ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2018)

Using artificial censoring to improve extreme tail quantile estimates

Yang Liu, Matías Salibián-Barrera, Ruben H. Zamar and James V. Zidek

University of British Columbia, Vancouver, Canada

[Received September 2016. Revised December 2017]

Summary. Under certain regularity conditions, maximum-likelihood-based inference enjoys several optimality properties, including high asymptotic efficiency. However, if the distribution of the data deviates slightly from the model proposed, the statistical properties of inference methods based on maximum likelihood can quickly deteriorate. We focus on the situation when the interest lies in one of the tails of the distribution, e.g. when we are estimating a high or low quantile. In this case, it may be natural, if slightly unorthodox, to consider models that fit well the corresponding tail of the sample, rather than its whole range. For example, if we are interested in estimating the fifth percentile, we can pretend that all observations above the 10th percentile have been censored and fit a parametric censored model to the lower tail of the sample. Such an approach, which we call 'artificial censoring', has been studied in the engineering literature. We study a data-dependent method to select the amount of artificial censoring and show that it compares favourably with the optimally chosen ('oracle') method, which is generally unavailable in practice. We also show that the artificial censoring approach can be applied to estimate tail dependence parameters in copula models, and that it performs well both in simulation and in real data studies.

Keywords: Artificial censoring; Copula; Daily log-return; Extreme quantile; Lumber strength; Peaks over threshold

1. Introduction

It is well known that, under certain regularity conditions, maximum-likelihood- (ML) based inference enjoys several optimality properties, including high asymptotic efficiency (see, for example, Casella and Berger (2002)). However, it is easy to show that, if the distribution of the data deviates slightly from the model proposed, the statistical properties of inference methods based on ML can quickly deteriorate. This paper focuses on the situation where interest lies in one of the tails of the distribution, e.g. where we are estimating a high or low quantile. In this case it may be natural, if slightly unorthodox, to consider models that fit well the corresponding tail of the sample, rather than its whole range. For instance if we are interested in estimating the fifth percentile, we can pretend that all observations above the 10th percentile have been censored and use methods based on censored data (see for example Lawless (2003) for a definition and detailed treatment) to obtain parameter estimates that can in turn be used to construct quantile estimates. More formally, assume that $X_1, X_2, \ldots, X_n$ are an independent and identically distributed random sample, from a family of distributions parameterized by the vector $\boldsymbol{\theta}$, e.g. a two-parameter Weibull distribution. The parameters $\boldsymbol{\theta}$ can be estimated by maximizing the following type II censored likelihood (Lawless, 2003), where the largest $n - m$ observations have

Address for correspondence: Yang Liu, Department of Statistics, University of British Columbia, 3182 Earth Sciences Building, 2207 Main Mall, Vancouver, British Columbia, V6T 1Z4, Canada.
E-mail: yang.liu@stat.ubc.ca

© 2018 Royal Statistical Society 0035–9254/18/67000

been artificially censored, i.e. by maximizing

$$L(X_1,\ldots,X_n;\boldsymbol{\theta}) = \{1 - F(X_{(m)};\boldsymbol{\theta})\}^{n-m} \prod_{i=1}^{m} f(X_{(i)};\boldsymbol{\theta}) \tag{1}$$

where $X_{(i)}, i = 1, 2, \ldots, n$, are the order statistics and $f(\cdot;\boldsymbol{\theta})$ and $F(\cdot;\boldsymbol{\theta})$ are the probability density function and cumulative distribution function (CDF) respectively. In other words, we 'pretend' that only the smallest m observations in the sample were observed and that the rest are right censored. This artifical censoring (AC) approach has been used, for example, in American Society for Testing and Materials (2015), which is an engineering wood standards document which specifies that, to estimate the fifth percentile of the lumber strength distribution, the upper 90% of the collected data should be artificially censored and a Weibull distribution used in equation (1) with $m = [n/10]$.

If the sample were in fact generated from the assumed parametric model, the AC approach above leads to a less efficient parameter estimate than the standard maximum likelihood estimate (MLE) for $\boldsymbol{\theta}$. However, in most practical applications, the parametric family under consideration might not include the actual distribution of the data. One of the main contributions of this paper is to show that, in this case, for the quantile estimates that are obtained from equation (1) to work well it is sufficient that the tail of the distribution be well approximated by one of $F(\cdot;\boldsymbol{\theta})$.

Although non-parametric quantile estimates do not suffer from the potential bias that is associated with an inadequately chosen model, their relatively large variance makes them a less attractive option. Another limitation of using empirical quantile estimates is that they cannot be extrapolated beyond the range of the data. For example, one cannot estimate non-parametrically the first percentile with a sample of size $n = 90$. The AC approach that is studied here addresses both of these limitations: it partially inherits the efficiency of model-based procedures, and one can use the estimated parameters $\boldsymbol{\theta}$ in $F(\cdot;\boldsymbol{\theta})$ to estimate any quantile, regardless of the range or size of the sample.

The idea of using only a portion of the sample to estimate tail characteristics of the distribution resembles the 'peaks-over-threshold' (POT) approach that is used in extreme value theory (Hill, 1975). The POT method assumes that, for a suitably chosen u, the conditional distribution of $Y = X - u$ given $X > u$ is

$$\begin{aligned}\Pr(X - u \leqslant y | X > u) &\approx H(y;\xi,\sigma)\\ &= 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \qquad y > 0,\end{aligned}$$

where $\xi > 0$ and $\sigma > 0$ are shape and scale parameters respectively. The distribution $H(\cdot;\xi,\sigma)$ is known as the generalized Pareto distribution (Coles, 2001), and its parameters can be estimated by maximizing the conditional log-likelihood

$$l(X_1,\ldots,X_n;\xi,\sigma) = \sum_{i=1}^{m} \log\{h(x_i - u;\xi,\sigma)\}, \tag{2}$$

where $h(y;\xi,\sigma) = \mathrm{d}H(y)/\mathrm{d}y$ and $x_1, x_2, \ldots, x_m$ are the m observations that are larger than the threshold u. Given parameter estimates $\hat{\xi}$ and $\hat{\sigma}$, the $(1 - 1/m)$-quantile can be estimated as

$$\hat{q}_{1-1/m} = u + \frac{\hat{\sigma}}{\hat{\xi}}[\{m\widehat{\Pr}(X > u)\}^{\hat{\xi}} - 1]. \tag{3}$$

This expression for the POT method is used to estimate the upper quantiles (i.e. the level of the 100-year return period event), instead of the lower quantiles of concern in the rest of this paper. To estimate lower quantiles, we can apply the above procedure to $-X$. Buch-Kromann (2009)

expanded the idea of conditional ML to the Champernowne transformed kernel density estimator (Buch-Larsen *et al.*, 2005) and applied it to operational risk estimation. The Champernowne transformation is a popular method in quantitative operational risk models (see for example Bolancé *et al.* (2012)) for estimating the loss distribution in insurance. As the Champernowne transformation involves a more complex parameterization of the density and its application to quantile estimation is based on a conditional likelihood similar to the POT approach, we focus only on a comparison of the AC and POT approaches in this paper. Both the AC and the POT approaches can be viewed as assigning a parametric model to only part of the data, but using differently formulated likelihoods; the AC approach is based on the censored likelihood whereas the POT approach utilizes a conditional likelihood.

In practice both of the quantile estimation methods described above require the user to select a 'tuning constant'—the proportion of censoring in AC, and the upper threshold u in the POT method. In this paper we shall show that, for AC, a bootstrap-based selection method for the censoring proportion performs quite well when compared with the theoretically optimal choice. The comparison was done via simulation studies because in practice the true optimum is unknown.

Selecting the threshold for the POT approach is generally done by examining a plot of the mean excess against a sequence of threshold candidates and finding the largest above which the mean excess is approximately a linear function of u. Coles (2001) provided a detailed description for this selection method. A similar idea is used in selecting the threshold for the non-parametric estimation of copula tail dependence parameters (Dobrić and Schmid, 2005). We shall compare the statistical properties of both the AC and the POT quantile estimates where the POT threshold is selected so that equations (1) and (2) have the same number of observations.

The idea of using only either the largest or smallest values in the sample, whichever is appropriate, to estimate a tail-related parameter can, in principle, be applied well beyond the domain of quantile estimation. Thus, for example, we shall apply this approach to estimate the tail dependence parameter, which measures the tendency of coextremes to occur together in the tail of the joint distribution of two random variables. Some copula models for bivariate distributions result in explicit formulae for the tail dependence parameter, which can then be estimated by using MLEs for the copula parameters. Note that this setting presents the usual trade-off between efficiency (when the copula model is correct) and potential bias (when the chosen copula is not appropriate), and our approach attempts to find an appropriate subset of the sample where the model proposed provides a good fit (thereby possibly reducing bias) and the resulting estimates inherit the intrinsic stability of model-based inference, which contributes to the efficiency of the approach.

The rest of this paper is organized as follows. Section 2 studies AC in estimating lower quantiles, including the effects of parametric modelling and the amount of censoring. We also compare the AC, POT and the empirical quantile approaches by using extensive simulation studies in Section 2. Section 3 describes empirical ways of choosing the degree of censoring. The AC approach is applied to estimate the tail dependence parameters of copula models in Section 4. Section 5 includes a discussion of our findings along with our conclusions. Additional mathematical details, data sets used and simulation results for this paper are provided in the on-line supporting materials. We also implement the methods in this paper in R package `extWeibQuant` (Liu, 2014), which is available from the Comprehensive R Archive Network.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Artificial censoring to estimate lower quantiles

Estimating tail quantiles is important in many applications. Return levels (upper and lower quantiles) are used in many areas, such as hydrology, geosciences and finance, as an indicator of the seriousness of certain events such as floods, earthquakes and financial crises and are vital in the planning and preparing for those events (Coles, 2001). We shall focus here on a structural engineering context, where the fifth percentile of the distribution of the maximum load that a piece of lumber can sustain is used to find the 'design value'. In other words, to avoid structural failure the load on such pieces should not be larger than this value (Durrans and Triche, 1997). Our data consist of two experimental samples of wood strength (specifically, the modulus of rupture MOR) data, collected in a laboratory at FPInnovations, Vancouver, British Columbia. These measurements can be viewed as the survival time of wooden boards when pressure is applied vertically to its grain. The sample sizes are 98 and 282, and we shall refer to these two samples as MOR1 and MOR2 respectively.

The industrial standards document (American Society for Testing and Materials, 2015) states that, to estimate the fifth percentile of the distribution of these MORs, the largest 90% of the data should be censored and a two-parameter Weibull family used in equation (1). In principle, the resulting quantile estimates may vary appreciably depending on which family of distributions is used. For example, one could choose another distribution that is commonly used for modelling the survival times, such as the log-normal, gamma or Gumbel distributions (Lawless, 2003). To investigate the sensitivity of AC-based quantile estimates to this choice, we fitted these four models (Weibull, log-normal, gamma and Gumbel) to MOR1 and MOR2 with 90% of the largest observations artificially censored. In Fig. 1, the resulting probability density functions are overlaid on the histogram for each data set. Note that the estimated densities based on all four distribution families are almost indistinguishable in the lower tail of both data sets. Furthermore, Table 1 tabulates the estimates of the fifth percentile for these four models together with their standard errors (SEs) estimated by using 5000 bootstrap samples (Efron, 1979). We see that, for each data set, the quantile estimates that are obtained from using these four families of distributions are statistically indistinguishable. Thus the choice of the parametric model (1)

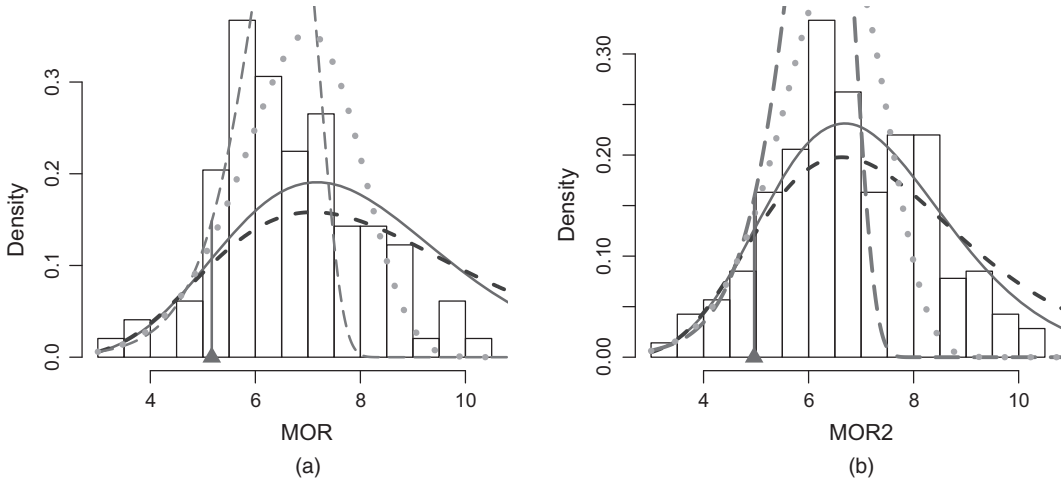

Fig. 1. Histograms of samples (a) MOR1 (▲, $C = 5.17$) and (b) MOR2 (▲, $C = 4.96$) together with four distributions fitted to the left tail with artificial censoring, where 90% of the largest observations are censored: |, 'censoring' thresholds $C = X_{(m)}$; · · · · · ·, Weibull; — — —, log-normal; ———, gamma; — — —, Gumbel

Table 1. Fifth-percentile estimates for two measures of the bending strength of lumber, MOR1 and MOR2, with various parametric distributions fitted with AC†

Sample	*Results for the following models*:			
	Weibull	*Log-normal*	*Gamma*	*Gumbel*
MOR1	4.51 [0.255]	4.47 [0.257]	4.48 [0.231]	4.53 [0.242]
MOR2	4.64 [0.141]	4.57 [0.138]	4.59 [0.138]	4.69 [0.140]

†The SEs of these estimates are calculated by the bootstrap method and are shown in the square brackets.

does not affect significantly the resulting quantile estimates based on our samples, and hence we shall follow the recommendation of the standard (American Society for Testing and Materials, 2015) and use a two-parameter Weibull family.

Another aspect of the AC methodology that may need to be validated is the proportion of censoring that is necessary to estimate a given quantile. For example, there are no details in American Society for Testing and Materials (2015) justifying the recommendation of censoring 90% of the data to estimate the fifth percentile. We carried out a panel of simulations to determine whether this censoring proportion is in fact optimal for estimating the fifth percentile, and whether other proportions should be used when estimating other quantiles. In addition, we investigated how the AC compares with the POT approach under various amounts of censoring or exceedance conditioning.

We compare quantile estimates by using their root-mean-square error (RMSE), which combines bias and efficiency. In our particular application, overestimation of the quantile carries a higher risk of structural damage whereas underestimation may lead to a waste of material and higher construction costs. Furthermore efficiency is important because load data in wood engineering are collected by breaking every board in a sample, which is destructive and expensive. Sample size can be an important factor influencing the bias–variance trade-off. Whereas sampling variability may be more prevalent for small sample sizes and more extreme quantiles (i.e. 1%), bias may be more important for larger samples and less extreme quantiles (5%). In this study we consider sample sizes 100, 300 and 500 for both first- and fifth-percentile estimation. These are realistic sample sizes used in estimating a lower quantile of the wood strength data (American Society for Testing and Materials, 2015; Durrans and Triche, 1997).

We generated data from eight parametric and non-parametric distributions mimicking our two real data sets, including the Weibull, log-normal, gamma and Gumbel distributions as before. Noting that the histograms in Fig. 1 suggest the possibility of bimodal distributions, we also included two-component mixtures of normal, log-normal and Weibull distributions. Finally, we simulated data from the kernel density estimate of the real data sets. Details about these choices can be found in section 1 of the on-line supporting material.

Denote by $r = m/n$ the proportion of observations that are not censored in the AC approach. The standard (American Society for Testing and Materials, 2015) specifies that $r = 0.1$ whereas $r = 1$ corresponds to the usual MLE. We compared the AC quantile estimator with $100(1-r)\%$ censoring with the POT estimator with the threshold $u = X_{[rn]}$, the $[rn]$th-order statistic of the sample, so that both approaches use the same numbers of observations in their 'likelihoods'. The empirical quantile is also included as a benchmark. We varied r betwen 0.1 and 1 for

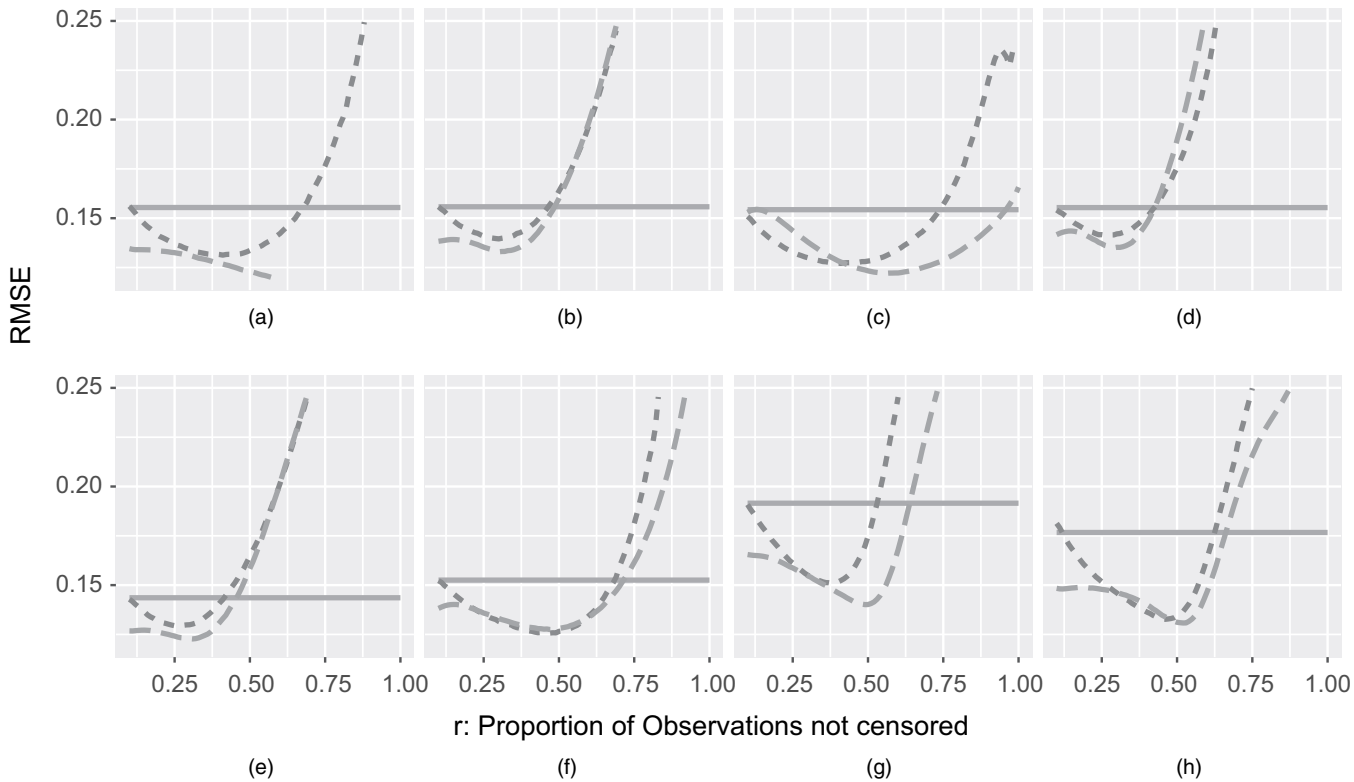

Fig. 2. RMSE of the fifth-percentile estimates with sample size 300 from using the AC approach (— —) and POT approach (– – –) with various amounts of censoring (proportion of observations used in the parametric part) (———, empirical quantile): (a) Weibull; (b) gamma; (c) Gumbel; (d) log-normal; (e) normal mixture; (f) log-normal mixture; (g) Weibull mixture; (h) kernel density estimate

estimating the fifth percentile and between 0.02 and 1 for the first-percentile estimates. For each combination of distribution and sample size we generated 10 000 independent and identically distributed samples and for each value of r computed both the AC- and the POT-based fifth- and first-percentile estimators.

Sections 2.1 and 2.2 report the results for the fifth- and first-percentile estimates respectively from the models mimicking the MOR2 data. The results from the models mimicking MOR1 were similar and thus have been omitted.

2.1. Bias, variance and root-mean-square error of fifth-percentile estimates

Fig. 2 depicts the RMSE for the fifth-percentile estimates with various rs obtained by using both AC and POT approaches. The curves of the RMSE are mostly 'V' shaped. The RMSE for the AC and POT approaches can be large when too few or too many observations are censored or conditioned. To illustrate the trends in the RMSE better, the curves of absolute bias and SE are provided in Figs 3 and 4 respectively.

Fig. 3 shows clearly that the biases of the POT and AC methods are extremely large when r is close to 1. This is because, when too few observations are censored or conditioned, the two approaches lose their focus on the lower tail. As the likelihood is misspecified in most of the models that we considered, the AC and POT likelihoods fail to provide reasonable approximations of the lower tail, which causes enormous bias. Clearly, the exception is the AC approach when the data are generated from a Weibull distribution. As the likelihood is correct in this case, the MLE is known to be asymptotically unbiased and most efficient.

Fig. 4 shows that the SE of the AC estimates generally decreases with increases in r, as more observations are included in the parametric part of the likelihood. In contrast, the SE curves of the POT estimates are mostly 'U' shaped. This can be explained by the fact that its quantile estimate depends on an empirical quantile as in equation (3), whose variance is large for both small and large r.

The combination of bias and SE explains the changes in the RMSE. When r is small (less than 0.3), the variances of both POT and AC approaches decrease whereas the biases remain similar, which results in a decreasing RMSE. In contrast, when $r > 0.3$, more non-tail observations are used for parametric estimation and that reduces the accuracy of the approximation to the tail, which in turn causes a sharp increase in the bias. The minimum RMSE in both approaches is achieved when $r \in (0.25, 0.5)$ for all except one of the models that we considered. The exception is the Weibull model, where the asymptotically most efficient MLE is guaranteed to achieve the smallest RMSE. From Fig. 2, we also learn that the RMSE of the quantile estimates from the POT and AC approaches is smaller than that of the empirical quantile when r is small (roughly smaller than 0.5). This advantage is gained by reducing the variance of the estimates (Fig. 4) while keeping the bias low (Fig. 3).

We find that the AC approach is usually relatively better than the POT approach when r is very small. The RMSE of the POT approach decreases faster than that of the AC approach when r increases. This illustrates a key property of the POT approach: its efficiency depends on both the parametric estimate in the tail and also its threshold, which is an empirical quantile. When r increases (the threshold increases), both the parametric estimates and the threshold quantile become more efficient. In contrast, the improvement in efficiency in the AC approach can only contribute through the parametric estimate.

This difference can also be seen in the RMSE curves with sample sizes 100 and 500 in the on-line supporting material section 3. When the sample size is 100, the difference between the POT and AC approaches by using 10% of the observations is larger than their difference when the sample size is 300, as the threshold quantile in the POT approach has an extra large variance.

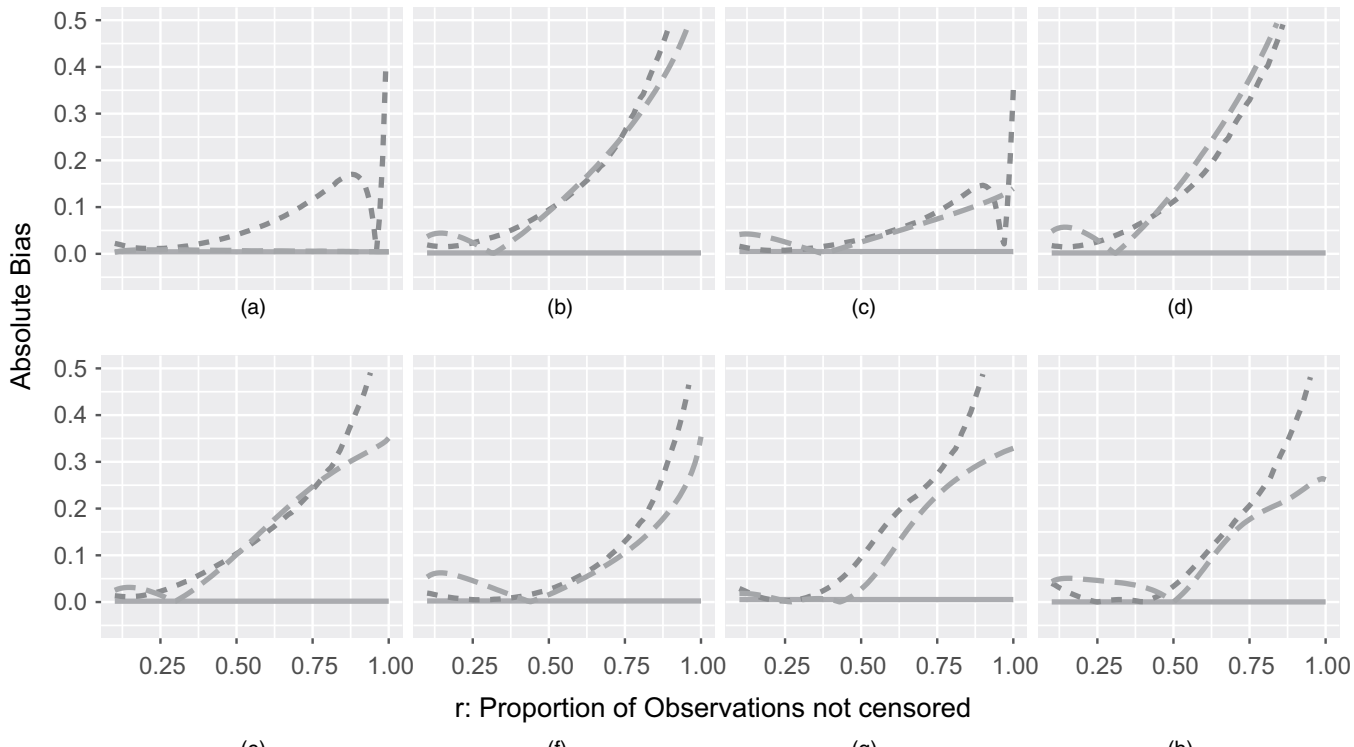

Fig. 3. Absolute bias of the fifth-percentile estimates with sample size 300 by using the AC approach (— —) and POT approach (– – –) with various amounts of censoring (proportion of observations used in the parametric part) (———, empirical quantile): (a) Weibull; (b) gamma; (c) Gumbel; (d) log-normal; (e) normal mixture; (f) log-normal mixture; (g) Weibull mixture; (h) kernel density estimate

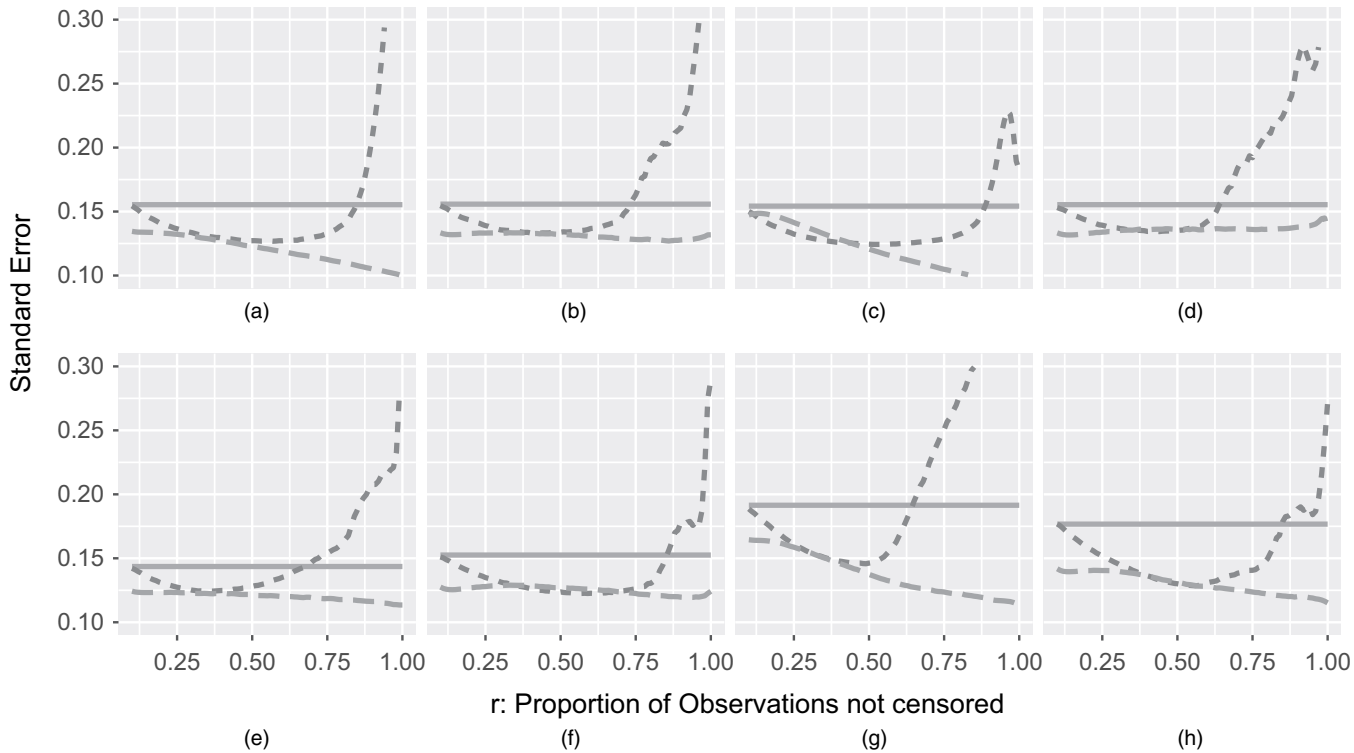

Fig. 4. Standard error of the fifth-percentile estimates with sample size 300 by using the AC approach (— —) and POT approach (– – –) with various amounts of censoring (proportion of observations used in the parametric part) (———, empirical quantile): (a) Weibull; (b) gamma; (c) Gumbel; (d) log-normal; (e) normal mixture; (f) log-normal mixture; (g) Weibull mixture; (h) kernel density estimate

This difference becomes smaller when the sample size is 500 and the threshold quantile becomes more efficient.

Summarizing the results for the RMSE assessment of the fifth-percentile estimates under all three sample sizes, we find that the minimum RMSE is rarely achieved when $r > 0.5$. The AC approach achieves a smaller minimum RMSE than the POT approaches under most settings that we considered. For those settings where the POT approach has a smaller minimum RMSE, as in the case of the log-normal mixture model with a sample size of 300, its advantage is very small. Moreover, when we consider an arbitrary fixed proportion, the RMSE of the AC method is often smaller, as its curve is usually below the curve of the POT approach. The exceptions are seen in the case of the Gumbel and log-normal mixture model, where the POT method is better when 15–30% of observations are used. Yet, the RMSE for the AC remains close to that of the POT approach even in those situations. This leads us to conclude that the AC approach is generally better than the POT approach for estimating the fifth percentile.

2.2. Root-mean-square error of first-percentile estimates

We use a sample size of 300 to illustrate the first-percentile curves in Fig. 5 and the curves for sample sizes of 100 and 500 are included in on-line supporting material section 3. These plots show that the curves for the AC estimates of the first-percentile estimates are still V shaped like those for the fifth percentiles. But those for the POT estimates are often U shaped, i.e. they remain almost the same for a wide range of thresholds from 20% to 50%, where the changes in bias and variance of the POT approach are tied. The balance is broken only when the threshold is larger than 50%, where the bias increases dramatically. In contrast, the bias of the AC approach for the first-percentile estimate increases abruptly at a relatively smaller threshold around 25% for four models—gamma, log-normal, normal mixture and log-normal mixture. This indicates that a relatively small amount of non-tail data can make the AC approach lose its focus on the tail, whereas the POT approach remains a good approximation under these circumstances.

However, as in the case of the fifth percentile, the minimum RMSE of the AC's first-percentile estimate is smaller than that of the POT approach. Also, when the proportion r is small, the AC approach has a much smaller RMSE than that of the POT approach. Thus the AC approach maintains its advantage in this way over the POT approach as long as the proportions of observations used are small, i.e. less than 10%.

To conclude, these simulations suggest that, given a reasonable r, the semiparametric POT and AC approaches can clearly outperform the MLE ($r = 1$) and the non-parametric empirical estimates. Their advantage is achieved by a good balance in their bias–variance trade-offs. The AC and POT results might be slightly biased when compared with the empirical quantile, but they are more efficient than it. The MLE may be more efficient than the AC and POT approaches, but at the potential cost of serious bias when the model is misspecified. Also, the AC approach can provide quantile estimates with smaller RMSEs than those of the POT approach in most of the cases that we considered.

3. Data-driven methods to select the censoring proportion

The simulation results that were described in the previous section show that the amount of censoring plays a crucial role in determining the statistical properties of the resulting AC-based quantile estimator. For example, a larger r than the 10% recommended by the industry standard can reduce the RMSE of the estimator. Here we explore two empirical approaches to choosing r. The first is based on visual graphical examination (Section 3.1) and, the second, a bootstrap estimate of the mean-square error (MSE) (Section 3.2). Section 3.3 compares the RMSE of the

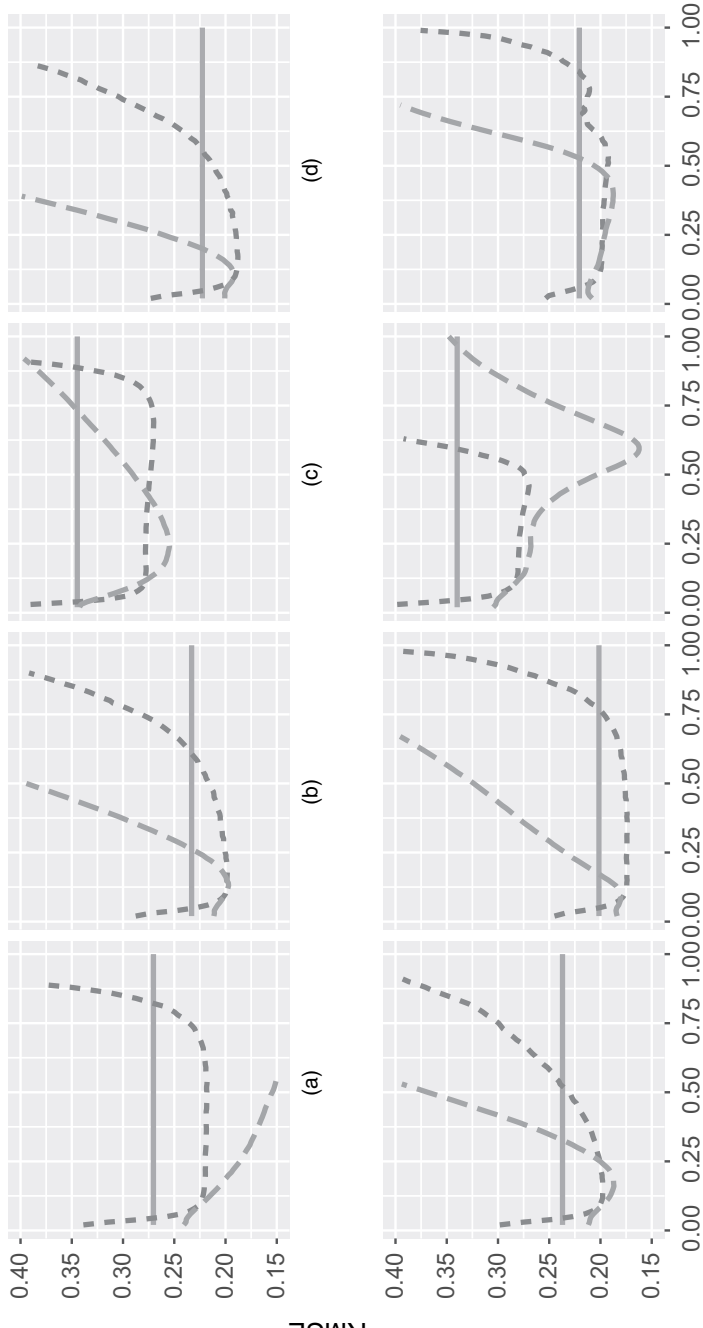

Fig. 5. RMSE of the first-percentile estimates with sample size 300 from the AC approach (— —) and POT approach (– – –) with various amounts of censoring (proportions of observations used in the parametric part) (———, empirical quantile): (a) Weibull; (b) gamma; (c) Gumbel; (d) log-normal; (e) normal mixture; (f) log-normal mixture; (g) Weibull mixture; (h) kernel density estimate

AC quantile estimators based on these two data-driven r selection methods with the 'oracle' obtained when r is set to its optimal value based on our simulations from the previous section.

3.1. Amount of censoring selection via graphical examination

The threshold u that is required to use the POT approach can be chosen visually on the basis of the plot of a sequence of threshold candidates $u_1 < u_2 < u_3 \ldots$, *versus* the corresponding mean excesses (the mean of the observations that are larger than u_i). Coles (2001) recommended selecting the largest candidate u_i such that the mean excess is approximately linear in u for $u > u_i$. Similarly, Dobrić and Schmid (2005) selected the threshold for the non-parametric tail dependence parameter estimates by locating the 'elbow' of the plot of threshold candidates $u_1 < u_2 < u_3 \ldots$ *versus* the resulting estimates. Following these ideas, Fig. 6 contains a plot of the AC fifth-percentile estimates as a function of the proportion of non-censoring r for both our real data sets. For both data sets, the quantile estimates begin to decrease dramatically when r is larger than 0.4. If we refer to Fig. 2, we see that the the RMSE of the AC estimates with $r = 0.4$ is indeed smaller than those obtained when $r = 0.1$ for most of our simulation settings. This observation suggests that we transform this visual procedure into the following *algorithm 1*.

(a) Predetermine a set of increasing threshold candidates $0 < r_1 < r_2 < r_3 < \ldots < r_i, \ldots, r_K \leqslant 1$. For each i, calculate the AC quantile estimate $\tilde{q}(r_i)$ with threshold r_i.
(b) Smooth the $\tilde{q}(r_i)$, $i = 1, \ldots, K$, with the locally weighted polynomial regression smooth LOWESS. Denote the smoothed estimates at r_i with s_i.
(c) Calculate the numerical absolute derivatives $d_i = |s_{i+1} - s_i|/(r_{i+1} - r_i)$ for $i = 1, 2, \ldots, K - 1$ and then rescale d_i into $(0, 1)$, which are denoted by $\tilde{d}_i$.
(d) Find the first $\tilde{d}_i > 0.15$ and choose r_{i-1} as the threshold. If $\tilde{d}_1 > 0.15$, choose r_1 as the threshold.

In step 2, we use the default choices of the function `lowess` in R (R Core Team, 2015). The threshold '0.15' in step (d) was chosen so that the algorithm choices when applied to our MOR1 and MOR2 data sets (shown by the broken vertical lines in Fig. 6) correspond to $r = 0.40$. These algorithm settings can also reproduce our visual choice reasonably well in a few simulated data sets. We refer to this method as graphical artificial censoring (GAC).

3.2. Amount of censoring selection via the bootstrap

An alternative approach to the visual method that was described above is to select the threshold that results in a quantile estimator that minimizes a bootstrap estimate of its RMSE. This idea has been proposed in the context of the POT estimators (see for example Hall (1990) and Qi (2008)). For the AC quantile estimators, we would ideally like to find the value $\hat{r}$ that minimizes the true MSE of the quantile estimator:

$$E_F\{\tilde{q}(r) - q_\alpha\}^2,$$

where F is the true distribution of the data, q_α is the α-quantile of interest and $\tilde{q}(r)$ is the AC estimate. Since the true distribution F is unknown in practice, we replace F by its empirical counterpart $\hat{F}_n$ and the true quantile q_α by its empirical estimate $\hat{q}_\alpha$. A bootstrap estimator of the MSE above can then be chosen to be

$$E_{F_n}\{\tilde{q}^*(r) - \hat{q}_\alpha\}^2.$$

This MSE estimator can be approximated as follows: for a fixed r, obtain a bootstrap sample (i.e. a sample from the empirical distribution) and compute the AC quantile estimate. Repeat

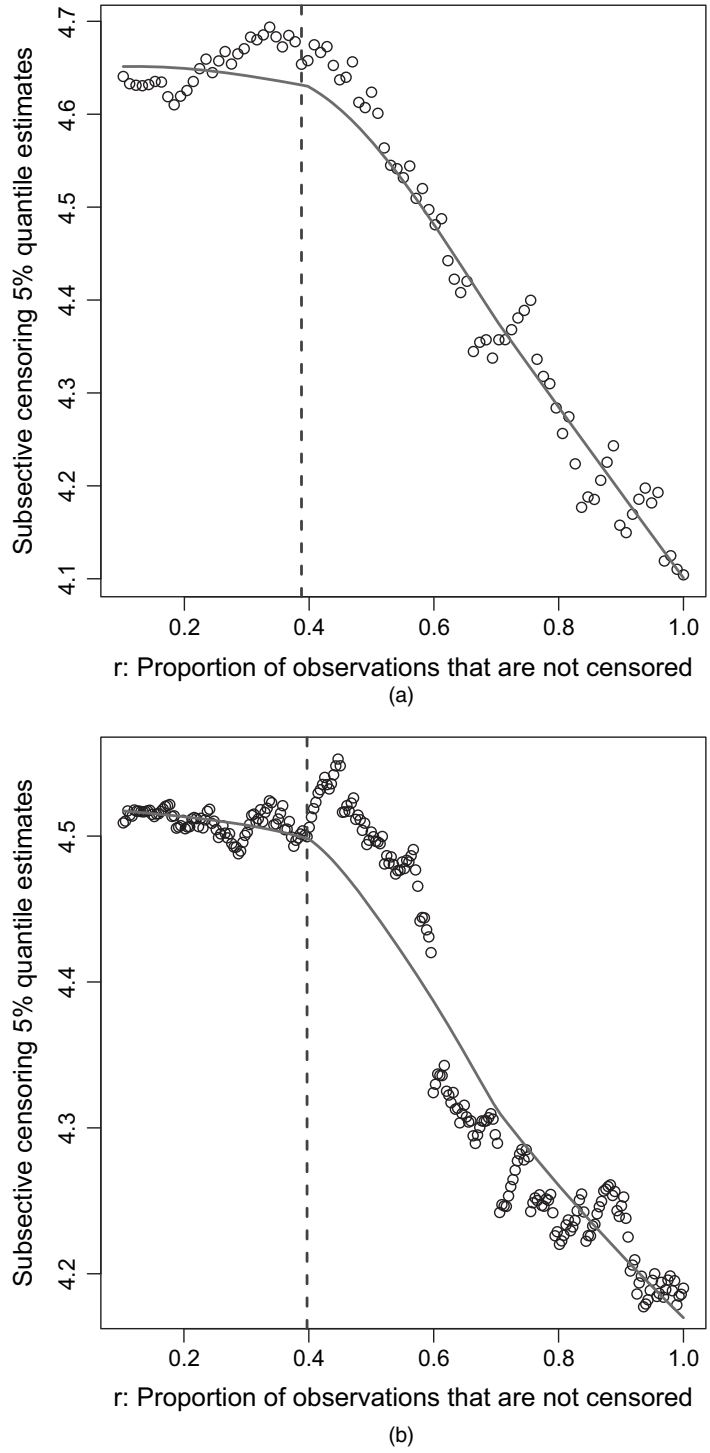

Fig. 6. Fifth-percentile estimate from the AC approach with various censoring thresholds r for the two real data sets (a) MOR1 and (b) MOR2: ———, curve smoothed by the LOWESS function in R (R Core Team, 2015); ¦, selected threshold

this process a large number B of times (e.g. $B = 5000$) and denote the resulting quantile estimates by $\tilde{q}_i^*(r)$, $1 \leqslant i \leqslant B$. Then,

$$E_{F_n}\{\tilde{q}^*(r) - \hat{q}_\alpha\}^2 \approx \frac{1}{B}\sum_{i=1}^{B}\{\tilde{q}_i^*(r) - \hat{q}_\alpha\}^2. \tag{4}$$

We use the above bootstrap procedure to estimate the MSE of the AC quantile estimators under a predetermined set of r candidates and select the estimator with the smallest bootstrap MSE. We refer to the AC quantile estimators by using a threshold selected via approximation (4) as bootstrap artificial censoring (BAC) estimators. In the following section we use a simulation study to compare their performance with that of their GAC and oracle counterparts. This serves as an empirical assessment of the bootstrap procedure, and we shall address the asymptotic properties of this bootstrap estimation in future work. On the basis of studies of bootstrap estimators in similar tail estimation problems, such as Hall (1990), Hall and Weissman (1997) and Gámiz *et al.* (2016), we conjecture that estimator (4) is asymptotically consistent.

3.3. Simulation comparison

We simulated 10 000 data sets from the same eight distributions that were used above, using samples of size $n = 300$. For the graphical threshold selection method (GAC) we considered the candidate thresholds $r = 0.1, 0.1 + 1/n, 0.1 + 2/n, \ldots, 1$, whereas for the bootstrap threshold selection method (BAC) we considered $r = 0.1, 0.2, 0.3, 0.4, 0.5$, as the minimum RMSE is rarely achieved with $r > 0.5$ in Fig. 2. We used $B = 5000$ bootstrap replicates to estimate the MSE in estimator (4).

The RMSE, bias and SE of the oracle AC estimator and our two data-driven proposals, GAC and BAC, are summarized in Tables 2, 3 and 4 respectively. The empirical quantile is introduced as a benchmark for this comparison. The GAC and BAC estimates have smaller RMSEs than does the empirical quantile mainly because of a smaller variance (Table 4). Also GAC and BAC achieve similar RMSEs, biases and SEs to those of the oracle AC estimates in all the eight models that we considered. The largest difference between the RMSEs is just 0.03. In some cases, such as the gamma, and normal mixture models, the GAC and BAC estimates achieve smaller biases than those of the oracle AC estimator. The SEs of the GAC and BAC quantile estimators are usually larger than those of the oracle counterparts. This is to be expected as the oracle AC approach uses a fixed threshold r, whereas the GAC and BAC methods introduce extra variability into the quantile estimates as the threshold r is random in these two procedures. Overall we find that the RMSE of GAC and BAC are comparable with that of the oracle estimate.

Table 2. RMSE of the empirical quantile EMP, oracle AC, GAC and BAC estimates of the fifth percentile given a sample size of 300†

Model	*EMP*	*Oracle estimate*	*GAC estimate*	*BAC estimate*
Weibull	0.15	0.10	0.13	0.13
Log-normal	0.16	0.14	0.15	0.14
Gamma	0.16	0.13	0.13	0.14
Gumbel	0.15	0.12	0.15	0.15
Normal mixture	0.15	0.12	0.13	0.13
Log-normal mixture	0.15	0.13	0.13	0.14
Weibull mixture	0.19	0.14	0.17	0.16
Kernel density estimate	0.18	0.13	0.14	0.14

†Note here that the bias and SE for the oracle estimates are the bias and SE at the oracle proportion r that minimizes the RMSE, not the minimum bias or SE.

Table 3. Absolute bias of the empirical quantile EMP, oracle AC, GAC and BAC estimates of the fifth percentile given a sample size of 300†

Model	*EMP*	*Oracle estimate*	*GAC estimate*	*BAC estimate*
Weibull	0.00	0.00	0.01	0.01
Log-normal	0.00	0.01	0.04	0.02
Gamma	0.00	0.04	0.01	0.01
Gumbel	0.01	0.01	0.04	0.03
Normal mixture	0.00	0.04	0.00	0.00
Log-normal mixture	0.00	0.01	0.04	0.04
Weibull mixture	0.01	0.02	0.01	0.01
Kernel density estimate	0.00	0.01	0.04	0.04

†Note here that the bias and SE for the oracle estimates are the bias and SE at the oracle proportion r that minimizes the RMSE, not the minimum bias or SE.

Table 4. SE of the empirical quantile EMP, oracle AC, GAC and BAC estimates of the fifth percentile given a sample size of 300†

Model	*EMP*	*Oracle estimate*	*GAC estimate*	*BAC estimate*
Weibull	0.15	0.10	0.13	0.13
Log-normal	0.16	0.13	0.14	0.14
Gamma	0.16	0.13	0.13	0.14
Gumbel	0.15	0.12	0.14	0.14
Normal mixture	0.15	0.12	0.13	0.13
Log-normal mixture	0.15	0.13	0.13	0.13
Weibull mixture	0.19	0.14	0.17	0.16
Kernel density estimate	0.18	0.13	0.14	0.14

†Note here that the bias and SE for the oracle estimates are the bias and SE at the oracle proportion r that minimizes the RMSE, not the minimum bias or SE.

4. Subjective censoring in estimating copula tail properties

Besides the quantiles, another important extreme tail quantity is the tail dependence parameter in copula models, which measures the tendency of extreme co-movements in the lower or upper tail of two variables. This parameter is important in management of risk during extreme events (Dobrić and Schmid, 2005). Since the idea of artificial censoring works well in estimating the lower quantile, we therefore explore its use in estimating the tail dependence parameter.

A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform and it is used to describe the dependence between random variables. To be more precise, let X and Y denote two random variables with a joint distribution function

$$F_{X,Y}(x, y) = \Pr(X \leqslant x, Y \leqslant y)$$

and denote the marginal distributions of X and Y by F_X and F_Y respectively. A copula is a distribution function $C : [0, 1]^d \to [0, 1]$ with uniform $[0, 1]$ margins that satisfies

$$F_{X,Y}(x, y) = C\{F_X(x), F_Y(y); \delta\}.$$

Equivalently, $C(u, v)$ is the joint distribution function of the random variables $U = F_X(X)$ and $V = F_Y(Y)$,

$$C(u, v; \delta) = \Pr(U \leqslant u, V \leqslant v), \tag{5}$$

where δ presents the possible parameters in the copula model. For example, the Galambos copula function is defined as

$$C(u, v; \delta) = uv \exp[\{-\log(u)\}^{-\delta} + \{-\log(v)\}^{-\delta}]^{-1/\delta}.$$

Thus the copula describes the dependence between the two variables in a manner that does not depend on the marginal distributions. See Joe (2014) for a more detailed introduction to copulas.

The upper tail dependence parameter λ_U for a bivariate copula is defined as

$$\lambda_U = \lim_{u \to 0} \frac{\bar{C}(1-u, 1-u)}{u}, \tag{6}$$

provided that the limit on the right-hand side exists and $\bar{C}(u, v; \delta) = \Pr(U > u, V > v)$ is the survival copula (Joe, 2014). The lower tail dependence parameter can be defined similarly, i.e. $\lambda_L = \lim_{u \to 0} C(u, u)/u$. λ_U and λ_L describe the extreme co-movements of X and Y in the tails. For example, financial risk managers are concerned about the tendency of two assets to perform extreme co-movements, i.e. whether an extreme loss in one asset entails an extreme loss in another and vice versa (Dobrić and Schmid, 2005). The tail dependence parameters are essential in managing risk as global dependence measures, such as the correlation (Pearson, Spearman or Kendall's τ), are known to be unsuitable for measuring extreme co-movements in the lower tails (Embrechts *et al.*, 2002).

The tail dependence parameter can be estimated on the basis of the MLE of parameters when the copula has an explicit expression for the tail dependence parameter, as in the Gumbel, Joe and Galambos copulas (Joe, 2014). However, not all copula families have this property and there is always the issue of model misspecification. To overcome the limitation of parametric estimates, several non-parametric estimates have been proposed and studied in Dobrić and Schmid (2005) and their performances are very similar. Yet what we have learned from estimating the extreme quantiles suggests that those non-parametric estimates may be inefficient. This motivates the use of artificial censoring here.

As in the case of quantile estimates, non-parametric and fully parametric estimates lack efficiency and robustness to model misspecification respectively. Therefore, we work with the idea of AC to find 'semiparametric' estimates that achieve a good balance between minimizing the bias and variance. Since the lower and upper tails are formally interchangeable by transforming the data, we need to work with only the upper tail dependence parameter in this paper. To set up artificial censoring, we first partition the domain of the random variables U and V in equation (5), $[0, 1]^2$, into the following four quadrants:

$$\begin{aligned}
\mathcal{S}_1 &= \{(u, v) | u > T_u, v > T_v\}, \\
\mathcal{S}_2 &= \{(u, v) | u \leqslant T_u, v > T_v\}, \\
\mathcal{S}_3 &= \{(u, v) | u \leqslant T_u, v < T_v\}, \\
\mathcal{S}_4 &= \{(u, v) | u > T_u, v \leqslant T_v\},
\end{aligned}$$

where $T_u, T_v \in (0, 1)$ decide the amount of censoring in the two dimensions (i.e. the proportion r in two dimensions). We treat observations that fall in $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$ as censored and observations in $\mathcal{S}_1$ as observed. For an independent and identically distributed bivariate sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we first transform the observations with their empirical CDFs $\hat{F}_n$

and $\hat{H}_n$ respectively,

$$u_i = \hat{F}_n(x_i),$$
$$v_i = \hat{H}_n(y_i),$$

so that (u_i, v_i) has uniform $[0, 1]$ margins. We denote the transformed sample by $\mathbf{W}$ and the copula CDF and probability density function by $C(\cdot, \cdot; \delta)$ and $c(\cdot, \cdot; \delta)$, where δ is the parameter(s) in the copula model. Let m_k, $k = 1, 2, \ldots, 4$, be the count of observations in $\mathcal{S}_k$, so that the AC copula likelihood is

$$\begin{aligned} L(\delta; \mathbf{W}) &= A \left\{ \prod_{i=1}^{m_1} c(u_i, v_i; \delta) \right\} \Pr(u < T_u, v > T_v)^{m_2} \Pr(u < T_u, v < T_v)^{m_3} \Pr(u > T_u, v < T_v)^{m_4} \\ &= A \left\{ \prod_{i=1}^{m_1} c(u_i, v_i; \delta) \right\} \{T_u - C(T_u, T_v; \delta)\}^{m_2} C(T_u, T_v; \delta)^{m_3} \{T_v - C(T_u, T_v; \delta)\}^{m_4}, \end{aligned} \tag{7}$$

where

$$A = \binom{n}{m_1} \binom{n - m_1}{m_2} \binom{n - m_1 - m_2}{m_3}$$

is a constant that is invariant to the specification of the parameter δ in the likelihood. To estimate δ, we can numerically maximize the log-likelihood,

$$\begin{aligned} l(\delta; \mathbf{W}) = &\sum_{i=1}^{m_1} \log\{c(u_i, v_i; \delta)\} + m_2 \log\{T_u - C(T_u, T_v; \delta)\} + m_3 \log\{C(T_u, T_v; \delta)\} \\ &+ m_4 \log\{T_v - C(T_u, T_v; \delta)\}. \end{aligned}$$

The likelihood (7) can be viewed as a bivariate extension of equation (1). The censoring thresholds T_u and T_v control how many observations are censored. It is straightforward to show that, when $T_u = T_v = 0$, the above likelihood becomes the uncensored likelihood and maximizing it yields the MLE. Yet, the AC approach loses its dependence on the upper tail when $1 - T_u$ and $1 - T_v$ increase, as more observations from the non-tail area are used, which might lead to bias in the tail dependence parameter estimates. Here we do not have a specific rule to choose the censoring thresholds T_u and T_v and we set them to be $T_u = T_v = 1 - 1/\sqrt{n}$: the same cut-offs as used in the non-parametric estimate introduced in what follows.

In this paper, we consider four one-parameter copula models, Galambos, Gumbel, Joe and MTCJ (Joe, 2014), for likelihood (7), which are chosen on the basis of their popularity and simplicity in parameterization. Some brief introduction is included in on-line supporting material section 2 and a full treatment can be found in Joe (2014). For the Galambos and MTCJ copulas, the tail dependence parameter is

$$\lambda_{\mathrm{U}} = 2^{-1/\delta},$$

whereas, for the Gumbel and Joe copulas, it is

$$\lambda_{\mathrm{U}} = 2 - 2^{1/\delta}.$$

These formulae are used to calculate the tail dependence parameter estimates based on the parameter estimates of δ, which can be either the full MLE without any censoring (denoted by MLE) or the artificially censored MLE (denoted by SC). For the non-parametric estimates, we consider the following type I estimate from Dobrić and Schmid (2005):

$$\hat{\lambda}_{\mathrm{U}} = \frac{1}{k} \sum_{i=1}^{n} I\left(u_i > 1 - \frac{k}{n}, v_i > 1 - \frac{k}{n}\right) = \left(\frac{k}{n}\right)^{-1} \frac{1}{n} \sum_{i=1}^{n} I\left(u_i > 1 - \frac{k}{n}, v_i > 1 - \frac{k}{n}\right). \tag{8}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n} I\left(u_i > 1-\frac{k}{n}, v_i > 1-\frac{k}{n}\right)$$

is an empirical estimate of $\bar{C}(1-u, 1-u)$ at $u = 1/\sqrt{n}$. With this definition, the empirical estimate was shown to be marginally better than the other two types (Dobrić and Schmid, 2005). They also showed that choosing $k = \sqrt{n}$ ensures the consistency of the non-parametric estimates and we follow this recommendation in this paper.

4.1. Simulation study

In this simulation study, we generated bivariate data from the four copula models. For each data set, we calculated the tail dependence parameter estimates from the non-parametric approach, the fully parametric approach (MLE) and the AC approach. The MLE and AC approaches treated all four copula models in the likelihood. For example, we first generated a data set from the Galambos copula and then fitted the Galambos, Gumbel, Joe and MTCJ copulas to this data set with either ML estimation or AC, and calculated the dependence parameter estimates for each fitted model, which results in eight estimates besides the non-parametric estimate. The sample size was chosen to be 1000 and we simulated 10 000 replicates under each model. Simulations were performed at various parameter values of each model and similar conclusions regarding the comparison of these different estimates were found. We therefore present only the result for one set of models of randomly chosen parameters.

Table 5 tabulates the RMSE of the tail dependence parameter estimates obtained by all approaches. A comparison of the second column with the rest of Table 5 shows that the non-parametric estimate is inferior to those parametric estimates as its RMSE is always larger than those of the MLE and AC estimates. When comparing the MLE and AC approaches, it is not surprising that the MLE is strikingly better than the AC estimate when the model is correctly specified. However, the RMSEs of the MLEs can be 2–4 times larger than those of the AC estimates under misspecified models. For example, when the data are generated from the Joe copula but the likelihood uses the Galambos copula, the RMSE of the ML approach (8.67) is around four times that of the RMSE of the AC approach (2.10).

To compare these estimates further and to study the decomposition of the RMSE, we plot the boxplots of the nine quantile estimates when the data are generated from the Joe copula, i.e. the third row of Table 5 (Fig. 7). Similar plots are found for the other data-generating

Table 5. RMSE (×100) of various tail dependence estimates under different copula models†

Model	*Non-parametric estimate*	*Galambos*		*Gumbel*		*Joe*		*MTCJ*	
		MLE	*AC*	*MLE*	*AC*	*MLE*	*AC*	*MLE*	*AC*
Galambos	15.99	0.44	2.48	0.44	2.50	4.71	2.47	4.58	2.44
Gumbel	15.50	0.64	3.47	0.64	3.52	6.10	3.46	5.80	3.42
Joe	16.22	8.67	2.10	8.64	2.12	0.41	2.09	0.44	2.08
MTCJ	15.78	9.93	3.12	9.92	3.18	0.72	3.10	0.65	3.05

†The row name indicates the data-generating model and the column names indicate the model used in the ML estimation (or AC estimation).

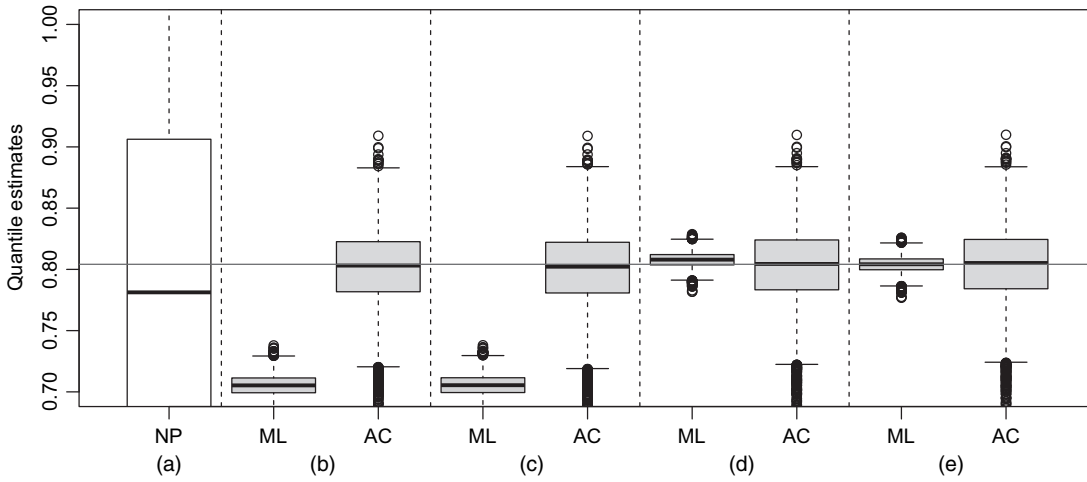

Fig. 7. Boxplot of the quantile estimates when the data were generated from the Joe copula: (a) non-parametric estimates; (b) Galambos copula; (c) Gumbel copula; (d) Joe copula; (e) MTCJ copula

models and have thus been omitted. Clearly, the non-parametric approach is inefficient with huge variances. This is caused by the fact that there are few observations in the extreme upper tail and their contribution to the estimate in equation (8) is just a 0–1 indicator. The vast amount of information in the non-tail part of the sample and the numerical value of (u_i, v_i) are completely missed in non-parametric estimation. Meanwhile, the ML and AC estimates take advantage of this information and are thus more efficient, as shown by the boxes in Fig. 7.

In the comparison of the ML and AC approaches, the AC approach is found to have much larger variances; it censors 92% observations on average in our simulations, i.e. only 8% of the observations fall in $\mathcal{S}_1$. This loss of efficiency explains why the RMSE of the AC approach is around five times larger than those of the MLE when the likelihood model is correctly specified and the RMSE is dominated by the variance. However, when the likelihood model is misspecified, the MLE incurs a huge bias in the tail dependence estimate, which causes it to have a 3–5 times larger RMSE.

Another interesting finding in this simulation is that the MLE with MTCJ likelihood performs reasonably well when the data are generated from the Joe copula (see Fig. 7(e)), or vice versa (see the last four column entries for the last row of Table 5). A similar result obtains for the Galambos and Gumbel copulas. This may tell us that the Galambos and Gumbel copulas, or Joe and MTCJ copulas, are 'exchangeable' in the extreme tail. However, this conclusion is somewhat speculative and will be explored further in future work.

4.2. Case-study of stock market data

To illustrate our AC approach, we estimated the tail dependence parameters for the stocks of the following 10 companies:

(a) the technology companies Apple Inc, AAPL, and Yahoo! Inc., YHOO;
(b) the energy companies Chevron, CVX, British Petroleum, BP, and Exxon Mobil Corporation, XOM;
(c) one car company, the Ford Motor Company, FORD;
(d) the pharmaceutical companies Merck & Co., Inc., MRK, Eli Lilly and Co., LLY, GlaxoSmithKline PLC, GSK, and Pfizer, PFE.

The daily closing prices of these equities during the period from January 1st, 2001, to February 1st, 2016, were downloaded from Yahoo Finance (`http://finance.yahoo.com/`), which has 3792 trading days in total. As the simultaneous extreme declines are usually more important than the simultaneous rises, we calculated negative log-returns,

$$\log(\text{closing price in day } i/\text{closing price in day } i+1),$$

so that the upper tail dependence parameter measures the stocks' co-movement in the declines. The univariate marginal distribution of each stock's negative log-return was estimated by the empirical CDF and we calculated the non-parametric and the AC estimates for each bivariate pair of stocks. The threshold k in both estimates was fixed at $\sqrt{n}$, where n is the sample size. The non-parametric estimates are tabulated in Table 6. For the AC approach, all four copulas were compared by their deviance and only the estimates from the model with the smallest deviance were reported, which are summarized in Table 7.

Table 6. Non-parametric estimates of lower tail dependence of stock daily log-returns†

Company	*AAPL*	*YHOO*	*CVX*	*BP*	*XOM*	*FORD*	*MRK*	*LLY*	*GSK*	*PFE*
AAPL		0.15	0.16	0.13	0.13	0.10	0.10	0.11	0.11	0.11
YHOO	0.15		0.11	0.15	0.11	0.06	0.05	0.13	0.10	0.11
CVX	0.16	0.11		0.47	0.61	0.23	0.27	0.34	0.34	0.37
BP	0.13	0.15	0.47		0.47	0.21	0.23	0.29	0.32	0.26
XOM	0.13	0.11	0.61	0.47		0.19	0.26	0.35	0.29	0.34
FORD	0.10	0.06	0.23	0.21	0.19		0.13	0.16	0.16	0.21
MRK	0.10	0.05	0.27	0.23	0.26	0.13		0.40	0.26	0.34
LLY	0.11	0.13	0.34	0.29	0.35	0.16	0.40		0.32	0.37
GSK	0.11	0.10	0.34	0.32	0.29	0.16	0.26	0.32		0.31
PFE	0.11	0.11	0.37	0.26	0.34	0.21	0.34	0.37	0.31	

†The companies in different sectors are separated by the vertical or horizontal lines.

Table 7. Lower tail dependence of stock daily log-returns, estimated by the AC approach†

Company	*AAPL*	*YHOO*	*CVX*	*BP*	*XOM*	*FORD*	*MRK*	*LLY*	*GSK*	*PFE*
AAPL		0.13	0.15	0.12	0.12	0.09	0.08	0.10	0.10	0.11
YHOO	0.13		0.10	0.13	0.10	0.05	0.03	0.11	0.09	0.10
CVX	0.15	0.10		0.54	0.68	0.21	0.28	0.35	0.36	0.38
BP	0.12	0.13	0.54		0.51	0.20	0.23	0.30	0.34	0.28
XOM	0.12	0.10	0.68	0.51		0.18	0.26	0.37	0.31	0.36
FORD	0.09	0.05	0.21	0.20	0.18		0.12	0.15	0.17	0.21
MRK	0.08	0.03	0.28	0.23	0.26	0.12		0.41	0.28	0.36
LLY	0.10	0.11	0.35	0.30	0.37	0.15	0.41		0.34	0.39
GSK	0.10	0.09	0.36	0.34	0.31	0.17	0.28	0.34		0.33
PFE	0.11	0.10	0.38	0.28	0.36	0.21	0.36	0.39	0.33	

†The companies in different sectors are separated by the vertical or horizontal lines.

The estimates from the non-parametric approach and our AC approach are overall quite similar. Dobrić and Schmid (2005) showed that the lower tail dependence for companies in the same industry is higher than those of companies belonging to different branches, which is further illustrated in our case-study. Table 7 shows that the lower tail dependence is strongest among the companies in the energy sector; for example, XOM–CVX has the largest tail dependence of 0.68. This can be explained by the fact that the profits from energy companies are from almost identical products and their shares fall simultaneously when the oil price drops dramatically. In contrast, the tail dependence between the two technology companies Apple and Yahoo are very small, as they have very different businesses or sources of profits. The degree of overlap between the pharmaceutical companies is smaller than between the energy companies, but larger than for the technology companies, i.e. they compete in certain areas, but each has its own unique focus. This is also correctly reflected by our tail dependence parameter estimates. The estimates of the tail dependence parameters between the pharmaceutical companies are smaller than those between the energy companies, but larger than between the technology companies. For companies in different sectors, their tail dependences parameter estimates are very small. For example, the tail dependences between Ford and any of the other companies that we considered are all close to zero. This illustrative example indicates that our AC approach provides reasonable estimates of the lower tail dependence parameters.

5. Discussion

In this paper we study what we call an AC approach that attempts to reduce the bias that is introduced by potential model misspecification and, at the same time, partially to retain part of the efficiency that is inherent in model-based inference methods. When applied to quantile estimation, the essential idea is to use likelihood-based methods but restricting them to the tail of interest by deliberately censoring the rest of the observations. Our simulations show that the AC quantile estimators are almost as unbiased as their non-parametric estimators, but remarkably more efficient than them. Although the MLE without any censoring is more efficient when the model is correctly chosen, the AC approach is not affected by large biases when the model is misspecified. We also show that model choice is not very important for the AC approach since different parametric models have similar tail behaviours. Our simulations show that the AC quantile estimators in general have smaller RMSEs than those of the POT estimators. For estimating the fifth percentiles, the AC approach is nearly uniformly better than the POT approach. When the interest lies in the first percentile, the AC approach is better than the POT approach when non-tail observations (e.g. the largest 90% values in the sample) are censored.

Furthermore, we explored two data-driven methods to estimate the appropriate proportion of censoring: one based on graphical examination and the other based on a bootstrap estimator of the RMSE. We find in our simulations that the quantile estimates from these two methods have similar RMSEs to those of the oracle AC estimates. There is room for improvement in our bootstrap estimates of the RMSE of the AC quantile estimator. Hall (1990) suggested that special care is needed when bootstrapping an MSE estimator, such as bootstrapping an explicit asymptotic expansion of the original estimator (Danielsson *et al.*, 2001) or using a bootstrap sample that is different from the original sample size (Ferreira *et al.*, 2003). Our initial investigation showed that bootstrapping with different sample sizes did not improve the performance of the naive bootstrap estimator (Liu, 2012). We conjecture that the threshold selection methods that were developed above for quantile estimation can be further extended to the same problem in tail dependence parameter estimation. That large topic is the subject of future work.

References

American Society for Testing and Materials (2015) *ASTM D5457-15, Standard Specification for Computing Reference Resistance of Wood-based Materials and Structural Connections for Load and Resistance Factor Design*. West Conshohocken: ASTM International.

Bolancé, C., Guillén, M., Gustafsson, J. and Nielsen, J. P. (2012) *Quantitative Operational Risk Models*. Boca Raton: CRC Press.

Buch-Kromann, T. (2009) Comparison of tail performance of the Champernowne transformed kernel density estimator, the generalized Pareto distribution and the g-and-h distribution. *J. Oper. Risk*, **4**, 43–67.

Buch-Larsen, T., Nielsen, J. P., Guillén, M. and Bolancé, C. (2005) Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, **39**, 503–516.

Casella, G. and Berger, R. L. (2002) *Statistical Inference*. Pacific Grove: Duxbury.

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Berlin: Springer.

Danielsson, J., de Haan, L., Peng, L. and de Vries, C. G. (2001) Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multiv. Anal.*, **76**, 226–248.

Dobrić, J. and Schmid, F. (2005) Nonparametric estimation of the lower tail dependence λ l in bivariate copulas. *J. Appl. Statist.*, **32**, 387–407.

Durrans, S. R. and Triche, M. H. (1997) Parameter and quantile estimation for the distributions of failure strength of structural lumber. *Forst Prods J.*, **47**, 80–88.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Embrechts, P., McNeil, A. and Straumann, D. (2002) Correlation and dependence in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, vol. 1 (ed. M. A. H. Dempster), pp. 176–223. Cape Town: Cambridge University Press.

Ferreira, A., de Haan, L. and Peng, L. (2003) On optimising the estimation of high quantiles of a probability distribution. *Statistics*, **37**, 401–434.

Gámiz, M. L., Mammen, E., Miranda, M. D. M. and Nielsen, J. P. (2016) Double one-sided cross-validation of local linear hazards. *J. R. Statist. Soc.* B, **78**, 755–779.

Hall, P. (1990) Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multiv. Anal.*, **32**, 177–203.

Hall, P. and Weissman, I. (1997) On the estimation of extreme tail probabilities. *Ann. Statist.*, **25**, 1311–1326.

Hill, B. M. (1975) A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, **3**, 1163–1174.

Joe, H. (2014) *Dependence Modeling with Copulas*. Boca Raton: CRC Press.

Lawless, J. F. (2003) *Statistical Models and Methods for Lifetime Data*. Hoboken: Wiley-Interscience.

Liu, Y. (2012) Lower quantile estimation of wood strength data. *Master's Thesis*. University of British Columbia, Vancouver.

Liu, Y. (2014) extWeibQuant: estimate lower extreme quantile with the censored Weibull MLE and censored Weibull mixture. *R Package Version 1.1*. University of British Columbia, Vancouver. (Available from `https://CRAN.R-project.org/package=extWeibQuant`.)

Qi, Y. (2008) Bootstrap and empirical likelihood methods in extremes. *Extremes*, **11**, 81–97.

R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supporting materials: Using artificial censoring to improve extreme tail quantile estimates'.