Repeated out of Sample Fusion in the Estimation of Small Tail Probabilities

Benjamin Kedem, Lemeng Pan, Paul Smith, and Chen Wang

Department of Mathematics
University of Maryland, College Park

May 2018
Abstract

Often, it is required to estimate the probability that a quantity such as toxicity level, plutonium, temperature, rainfall, damage, wind speed, wave size, earthquake magnitude, risk, etc., exceeds an unsafe high threshold. The probability in question is then very small. To estimate such a probability, information is needed about large values of the quantity of interest. However, in many cases, the data only contain values below or even far below the designated threshold, let alone exceedingly large values. It is shown that by repeated fusion of the data with externally generated random data, more information about small tail probabilities is obtained with the aid of certain new statistical functions. This provides relatively short, yet reliable interval estimates based on moderately large samples. A comparison of the approach with a method from extreme values theory (Peaks over Threshold, or POT), using both artificial and real data, points to the merit of repeated out of sample fusion.

Keywords: Density ratio model, semiparametric, coverage, out of sample fusion, peaks-over-threshold, B-curve.

MSC 2000: Primary 62F40; Secondary 62F25
1 Introduction

We wish to estimate a small tail probability $p$ of exceeding a high threshold $T$ from a moderately large random sample $X_1, ..., X_{n_0}$. This is done by fusing or combining the sample repeatedly with computer generated uniform samples. The number of fusions can be as large as we wish. For example 10,000 or 100,000 or 1,000,000 or more fusions. Throughout the paper the sample size $n_0$ is moderately large (100, 120, or 200), and, since in many cases the data only contain values below or even far below the designated threshold, it is assumed that the measurements $X_i$ are all below $T$.

Fusing a given sample repeatedly with computer generated data is referred to as repeated out of sample fusion (ROSF).

The large number of fusions results in what is called a $B$-curve defined in Section 2. The $B$-curve is monotonically increasing and it contains a point whose ordinate is very close to $p$ with a high probability. In fact, as the number of fusions increases the ordinate of that point essentially coincides with $p$. The goal is to “capture” that point.

Estimating $p$ is equivalent to “capturing” the said point on the $B$-curve, and this paper provides an iterative algorithm for doing so. The consequential interval estimates of $p$ are quite precise. A comparison with peaks-over-threshold (POT) from extreme value theory indicates that ROSF can bring about a substantial gain in reliability as well as in precision across a fairly wide range of tail behavior, given moderately large samples $X_0$.

The question then is how to tie or connect the real data and the generated random data to obtain useful reliable interval estimates for small tail probabilities. Connecting the real and artificial data can be approached by means of their respective distributions under the so called density ratio model framework. This paper describes ROSF together with an iterative method (IM) in the estimation of small tail probabilities against the backdrop of the density ratio model.

1.1 A Note about Extremes

The estimation of small tail probabilities has been around for a long time, at least since the celebrated work of Fisher and Tippette (1928) on the extremes of random samples. Of the various statistical methods dealing with this estimation problem, the block maxima (BM) and peaks-over-threshold (POT) are two widespread methods discussed, for example, in Beirlant et al. (2004) and more recently in Ferreira and de Hann (2015), among many others.
BM and POT might not be sufficiently reliable when the data sets are not large enough as both approaches entail a reduction in the number of observations. Specifically, by the POT method only observations above a sufficiently high threshold are used, and by the BM method the data are first divided into blocks and only the maximum from each block is used in estimation. Thus, if the data size is not large to begin with, a further data reduction might reduce considerably the reliability of the estimation results.

ROSF is of an entirely different nature in that it is not based on extreme value theory. It is an augmentation method where a reference sample is combined many times with additional data, albeit “fake” artificial data. Hence, unlike BM and POT there is no loss of observations.

ROSF has been introduced and applied in the estimation of small tail probabilities in connection with food safety in Kedem et al. (2016). A large number of experiments show that ROSF accommodates a fairly wide range of tail behavior, including that of gamma, lognormal, inverse Gaussian, Pareto, and Weibull, and that of environmental variables whose distributions possess very long tails, including lead intake, mercury, and chlorophenol compounds.

The special case where only a single fusion with artificial data occurs is dubbed in Wen (2013) and in Katzoff et al. (2014) as out of sample fusion (OSF). In connection with importance sampling, Fokianos and Qin (2008) use this idea in estimating the normalizing constant of a parametric probability distribution. Similarly, Fithian and Wager (2015) study heavy-tailed distributions given a relatively small sample, and a much larger background sample from another distribution assuming that the tails of the two distributions are connected by an exponential tilt model.

The relative efficiency of BM and POT has been discussed in Ferreira and de Hann (2015) and references therein. Under certain conditions the two methods are quite similar. In this paper we shall compare ROSF together with its IM companion with POT only, using moderately large samples.

2 ROSF and the B-Curve

We are in pursuit of a small tail probability $p$. It is shown how to construct a curve which contains with a high probability a point whose ordinate is $p$.

Let $X_0$ denote a given reference sample $X_1, ..., X_{n_0}$ from some reference distribution $g$, and suppose we wish to estimate a small tail probability $p$ of that distribution. The variable $X \sim g$ could represent quantities such as earthquake magnitude, radioactive contamination, claim amounts, financial returns, poverty levels, wealth, temperature, and so on, and the interest is
in the tail probability \( p = P(X > T) \) for some relatively high threshold \( T \). Further, suppose we have a way to fuse or combine the reference sample \( X_0 \) with a computer-generated sample \( X_1 \). Then \( X_0 \) can be fused again with another independent computer generated sample which we again denote by \( X_1 \) (\( X_1 \) is used generically), and so on. All these \( X_1 \) samples are independent and are generated in an identical manner and all have the same size \( n_1 \). We refer to these computer-generated samples as fusion samples. Observe that the fused or combined samples all have size \( n_0 + n_1 \).

Here is how B-curves are constructed. We fuse the given reference sample \( X_0 \) with a computer-generated fusion sample \( X_1 \) from \( g_1 \) and get in a certain way a confidence interval \([0, B_1]\) for the small tail probability \( p \). Since \( p \) is small we take the lower bound to be 0, and compute the upper bound \( B_1 \). We fuse the given reference sample \( X_0 \) again with another artificial fusion sample \( X_1 \) from \( g_1 \) and get in the same manner as before another confidence interval \([0, B_2]\) for \( p \). This process is repeated many times to produce a long sequence of confidence intervals \([0, B_i]\), \( i = 1, 2, ..., n \). Conditional on \( X_0 \), the sequence of upper bounds \( B_1, B_2, ..., B_n \) is then an independent and identically distributed sequence of random variables from some distribution \( F_B \). It is assumed that

\[
P(B_1 > p) = 1 - F_B(p) > 0. \tag{1}
\]

Let \( B_{(1)}, B_{(2)}, ..., B_{(n)} \) be the sequence of order statistics from smallest to largest. Then, as \( n \to \infty \), \( B_{(1)} \) decreases and \( B_{(n)} \) increases. Hence, as the number of fusions \( n \) increases the plot consisting of the pairs

\[(1, B_{(1)}), (2, B_{(2)}), ..., (n, B_{(n)}) \tag{2}\]

contains a point whose ordinate is \( p \) with probability approaching 1. It follows that as \( n \to \infty \), there is a \( B_{(j)} \) which essentially coincides with \( p \). The plot of points consisting of the pairs \((j, B_{(j)})\) in (2) is referred to as the B-curve. Typical B-curves corresponding to the tail probability \( p = P(X > T) = 0.001 \) for various reference samples \( X_0 \) from the indicated distributions or data are shown in Figure 1. Notice that to get \( p = 0.001 \), in each case the threshold \( T \) must change accordingly, and that in each plot there is a \( B_{(j)} \) nearest or closest to \( p = 0.001 \).

A key fact of the present approach is that since the fusions can be repeated indefinitely we can approximate the distribution of the \( B \) upper bounds arbitrarily closely.

Let \( \hat{F}_B \) be the empirical distribution obtained from the sequence of upper bounds \( B_1, B_2, ..., B_n \). Then from the Glivenko-Cantelli Theorem, \( \hat{F}_B \) converges to \( F_B \) almost surely uniformly as \( n \) increases. Since the fusion process
Figure 1: Typical B-Curves from $B_{(1)}, \ldots, B_{(10,000)}$ containing a point corresponding to $p = 0.001$. Clockwise from top left: Gamma(1,0.01), LN(1,1), Lead exposure, Mercury. $T = 690.7755, 59.7538, 25.00, 22.41$, respectively, $n_0 = n_1 = 100$. Histograms representing the distributions are shown in Figure 2.
Figure 2: Histograms representing distributions with long right tails. The lead intake data are discussed in Kedem et al. (2016). The mercury data source is NOAA's National Status and Trends Data https://products.coastalscience.noaa.gov/nsandt_data/data.aspx
can be repeated as many times as we wish, our key idea, \( F_B \) is known for all practical purposes. Assume then that \( F_B \) was obtained from numerous fusions, for example 10,000 fusions. Then, under (1), from a random sample \( B_1, \ldots, B_K \), the probability that the maximum \( B_{(K)} \) exceeds \( p \),

\[
P(B_{(K)} > p) = 1 - F^K_B(p)
\]

increases with \( K \). It follows that for all \( K > K_0 \), for some sufficiently large \( K_0 \), we have for a small \( \alpha > 0 \) the inequality

\[
1 - F^K_B(p) \geq 1 - \alpha
\]

or

\[
0 < p \leq F^{-1}_B(\alpha^{1/K}).
\]

The interval (5) covers \( p \) with at least 100(1 - \( \alpha \))% confidence, and it has been applied in food safety in Kedem et al. (2016). Experimental results indicate that in many cases \( K = 100 \) is a conservative choice and that at times a much smaller \( K \) suffices. However, when \( \max(X_0) \) is small relative to \( T \) a larger \( K \) is needed, for example \( K = 300 \) or even larger. As a byproduct, the present paper provides a way for choosing \( K \).

### 2.1 Getting Upper Bounds by Data Fusion

Clearly, the preceding argument is quite general, and the effectiveness of the procedure will depend on the quality of the \([0, B_i]\) confidence intervals. In this section we describe a particular way of generating these confidence intervals, which amounts to data fusion of the real and computer-generated data (“augmented reality” as it were) under the density ratio model.

In general, by “fusion” or “data fusion” we mean the combined data from \( m + 1 \) sources where each source is governed by a probability distribution. In the spirit of augmented reality, random data generating computer algorithms are perfectly legitimate data sources. Using the combined data, semiparametric statistical inference can be ensued under the density ratio model assumption (Kedem, et al. 2017).

Recall that the reference random sample \( X_0 \) of size \( n_0 \) follows an unknown reference distribution with probability density \( g \), and let \( G \) be the corresponding cumulative distribution function (cdf).

Let

\[
X_1, \ldots, X_m,
\]
be additional computer-generated random samples where \( X_j \sim g_j, G_j \), with size \( n_j, j = 1, \ldots, m \). For now \( m \geq 1 \) but later we specialize to \( m = 1 \) only. We refer to the vector \( t = (t_1, \ldots, t_n)' = (X'_0, X'_1, \ldots, X'_m)' \),

\[ (6) \]

of size \( n = n_0 + n_1 + \cdots + n_m \) as the fused data. We further assume the density ratio model (Qin and Zhang 1997, Lu 1997)

\[ \frac{g_j(x)}{g(x)} = \exp(\alpha_j + \beta_j' h(x)), \quad j = 1, \ldots, m, \]

\[ (7) \]

where \( \beta_j \) is a \( p \times 1 \) parameter vector, \( \alpha_j \) is a scalar parameter, and \( h(x) \) is a \( p \times 1 \) vector valued distortion or tilt function. Clearly, to generate the \( X_j \) samples we must know the corresponding \( g_j \). However, beyond the generating process, we do not make use of this knowledge. Thus, by our estimation procedure, none of the probability densities \( g, g_1, \ldots, g_m \) and the corresponding \( G_j \)'s, and none of the parameters \( \alpha \)'s and \( \beta \)'s are assumed known, but, strictly speaking, the so called tilt function \( h \) must be a known function.

Since all the probability distributions are connected by the density ratio model \((7)\), each distribution pair \( g_j, G_j \) is estimated from the entire fused data \( t \) and not just from \( X_j \) only. The same holds for the reference pair \( g, G \). Thus, for example, the reference \( G \) is estimated from the entire fused data \( t \) with \( n \) observations and not just from the reference sample \( X_0 \) with \( n_0 =< n \) observations.

Under the assumption that the density ratio model \((7)\) holds, the maximum likelihood estimate of \( G(x) \) based on the fused data \( t \) is given in \((15)\) in Section A.1 in the Appendix, along with its asymptotic distribution described in Theorem A.1. From the theorem we obtain confidence intervals for \( p = 1 - G(T) \) for any threshold \( T \) using \((18)\). In addition, from \((15)\) we get the point estimate \( \hat{p} = 1 - \hat{G}(T) \) as well.

Obviously, the density ratio model per se need not hold, and even if it does for some tilt function \( h \), the validity or goodness of any choice of \( h \) is uncertain. Furthermore, if \( \max(X_0) \) is much smaller than the threshold \( T \) then \( \hat{p} \) from \((15)\) could be just too small. However, for the implementation of ROSF, the density ratio model need not hold precisely and any reasonable choice of \( h \) suffices as long as \((7)\) holds, which is a mild requirement. Experience shows that the “gamma tilt” \( h(x) = (x, \log x) \) is a sensible choice for skewed data similar to those shown in Figure 2. Similarly, the “lognormal tilt” \( h(x) = (\log x, (\log x)^2) \) is another useful choice.
Our strategy then is to obtain interval estimates for small $p = 1 - G(T)$ for a relatively large $T$ using numerous upper bounds from (18), obtained by ROSF, call the upper bounds $B_i$, and take the lower bounds as 0. This is the “certain way” referred to in the previous section by which we obtain the $[0, B_i]$ confidence intervals. When assumption (11) holds, many of the $B_i$ will be greater than $p$ as their number increases, but some will not. Hence, the ordered $B_i$ engulf or surround $p$ with probability approaching one as the number of fusions increase. This is illustrated in Figure 1 with 10,000 fusions.

Thus, $[0, B_1]$ is obtained from the first fusion of $X_0$ with a set of $m$ computer-generated samples. Then $[0, B_2]$ is obtained by fusing $X_0$ again but with a different independent set of $m$ computer-generated samples, and so on. From each fusion we obtain a point estimate $\hat{p} = 1 - \hat{G}(T)$ using (15) and an upper bound $B_i$ from (18). Since this fusion process is repeated numerous times, we obtain both numerous point estimates $\hat{p}$’s and numerous upper bounds $B_i$’s. In general, as the number of fusions increases, the set of pairs $(j, B_{(j)})$ engulfs the desired point on the $B$-curve with probability approaching one. That is, with a large number of fusions the ordered $B_{(j)}$ engulf $p$ with a high probability. This, in general, cannot be said about the ordered $\hat{p}$’s unless the number of fusions is exceedingly large. See Section 3.1.2 for a case where the $\hat{p}$’s from (18) are too small.

In this paper $m = 1$ only, and the fusion samples are uniform random samples supported over a wide range which covers $T$, that is, $\max(X_1) > T$. But why uniform? First, when the density ratio model holds for some $g$ and $g_1$, then it also holds approximately by taking $g_1$ as a uniform density supported over a sufficiently wide range. Second, and more to the point, ROSF requires only the mild assumption (1). Experience shows that assumption (1) holds well when fusing $X_0$ with uniform samples using the tilt function $h(x) = (x, \log x)$ across a wide range of tail types. Evidently, the B-curves used in this paper provide further support for the validity of assumption (1).

To summarize, Under assumption (1), the B-curves are constructed from ordered upper bounds $B_{(j)}$ (18) for $p = P(X > T)$ obtained from a large number of repeated fusions of $X_0$ with random uniform samples $X_1$ where the upper limit of the uniform distribution exceeds $T$. Throughout the paper, $\max(X_0) < T$ and $h = (x, \log x)$. 
3 Capturing a Point on the B-Curve

Due to a large number of fusions \( n \), \( F_B \) is known for all practical purposes and with probability close to 1

\[
B_{(1)} < p < B_{(n)}. \tag{8}
\]

In general, even for \( n = 1,000 \), \( B_{(1,000)} \) is much larger than the true \( p \) and \( B_{(1)} \) is very close to 0. The goal is to find \( B_{(j)} \) close to \( p \).

It follows, by the monotonicity of the B-curve and (8), that as \( j \) decreases (for example from \( n = 10,000 \)), the \( B_{(j)} \) approach \( p \) from above so that there is a \( B_{(j)} \) very close to \( p \). Thus, the B-curve establishes a relationship between \( j \) and \( p \).

From a basic fact about order statistics it is known that

\[
P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{n}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k}. \tag{9}
\]

Therefore, as (9) is monotone decreasing, the \textit{smallest} \( p \) which satisfies the inequality

\[
\sum_{k=0}^{j-1} \binom{n}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k} \leq 0.95 \tag{10}
\]

provides another relationship between \( j \) and \( p \). Note that if “\( > \)” is used instead of “\( \leq \)” in (10) then the solution of (10) is \( p = 0 \). This is so since (9) is a steep monotone decreasing step function of the type shown in Figures 3,4.

Iterating between these two monotone relationships is what was referred to earlier as the iterative method (IM). The iterative method provides our \( p \) estimates. In general, the iteration process could start with a sufficiently large \( j \) suggested by the B-curve. With that \( j \equiv j_1 \) we look for the smallest \( p \equiv p_{j_1} \) satisfying (10). Next we find a \( B_{(j_2)} \) on the B-curve closest to \( p_{j_1} \). This gives a new \( j \equiv j_2 \) and the previous steps are repeated until convergence occurs and we keep getting the same \( p \). This is our point estimate from the iteration process and it is different than \( \hat{p} \) obtained from (15) in the Appendix.

In symbols, with \( B_{(j_k)} \)'s from the B-curve, and \( p_{(j_k)} \)'s the smallest \( p \)'s satisfying (10) with \( j = j_k \), and \( B_{(j_{k+1})} \) closest to \( p_{(j_k)} \), \( k = 1,2,... \),

\[
B_{(j_1)} \rightarrow p_{(j_1)} \rightarrow B_{(j_2)} \rightarrow \cdots \rightarrow B_{(j_k)} \rightarrow p_{j_k} \rightarrow B_{(j_{k+1})} \rightarrow p_{j_k} \rightarrow B_{(j_{k+1})} \rightarrow p_{j_k} \cdots
\]
so that \( p_{j_k} \) keeps giving the same \( B_{(j_{k+1})} \) (and hence the same \( j_{k+1} \)) and vice versa. This can be expressed more succinctly as,

\[
\begin{align*}
\text{j}_1 & \rightarrow p_{(j_1)} \rightarrow \text{j}_2 \rightarrow p_{(j_2)} \rightarrow \cdots \text{j}_k \rightarrow p_{j_k} \rightarrow \text{j}_{k+1} \rightarrow p_{j_{k+1}} \rightarrow \text{j}_{k+2} \rightarrow p_{j_{k+2}} \rightarrow \cdots
\end{align*}
\]

As will be illustrated in Section [3.1] under some computational conditions this iterative process results in a contraction in a neighborhood of the true \( p \). In a small neighborhood of the true \( p \) the \( B_{(j_k)} \) can move either up or down, an example of which is given in the lead example in Section [3.1.4].

Computationally, the iteration process depends on \( n \) and the increments of \( p \) at which (10) is evaluated. In practice, due to computational limitations of large binomial coefficients the iteration is done as follows. After \( F_B \) is obtained from a large number of fusions, 1000 \( B_{(j)} \)'s are sampled from, say, \( n = 10,000 \) \( B_{(j)} \)'s to obtain an approximate B-curve. Next, the binomial coefficients \( \binom{n}{k} \) are replaced by \( \binom{1000}{k} \). We then iterate between an approximate B-curve and approximate (10) with \( n = 1000 \) (as in (11) below) until convergence occurs, in which case an estimate for \( p \) is obtained. This procedure can be repeated many times by sampling repeatedly many different sets of 1000 \( B_{(j)} \)'s to obtain many point estimates from which interval estimates can then be constructed. This iteration process is illustrated next.

### 3.1 Illustrations of an Iterative Process

The following illustrations deal with two lognormal and two real data examples. The four cases underscore the fact that ROSF is used with a gamma tilt function while the data, at least in the lognormal cases, are not gamma distributed. Running 10,000 fusions takes about 5 minutes in R which translates to about 8 hours for 1,000,000 fusions. In what follows the \( p \)-increments at which (11) is evaluated are chosen mostly as \( \mathcal{O}(B) \). In all cases the maxima of the approximate B-curves were larger than the true \( p \).

#### 3.1.1 Lognormal(1,1)

In this example \( X_0 \) is a LN(1,1) sample where \( \max(X_0) = 25.17781 \). With \( T = 59.75377 \) the true tail probability to be estimated is \( p = 0.001 \), using \( n_0 = n_1 = 100 \) and \( h = (x, \log x) \). The generated fusion samples \( X_1 \) are from Unif(0,100), \( 100 > T \), and \( F_B \) was obtained from 10,000 fusions.

We first sample 1000 from 10,000 \( B_{(j)} \)'s to get an approximate B-curve, and then iterate between it and the smallest \( p \) such that

\[
\sum_{k=0}^{j-1} \binom{1000}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k} \leq 0.95 \tag{11}
\]
evaluated at increments of $p = 0.0001$ ($\bar{B} = 0.00060$). Starting with $j = 1000$, the sequence $(j, p_j)$ is

$$
1000 \rightarrow 0.0035 \rightarrow 996 \rightarrow 0.0028 \rightarrow 985 \rightarrow 0.0022 \rightarrow 968 \rightarrow 0.0019 \rightarrow 951 \rightarrow 0.0017 \rightarrow 937 \rightarrow 0.0016 \rightarrow 929 \rightarrow 0.0015 \rightarrow 915 \rightarrow 0.0014 \rightarrow 905 \rightarrow 0.0013 \rightarrow 888 \rightarrow 0.0012 \rightarrow 871 \rightarrow 0.0012 \cdots
$$

so that convergence occurs at $\hat{p} = 0.0012$ as $0.0012$ gives $j = 871$ again and again. This also suggests $K = 20$ in (5) which gives 0.0012 as an upper bound for $p$. The left side of (11) for $j = 871$ is the step function shown in Figure 3.

Repeating this with a different LN(1,1) reference sample $X_0$ such that $\max(X_0) = 28.27287$, and fusing 10,000 times with $X_1$ from Unif(0,80), $80 > T$, gives with $p$-increments of 0.0002 ($\bar{B} = 0.00031$) the $(j, p_j)$ sequence,

$$
1000 \rightarrow 0.003 \rightarrow 995 \rightarrow 0.0024 \rightarrow 991 \rightarrow 0.002 \rightarrow 986 \rightarrow 0.0018 \rightarrow 977 \rightarrow 0.0016 \rightarrow 965 \rightarrow 0.0014 \rightarrow 954 \rightarrow 0.0012 \rightarrow 941 \rightarrow 0.001 \rightarrow 923 \rightarrow 0.001 \cdots
$$

so that $\hat{p} = 0.001$ is equal to the true $p$.

Now, convergence might be problematic when $\max(X_0)$ is small relative to $T$. In that case an augmentation of the data is helpful. Thus, repeating the previous illustration with a LN(1,1) sample where $\max(X_0) = 16.92843$, the latter is somewhat small relative to $T = 59.75377$. Indeed, $n_0 = n_1 = 100$ and $p$-increments of 0.0001 (although $\bar{B} = 4.661 \times 10^{-5}$, 0.00005 was not useful), gave an imprecise $\hat{p}$ for the true tail probability $p = 0.001$:

$$
1000 \rightarrow 0.001 \rightarrow 999 \rightarrow 0.0008 \rightarrow 995 \rightarrow 0.0006 \rightarrow 992 \rightarrow 0.0005 \rightarrow 989 \rightarrow 0.0005 \cdots
$$

Augmenting the sample with 20 additional LN(1,1) observations resulted in a larger $\max(X_0) = 31.7835$ and $n_0 = n_1 = 120$. We have with $X_1$ from Unif(0,100), $100 > T$, and $p$-increment=0.0001 (now $\bar{B} = 0.0003211$)

$$
1000 \rightarrow 0.0038 \rightarrow 998 \rightarrow 0.0031 \rightarrow 995 \rightarrow 0.0028 \rightarrow 991 \rightarrow 0.0024 \rightarrow 987 \rightarrow 0.0022 \rightarrow 980 \rightarrow 0.0019 \rightarrow 970 \rightarrow 0.0016 \rightarrow 959 \rightarrow 0.0014 \rightarrow 951 \rightarrow 0.0013 \rightarrow 946 \rightarrow 0.0012 \rightarrow 938 \rightarrow 0.0011 \rightarrow 932 \rightarrow 0.001 \rightarrow 926 \rightarrow 0.001 \cdots
$$

so that with the augmented data $\hat{p} = 0.001$ has been rendered precise.
3.1.2 Lognormal\((0,1)\)

Here \(X_0\) is a LN\((0,1)\) sample where \(\max(X_0) = 5.77902\), which is small relative to \(T = 21.98218\). Instead of addition of more data, we opt for more fusions.

In this example \(\text{[11]}\) is evaluated as a function of \(p\) using increments of 0.0002. The true tail probability is \(p = 0.001\), and \(F_B\) was obtained from 1,000,000 fusions of \(X_0\) with \(X_1\) from Unif\((0,40\)\), 40 \(> T\). Again \(n_0 = n_1 = 100\) and \(h = (x, \log x)\). In this example the largest point estimate of \(p\) from one million point estimates (obtained from \(\text{[11]}\) in the Appendix) was only 0.0004186, much lower than the true \(p\), and the reason why we use the \(B_j\) upper bounds. Again, first 1000 \(B_{(j)}\)'s were sampled at random from 1,000,000 \(B_{(j)}\)'s to get an approximate B-curve with 1000 points \((j, B_{(j)})\). Starting with the maximum \(j = 1000\) the sequence \((j, p_j)\) along \(p\)-increments of 0.00002 \((B = 0.000065)\) is:

\[
1000 \rightarrow 0.001 \rightarrow 1000 \rightarrow 0.001 \rightarrow 1000 \rightarrow 0.001 \cdots
\]

so that the convergence gives the exact \(p = 0.001\). This suggests \(K = 2100\) in \(\text{[5]}\) giving an upper bound of 0.0011. Note that \(K\) is large as \(\max(X_0)\) is small relative to \(T\).
Repeating this with 100,000 fusions, \( \max(X_0) = 7.510843 \), and \( X_1 \) from Unif(0,30), \( 30 > T \), gives the \((j,p_j)\) sequence along p-increments of 0.0002 (\( \bar{B} = 0.00014 \))

\[ 1000 \to 0.0016 \to 998 \to 0.0014 \to 997 \to 0.0012 \to 994 \to 0.001 \to 991 \to 0.001 \cdots \]

which again converges to the exact \( p = 0.001 \). This suggests \( K = 300 \) in (5), giving an upper bound of 0.0011.

3.1.3 Mercury

Here \( X_0 \) is a sample of size \( n_0 = 100 \) from the mercury data whose histogram is shown in Figure 2. Again \( n_1 = 100 \) and \( h = (x, \log x) \). The mercury data consist of 8266 observations for which \( T = 22.41 \) gives \( p = 0.001088797 \approx 0.001 \). We have \( \max(X_0) = 9.09 \), \( X_1 \sim \text{Unif}(0,40), \ 40 > T \), and \( F_B \) was obtained from 1,000,000 fusions. Sampling 1000 \( B(j) \)'s from 1,000,000 \( B(j) \)'s, the \((j,p_j)\) sequence along p-increments of 0.0002 (\( \bar{B} = 0.00096 \)) is:

\[ 1000 \to 0.0052 \to 996 \to 0.0046 \to 991 \to 0.0042 \to 981 \to 0.0038 \to 966 \to 0.0034 \to 949 \to 0.0032 \to 942 \to 0.0030 \to 911 \to 0.0026 \to 895 \to 0.0024 \to 879 \to 0.0022 \to 851 \to 0.0020 \to 829 \to 0.0018 \to 801 \to 0.0016 \to 768 \to 0.0014 \to 732 \to 0.0014 \cdots \]

converging to \( \hat{p} = 0.0014 \), exactly what we get with \( K = 10 \) in (5).

Starting with a \( B(j) \) closer to the true \( p = 0.001 \) we get an upward convergence,

\[ 637 \to 0.001 \to 651 \to 0.001 \]

In a different run with only 10,000 \( B(j) \), \( \max(X_0) = 11 \), \( X_1 \sim \text{Unif}(0,50), \ 50 > T \), and sampling 1000 \( B(j) \)'s from 10,000 \( B(j) \)'s, the \((j,p_j)\) sequence along p-increments of 0.0001 (although \( \bar{B} = 0.001331 \)) is:

\[ 745 \to 0.0019 \to 722 \to 0.0017 \to 695 \to 0.0015 \to 657 \to 0.0013 \to 634 \to 0.0012 \to 617 \to 0.0011 \to 606 \to 0.001 \to 589 \to 0.001 \cdots \]

**Mercury: a Higher Probability.** Consider the higher probability \( p = 0.01004113 \) corresponding to \( T = 9.375 \), and a mercury sample \( X_0 \) of size
\[ n_0 = 100 \text{ where } \max(X_0) = 7.77 < T, \text{ and } X_1 \sim \text{Unif}(0,20), \ 20 > T, \ n_1 = 100. \]

Out of 10,000 fusions with \( h(x) = (x, \log x) \), giving a different \( F_B \) than the previous one, the maximum probability estimate (out of 10,000) using (15) is 0.003738044, far below the true \( p = 0.01004113 \), and the maximum likelihood estimate based on \( X_0 \) only is 0. On the other hand, sampling 1000 \( B_{(j)} \)'s from 10,000 \( B_{(j)} \)'s, the ROSF iterative \((j, p_j)\) sequence along \( p \)-increments of 0.001 \((B = 0.002686784)\) is:

\[
1000 \rightarrow 0.011 \rightarrow 1000 \rightarrow 0.011 \cdots
\]

while starting from \( B_{(999)} \) yields

\[
999 \rightarrow 0.01 \rightarrow 996 \rightarrow 0.01 \cdots
\]

so that \( \hat{p} \approx p. \ K = 800 \) in (5) gives 0.0102 as an upper bound.

### 3.1.4 Lead Intake

Here \( X_0 \) is a sample of size \( n_0 = 100 \) from the lead intake data whose histogram is shown in Figure 2. Again \( n_1 = 100 \) and \( h = (x, \log x) \). The lead data consist of 3000 observation for which \( T = 25 \) gives \( p = 0.001 \). We have \( \max(X_0) = 11.55768, \ X_1 \sim \text{Unif}(0,40), \ 40 > T, \) and \( F_B \) was obtained from 10,000 fusions. Again 1000 \( B_{(j)} \)'s were sampled from 10,000 \( B_{(j)} \)'s and the \((j, p_j)\) sequence was observed along \( p \)-increments of 0.0001. In this example the iteration process starts with \( j = 400 \) giving \( p_{400} \) not far from the true \( p = 0.001 \). We have:

\[
400 \rightarrow 0.0017 \rightarrow 371 \rightarrow 0.0016 \rightarrow 351 \rightarrow 0.0015 \rightarrow 327 \rightarrow 0.0014 \rightarrow 302 \rightarrow 0.0013 \rightarrow 278 \rightarrow 0.0012 \rightarrow 252 \rightarrow 0.0011 \rightarrow 229 \rightarrow 0.0011 \cdots
\]

Thus, the sequence \( p_j \) converges to \( \hat{p} = 0.0011 \). This corresponds to \( K = 2 \) in (5).

Figure 4 shows the step function (9) for \( n = 1000 \) when convergence occurs at \( j = 229 \). Observe that \( \hat{p} = 0.0011 \) is the smallest \( p \) satisfying (11), giving a point on the cord corresponding to the pair \((0.0011, 0.3648204)\).

Now, let us see what happens in neighborhood of true \( p=0.001 \). We have:

\[
201 \rightarrow 0.001 \rightarrow 203 \rightarrow 0.001 \cdots
\]

and the convergence is upward. With

\[
205 \rightarrow 0.001 \rightarrow 203 \rightarrow 0.001 \cdots
\]
the convergence is downward. This shows that in a neighborhood of the true $p$ the $B(j_k)$ can change course to lock on the true, or approximately true, $p$ from above or from below.

![Step function](image)

**Figure 4:** Step function (11) from lead intake $X_0$ fused with $X_1 \sim \text{Unif}(0, 40)$ data for $j = 229$ and containing a point corresponding to $\hat{p} = 0.0011$ whose ordinate is $0.3648204 < 0.95$.

**Lead Intake: a Higher Probability.** Consider the higher probability $p = 0.01$ corresponding to $T = 10$, and a lead intake sample $X_0$ of size $n_0 = 100$ where $\max(X_0) = 6.875607 < T$, and $X_1 \sim \text{Unif}(0, 20)$, $20 > T$, $n_1 = 100$. Out of 10,000 fusions with $h(x) = (x, \log x)$, giving $F_B$, the maximum probability estimate (out of 10,000) using (15) is $0.003550$, far below the true $p = 0.01$, and the maximum likelihood estimate based on $X_0$ only is $0$. On the other hand, sampling 1000 $B(j_k)$’s from 10,000 $B(j_k)$’s, the IM iterative $(j, p_j)$ sequence along $p$-increments of $0.001$ ($\bar{B} = 0.003516579$) is:

$1000 \rightarrow 0.01 \rightarrow 999 \rightarrow 0.009 \rightarrow 998 \rightarrow 0.009 \cdot \cdot \cdot$

so that $\hat{p} = 0.009$. We note that $K = 5000$ in (5) gives $0.01$. 

15
3.2 Explaining the Convergence

Clearly, the $p_j$ sequence need not converge to a point in a neighborhood of the true $p$. However, as we have seen from the previous illustrations, iterating between the two monotone relationships, the B-curve and (11) along certain $p$-increments, we do get in many cases $p$ estimates in neighborhoods of the true $p$. This can be explained as follows.

Suppose that, for some $j$, $B(j)$ lands in a neighborhood of the true $p$. Observe that as $n \to \infty$, the ordered $B_j$ become ever more dense so that the absolute difference $|B(j\pm k) - B(j)|$ becomes arbitrarily small for $B(j\pm k)$ in that neighborhood. Therefore, the smallest $p$’s which satisfy

$$P(B(j) > p) \leq 0.95$$

(12)

(e.g. as in (11)) along certain $p$-increments in that neighborhood are equal or nearly equal for entire stretches of adjacent $B(j)$, thus increasing the probability that two successive $p_j$ in the iteration process are equal, in which case convergence occurs in a neighborhood of the true $p$.

This can be illustrated with $X_0 \sim LN(0,1)$, $\max(X_0) = 9.274283$, $X_1 \sim \text{Unif}(0,30)$, $30 > T = 21.98218$, $p = 0.001$, $n_0 = n_1 = 100$, $p$-increments of 0.0001 ($\bar{B} = 0.000414989$), and 30,000 $B_j$. Sampling 1000 $B_j$, successive $B(j)$ in a neighborhood of $p = 0.001$ give:

For $957 \leq j \leq 950$, the smallest $p_j$ which satisfies (11) is 0.0011.
For $949 \leq j \leq 938$, the smallest $p_j$ which satisfies (11) is 0.0010.
For $937 \leq j \leq 924$, the smallest $p_j$ which satisfies (11) is 0.0009.

Hence, over the stretch $j = 957$ to $j = 924$ there are 34 consecutive $p_j$’s which are markedly close to the true $p = 0.001$. As a result, in this stretch, starting with $j = 957$ along increments of 0.0001 the next $j$ in the iteration process is $j = 950$ and we have two equal successive $p_j$,

$$957 \rightarrow 0.0011 \rightarrow 950 \rightarrow 0.0011,$$

whereas starting with $j = 949$ the next $j$’s in the iteration process are $j = 937$ and $j = 926$, and again there are two equal successive $p_j$,

$$949 \rightarrow 0.001 \rightarrow 937 \rightarrow 0.0009 \rightarrow 926 \rightarrow 0.0009.$$

Thus, entering a neighborhood of the true $p$, the iteration method (IM) produces further $p$’s which, as $n$ increases, tend to stay in that neighborhood leading to convergence. We have seen this tendency throughout the previous illustrations, and we see more of it from the tables in the next section.
4 Comparison: ROFS vs POT

Against the background provided in the previous sections, we compare two very different ways to obtain interval estimates for small tail probabilities: POT based on extreme value theory, and an iterative process based on repeated fusion of a given reference sample with external computer-generated uniformly distributed samples. The comparison is based on confidence interval coverage, width, and on the mean absolute error (MAE) which measures the discrepancy between \( \hat{p} \) and the true tail probability \( p \). In Tables 1 to 11, \( p \) is relatively small, \( p = 0.001 \), whereas in the last three tables 12, 13, 14 \( p \) is smaller, \( p = 0.0001 \).

Throughout the comparison the sample sizes are \( n_0 = n_1 = 100 \) or \( n_0 = n_1 = 200 \), and \( h(x) = (x, \log x) \). Thus, in the present comparison the reference \( X_0 \) and the fusion samples \( X_1 \) have size \( n_0 = 100 \) or \( n_0 = 200 \).

To save computation time, in each case of the iteration process \( F_B \) was obtained from 1000 fusions, and the starting \( j \) is such that \( B(j) \) is approximately equal to the 3rd quartile of the observed 1000 \( B \)'s.

Remark: Starting at the 3rd quartile is computationally sensible as the corresponding \( B(j) \) is usually in a neighborhood above \( p \). In most cases subsequent \( B(j) \) do enter a neighborhood of \( p \) and convergence occurs, as explained earlier. Starting too low might lead to convergence to a point lower than the true \( p \).

The following tables are the result of 500 runs. In each run the iteration method (IM) was repeated 500 times.

From the mean residual life (MRL) plots we obtained the thresholds \( u \) needed for the POT method. In all cases reported in the tables, the MRL plots suggest the use of the largest 20% of the reference data \( X_0 \) for fitting the generalized Pareto (GP) distribution. We have noticed a deterioration in the POT results when using 30%, 15% or 10% of \( X_0 \). The simulation details are given in Section A.2 in the Appendix.

An interesting picture emerges from Tables 1 to 14. For moderately large sample sizes of \( n_0 = 100 \) and \( n_0 = 200 \), regardless of the tail type, as \( N \), the number of \( \hat{p} \)'s used in forming the CI for the true \( p \) grows the iteration process gives reliable and relatively narrow confidence intervals, whereas the POT gives unacceptable coverage and in many cases wider CI's as well. The POT coverage increases significantly going from \( n_0 = 100 \) to \( n_0 = 200 \), how-
ever, it seems that for the method to “fire up” larger samples are needed. Regarding ROSF, the choice of N = 50 seems prudent across all cases, and with n₀ = 200 shorter CIs achieve coverage similar to that from n₀ = 100. In all cases the MAE from the iteration process is much smaller than that obtained from POT.

4.1 Comparison Tables

The following tables compare ROSF and POT for p = 0.001 and p = 0.0001.

Table 1: $X₀ \sim \text{Weibull}(1,2) : p = 1 - G(T) = 0.001, T = 13.81551, X₁ \sim \text{Unif}(0,16), n₀ = n₁, h(x) = (x, \log x). p$-increment 0.00005.

| Method     | N   | Coverage | CI Length | MAE     | Coverage | CI Length | MAE     |
|------------|-----|----------|-----------|---------|----------|-----------|---------|
| POT        | -   | 82.7%    | 0.00431   | 0.00131 | 87.8%    | 0.00333   | 0.00083 |
| 5          | 43.2% | 0.00040 | 0.00068   | 52.4%   | 0.00042 | 0.00051   |
| 10         | 65.2% | 0.00083 | -         | 72.7%   | 0.00091 | -         |
| 25         | 84.2% | 0.00159 | -         | 85.6%   | 0.00154 | -         |
| ROSF & IM  | 50  | 92.5%    | 0.00287   | -       | 92.8%    | 0.00231   | -       |
| 100        | 100%  | 0.00381 | -         | 100%    | 0.00321 | -         |
| 300        | 100%  | 0.00506 | -         | 100%    | 0.00402 | -         |

Table 2: $X₀ \sim \text{Pareto}(1,4) : p = 1 - G(T) = 0.001, T = 5.623413, X₁ \sim \text{Unif}(1,8), n₀ = n₁, h(x) = (x, \log x). p$-increment 0.0001.

| Method     | N   | Coverage | CI Length | MAE     | Coverage | CI Length | MAE     |
|------------|-----|----------|-----------|---------|----------|-----------|---------|
| POT        | -   | 81.8%    | 0.00419   | 0.00121 | 84.5%    | 0.00357   | 0.00070 |
| 5          | 59.1% | 0.00068 | 0.00052   | 62.4%   | 0.00066 | 0.00041   |
| 10         | 66.7% | 0.00093 | -         | 74.8%   | 0.00091 | -         |
| 25         | 84.1% | 0.00154 | -         | 86.1%   | 0.00148 | -         |
| ROSF & IM  | 50  | 96.2%    | 0.00232   | -       | 97.8%    | 0.00231   | -       |
| 100        | 100%  | 0.00272 | -         | 100%    | 0.00269 | -         |
| 300        | 100%  | 0.00397 | -         | 100%    | 0.00377 | -         |
**Table 3:** \( X_0 \sim \text{Gamma}(3,1) : p = 1 - G(T) = 0.001, T = 11.22887, X_1 \sim \text{Unif}(0,20), n_0 = n_1, h(x) = (x, \log x). \) \( p \)-increment 0.00005.

| Method   | \( n_0 = 100 \) | \( n_0 = 200 \) |
|----------|-----------------|-----------------|
| POT      | Coverage CI Length MAE | Coverage CI Length MAE |
|          | 86.1% 0.00321 0.00021 & 0.00081 | 86.1% 0.00321 0.00021 & 0.00081 |
| 5        | 41.2% 0.00057 0.00054 & 0.00043 | 41.2% 0.00057 0.00054 & 0.00043 |
| 10       | 49.6% 0.00093 - & - | 49.6% 0.00093 - & - |
| 25       | 73.2% 0.00137 - & - | 73.2% 0.00137 - & - |
| ROSF & IM | 50 93.4% 0.00188 - & - | 50 93.4% 0.00188 - & - |
| 100      | 100% 0.00256 - & - | 100% 0.00256 - & - |
| 300      | 100% 0.00338 - | 100% 0.00338 - |

**Table 4:** \( X_0 \sim \text{F}(2,12) : p = 1 - G(T) = 0.001, T = 12.97367, X_1 \sim \text{Unif}(0,16), n_0 = n_1, h(x) = (x, \log x). \) \( p \)-increment 0.00005.

| Method   | \( n_0 = 100 \) | \( n_0 = 200 \) |
|----------|-----------------|-----------------|
| POT      | Coverage CI Length MAE | Coverage CI Length MAE |
|          | 87.9% 0.00292 0.00082 & 0.00082 | 87.9% 0.00292 0.00082 & 0.00082 |
| 5        | 43.1% 0.00066 0.00051 & 0.00031 | 43.1% 0.00066 0.00051 & 0.00031 |
| 10       | 54.2% 0.00094 - & - | 54.2% 0.00094 - & - |
| 25       | 78.5% 0.00136 - & - | 78.5% 0.00136 - & - |
| ROSF & IM | 50 96.1% 0.00217 - & - | 50 96.1% 0.00217 - & - |
| 100      | 100% 0.00289 - & - | 100% 0.00289 - & - |
| 300      | 100% 0.00344 - & - | 100% 0.00344 - & - |

**Table 5:** \( X_0 \sim \text{IG}(2,40) : p = 1 - G(T) = 0.001, T = 3.835791, X_1 \sim \text{Unif}(0,8), n_0 = n_1, h(x) = (x, \log x). \) \( p \)-increment 0.00005.

| Method   | \( n_0 = 100 \) | \( n_0 = 200 \) |
|----------|-----------------|-----------------|
| POT      | Coverage CI Length MAE | Coverage CI Length MAE |
|          | 83.1% 0.00321 0.00123 & 0.00092 | 83.1% 0.00321 0.00123 & 0.00092 |
| 5        | 41.2% 0.00057 0.00054 & 0.00043 | 41.2% 0.00057 0.00054 & 0.00043 |
| 10       | 49.6% 0.00093 - & - | 49.6% 0.00093 - & - |
| 25       | 73.2% 0.00137 - & - | 73.2% 0.00137 - & - |
| ROSF & IM | 50 93.4% 0.00188 - & - | 50 93.4% 0.00188 - & - |
| 100      | 100% 0.00256 - | 100% 0.00256 - |
| 300      | 100% 0.00338 - | 100% 0.00338 - |

**Table 6:** \( X_0 \sim \text{IG}(4,5) : p = 1 - G(T) = 0.001, T = 28.95409, X_1 \sim \text{Unif}(0,35), n_0 = n_1, h(x) = (x, \log x). \) \( p \)-increment 0.00005.

| Method   | \( n_0 = 100 \) | \( n_0 = 200 \) |
|----------|-----------------|-----------------|
| POT      | Coverage CI Length MAE | Coverage CI Length MAE |
|          | 84.3% 0.00412 0.00123 & 0.000103 | 84.3% 0.00412 0.00123 & 0.000103 |
| 5        | 73.6% 0.00106 0.00052 & 0.00041 | 73.6% 0.00106 0.00052 & 0.00041 |
| 10       | 89.2% 0.00148 - & - | 89.2% 0.00148 - & - |
| 25       | 97.5% 0.00217 - & - | 97.5% 0.00217 - & - |
| ROSF & IM | 50 100% 0.00265 - & - | 50 100% 0.00265 - & - |
| 100      | 100% 0.00345 - & - | 100% 0.00345 - & - |
| 300      | 100% 0.00372 - | 100% 0.00372 - |
Table 7: \( X_0 \sim \text{LN}(0,1) : p = 1 - G(T) = 0.001, T = 21.98218, X_1 \sim \text{Unif}(1,60), n_0 = n_1, h(x) = (x, \log x), p\text{-increment 0.00005.} \)

| Method     | N   | Coverage | CI Length | MAE | Coverage | CI Length | MAE |
|------------|-----|----------|-----------|-----|----------|-----------|-----|
| POT        | 5   | 81.5%   | 0.00121   | 0.00047 | 83.6%   | 0.00108   | 0.00039 |
|            | 10  | 88.7%   | 0.00169   | -     | 90.4%   | 0.00141   | -   |
|            | 25  | 95.3%   | 0.00191   | -     | 98.1%   | 0.00173   | -   |
| ROSF & IM  | 50  | 100%    | 0.00234   | -     | 100%    | 0.00244   | -   |
|            | 100 | 100%    | 0.00267   | -     | 100%    | 0.00283   | -   |
|            | 300 | 100%    | 0.00301   | -     | 100%    | 0.00283   | -   |

\( n_0 = 100 \) \( n_0 = 200 \)

Table 8: \( X_0 \sim \text{LN}(1,1) : p = 1 - G(T) = 0.001, T = 59.75377, X_1 \sim \text{Unif}(1,140), n_0 = n_1, h(x) = (x, \log x), p\text{-increment 0.0001.} \)

| Method     | N   | Coverage | CI Length | MAE | Coverage | CI Length | MAE |
|------------|-----|----------|-----------|-----|----------|-----------|-----|
| POT        | 5   | 43.7%   | 0.00078   | 0.00069 | 53.2%   | 0.00061   | 0.00052 |
|            | 10  | 56.9%   | 0.00109   | -     | 68.1%   | 0.00099   | -   |
|            | 25  | 79.6%   | 0.00143   | -     | 89.7%   | 0.00121   | -   |
| ROSF & IM  | 50  | 89.1%   | 0.00147   | -     | 100%    | 0.00164   | -   |
|            | 100 | 100%    | 0.00199   | -     | 100%    | 0.00192   | -   |
|            | 300 | 100%    | 0.00243   | -     | 100%    | 0.00234   | -   |

\( n_0 = 100 \) \( n_0 = 200 \)

Table 9: \( X_0 \sim \text{Mercury} : p = 1 - G(T) = 0.001, T = 22.41, X_1 \sim \text{Unif}(0,50), n_0 = n_1, h(x) = (x, \log x), p\text{-increment 0.0001.} \)

| Method     | N   | Coverage | CI Length | MAE | Coverage | CI Length | MAE |
|------------|-----|----------|-----------|-----|----------|-----------|-----|
| POT        | 5   | 34.5%   | 0.00073   | 0.00048 | 49.9%   | 0.00063   | 0.00045 |
|            | 10  | 66.7%   | 0.00095   | -     | 76.7%   | 0.00096   | -   |
|            | 25  | 84.9%   | 0.00157   | -     | 96.7%   | 0.00145   | -   |
| ROSF & IM  | 50  | 97.5%   | 0.00215   | -     | 100%    | 0.00197   | -   |
|            | 100 | 100%    | 0.00259   | -     | 100%    | 0.00238   | -   |
|            | 300 | 100%    | 0.00337   | -     | 100%    | 0.00313   | -   |

\( n_0 = 100 \) \( n_0 = 200 \)

Table 10: \( X_0 \sim \text{Lead Intake} : p = 1 - G(T) = 0.001, T = 25, X_1 \sim \text{Unif}(0,30), n_0 = n_1, h(x) = (x, \log x), p\text{-increment 0.0001.} \)

| Method     | N   | Coverage | CI Length | MAE | Coverage | CI Length | MAE |
|------------|-----|----------|-----------|-----|----------|-----------|-----|
| POT        | 5   | 51.1%   | 0.00095   | 0.00066 | 49.6%   | 0.00088   | 0.00058 |
|            | 10  | 69.3%   | 0.00151   | -     | 78.1%   | 0.00153   | -   |
|            | 25  | 88.4%   | 0.00189   | -     | 93.7%   | 0.00179   | -   |
| ROSF & IM  | 50  | 100%    | 0.00247   | -     | 100%    | 0.00229   | -   |
|            | 100 | 100%    | 0.00289   | -     | 100%    | 0.00268   | -   |
|            | 300 | 100%    | 0.00346   | -     | 100%    | 0.00317   | -   |

\( n_0 = 100 \) \( n_0 = 200 \)
Table 11: \( X_0 \sim URX3TB : p = 1 - G(T) = 0.001, T = 9.50, X_1 \sim Unif(0,12), \) \( n_0 = n_1, h(x) = (x, \log x). \) p-increment 0.0001. Data source for URX3TB - 2,4,6-trichlorophenol (ug/L): [https://wwwn.cdc.gov/nchs/nhanes](https://wwwn.cdc.gov/nchs/nhanes)

| Method  | \( n_0 = 100 \) | \( n_0 = 200 \) |
|---------|----------------|----------------|
| POT     | Coverage | CI Length | MAE  | Coverage | CI Length | MAE  |
|         | 81.1%    | 0.00433    | 0.00143 | 87.1%    | 0.00076   | 0.00129 |
|         | 5        | 38.9%      | 0.00078  | 0.00055  | 42.6%      | 0.00071  | 0.00044  |
|         | 10       | 54.3%      | 0.00094  | -        | 61.8%      | 0.00092  | -        |
|         | 25       | 72.1%      | 0.00131  | -        | 81.7%      | 0.00125  | -        |
|         | 100      | 100%       | 0.00241  | -        | 100%       | 0.00235  | -        |
|         | 300      | 100%       | 0.00264  | -        | 100%       | 0.00259  | -        |
| ROSF & IM | 50   | 89.1%      | 0.00179  | -        | 96.9%      | 0.00177  | -        |

Table 12: \( X_0 \sim F(2,12) : p = 1 - G(T) = 0.0001, T = 21.84953, X_1 \sim Unif(0,25), \) \( n_0 = n_1, h(x) = (x, \log x). \) p-increment 0.00001.

| Method  | \( n_0 = 100 \) | \( n_0 = 200 \) |
|---------|----------------|----------------|
| POT     | Coverage | CI Length | MAE  | Coverage | CI Length | MAE  |
|         | 71.4%    | 0.00062    | 0.00052 | 81.6%    | 0.00053   | 0.00045 |
|         | 5        | 45.2%      | 0.00021  | 0.00022  | 49.1%      | 0.00017  | 0.00019  |
|         | 10       | 67.2%      | 0.00033  | -        | 77.1%      | 0.00026  | -        |
|         | 25       | 88.5%      | 0.00045  | -        | 89.3%      | 0.00037  | -        |
|         | 100      | 100%       | 0.00082  | -        | 100%       | 0.00069  | -        |
|         | 300      | 100%       | 0.00105  | -        | 100%       | 0.00087  | -        |
| ROSF & IM | 50   | 95.2%      | 0.00059  | -        | 96.3%      | 0.00052  | -        |

Table 13: \( X_0 \sim LN(0,1) : p = 1 - G(T) = 0.0001, T = 41.22383, X_1 \sim Unif(1,60), \) \( n_0 = n_1, h(x) = (x, \log x). \) p-increment 0.00001.

| Method  | \( n_0 = 100 \) | \( n_0 = 200 \) |
|---------|----------------|----------------|
| POT     | Coverage | CI Length | MAE  | Coverage | CI Length | MAE  |
|         | 72.1%    | 0.00064    | 0.00045 | 82.6%    | 0.00047   | 0.00039 |
|         | 5        | 55.2%      | 0.00021  | 0.00021  | 69.1%      | 0.00017  | 0.00017  |
|         | 10       | 77.2%      | 0.00033  | -        | 89.1%      | 0.00020  | -        |
|         | 25       | 98.5%      | 0.00041  | -        | 99.3%      | 0.00034  | -        |
|         | 100      | 100%       | 0.00066  | -        | 100%       | 0.00057  | -        |
|         | 300      | 100%       | 0.00113  | -        | 100%       | 0.00094  | -        |
| ROSF & IM | 50   | 100%       | 0.00083  | -        | 100%       | 0.00079  | -        |

Table 14: \( X_0 \sim Mercury : p = 1 - G(T) = 0.0001, T = 39.60, X_1 \sim Unif(0,80), \) \( n_0 = n_1, h(x) = (x, \log x). \) p-increment 0.00001.

| Method  | \( n_0 = 100 \) | \( n_0 = 200 \) |
|---------|----------------|----------------|
| POT     | Coverage | CI Length | MAE  | Coverage | CI Length | MAE  |
|         | 62.4%    | 0.00059    | 0.00049 | 73.4%    | 0.00051   | 0.00042 |
|         | 5        | 53.1%      | 0.00019  | 0.00023  | 64.2%      | 0.00016  | 0.00019  |
|         | 10       | 71.8%      | 0.00025  | -        | 79.8%      | 0.00021  | -        |
|         | 25       | 88.3%      | 0.00037  | -        | 91.5%      | 0.00033  | -        |
|         | 100      | 100%       | 0.00083  | -        | 100%       | 0.00079  | -        |
|         | 300      | 100%       | 0.00113  | -        | 100%       | 0.00094  | -        |

21
5 Discussion

The numerous number of fusions of a given reference sample with computer generated samples can give rise to many observables including the upper bounds for a tail probability $p$ that were used in the paper. The upper bounds, obtained from the combined real and artificial data, were mostly much larger than $p$, some were less than $p$, but some among the multitude of upper bounds essentially coincided with $p$ and they were identified to a reasonable degree of approximation using an iteration procedure.

Evidently, the repeated fusion of a sample with generated uniform random data allowed us to gain information about the tail behavior beyond the threshold using the notion of B-curves coupled with a well known formula from order statistics.

The ideas presented in this paper can be extended in a number of ways. For example, using “fake” data from distributions other than uniform, and using different fusion mechanisms other than the semiparametric method used in the paper.

Reliable estimation of tail probabilities is important in numerous fields from finance to geophysics to meteorology to the design of ships and to optics [10], [12].
A Appendix

A.1 Asymptotic Distribution of $\hat{G}(x)$

Define $\alpha_0 \equiv 0, \beta_0 \equiv 0, w_j(x) = \exp(\alpha_j + \beta_j' h(x)), \rho_i = n_i/n_0, j = 1, \ldots, m$.

Maximum likelihood estimates for all the parameters and $G(x)$ can be obtained by maximizing the empirical likelihood over the class of step cumulative distribution functions with jumps at the observed values $t_1, \ldots, t_m$ (Owen 2001). Let $p_i = dG(t_i)$ be the mass at $t_i$, for $i = 1, \ldots, n$. Then the empirical likelihood becomes

$$L(\theta, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{m} \exp(\alpha_1 + \beta_1' h(x_{1j})) \cdots \prod_{j=1}^{m} \exp(\alpha_m + \beta_m' h(x_{mj})).$$

(13)

Maximizing $L(\theta, G)$ subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, \ldots, \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0$$

(14)

we obtain the desired estimates. In particular,

$$\hat{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n} \frac{I(t_i \leq t)}{1 + \rho_1 \exp(\hat{\alpha}_1 + \hat{\beta}_1' h(t_i)) + \cdots + \rho_m \exp(\hat{\alpha}_m + \hat{\beta}_m' h(t_i))},$$

(15)

where $I(t_i \leq t)$ equals one for $t_i \leq t$ and is zero, otherwise. Similarly, $\hat{G}_j$ is estimated by summing $\exp(\hat{\alpha}_j + \hat{\beta}_j' h(t_i)) dG(t_i)$.

The asymptotic properties of the estimators have been studied by a number of authors including Qin and Zhang (1997), Lu (2007), and Zhang (2000).

Define the following quantities: $\mathbf{\rho} = \text{diag}\{\rho_1, \ldots, \rho_m\}$,

$$A_j(t) = \int \frac{w_j(y)I(y \leq t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y), \quad B_j(t) = \int \frac{w_j(y)h(y)I(y \leq t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y),$$

$$\bar{A}(t) = (A_1(t), \ldots, A_m(t))', \quad \bar{B}(t) = (B_1(t), \ldots, B_m(t))'.$$

Then the asymptotic distribution of $\hat{G}(t)$ for $m \geq 1$ is given by the following result due to Lu (2007).

Theorem A.1 Assume that the sample size ratios $\rho_j = n_j/n_0$ are positive and finite and remain fixed as the total sample size $n = \sum_{j=0}^{m} n_j \to \infty$. The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges to a zero-mean Gaussian process in the
space of real right continuous functions that have left limits with covariance matrix given by

\[
\text{Cov}\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = \\
\left(\sum_{k=0}^{m} \rho_k\right) \left( G(t \wedge s) - G(t)G(s) - \sum_{j=1}^{m} \rho_j A_j(t \wedge s) \right) \\
+ \left( \hat{A}'(s) \rho, \hat{B}'(s) (\rho \otimes I_p) \right) S^{-1} \left( \begin{pmatrix} \rho \hat{A}(t) \\ (\rho \otimes I_p) \hat{B}(t) \end{pmatrix} \right) .
\]  

(16)

where \( I_p \) is the \( p \times p \) identity matrix, and \( \otimes \) denotes Kronecker product.

For a complete proof see Lu (2007). The proof for \( m = 1 \) is given in Zhang (2000).

Denote by \( \hat{V}(t) \) the estimated variance of \( \hat{G}(t) \) as given in (16). Replacing parameters by their estimates, a \( 1 - \alpha \) level pointwise confidence interval for \( G(t) \) is approximated by

\[
\left( \hat{G}(t) - z_{\alpha/2} \sqrt{\hat{V}(t)}, \hat{G}(t) + z_{\alpha/2} \sqrt{\hat{V}(t)} \right),
\]

(17)

where \( z_{\alpha/2} \) is the upper \( \alpha/2 \) point of the standard normal distribution. Hence, a \( 1 - \alpha \) level pointwise confidence interval for \( 1 - G(T) \) for any \( T \), and in particular for relatively large thresholds \( T \) is approximated by

\[
\left( 1 - \hat{G}(t) - z_{\alpha/2} \sqrt{\hat{V}(t)}, 1 - \hat{G}(t) + z_{\alpha/2} \sqrt{\hat{V}(t)} \right).
\]

(18)

A.2 Simulation Description

The following steps were followed. There were 500 runs. In each run the iteration method (IM) was repeated 500 times.

First, a reference \( X_0 \) was obtained.

POT:

The POT procedure was applied to get both an estimate \( \hat{p} \) and a confidence interval (CI). The MRL plots suggest the use of the largest 20% of the reference data \( X_0 \) for fitting the generalized Pareto (GP) distribution.

ROSF/IM:
was fused with $X_1$ 1000 times (ROSF) to get $F_B$ and then $\hat{p}$ (IM).

$X_0$ was fused again with different $X_1$ 1000 times to get $F_B$ and $\hat{p}$.

This was repeated 500 times.

The iterative method thus gave 500 $\hat{p}$’s. We then chose at random $N$ $\hat{p}$’s from 500 $\hat{p}$’s to construct a CI for the true $p$ as $(\min(\hat{p}), \max(\hat{p}))$.

This is run 1.

The above steps were repeated, for both POT and ROSF/IM each time with a different $X_0$, 500 times (runs) to obtain coverage and average CI length. In the tables, CI length is an average length from 500 intervals.

Since there are 500 runs, POT gave 500 $\hat{p}$’s. Regarding IM, a single $\hat{p}$ was chosen at random (out of 500 $\hat{p}$’s) from each of the 500 runs. The mean absolute error (MAE) was obtained in both cases from the mean of 500 absolute differences $\sum |\hat{p}_i - p|/500$, where $p = 0.001$ or $p = 0.0001$. In the iterative method, in each table the MAE is reported once on the line corresponding to $N = 5$.

References

[1] Beirlant, J., Goegebeur, Y., Teugels, J.L., and Segers, J. Statistics of extremes : theory and applications. Wiley: Hoboken, 2004.

[2] Ferreira, A. and De Haan, L. On the block maxima method in extreme value theory: PWM estimators. The Annals of Statistics 2015, 43: 276-298.

[3] Fisher, R.A. and Tippette, L.H.C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society 1928, 24: 180-190.

[4] Fithian, W. and Wager, S. Semiparametric exponential families for heavy-tailed data. Biometrika 2015, 102: 486-493.

[5] Fokianos, K. and Qin J. A Note on Monte Carlo Maximization by the Density Ratio Model. Journal of Statistical Theory and Practice 2008; 2: 355-367.

[6] Katzoff, M., Zhou, W., Khan, D., Lu, G., and Kedem, B. Out of sample fusion in risk prediction. Journal of Statistical Theory and Practice 2014; 8: 444-459.
[7] Kedem, B., Pan, L., Zhou, W., and Coelho, C.A. Interval estimation of small tail probabilities – application in food safety. *Statistics in Medicine* 2016; **35**: 3229-3240.

[8] Kedem, B., De Oliveira, V., and Sverchkov, M. (2017). *Statistical Data Fusion*, World Scientific, Singapore.

[9] Lu, G. Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting. Ph.D. Dissertation, University of Maryland, College Park, 2007.

[10] Pelinovsky, E. and Kharif, C., Editors. *Extreme Ocean Waves*, Springer, New York, 2008

[11] Qin, J. and Zhang, B. A Goodness of fit test for logistic regression models based on case-control data. *Biometrika* 1997; **84**: 609-618.

[12] Solli, D.R., Ropers, C., Koonath, P., and Jalali, B. Optical rogue waves. *Nature* 2007; **450**: 1054-1057.

[13] Zhang, B. A goodness of fit test for multiplicative-intercept risk models based on case-control data. *Statistica Sinica* 2000; **10**: 839-865.

[14] Zhou, W. Out of Sample Fusion. Ph.D. Dissertation, University of Maryland, College Park, 2013.

[15] Centers for Disease Control and Prevention, [http://wwwn.cdc.gov/nchs/nhanes/2007-2008/PP_E.htm](http://wwwn.cdc.gov/nchs/nhanes/2007-2008/PP_E.htm)