Mathematics and Statistics 7(5): 172-181, 2019
DOI: 10.13189/ms.2019.070503 http://www.hrpub.org

Estimation of Small Tail Probabilities by Repeated Fusion

Benjamin Kedem*, **Lemeng Pan**, **Paul J. Smith Chen Wang**

Department of Mathematics, University of Maryland, College Park, 20742, Maryland, United States

Received September 2, 2019; Revised September 29, 2019; Accepted October 07, 2019

Abstract It is shown how to estimate any threshold probability from data below or even far below the threshold through repeated fusion of the data with externally generated random samples. This is referred to as repeated out of sample fusion (ROSF). A comparison of the approach with peaks-over-threshold (POT) across different tail types shows that ROSF provides more precise point and interval estimates based on moderately large samples.

Keywords Density Ratio Model, Semiparametric, Optimization, Iterative, Order Statistics.

1 Introduction

This paper presents a novel statistical idea, that of "Down-Up" sequences which "capture" small tail probabilities with surprising precision without knowing the underlying probability distributions. We describe the idea, its implementation, and its usefulness in the estimation of small tail probabilities using limited amounts of data. The gist of the paper is expressed in Tables 1 to 13 showing the Down-Up shifts occurring in *small neighborhoods* of the true tail probabilities in question, either 0.001 or 0.0001 from samples of size 100 only.

The paper is an extension of [1] which dealt with interval estimation of small tail probabilities with applications in food safety. Here we shall deal with the corresponding problem of point estimation.

The problem is motivated by the following generic problem. Consider a sample of rogue ocean waves none of which exceeds $T = 200$ feet in height, and yet we wish to estimate the small chance of exceeding T from a moderately large sample $\boldsymbol{X}_0 = (X_1, ..., X_{n_0})$, referred to as a *reference* sample. Similar problems pertain to insurance claims, food safety, and environmental risks such as radiation levels, where the problem is to estimate the probability of exceeding catastrophic levels from limited amounts of data.

Accordingly, we consider a moderately large random sample $\boldsymbol{X}_0$ where all the observations are much smaller than a high threshold T, that is $\max(\boldsymbol{X}_0) << T$. Based on the sample we wish to estimate the tail probability p of exceeding T without knowing the underlying distribution. However, as is, the sample may not contain sufficient amount of information to tackle the problem. To gain more information, the problem is approached by combining or fusing the sample *repeatedly* with externally generated computer data. That is, *augmented reality* as it were.

1.1 Repeated Out of Sample Fusion

Let $\boldsymbol{X}_i$ denote the ith computer generated sample of size $n_1 = n_0$. Then the *fused* or *combined* samples are the *fusions* or *augmentations*

$$(\boldsymbol{X}_0, \boldsymbol{X}_1), (\boldsymbol{X}_0, \boldsymbol{X}_2), (\boldsymbol{X}_0, \boldsymbol{X}_3)... \tag{1}$$

where $\boldsymbol{X}_0$ is a real reference sample and the $\boldsymbol{X}_i$ are different independent computer-generated 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. The computer-generated samples $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,... are independent and are generated in an identical manner, and all have the same size n_1. They are referred to as *fusion samples*.

The question then is how to tie or connect the real data and the generated random data to obtain useful reliable estimates for small tail probabilities. As in [1], Connecting or fusing the real and artificial data can be approached by means of their respective probability distributions under the so called *density ratio model* framework, discussed briefly in Section 3 and in the Appendix.

In that way we can extract information about tail probabilities possibly not available in the original reference sample by itself. As we shall see, the consequential estimates as well as interval estimates of p are quite precise.

Thus, the paper describes what we call *repeated out of sample fusion* (ROSF) and a related iterative method (IM) in the estimation of small tail probabilities, against the backdrop of the density ratio model.

A comparison with peaks-over-threshold (POT) from extreme value theory, discussed in [2] and [3], 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 $\boldsymbol{X}_0$.

ROSF has been discussed in [1],[4]. Unlike the bootstrap, additional information is sought repeatedly from outside the sample. Related ideas concerning a single fusion are studied in [5],[6],[7],[8].

2 A Sequence of Upper Bounds

Suppose $\boldsymbol{X}_0 = (X_1, ..., X_{n_0})$ is a reference sample from some unknown reference probability density g, and let g_1 denote the probability density governing the computer generated fusion samples $\boldsymbol{X}_i$, $i = 1, ..., n$. The problem is to estimate a small tail probability $p = P(X > T)$ for some relatively high threshold T. The idea is to produce numerous upper bounds to engulf p. When that happens, some upper bounds take values in a neighborhood of p. Getting these upper bounds is tantamount to getting p.

We fuse the given reference sample $\boldsymbol{X}_0$ with a computer-generated fusion sample $\boldsymbol{X}_1$ from g_1 and get in a certain way, described in the next section, a confidence interval for the small tail probability p. Let B_1 denote the upper bound of that interval. We fuse the given reference sample $\boldsymbol{X}_0$ again with another artificial fusion sample $\boldsymbol{X}_2$ from g_1, independent of $\boldsymbol{X}_1$, and get in the same manner another upper bound B_2 for p. This process is repeated many times to produce a long sequence of confidence intervals and hence a long sequence of upper bounds B_i, $i = 1, 2, ...$ Conditional on $\boldsymbol{X}_0$, the sequence of upper bounds $B_1, B_2, ...$ is then an independent and identically distributed sequence of random variables from some distribution F_B. It is assumed that

$$P(B_1 > p) = 1 - F_B(p) > 0. \tag{2}$$

Let $B_{(1)}, B_{(2)}, ..., B_{(n)}$ be a sequence of order statistics from smallest to largest. Hence, as $n \to \infty$, $B_{(1)}$ decreases and $B_{(n)}$ increases so that

$$B_{(1)} < p < B_{(n)}. \tag{3}$$

It follows that the ordered sequence of upper bounds $B_{(1)}, B_{(2)}, ..., B_{(n)}$ contains a $B_{(j)}$ closest to p. In fact, as $n \to \infty$, that $B_{(j)}$ essentially coincides with p. We will show how to obtain $B_{(j)}$'s in a neighborhood of p.

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 F_B 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 number of fusions can be as large as we wish, *our key idea*, F_B is known for all practical purposes.

2.1 Confidence Interval for p

Assume then that F_B was obtained from numerous fusions (in some of our simulations we use $10,000$ fusions) and that

$$P(B_1 > p) = 1 - F_B(p) > 0. \tag{4}$$

Then, from a random sample $B_1, ..., B_N$, the maximum $B_{(N)}$ entails that

$$P(B_{(N)} > p) = 1 - F_B^N(p) \tag{5}$$

increases with N. It follows that, conditional on the given sample, for all $N > N_0$, for some sufficiently large N_0, we have the inequality

$$1 - F_B^N(p) \geq 0.95 \tag{6}$$

or

$$0 < p \leq F_B^{-1}(0.05^{1/N}). \tag{7}$$

The interval (7) covers p with at least 95% confidence, and is refinement of the unduly large interval (3) for large n. As has been shown in [1], $N = 100$ is a conservative choice for N, and that in many cases a much smaller N suffices. Clearly the smaller N is the narrower is the confidence interval (7). In general, the confidence interval (7) gives an idea as to the magnitude of p, and we could search for the latter within $[0, F_B^{-1}(0.05^{1/N})]$.

3 Getting Upper Bounds by Data Fusion

This section describes a particular way of generating upper bounds for tail probabilities p by data fusion of the real $\boldsymbol{X}_0$ and additional computer-generated data (augmented reality) under the density ratio model defined in (9) below.

In general, by "fusion" or "data fusion" we mean the combined data from $m + 1$ sources, $m \geq 1$, where each source is governed by a probability distribution. As an example, consider the semiparametric extension of one-way ANOVA discussed in [9]. Note that in (1) $m = 1$. In the spirit of augmented reality, computer algorithms which generate random data are perfectly legitimate data sources. Using the combined data, semiparametric statistical inference can be ensued under the density ratio model assumption as described in detail in [4],[10],[11].

Recall that the reference random sample $\boldsymbol{X}_0$ of size n_0 follows an unknown reference distribution with probability density g, and let G be the corresponding cumulative distribution function (cdf).

Let

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_m,$$

be additional computer-generated random samples where $\boldsymbol{X}_j \sim g_j, G_j$, with size n_j, $j = 1, ..., m$. As in the Appendix, for now $m \geq 1$ but later we specialize to $m = 1$ only as in (1). The augmentation of $m + 1$ samples

$$\boldsymbol{t} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_m), \tag{8}$$

of size $n_0 + n_1 + \cdots + n_m$ gives the fused data. The density ratio model stipulates that

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)), \quad j = 1, \ldots, m, \tag{9}$$

where $\boldsymbol{\beta}_j$ is an $r \times 1$ parameter vector, α_j is a scalar parameter, and $\boldsymbol{h}(x)$ is an $r \times 1$ vector valued distortion or tilt function.

Clearly, to generate the $\boldsymbol{X}_j$ 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 $g, g_1, ..., g_m$ and the corresponding G_j's, and none of the parameters α's and $\boldsymbol{\beta}$'s are assumed known, but, strictly speaking, the so called tilt function $\boldsymbol{h}$ must be a known function. However, in the present application the requirement of a known $\boldsymbol{h}$ is apparently mitigated as accentuated by assumption (4) above, which may hold for misspecified $\boldsymbol{h}$, and by *numerous* examples with many different tail types.

Since all the probability distributions are connected by the density ratio model (9), each distribution pair g_j, G_j is estimated from the entire fused data $\boldsymbol{t}$ and not just from $\boldsymbol{X}_j$ only. The same holds for the reference pair g, G. Thus, for example, the reference G is estimated from the entire fused data $\boldsymbol{t}$ with $n_0 + n_1 + \cdots + n_m$ observations and not only from the reference sample $\boldsymbol{X}_0$ with n_0 observations. Estimates of all the α_j's and $\boldsymbol{\beta}_j$'s and of the pair g, G are obtained by maximizing the *empirical likelihood* subject to several constraints as described in the Appendix. Thus, at its core, the density ratio model involves semiparametric optimization.

Under the density ratio model (9), the maximum likelihood estimate of $G(x)$ based on the fused data $\boldsymbol{t}$ 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). We shall use the upper bounds from these confidence intervals. In addition, from (15) we get point estimates for p as well. However, these point estimates are not used in the paper, as in many cases they underestimates p.

Our data analysis in Section 4 regarding many different tail types, and additional examples in [4], indicate that for the implementation of ROSF, the density ratio model need not hold precisely, and that the "gamma tilt" $\boldsymbol{h}(x) = (x, \log x)$ is a sensible choice for skewed data. The real issue is assumption (4) which is a mild requirement.

When assumption (4) holds, as their number increases, many of the B_i obtained from (18) will be greater than p but some will be smaller than p. Hence, the *ordered* $B_{(i)}$ engulf or surround p with probability approaching one. That is, as the number of fusions increases, the ordered $B_{(j)}$ engulf p with probability approaching one.

In this paper $m = 1$ only, and the fusion samples are uniform random samples supported over a wide range which covers T. The reason for uniform samples is that 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. Examples of this are the entries in Table 4.1 in [4] regarding normal and uniform data. Moreover, numerous experiments, and the results reported in Tables 1 to 13 and 15 to 25, suggest that the uniform choice is sensible.

To summarize, numerous examples with skewed data suggest that the confidence intervals (18) are still useful in conjunction with $\boldsymbol{h}(x) = (x, \log x)$ even when the density ratio model does not hold in a strict sense. In that case, the reference sample $\boldsymbol{X}_0$ is fused repeatedly with identically distributed independent random uniform samples $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,..., as in (1), where the upper limit of the uniform support exceeds T. Repeated fusion gives upper bounds $B_1, B_2, ...$ for $p = P(X > T)$ using (18). Conditional on $\boldsymbol{X}_0$, the upper bounds B_i are independent and identically distributed random variables from some distribution F_B.

4 Capturing p by Upper Bound Sequences

As noted earlier, due to a large number of fusions n, F_B is known for all practical purposes and with probability close to one $B_{(1)} < p < B_{(n)}$. 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 $B_{(j)}$ and (3), 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. Likewise, the $B_{(j)}$ can approach p from below. This establishes a relationship between j and p.

Another relationship between j and p is obtained from a basic fact about order statistics where it is known that

$$P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{n}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k}. \quad (10)$$

Suppose now that the probabilities

$$P(B_{(j_1)} > p_{j_1}),\ P(B_{(j_2)} > p_{j_2}), ...$$

are sufficiently high probabilities, and that from the sequence of upper bounds we get the close approximations

$$p_{j_1} \doteq B_{(j_2)},\ p_{j_2} \doteq B_{(j_3)}....$$

Then with a high probability we get a decreasing "Down" sequence

$$B_{(j_1)} > B_{(j_2)} > B_{(j_3)} \cdots .$$

Replacing the "sufficiently high probabilities" by "sufficiently low probabilities", then a dual argument leads to an increasing "Up" sequence

$$B_{(j_1')} < B_{(j_2')} < B_{(j_3')} \cdots .$$

Thus, when the probabilities (10) are sufficiently high the $B_{(j_k)}$ decrease, and when the probabilities (10) are sufficiently low the $B_{(j_k)}$ increase. In particular, this "Down-Up" phenomenon occurs in a neighborhood of the true p, where a *transition or shift* occurs from "Down" to "Up" or vice versa, resulting in a "capture" of p. Thus, allowing for high and low probabilities by bounding (10) by a sufficiently high probability, we have.

Proposition: *Assume that the samples size n_0 of $\boldsymbol{X}_0$ is large enough, and that the number of fusions n is sufficiently*

large so that $B_{(1)} < p < B_{(n)}$. Consider the smallest $p_j \in (0,1)$ which satisfy the inequality

$$P(B_{(j)} > p_j) = \sum_{k=0}^{j-1} \binom{n}{k} [F_B(p_j)]^k [1 - F_B(p_j)]^{n-k} \leq 0.95, \quad (11)$$

where the p_j are evaluated along appropriate numerical increments. Then, (11) produces "Down" and "Up" sequences depending on the $B_{(j)}$ relative to p_j. In particular, in a neighborhood of the true tail probability p, with a high probability, there are "Down" sequences which converge from above and "Up" sequences which converge from below to points close to p.

This will be demonstrated copiously across different tail types using an approximation to (11). We note that (10) is a steep monotone decreasing step function so that if "$>$" is used instead of "$\leq$" in (11) then the solution of (11) is $p = 0$, and that replacing 0.95 by 0.99 in (11) gives similar results.

Iterating between these two monotone relationships, the $B_{(j)}$ relative to p, and (11), is what was referred to earlier as the iterative method (IM). The iterative method provides our p estimates. The iterations could start with a sufficiently large j, or, alternatively, with a sufficiently small j, until the Down and Up sequences converge to the same or very close points. The average of these points, or an approximation thereof, can serve as a point estimate from the iterative process and it is different than the p-estimate obtained from (15) in the Appendix.

In general, starting with any j, convergence occurs by monotonicity and we keep getting the same point.

In symbols, with $B_{(j_k)}$'s from the sequence of ordered upper bounds, and $p_{(j_k)}$'s the smallest p's satisfying (11) with $j = j_k$, and $B_{(j_{k+1})}$ closest to $p_{(j_k)}$, $k = 1, 2, ...,$

$$B_{(j_1)} \to p_{(j_1)} \to B_{(j_2)} \to \cdots B_{(j_k)} \to p_{j_k} \to B_{(j_{k+1})} \to p_{j_k} \to B_{(j_{k+1})} \to p_{j_k} \cdots$$

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,

$$j_1 \to p_{(j_1)} \to j_2 \to p_{(j_2)} \to \cdots j_k \to p_{j_k} \to j_{k+1} \to p_{j_k} \to j_{k+1} \to p_{j_k} \cdots$$

As will be illustrated in Section 4.2, under some computational conditions this iterative process results in a contraction in a neighborhood of the true p.

4.1 Computational Considerations

Computationally, the iterative process depends on n and the increments of p at which (11) 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, say $n = 10,000$ fusions (which give 10,000 B's), then 1000 B's are sampled at random from the original $n = 10,000$ B's. Next, the binomial coefficients $\binom{n}{k}$ are replaced by $\binom{1000}{k}$. We then iterate between $B_{(j)}$ approximatiopns of p and approximate (11) with $n = 1000$ as in

$$\sum_{k=0}^{j-1} \binom{1000}{k} [F_B(p_j)]^k [1 - F_B(p_j)]^{n-k} \leq 0.95 \quad (12)$$

until a "Down-Up" convergence occurs, in which case an estimate for p is obtained as the Down-Up shift point. The iterative process is illustrated in the next section. This procedure can be repeated many times by sampling repeatedly many different sets of 1000 B's to obtain many point estimates $\hat{p}$ from which interval estimates can then be constructed, as well as variance estimates.

Running 10,000 fusions takes about 5 minutes in R which translates into about 8 hours for 1,000,000 fusions. In what follows the p-increments at which (12) is evaluated are 0.0001 when $p = 0.001$ and 0.000015 when $p = 0.0001$.

4.2 Illustrations of an Iterative Process

The Down-Up convergence results together with the number of iterations are summarized in Tables 1 to 10 for $p = 0.001$ and also for $p = 0.0001$. Due to insufficiently large data sets, the Down-Up convergence results for real data in Tables 11 – 13 do not deal with the smaller $p = 0.0001$. The cdf F_B was obtained from 10,000 B's (the result of 10,000 fusions), and each entry in the tables was obtained from a *different* sample of 1,000 B's sampled at random from 10,000 B's.

With 10,000 fusions, the interval $(B_{(1)}, B_{(10,000)})$ contains p with probability close to 1, and gives an idea as to the magnitude of p.

4.2.1 Some Typical Down-Up Sequences

It is instructive first to realize some typical Down-Up sequences. Let $\boldsymbol{X}_0$ be a LN(1,1) sample where $\max(\boldsymbol{X}_0) = 32.36495$. With $T = 59.75377$ the true tail probability to be estimated is $p = 0.001$, using $n_0 = n_1 = 100$ and $\boldsymbol{h} = (x, \log x)$, p-increment 0.0001. The generated fusion samples $\boldsymbol{X}_1$ are from Unif(0,80), $80 > T$, and F_B was obtained from 10,000 fusions.

Typical convergent Down-Up sequences (j, p_j) are given next. Again, each sequence was derived from a different B-sample of size 1000 drawn from 10,000 B's. More examples are given in [12].

Down: $800 \to 0.001299466 \to 775 \to 0.001199466 \to 743 \to 0.001099466 \to 712 \to 0.0009994658 \to 680 \to 0.0009994658 \cdots$.

Up: $670 \to 0.0008994658 \to 675 \to 0.0009994658 \to 711 \to 0.0009994658 \cdots$.

Thus a Down-Up shift occurs at 0.0009994658 very close to the true $p = 0.001$. More details about this example are given in Table 3.

We note that the number of Down-Up iterations decreases dramatically in a neighborhood of the true p. As seen from Tables 1 to 13 below, in many cases 1 or 2 iterations in a neighborhood of p suffice. This reduction can serve as a telltale sign that convergence took place.

We further note that the Gamma cases in Tables 1 and 2 are nearly specified whereas this cannot be said about the cases in Tables 3 to 13. However, the tables portray a very similar picture for both real and simulated data, for (nearly) specified or misspecified cases, giving precision on the order of 10^{-5} or better for $p = 0.001$ and order of 10^{-6} for $p = 0.0001$, where $n_0 = n_1 = 100$. The results in the tables were obtained from 10,000 out of sample fusions, and in all cases it has been observed that $p \in (B_{(1)}, B_{(10,000)})$. *Notably, as seen from Tables 1 to 13, the Down-Up shift points are close to p.*

4.2.2 Gamma(1,0.05)

Table 1. $\mathbf{p = 0.001}$, $\boldsymbol{X}_0 \sim Gamma(1, 0.05)$, $\boldsymbol{X}_1 \sim Unif(0, 170)$, $\max(\boldsymbol{X}_0) = 73.0467$, $T = 138.1551$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.001087173.**

Starting j	Convergence to	Iterations	
1000	0.002887173	13	Down
400	0.001487173	1	Down
300	0.001287173	1	Down
230	0.001187173	1	Down
215	0.001087173	1	Down
210	0.001087173	1	Down
200	0.001087173	1	Up
180	0.001087173	1	Up
150	0.000987172	1	Up
140	0.000987172	1	Up

A sensible estimate of $p = 0.001$ is the average from the last 7 entries in Table 1 which gives $\hat{p} = 0.001072887$ with absolute error of 7.2887×10^{-5}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0.0001872, 0.0063686)$. Interestingly, with $N = 100$, $[0, F_B^{-1}(0.05^{1/N})] = [0, 0.0038]$.

Table 2. $\mathbf{p = 0.0001}$, $\boldsymbol{X}_0 \sim Gamma(1, 0.05)$, $\boldsymbol{X}_1 \sim Unif(0, 210)$, $\max(\boldsymbol{X}_0) = 77.61753$, $T = 184.2068$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.000015. **Shift at 0.0001039967.**

Starting j	Convergence to	Iterations	
888	0.0003439967	2	Down
577	0.0001339967	4	Down
450	0.0001189967	3	Down
350	0.0001189967	1	Down
310	0.0001039967	1	Down
300	0.0001039967	1	Down
290	0.0001039967	1	Up
280	0.0001039967	1	Up
270	0.0001039967	1	Up
260	0.0001039967	1	Up

A sensible estimate of $p = 0.0001$ is the value in the last 6 entries in Table 2 which gives $\hat{p} = 0.0001039967$ with absolute error of 3.9967×10^{-6}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.001628)$. Interestingly, with $N = 100$, $[0, F_B^{-1}(0.05^{1/N})] = [0, 0.0006]$.

4.2.3 Lognormal(1,1)

A sensible estimate of $p = 0.001$ is the average from the last 6 entries in Table 3 which gives $\hat{p} = 0.000999465$ with absolute error of 5.35×10^{-7}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.005572)$.

Table 3. $\mathbf{p = 0.001}$, $\boldsymbol{X}_0 \sim LN(1, 1)$, $\boldsymbol{X}_1 \sim Unif(0, 80)$, $\max(\boldsymbol{X}_0) = 32.36495$, $T = 59.75377$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.000999465.**

Starting j	Convergence to	Iterations	
1000	0.001199466	20	Down
950	0.001099466	12	Down
900	0.000999465	9	Down
800	0.000999465	4	Down
750	0.000999465	2	Down
700	0.000999465	1	Down
680	0.000999465	1	Up
680	0.000999465	1	Up
670	0.000999465	2	Up

Table 4. $\mathbf{p = 0.0001}$, $\boldsymbol{X}_0 \sim LN(1, 1)$, $\boldsymbol{X}_1 \sim Unif(0, 130)$, $\max(\boldsymbol{X}_0) = 44.82807$, $T = 112.058$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.000015. **Shift at 0.0001045544.**

Starting j	Convergence to	Iterations	
800	0.0001945544	22	Down
500	0.0001795544	9	Down
300	0.0001345544	4	Down
200	0.0001195544	1	Down
170	0.0001045544	1	Down
160	0.0001045544	1	Down
155	0.0001045544	1	Up
152	0.0001045544	1	Up
150	0.0001045544	2	Up

A sensible estimate of $p = 0.0001$ is the average from the last 5 entries in Table 4 which gives $\hat{p} = 0.0001045544$ with absolute error of 4.5544×10^{-6}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.003729)$.

4.2.4 Lognormal(0,1)

Table 5. $\mathbf{p = 0.001}$, $\boldsymbol{X}_0 \sim LN(0, 1)$, $\boldsymbol{X}_1 \sim Unif(0, 50)$, $\max(\boldsymbol{X}_0) = 11.86797$, $T = 21.98218$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.000999444.**

Starting j	Convergence to	Iterations	
1000	0.001099445	18	Down
900	0.001099445	5	Down
820	0.001099445	1	Down
800	0.000999444	2	Down
790	0.000999444	1	Down
780	0.000999444	1	Up
770	0.000999444	1	Up
760	0.001099445	3	Up

A sensible estimate of $p = 0.001$ is the average from the last 5 entries in Table 5 which gives $\hat{p} = 0.001019444$ with absolute error of 1.9444×10^{-5}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.005651)$.

A sensible estimate of $p = 0.0001$ is the average from the last 4 entries in Table 6 which gives $\hat{p} = 0.0001042241$ with absolute error of 4.2241×10^{-6}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.004288)$.

4.2.5 f(2,7)

A sensible estimate of $p = 0.001$ occurs at the Down-Up shift in Table 7 which gives $\hat{p} = 0.001003351$ with absolute error of 3.351×10^{-6}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.0104048)$. Interestingly, with $N = 100$, $[0, F_B^{-1}(0.05^{1/N})] = [0, 0.0066]$.

Table 6. $\mathbf{p = 0.0001}$, $\boldsymbol{X}_0 \sim LN(0,1)$, $\boldsymbol{X}_1 \sim Unif(0,70)$, $\max(\boldsymbol{X}_0) = 13.77121$, $T = 41.22383$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.000015. **Shift at 0.0001042241.**

Starting j	Convergence to	Iterations	
900	0.0002392241	27	Down
800	0.0001042241	25	Down
700	0.0001042241	17	Down
500	0.0001192241	5	Down
360	0.0001042241	1	Down
355	0.0001042241	1	Up
350	0.0001042241	1	Up
350	0.0001042241	1	Up

Table 7. $\mathbf{p = 0.001}$, $\boldsymbol{X}_0 \sim f(2,7)$, $\boldsymbol{X}_1 \sim Unif(0,50)$, $\max(\boldsymbol{X}_0) = 12.25072$, $T = 21.689$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.001003351.**

Starting j	Convergence to	Iterations	
500	0.001103351	9	Down
450	0.001003351	8	Down
400	0.001003351	6	Down
300	0.001003351	3	Down
210	0.001003351	1	Up
190	0.000903350	1	Up
180	0.000903350	1	Up
170	0.000903350	1	Up

A sensible estimate of $p = 0.0001$ occurs at the Down-Up shift in Table 8 which gives $\hat{p} = 0.0001041104$ with absolute error of 4.1104×10^{-6}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.003872)$.

4.2.6 Weibull(0.8,2)

In the 3rd entry there was an immediate convergence. A sensible estimate of $p = 0.001$ is the average from the last 5 entries in Table 9 which gives $\hat{p} = 0.001039261$ with absolute error of 3.9261×10^{-5}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.002953)$. Interestingly, with $N = 100$, $[0, F_B^{-1}(0.05^{1/N})] = [0, 0.0012]$.

A sensible estimate of $p = 0.0001$ is the average from the last 5 entries in Table 10 which gives $\hat{p} = 0.0001046393$ with absolute error of 4.6393×10^{-6}. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.004311)$.

4.2.7 2,4,6-trichlorophenol (ug/L)

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States.

We use NHANES trichlorophenol (dubbed urx3tb) data from [13]. There are 2604 observations of which the proportion exceeding $T = 9.5$ is $p = 0.001152074$. Consider the 2604

Table 8. $\mathbf{p = 0.0001}$, $\boldsymbol{X}_0 \sim f(2,7)$, $\boldsymbol{X}_1 \sim Unif(0,70)$, $\max(\boldsymbol{X}_0) = 14.62357$, $T = 45.13234$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.000015. **Shift at 0.0001041104.**

Starting j	Convergence to	Iterations	
750	0.0001341104	2	Down
740	0.0001041104	4	Down
730	0.0001041104	3	Down
660	0.0001041104	1	Down
650	0.0001041104	1	Up
645	0.0001041104	1	Up
640	0.0001041104	2	Up

Table 9. $\mathbf{p = 0.001}$, $\boldsymbol{X}_0 \sim Weibull(0.8,2)$, $\boldsymbol{X}_1 \sim Unif(0,40)$, $\max(\boldsymbol{X}_0) = 8.081707$, $T = 22.39758$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.000999262.**

Starting j	Convergence to	Iterations	
1000	0.001899263	2	Down
1000	0.001099263	7	Down
950	0.000999262	1	Immediate
950	0.000999262	1	Up
940	0.001099263	3	Up
940	0.000999262	2	Up

Table 10. $\mathbf{p = 0.0001}$, $\boldsymbol{X}_0 \sim Weibull(0.8,2)$, $\boldsymbol{X}_1 \sim Unif(0,50)$, $\max(\boldsymbol{X}_0) = 12.20032$, $T = 32.09036$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.000015. **Shift at 0.0001046393.**

Starting j	Convergence to	Iterations	
700	0.0002096393	20	Down
400	0.0001196393	10	Down
300	0.0001946393	1	Down
200	0.0001046393	4	Down
130	0.0001046393	1	Down
125	0.0001046393	1	Up
120	0.0001046393	1	Up
115	0.0001046393	1	Up

observations as a population and the problem is to estimate p from a sample $\boldsymbol{X}_0$ of size $n_0 = 100$ where $\max(\boldsymbol{X}_0) < T$.

we see from Table 11 that the shift from down to up occurs at $\hat{p} = 0.00119882$ close to p with absolute error 4.6746×10^{-5}. In this case the median 0.001091 from 10,000 B's provides an approximation to p . In general, however, the 3rd quartile (here 0.003386) is a more prudent first guess. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.007059)$.

Table 11. $\mathbf{p = 0.001152074}$, $\boldsymbol{X}_0$ a trichlorophenol sample. $\boldsymbol{X}_1 \sim Unif(0,30)$, $\max(\boldsymbol{X}_0) = 4.6$, $T = 9.5$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.00119882.**

Starting j	Convergence to	Iterations	
800	0.00119882	15	Down
700	0.00189882	8	Down
600	0.00119882	4	Down
590	0.00119882	5	Down
530	0.00119882	1	Up
520	0.00109882	1	Up
515	0.00109882	1	Up

4.2.8 Mercury (mg/kg)

The National Oceanic & Atmospheric Administration (NOAA) monitors the status of total mercury and methyl mercury in the coastal waters using oysters and sediments. We have used mercury data from NOAA's National Status and Trends Data [14].

The mercury data consist of 8,266 observations of which the proportion exceeding $T = 22.41$ is $p = 0.001088797$.

With 10,000 fusions, we see from Table 12 that the shift from down to up occurs at $\hat{p} = 0.001099501$ very close to the true $p = 0.001088797$. The median from 10,000 B's is 0.001704 giving an idea as to the magnitude of the true p. We note that $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0, 0.008299)$.

4.2.9 Application to Food Safety: Lead Exposure Risk (μg/week/kg)

A data set of 3,000 lead risk from lead exposure observations has been constructed in [1]. For $T = 25$ the tail probability is

Table 12. $\mathbf{p = 0.001088797}$, $\boldsymbol{X}_0$ a mercury sample. $\boldsymbol{X}_1 \sim Unif(0, 50)$, $\max(\boldsymbol{X}_0) = 11.9$, $T = 22.41$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.0001. **Shift at 0.001099501.**

Starting j	Convergence to	Iterations	
800	0.001199501	14	Down
700	0.001199501	11	Down
500	0.001199501	5	Down
400	0.001099501	1	Down
390	0.001099501	1	Up
380	0.001099501	1	Up
375	0.001199501	2	Up
360	0.001099501	2	Up

exactly $p = 0.001$, and with $N = 100$, we obtain on average from several cases corresponding to different $\max(\boldsymbol{X}_0)$ values the interval estimate of $[0, F_B^{-1}(0.05^{1/N})] = [0, 0.002675]$. The Down-Up sequences are shown in Table 13, where the shift occurs at $\hat{p} = 0.001004282$ very close to true p with an absolute error of 4.282×10^{-6}. We note that $\bar{B} = 0.0045076$ and $\hat{p} \in (B_{(1)}, B_{(10,000)}) = (0.0001943, 0.0120257)$.

Table 13. $\mathbf{p = 0.001}$, $\boldsymbol{X}_0$ a lead exposure sample. $\boldsymbol{X}_1 \sim Unif(0, 40)$, $\max(\boldsymbol{X}_0) = 12.12383$, $T = 25$, $n_0 = n_1 = 100$, $h = (x, \log x)$, p-increment 0.00009. **Shift at 0.001004282.**

Starting j	Convergence to	Iterations	
100	0.001184282	6	Down
60	0.001184282	3	Down
50	0.001184282	2	Down
40	0.001004282	3	Down
29	0.001004282	1	Down
28	0.001004282	1	Down
27	0.001004282	1	Down
26	0.001004282	1	Up
26	0.001094282	2	Up
25	0.001094282	3	Up
24	0.001004282	2	Up
20	0.001094282	4	Up

5 Variability of Point Estimates

Clearly, the iterative method (IM) can be repeated many times with different B-samples of size 1,000 taken from, say, 10,000 B's (10,000 is our default) to produce tail probability estimates as above from which variance approximations can be obtained. The following Table 14 shows typical variance approximations, obtained from single convergent sequences where the starting j corresponds to the 3rd quartile of the sampled 1,000 B's. Each entry was obtained from 1,000 runs for both $n_0 = n_1 = 100$ and $n_0 = n_1 = 200$. There is improvement in precision going from samples of size 100 to 200. In all cases reported here, and many other additional cases, $\sigma_{\hat{p}} = \mathrm{O}(10^{-4})$ for $p \doteq 0.001$. In the tables $\bar{p}$ is the average estimate of p from 1,000 runs.

6 Comparison: ROSF vs POT

From a practical view point, some comparison is needed to assess the relative merit of ROSF/IM. We provide in what follows a limited comparison against a well known method for $n_0 = 100$ across different tail types. An additional comparison for $n_0 = 200$ can be found in [12]. However, a more extensive comparison is warranted and will be dealt with elsewhere.

Table 14. SD($\hat{p}$) for $p = 1 - G(T) \doteq 0.001$, $n_0 = n_1$, $h(x) = (x, \log x)$.

$\boldsymbol{X}_0 \sim$ Urx3t, $T = 9.5$, $\boldsymbol{X}_1 \sim Unif(0, 12)$	$\bar{p}$	$\sigma_{\hat{p}}$
$n_0 = 100$, $\max(\boldsymbol{X}_0) = 8.8$	0.0011539	0.0004269
$n_0 = 200$, $\max(\boldsymbol{X}_0) = 8.8$	0.0011216	0.0002871
$\boldsymbol{X}_0 \sim$ LN(0,1),$T = 21.98$, $\boldsymbol{X}_1 \sim Unif(1, 60)$	$\bar{p}$	$\sigma_{\hat{p}}$
$n_0 = 100$, $\max(\boldsymbol{X}_0) = 11.04102$	0.0011401	0.0004100
$n_0 = 200$, $\max(\boldsymbol{X}_0) = 11.04102$	0.0010598	0.0002823
$\boldsymbol{X}_0 \sim$ Weibull(1,2),$T = 13.82$,$\boldsymbol{X}_1 \sim$ Unif(0,16)	$\bar{p}$	$\sigma_{\hat{p}}$
$n_0 = 100$, $\max(\boldsymbol{X}_0) = 8.626444$	0.0011215	0.0001524
$n_0 = 200$, $\max(\boldsymbol{X}_0) = 8.673713$	0.0010768	0.0001372
$\boldsymbol{X}_0 \sim$ Pareto(1,4),$T = 5.62$,$\boldsymbol{X}_1 \sim$ Unif(1,8)	$\bar{p}$	$\sigma_{\hat{p}}$
$n_0 = 100$, $\max(\boldsymbol{X}_0) = 3.08099$	0.0011549	0.0002256
$n_0 = 200$, $\max(\boldsymbol{X}_0) = 4.14516$	0.0009966	0.0002034

Thus, against the background provided in the previous sections, we compare two very different ways to obtain estimates for small tail probabilities, the well known peaks over threshold (POT) based on extreme value theory, and the present 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 and width, and on the mean absolute error (MAE) which measures the discrepancy between $\hat{p}$ and the true tail probability p. In Tables 15 to 23, p is relatively small, $p = 0.001$ (or approximately so), whereas in the last two Tables 24 and 25, p is smaller, $p = 0.0001$.

Throughout the comparison the sample sizes are $n_0 = n_1 = 100$, and $\boldsymbol{h}(x) = (x, \log x)$. Thus, in the present comparison the reference $\boldsymbol{X}_0$ and the fusion samples $\boldsymbol{X}_1$ have size $n_0 = 100$.

To save computation time, in each case of the iteration process F_B was obtained from 1000 fusions, and we use in each case a single convergent Down sequence where the starting j is such that $B_{(j)}$ is approximately equal to the 3rd quartile of the observed 1000 B's. Starting at the 3rd quartile is computationally sensible as the corresponding $B_{(j)}$ most often converge to a point in a neighborhood of p as j decreases. See Tables 1 to 13 and more examples in [12].
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 $\boldsymbol{X}_0$ for fitting the generalized Pareto (GP) distribution. We have noticed a deterioration in the POT results when using 30%, 15% or 10% of $\boldsymbol{X}_0$. The simulation details are given in Section A.2 in the Appendix.

An interesting picture emerges from Tables 15 to 25. For the moderately large sample size of $n_0 = 100$, regardless of the tail type, already with $\mathbf{N = 50}$, that is, the number of $\hat{p}$'s used in forming the CI for the true p of the form $(\min(\hat{p}), \max(\hat{p}))$ (defined in Section A.2), the iterative process gives reliable and relatively narrow confidence intervals, whereas the POT gives unacceptable coverage and in most cases wider CI's and greater MAE as well. It is seen from [12], the POT coverage increases significantly going from $n_0 = 100$ to $n_0 = 200$, however, it seems that for the method to "fire up" larger samples are needed. Regarding ROSF, the choice of $\mathbf{N = 50}$ seems prudent across all cases.

6.1 Comparison Tables

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

Table 15. $X_0 \sim \mathbf{t}_{(1)} > 0 : p = 1 - G(T) = 0.001, T = 631.8645, X_1 \sim$ Unif(0,800), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	63.2%	0.00372	0.00149
ROSF & IM	**50**	98.2%	0.00213	0.00061

Table 16. $X_0 \sim$ **Weibull**$(1, 2) : p = 1 - G(T) = 0.001, T = 13.81551, X_1 \sim$ Unif(0,16), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	82.7%	0.00431	0.00131
ROSF & IM	**50**	92.5%	0.00287	0.00068

Table 17. $X_0 \sim$ **Pareto**$(1, 4) : p = 1 - G(T) = 0.001, T = 5.623413, X_1 \sim$ Unif(1,8), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	81.8%	0.00419	0.00121
ROSF & IM	**50**	96.2%	0.00232	0.00052

Table 18. $X_0 \sim$ **Gamma**$(3, 1) : p = 1 - G(T) = 0.001, T = 11.22887, X_1 \sim$ Unif(0,20), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	77.3%	0.00410	0.00096
ROSF & IM	**50**	93.4%	0.00188	0.00054

Table 19. $X_0 \sim$ **IG**$(2, 40) : p = 1 - G(T) = 0.001, T = 3.835791, X_1 \sim$ Unif(0,8), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	69.6%	0.00324	0.00123
ROSF & IM	**50**	100%	0.00289	0.00047

Table 20. $X_0 \sim$ **LN**$(0, 1) : p = 1 - G(T) = 0.001, T = 21.98218, X_1 \sim$ Unif(1,60), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	81.5%	0.00451	0.00111
ROSF & IM	**50**	100%	0.00234	0.00047

Table 21. $X_0 \sim$ **LN**$(1, 1) : p = 1 - G(T) = 0.001, T = 59.75377, X_1 \sim$ Unif(1,140), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	81.4%	0.00435	0.00117
ROSF & IM	**50**	89.1%	0.00187	0.00069

Table 22. $X_0 \sim$ Mercury : $p = 1 - G(T) = 0.001088797, T = 22.41, X_1 \sim$ Unif(0,50), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.0001. Data source [14].

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	85.3%	0.00455	0.00130
ROSF & IM	**50**	97.5%	0.00215	0.00048

Table 23. $X_0 \sim$ URX3TB : $p = 1 - G(T) = 0.001152074, T = 9.50, X_1 \sim$ Unif(0,12), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.0001. Data source for URX3TB - 2,4,6-trichlorophenol [13].

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	81.1%	0.00433	0.00143
ROSF & IM	**50**	89.1%	0.00179	0.00055

Table 24. Smaller probability. $X_0 \sim$ **F**$(2, 12) : p = 1 - G(T) = 0.0001, T = 21.84953, X_1 \sim$ Unif(0,25), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.00001.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	71.4%	0.00062	0.00052
ROSF & IM	**50**	95.2%	0.00059	0.00022

Table 25. Smaller Probability. $X_0 \sim$ **LN**$(0, 1) : p = 1 - G(T) = 0.0001, T = 41.22383, X_1 \sim$ Unif(1,60), $n_0 = n_1, \boldsymbol{h}(x) = (x, \log x)$. p-increment 0.00001.

		$n_0 = 100$		
Method	N	Coverage	CI Length	MAE
POT	-	72.1%	0.00064	0.00045
ROSF & IM	**50**	100%	0.00066	0.00021

7 Conclusion

The numerous number of fusions of a given reference sample with computer generated samples gives rise to different 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 rather closely using an iterative procedure.

We have illustrated that, across a fairly wide range of distributional tail types, repeated fusion of a reference sample with externally generated uniform random data allowed us to gain information about the tail behavior beyond the threshold using order statistics from upper bounds for p. In neighborhoods of the true p, the consequential Down-Up sequences tended to transition or shift at points close to p, providing surprisingly close estimates. We have seen that with sample sizes on the order of 100 we can in many cases estimate tail probabilities on the order of 1/10,000. It seems that larger samples are needed for much smaller tail probabilities, and that the method could fail when $\max(\boldsymbol{X}_0)$ is exceedingly small relative to the threshold T.

Throughout the paper the fusion samples were uniform samples whose support contained T. That is, the upper limit of the support exceeded T. But other than this, no guide for choosing the upper limits was provided. Experience, however, shows that different upper limits give similar results.

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 together with appropriate inferential methods other than the semiparametric method used in the paper. That is, exploring different ways of connecting $\boldsymbol{X}_0$ and the fusion samples, other than by means of their distributions as expressed by the density ratio model.

Reliable estimation of tail probabilities is important in numerous fields from finance to geophysics to meteorology to the design of ships and to optics; see [15] and [16].

A Appendix

The appendix addresses the density ratio model (9) for $m+1$ data sources discussed briefly in Section 3.

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 $t_1, \ldots, t_n$ [17]. Let $p_i = dG(t_i)$ be the mass at t_i, for $i = 1, \ldots, n$. Then the empirical likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta}, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \beta_1' h(x_{1j})) \cdots \prod_{j=1}^{n_m} \exp(\alpha_m + \beta_m' h(x_{mj})). \tag{13}$$

Maximizing $\mathcal{L}(\boldsymbol{\theta}, G)$ subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, \ \ldots, \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0 \tag{14}$$

we obtain the desired estimates. In particular,

$$\hat{G}(t) = \frac{1}{n_0} \times \tag{15}$$

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

where $I(t_i \le t)$ equals one for $t_i \le t$ 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 [11],[10],[18].

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

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

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

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

Then the asymptotic distribution of $\hat{G}(t)$ for $m \ge 1$ 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*

$$\text{Cov}\{\sqrt{n}(\hat{G}(t) - G(t)), \sqrt{n}(\hat{G}(s) - G(s))\} = \left(\sum_{k=0}^{m} \rho_k\right) \left(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^{m} \rho_j A_j(t \wedge s)\right) + \left(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes I_p)\right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes I_p)\bar{B}(t) \end{pmatrix} \tag{16}$$

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

For a complete proof see Lu [10]. The proof for $m = 1$ is given in [18].

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

$$\left(\hat{G}(t) - z_{\alpha/2}\sqrt{\hat{V}(t)},\ \ \hat{G}(t) + z_{\alpha/2}\sqrt{\hat{V}(t)}\right), \qquad (17)$$

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

$$\left(1-\hat{G}(t) - z_{\alpha/2}\sqrt{\hat{V}(t)}, 1-\hat{G}(t) + z_{\alpha/2}\sqrt{\hat{V}(t)}\right). \qquad (18)$$

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 $\boldsymbol{X}_0$ 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 $\boldsymbol{X}_0$ for fitting the generalized Pareto (GP) distribution.

ROSF/IM:
$\boldsymbol{X}_0$ was fused with $\boldsymbol{X}_1$ 1000 times (ROSF) to get F_B and then $\hat{p}$ (IM).
$\boldsymbol{X}_0$ was fused again with different $\boldsymbol{X}_1$ 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 $\boldsymbol{X}_0$, 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 $N = 50$.

Statistics in Medicine 2016; **35**: 3229-3240.

REFERENCES

[1] Kedem, B., Pan, L., Zhou, W., and Coelho, C.A. Interval estimation of small tail probabilities – application in food safety.

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

[3] 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.

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

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

[6] 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.

[7] Katzoff, M., Zhou, W., Khan, D., Lu, G., and Kedem, B. Out of sample fusion in risk prediction. *Journal of Statitsical Theory and Practice* 2014; **8**: 444-459.

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

[9] Fokianos, K., Kedem, B., Qin J. and Short, D.A. A Semiparametric Approach to the One-Way Layout. *Technometrics*, (2001); **43**: 56-65.

[10] 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.

[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] Kedem, B., Pan, L., Smith, P. and Wang, C. (2018). Repeated out of sample fusion in the estimation of small tail probabilities. arXiv:1803.10766v2 [stat.ME], May 2018.

[13] The National Health and Nutrition Examination Survey, Online available from https://wwwn.cdc.gov/nchs/nhanes

[14] National Oceanic & Atmospheric Administration, Coastal Science, Online available from https://products.coastalscience.noaa.gov/nsandt_data/data.aspx

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

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

[17] Owen, A.B. *Empirical Likelihood*. CRC Press. Boca Raton: 2001

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