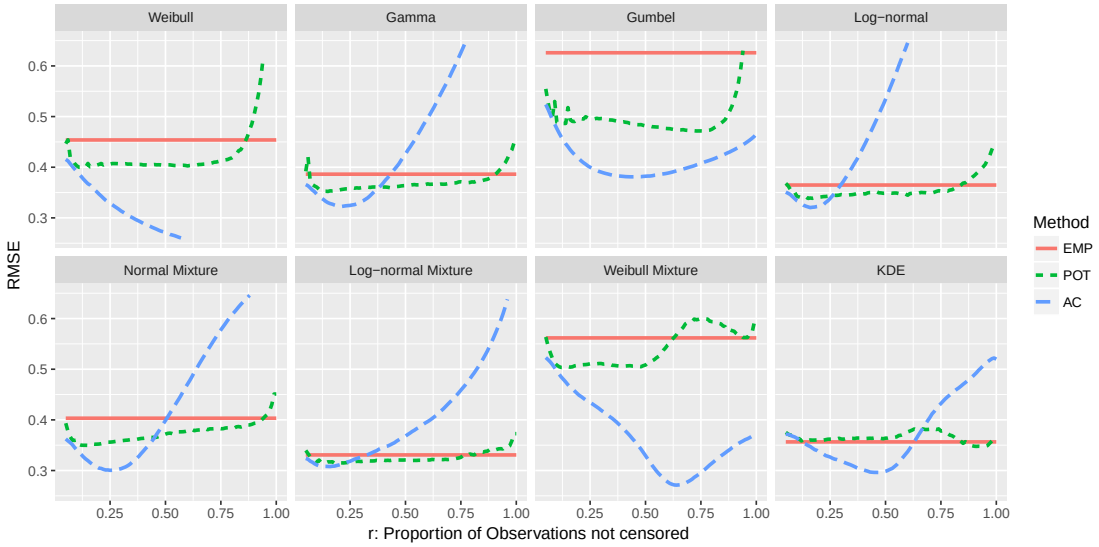


Fig. 4. RMSE of the first percentile estimates with sample size 100 from AC approach and POT with different amount of censoring (proportion of observations used in the parametric part).

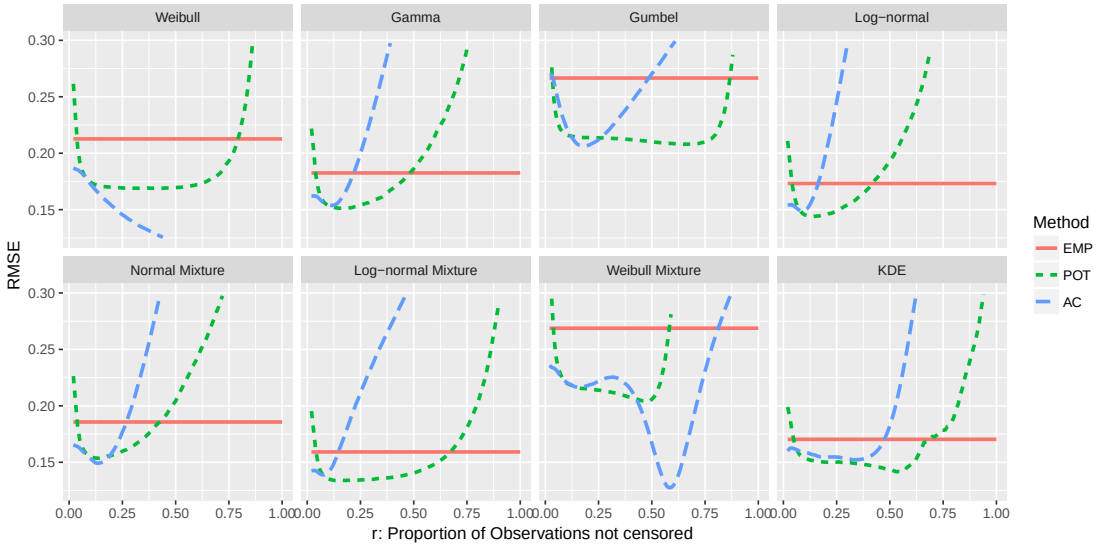


Fig. 5. RMSE of the first percentile estimates with sample size 500 from AC approach and POT with different amount of censoring (proportion of observations used in the parametric part).