Estimation of Tail Probabilities by Repeated Augmented Reality

Benjamin Kedem[1] and Saumyadipta Pyne[2,3]*

[1]Department of Mathematics and Institute for Systems Research
University of Maryland, College Park, MD

[2]Public Health Dynamics Laboratory and Department of Biostatistics,
Graduate School of Public Health,
University of Pittsburgh, Pittsburgh, PA
[3]Health Analytics Network, PA

*Email: spyne@pitt.edu

August 2020

Abstract

Synthetic data can enhance patterns in real data and thus provide insights into different phenomena. Here, the estimation of tail probabilities of rare events from a moderately large number of observations is considered. The problem is approached by a large number of augmentations or fusions of the real data with computer-generated synthetic samples. The tail probability of interest is approximated by subsequences created by a novel iterative process. The estimates are found to be quite precise.

Keywords: Repeated out of sample fusion, density ratio model, residential radon, upper bounds, iterative process, B-curve.

MSC 2000: Primary 62F40; Secondary 62F25

1 Introduction

The citation accompanying his U.S. National Medal of Science in 2002 honored C.R. Rao "as a prophet of new age for his pioneering contributions to the foundations of statistical theory and multivariate statistical methodology and their applications." When Professor Rao organized the 'International Conference on the Future of Statistics, Practice and Education' in Hyderabad (Indian School of Business, 12.29.04–01.01.05), one of us participated in it. Befitting this connection, we decided to contribute what we believe is a "futuristic" application of augmented reality in honor of Professor C.R. Rao's 100^{th} birthday on September 10, 2020.

In its February 4^{th} 2017 edition, *The Economist* noted the promise of augmented reality, claiming that "Replacing the real world with a virtual one is a neat trick. Combining the two could be more useful." We concur. Combining real data with synthetic data, i.e., *augmented reality* (AR), opens up new perspectives regarding statistical inference. Indeed, augmentation of real data with virtual information is an idea that has already found applications in fields such as robotics, medicine, and education.

In this article, we advance the notion of *repeated* augmented reality in the estimation of very small tail probabilities even from moderately sized samples. Our approach, much like the bootstrap, is computationally intensive and could not have been viable without the computing power of modern systems. However, rather than looking repeatedly *inside* the sample, we look repeatedly *outside* the sample. Fusing a given sample repeatedly with computer-generated data is referred to as *repeated out of sample fusion*

(ROSF) in Kedem, et al. (2016), (2019). Related ideas concerning a single fusion are studied in Fithian and Wager (2015), Fokianos and Qin (2008), Katzoff, et al. (2014), and Zhou (2013).

In 1984, the so-called "Watras incident" drew intense media and congressional attention in the U.S. to the problem of residential exposure to radon, a known carcinogenic gas. Radon at the Boyertown home of Diane and Stanley Watras, the latter a construction engineer, located in Berks county, on the Reading Prong geological formation in Pennsylvania, was recorded as almost 700 times the safe level, which is a lung cancer risk equivalent of smoking 250 packs of cigarettes per day! As noted by George (2015), this news caused a major alarm and led the U.S. Environmental Protection Agency to establish a radon measurement program. In this regard, the present article will review the underpinnings of ROSF in estimation of small tail exceedance probabilities. We will illustrate its application using residential radon level data from Beaver County, Pennsylvania.

1.1 The Problem

Consider a random variable $X \sim g$ and the corresponding moderately large random sample $\boldsymbol{X}_0 = (X_1, ..., X_{n_0})$ where all the observations are smaller than a high threshold T, that is $\max(\boldsymbol{X}_0) < T$. We wish to estimate $p = P(X > T)$ without knowing the distribution g. 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. This leads us to ROSF.

1.2 The Approach

Let $\boldsymbol{X}_i$ denote the ith computer-generated sample of size $n_i = n_1 = n_0$. Then the fused samples are the *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 supported on $(0, U)$, where $U > T$. The number of fusions can be as large as we wish. From each pair $(\boldsymbol{X}_0, \boldsymbol{X}_j)$, under a mild condition, we get in a certain way an upper bound B_j for p. Let $\{B_{(j)}\}$ be the sequence of order statistics. Then the sorted pairs

$$(1, B_{(1)}), (2, B_{(2)}), (3, B_{(3)}), \ldots (n, B_{(N)})$$

produce a monotone curve, referred to as the B-curve, which for large N, contains a point "$\bullet$" as in Figure 1. As N increases, the ordinate of the point essentially coincides with p with probability approaching one. It follows that the sequence

$$B_{(1)}, B_{(2)}, B_{(3)}, \ldots B_{(N)}$$

contains subsequences which approach p. The subsequences can be obtained by an iterative process to be described in Section 3.

1.3 Illustrations of an Iterative Process

Deferring details to later sections, it is helpful to shed light early on and introduce our iterative method which produces estimates of tail probabilities,

using reference samples $\boldsymbol{X}_0$ from $F(2,7)$ and $LN(1,1)$ distributions.

In the first illustration, $\boldsymbol{X}_0$ is a random sample from $F(2,7)$, $T = 21.689$, giving $p = 0.001$. Here, $n_0 = n_1 = 100$, $\max(\boldsymbol{X}_0) = 12.25072$, and the computer-generated samples consist of independent $Unif(0,50)$. With $N = 10,000$ fusions, and starting from $j = 450$, our iterative process (9) bellow produces a converging subsequence which approaches p from above, a "Down" subsequence:

$$450 \to 0.001703351 \to 438 \to 0.001603351 \to 407 \to 0.001503351 \to$$
$$369 \to 0.001403351 \to 341 \to 0.001303351 \to 312 \to 0.001203351 \to$$
$$278 \to 0.001103351 \to 246 \to 0.001003351 \to 221 \to 0.001003351 \cdots$$

Starting from $j = 210$, our iterative process (9) produces an "Up" subsequence which converges by a single iteration giving:

$$210 \to 0.001003351 \to 219 \to 0.001003351 \to 219 \to 0.001003351 \cdots$$

In the second illustration, $\boldsymbol{X}_0$ is a random sample from $LN(1,1)$, $T = 59.75377$, giving $p = 0.001$. Here $n_0 = n_1 = 200$, $\max(\boldsymbol{X}_0) = 33.63386$, and the computer-generated samples consist of independent $Unif(0,80)$. With $N = 10,000$ fusions, and starting from $j = 800$, our iterative process (9) below produces a converging "Down" subsequence which approaches p from above by a single iteration:

$$800 \to 0.001000281 \to 788 \to 0.001000281 \to 788 \to 0.001000281 \cdots$$

And staring from $j = 790$, our iterative process (9) produces an "Up" subsequence which converges by a single iteration giving:

$$790 \to 0.001000281 \to 815 \to 0.001000281 \to 815 \to 0.001000281 \cdots$$

Notice that the "Down-Up" convergence in both illustrations is remarkably close to the true $p = 0.001$. We have had quite a few similar results where the tail behavior differed markedly. The computation here required an important parameter called "p-increment" which in the present examples was 0.0001. We shall deal with this numerical issue soon.

1.4 A Useful Feature

A useful feature of the present article is the realization that we can come up with educated guesses about to the magnitude of p from the value of $\max(\boldsymbol{X}_0)$ relative to T. And this in turn suggests a set of discrete points in the interval $(\min(B_j), \max(B_j))$ at which p-estimates are evaluated. The difference between any two discrete points is the "p-increment" mentioned above.

2 Getting Upper Bounds for p by Data Fusion

Recall that $\boldsymbol{X}_0 = (X_1, ..., X_{n_0})$ is a reference sample from some reference probability density (pdf) $g(x)$ and let $G(x)$ denote the corresponding distribution function (CDF) . Since we shall deal with radon data, we assume

that $x \in (0, \infty)$. The goal is to estimate a small tail probability

$$p = P(X > T) = 1 - G(T) = \int_T^\infty g(x)dx.$$

Let $\boldsymbol{X}_1$ be a computer-generated random sample of size n_1 and assume $\boldsymbol{X}_1 \sim g_1, G_1$. The augmentation

$$\boldsymbol{t} = (t_1, \ldots, t_{n_0+n_1}) = (\boldsymbol{X}_0, \boldsymbol{X}_1), \tag{2}$$

of size $n_0 + n_1$ gives the fused data from $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$. We shall assume the density ratio model (Qin and Zhang 1997, Lu 1997)

$$\frac{g_1(x)}{g(x)} = \exp(\alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{h}(x)) \tag{3}$$

where α_1 is a scalar parameter, $\boldsymbol{\beta}_j$ is an $r \times 1$ vector parameter, and $\boldsymbol{h}(x)$ is an $r \times 1$ vector valued function. Clearly, to generate $\boldsymbol{X}_1$ we must know the corresponding g_1. 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 and the corresponding G, G_1, and none of the parameters α_1 and $\boldsymbol{\beta}_1$ 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 weakened considerably by the mild assumption (4) below, which may hold even for misspecified $\boldsymbol{h}$, as numerous examples with many different tail types show. Accordingly, based on numerous experiments, some of which discussed in Kedem et al. (2019), we assume the "gamma tilt" $h(x) = (x, \log x)$. Further justification for the

gamma tilt is provided by our data analysis below.

Under the density ratio model (11), the maximum likelihood estimate of $G(x)$ based on the fused data $\boldsymbol{t} = (\boldsymbol{X}_0, \boldsymbol{X}_1)$ is given in (14) 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 (17). In particular we get an upper bound B_1 for p. In the same way, from additional independent computer-generated samples $\boldsymbol{X}_2, \boldsymbol{X}_3, ..., \boldsymbol{X}_N$ we get upper bounds for p from the pairs $(\boldsymbol{X}_0, \boldsymbol{X}_2), (\boldsymbol{X}_0, \boldsymbol{X}_3), ...(\boldsymbol{X}_0, \boldsymbol{X}_N)$. Thus, conditional on $\boldsymbol{X}_0$, 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

$$0 < F_B(p) < 1 \tag{4}$$

so that

$$P(B_1 > p) = 1 - F_B(p) > 0.$$

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

$$(1, B_{(1)}), (2, B_{(2)}), ..., (N, B_{(N)}) \tag{5}$$

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 (5) is referred to as the *B-curve*.

We now make the following important observations.

a. Assumption (4) implies that as N increases,

$$B_{(1)} < p < B_{(N)} \tag{6}$$

with probability approaching one.

b. The point " $\bullet$ " moves down the B-curve when $max(\boldsymbol{X}_0)$ approaches T. The point " $\bullet$ " moves up the B-curve when $max(\boldsymbol{X}_0)$ decreases away from T.

Hence as N increases, the size of $max(\boldsymbol{X}_0)$ relative to T provides useful information as to the approximate magnitude of p. Specifically, the first quartile of $B_1, B_2, ..., B_N$ is a sensible guess of p as $max(\boldsymbol{X}_0)$ approaches T, and the third quartile, or even max(B), is a sensible approximation of p when $max(\boldsymbol{X}_0)$ is small. Otherwise the mean or median of $B_1, B_2, ..., B_N$ provides practical guesses of the approximate magnitude of p.

c. 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. Hence, as seen from **b**, F_B provides information about p.

Knowing F_B is a significant consequence of repeated out of sample fusion. Its implication is that the exact distribution of any $B_{(j)}$ is practically known.

3 Capturing p

For a sufficiently large number of fusions N, the monotonicity of the B-curve and (6) imply there are $B_{(j)}$ which approach p from above so that there is a $B_{(j)}$ very close to p. Likewise, the $B_{(j)}$ can approach p from below. Thus, the B-curve establishes a relationship between j and p. Another relationship between j and p is obtained from the well known distribution of order statistics,

$$P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p)]^k [1 - F_B(p)]^{N-k} \tag{7}$$

which can be computed since F_B is practically known for sufficiently large N. Iterating between these two relationships provides a way to approximate p as is described next. From (7) we can get the smallest p_j such that

$$P(B_{(j)} > p_j) = \sum_{k=0}^{j-1} \binom{N}{k} [F_B(p_j)]^k [1 - F_B(p_j)]^{N-k} \leq 0.95, \tag{8}$$

The 0.95 probability bound was chosen arbitrarily and can be replaced by other high probabilities. It is important to note that in practice, and in what

follows, the p_j in (8) are evaluated on a grid incrementally along specified small increments. Thus, with $B_{(j_k)}$'s from the B-curve, and $p_{(j_k)}$'s the smallest p's satisfying (8) with $j = j_k$, and $B_{(j_{k+1})}$ closest to $p_{(j_k)}$, $k = 1, 2, ...$, we have the iterative process,

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

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,

$$j_1 \to p_{(j_1)} \to j_2 \to p_{(j_2)} \to \cdots j_k \to p_{j_k} \to j_{k+1} \to p_{j_k} \to j_{k+1} \to p_{j_k} \cdots \quad (9)$$

In general, starting with any j, convergence occurs when for the first time $B_{(j_k)} = B_{(j_{k+1})}$ for some k and and we keep getting the same probability p_{j_k}. Clearly, the p_{j_k} sequence could decrease or increase producing "down" and "up" subsequences. For example, suppose 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 B-curve we get the closest 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

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

However, when the probabilities are sufficiently low it is possible for the closest $B_{(j)}$ approximations of the p_j to reverse course leading to an increasing "up" sequence

$$B_{(j'_1)} \leq B_{(j'_2)} \leq B_{(j'_3)} \cdots .$$

This "down-up" tendency has been observed numerous times with real and artificial data. It manifests itself clearly in the radon examples below. In particular, as was illustrated earlier in Section 1.3, this "down-up" phenomenon tends to occur 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. This is summarized in the following proposition.

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 (8) where the p_j are evaluated along appropriate numerical increments. Then, (8) 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.*

4 Illustrations Using Radon Exposure Data

We shall now demonstrate the proposition using radon data examples. Many additional examples were given in Kedem et al. (2019). All the examples point to a remarkable "down-up" patterns in a neighborhood of the true p, providing highly precise estimates of p. It should be noted that the number of iterations decreases as the $B_{(j)}$ approach p, suggestive of the fact that convergence is about to occur.

The iterative process (9) is repeated with different starting j's until a clear pattern emerges in which different successive j's give rise to Down-Up subsequences that converge to the same value, which is our estimate $\hat{p}$. The process may be repeated with different p-increments.

4.1 Computational Considerations

To enable computation with R, in (8) the binomial coefficients were evaluated with $N = 1000$, as if there were 1000 fusions only. However, there are no restrictions on the number of fusions and F_B was obtained throughout from $10,000$ fusions, and hence $10,000$ B's. Each entry in the following tables was obtained from a *different* sample of 1,000 B's sampled at random from 10,000 B's. More precisely, each entry was obtained from an approximate B-curve obtained from the sampled 1000 B's and an approximate (8) with $N = 1000$. Thus, for each entry, we iterated between an approximate B-curve and approximate (8) with $N = 1000$.

4.1.1 Choice of p-increment

An important consideration is the choice of the increments of p along which the probability (8) is evaluated. Certainly, any approximation of p must reside between consecutive B's. Hence, sensible p-increments are fractions of either the mean, median, 1st or 3rd quartiles, or even fractions of $\max(B) = B_{(10,000)}$. In the following example, the p-increments are of the order of magnitude approximately equal to one tenth of one of these quantities.

4.1.2 Beaver County Radon Tail Probabilities

Radon-222, or just radon, is a tasteless, colorless and odorless radioactive gas, which is a product of Uranium-238 and Radium-226, both of which are naturally abundant in the soil. Radon is known around the world as a carcinogen, and its exposure is the leading risk factor of lung cancer among non-smokers. Geological radon exposure takes place mostly through cracks and openings in the ground due to underlying geological formations. Approximately 40 percent of Pennsylvania (PA) homes have radon levels above the U.S. EPA action guideline of 4 pCi/L. Residential radon test levels were collected by PA Department of Environmental Protection (PADEP) statewide in the period from 1990 to 2007. See Zhang et al. (2020a) for a study of indoor radon concentrations from Beaver County and its neighboring counties in PA.

In the following examples, ROSF is applied to Beaver County radon data from 1989 to 2017, for various p-increments. There were 7425 radon observations, taken as a population, of which only 2 exceed 200. Hence, with $T =$

200 we wish to estimate the small probability $p = 2/7425 = 0.0002693603$. Throughout the examples, $\boldsymbol{X}_0$ is a reference random sample chosen without replacement from the 7425 radon observations. The generated $\boldsymbol{X}_1$ samples are from $Unif(0, 300)$ and $n_0 = n_1 = 500$.

In the tables below, "Down", "Up", "No j change", means that in the iterative process (9) there was a downward, or upward, or no change in j, respectively.

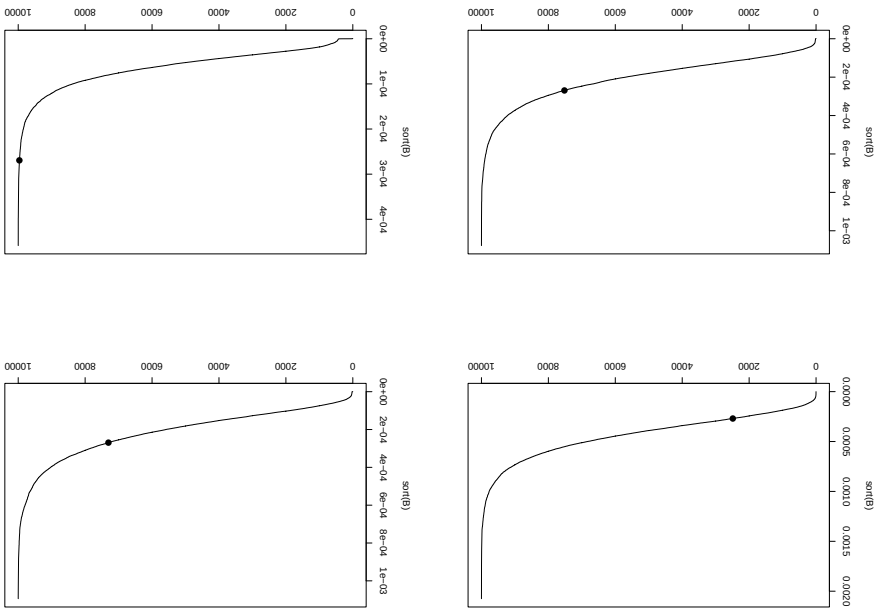

Figure 1: B-Curves, 10,000 B's, from residential radon sample $\boldsymbol{X}_0$. $p = 0.0002693603$, $\boldsymbol{X}_1 \sim Unif(0, 300)$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$. $\max(\boldsymbol{X}_0)$ values: top left 77.9, top right 107, bottom left 143, bottom right 193.7. The point "•" moves to the left as $\max(\boldsymbol{X}_0)$ increases relative to $T = 200$. The fusion samples are uniform with support covering T.

Figure 1 shows how the "•" moves along the B-curve as a function of the size of $\max(\boldsymbol{X}_0)$ relative to T. The figure should be referred to when reading the following examples. **Example 1:** $\max(\boldsymbol{X}_0) = 107$

Since 107 is close to $T/2$, the "•" point is in the "middle" of the B-curve, far removed from both ends. Hence we use as p-increment $Median/10 \approx 0.000018$. We observed that the 3rd quartile was 0.0002686, very close to the true p.

From Table 1, the shift from down to up occurs at $\hat{p} = 0.0002689389$ very close the the true $p = 0.0002693603$, giving an error of an order of 10^{-7}.

Example 2: $\max(\boldsymbol{X}_0) = 123.1$

Table 1: $\mathbf{p = 0.0002693603}$, $\boldsymbol{X}_1 \sim Unif(0, 300)$, $\max(\boldsymbol{X}_0) = 107$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$, p-increment 0.000018.

Starting j	Convergence to	Iterations	
1000	0.0007009389	3	Down
802	0.0002869389	1	Down
761	0.0002689389	1	Down
757	0.0002689389	1	Down
755	0.0002689389	1	Down
754	0.0002689389	1	Up
751	0.0002689389	1	Up
750	0.0002689389	1	Up
740	0.0002689389	1	Up
738	0.0002689389	1	Up

A different reference sample $\boldsymbol{X}_0$ was fused again 10,000 times with different $\boldsymbol{X}_1 \sim Unif(0, 300)$ independent samples. Since $\max(\boldsymbol{X}_0) = 123.1$, again we have, relative to $T = 200$, a "middle" "•" point suggesting a p-increment of one tenth of the mean of the B's. As the order of the mean was 10^{-4} we chose p-increment 0.00002, which is of the same order as that of $Mean(B)/10$.

From Table 2, the shift from Down to Up occurs at $\hat{p} = 0.0002601254$ not

far from $p = 0.0002693603$, giving an error on the order of 10^{-5}. **Example**

Table 2: $\mathbf{p = 0.0002693603}$, $\boldsymbol{X}_1 \sim Unif(0, 300)$, $\max(\boldsymbol{X}_0) = 123.1$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$, p-increment 0.00002.

Starting j	Convergence to	Iterations	
800	0.0003401254	18	Down
750	0.0003001254	18	Down
140	0.0002801254	2	Down
135	0.0002601254	1	Down
133	0.0002601254	1	Down
130	0.0002601254	1	Up
122	0.0002601254	1	Up
121	0.0002601254	1	Up
120	0.0002601254	1	Up
112	0.0002601254	1	Up

3: $\max(\boldsymbol{X}_0) = 193.7$

A different reference sample $\boldsymbol{X}_0$ was fused again 10,000 times with different $\boldsymbol{X}_1 \sim Unif(0, 300)$ independent samples. Since $\max(\boldsymbol{X}_0) = 193.7$, we have, relative to $T = 200$, a "•" point close to the lower end of the B-curve, suggesting a p-increment on the order of one tenth of the 1st quartile of the 10,000 B's. As the 1st quartile was 0.0002697 we chose a p-increment of 0.00001. A p-increment of 0.00002 gave identical results. We observe that the 1st quartile is very close to p.

From Table 3, the shift from Down to Up occurs at $\hat{p} = 0.0002600818$ not far from $p = 0.0002693603$, giving an error on the order of 10^{-5}. **Example 4:** $\max(\boldsymbol{X}_0) = 77.9$

A different reference sample $\boldsymbol{X}_0$ was fused again 10,000 times with different $\boldsymbol{X}_1 \sim Unif(0, 300)$ independent samples. Since $\max(\boldsymbol{X}_0) = 77.9$, we have, relative to $T = 200$, a "•" point close to the upper end of the B-curve, a difficult case, suggesting a p-increment on the order of one tenth of $\max(B)$ from 10,000 B's. As $\max(B) = 0.0004583$ we chose a p-increment of 0.00004583.

Table 3: $\mathbf{p = 0.0002693603}$, $\boldsymbol{X}_1 \sim Unif(0, 300)$, $\max(\boldsymbol{X}_0) = 193.7$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$, p-increment 0.00001.

Starting j	Convergence to	Iterations	
800	0.0003600818	21	Down
600	0.0002600818	19	Down
440	0.0002700818	9	Down
300	0.0002600818	4	Down
246	0.0002600818	1	Down
245	0.0002600818	1	Down
244	0.0002600818	1	Up
243	0.0002600818	1	Up
240	0.0002600818	1	Up
237	0.0002600818	1	Up
222	0.0002500818	1	Up
200	0.0002400818	1	Up

From Table 4, the shift from Down to Up occurs at $\hat{p} = 0.0002286204$ not far from $p = 0.0002693603$, giving an error on the order of 10^{-5}.

Table 4: $\mathbf{p = 0.0002693603}$, $\boldsymbol{X}_1 \sim Unif(0, 300)$, $\max(\boldsymbol{X}_0) = 77.9$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$, p-increment 0.00004583.

Starting j	Convergence to	Iterations	
1000	0.0002744504	2	Down
999	0.0002744504	1	Down
998	0.0002744504	1	Down
997	0.0002286204	2	Down
996	0.0002286204	1	Down
994	0.0002286204		No j change
993	0.0002286204	1	Up
992	0.0002286204	1	Up
991	0.0002286204	1	Up
990	0.0002286204	1	Up
989	0.0002286204	1	Up
988	0.0002286204	1	Up

Summary of ROSF applied to Beaver radon data.

Table 5 provides our estimates of $p = 0.0002693603$ from various random radon samples $\boldsymbol{X}_0$ of size $n_0 = 500$ fused repeatedly with independent $\boldsymbol{X}_1 \sim Unif(0, 300)$ of size $n_1 = 500$. In all cases $h(x) = (x, \log x)$. Some of the $\boldsymbol{X}_0$ samples are the same, but the p-increments are different still leading

to similar results. The mean and standard deviation of the $\hat{p}$ in the table are equal to $\bar{\hat{p}} = 0.0002606333$ and $1.052197e-05$, respectively. In general, variance estimates can be obtained by repeating ROSF again and again using different B-curves and different p-increments. Evidently the choice of $h(x) = (x, \log x)$ is a reasonable choice as the present radon analysis and many other examples with very diverse tail types indicate.

Table 5: $\mathbf{p = 0.0002693603}$, $\boldsymbol{X}_1 \sim Unif(0, 300)$, $T = 200$, $n_0 = n_1 = 500$, $h = (x, \log x)$.

$\max(\boldsymbol{X}_0)$	p-increment	$\hat{p}$	Error
77.9	0.00004583	0.0002286204	$4.073987e-05$
107.0	0.00002000	0.0002589389	$1.042137e-05$
107.0	0.00002500	0.0002739389	$4.578631e-06$
107.0	0.00003000	0.0002689389	$4.213694e-07$
107.0	0.00001800	0.0002689389	$4.213694e-07$
107.0	0.00002686	0.0002675389	$1.821369e-06$
107.0	0.00001175	0.0002574389	$1.192137e-05$
113.7	0.00002200	0.0002637656	$5.594700e-06$
123.1	0.00002000	0.0002601254	$9.234869e-06$
125.2	0.00002000	0.0002600310	$9.329269e-06$
130.7	0.00003000	0.0002639057	$5.454600e-06$
143.0	0.00002140	0.0002565210	$1.283927e-05$
193.7	0.00001000	0.0002600818	$9.278469e-06$
193.7	0.00002000	0.0002600818	$9.278469e-06$

5 Discussion

There are numerous situations where the interest is in the prediction of an observable exceeding a large or even a catastrophically large threshold level where the data at hand fall short of the the threshold. For example, consider the daily rainfall amount in a region where all the diurnal amounts fall short of a high threshold level, say, 10 inches in 24 hours, and yet for risk management it is important to obtain the chance that a future amount exceeds 10 inches in 24 hours, an extreme situation by all accounts. Similar problems involve annual flood levels, daily coronavirus counts, monthly

insurance claims, earthquake magnitudes, and so on, where the sample values are below certain high thresholds, and the interest is in very small tail probabilities. Furthermore, in many cases the given data could only be moderately large. In this paper, it has been shown how to approach such problems by a large number of augmentations or fusions of the given data with computer-generated external samples. From this we obtained a curve, called B-curve, containing a point whose ordinate was close to the tail probability of interest. Moreover, the magnitude of the largest sample value relative to a given high threshold provided rough guesses as to the true value of the tail probability. The rough guesses were needed for successful applications of our iterative method which produced accurate estimates of tail probabilities. The large number of fusions resulted in a large number of upper bounds $B_1, ..., B_N$, for a tail probability p, from some unknown CDF $F_B(x)$ where it was assumed that $0 < F_B(p) < 1$. The examples in this paper and many more in Kedem et al. (2019) indicate that the choice of the (mostly misspecified) tilt function $h(x) = (x, \log x)$ in the density ratio model did not go against that assumption. Clearly other tilts are possible as long as $F_B(p)$ is bounded away from 0 and 1.

The estimation of very small tail probabilities can be approached by extreme value methods. A well known method is referred to as peaks-over-threshold (Beirlant et al. 2004, Ferreira and DeHaan 2015), where as the name suggests, only values above a sufficiently high threshold are used. However, if the sample is not large to begin with, any reduction in the sample size, by discarding those values deemed not sufficiently large, reduces the sample size

and calls into question the applicability of the method. A comparison with ROSF is given in Kedem et al. (2019). The estimation of tail probabilities from fused residential radon data has been studied recently in Zhang et al. (2020 a,b) by using the density ratio model with variable tilts to fuse a given radon sample from a county of interest with radon samples from neighboring counties. We believe that with the advent of new generative models, realistic data synthesis, and faster computing capabilities, AR approaches of gaining better understanding of different phenomena will become increasingly popular in the future.

Acknowledgment: Research supported by a Faculty-Student Research Award, University of Maryland, College Park.

A Appendix

The appendix addresses the density ratio model (11) for $m+1$ data sources. Thus, we deal with the density ratio model more generally where $\boldsymbol{X}_0$ is fused with m computer-generated samples. Above we dealt with the special case of $m = 1$. Assume 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, $j = 1, ..., m$. The augmentation of $m + 1$ samples

$$\boldsymbol{t} = (t_1, \ldots, t_n) = (\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_m), \tag{10}$$

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

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)), \quad j = 1, \ldots, m, \tag{11}$$

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. 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.

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 $t_1, \ldots, t_n$ (Owen 2001). Let $p_i = dG(t_i)$ be the mass at t_i, for $i = 1, \ldots, n$. Then the empirical likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta}, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \beta_1' h(x_{1j})) \cdots \prod_{j=1}^{n_m} \exp(\alpha_m + \beta_m' h(x_{mj})). \tag{12}$$

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

$$\sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, \ldots, \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0 \tag{13}$$

we obtain the desired estimates. In particular,

$$\hat{G}(t) = \frac{1}{n_0} \cdot \sum_{i=1}^{n} \frac{I(t_i \leq t)}{1 + \rho_1 \exp(\hat{\alpha}_1 + \hat{\beta}_1' h(t_i)) + \cdots + \rho_m \exp(\hat{\alpha}_m + \hat{\beta}_m' h(t_i))}, \tag{14}$$

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: $\boldsymbol{\rho} = \text{diag}\{\rho_1, \ldots, \rho_m\}$,

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

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

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

$$\begin{aligned}\mathrm{Cov}\{\sqrt{n}(\hat{G}(t)-G(t)),\sqrt{n}(\hat{G}(s)-G(s))\} = \\ \left(\sum_{k=0}^{m}\rho_k\right)\left(G(t\wedge s)-G(t)G(s)-\sum_{j=1}^{m}\rho_j A_j(t\wedge s)\right) \\ +\Big(\bar{A}'(s)\boldsymbol{\rho},\bar{B}'(s)(\boldsymbol{\rho}\otimes I_p)\Big)S^{-1}\begin{pmatrix}\boldsymbol{\rho}\bar{A}(t)\\(\boldsymbol{\rho}\otimes I_p)\bar{B}(t)\end{pmatrix}. \qquad (15)\end{aligned}$$

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 (15). Replacing parameters by their estimates, a $1-\alpha$ level pointwise confidence interval for $G(t)$ is approximated by

$$\left(\hat{G}(t)-z_{\alpha/2}\sqrt{\hat{V}(t)},\ \ \hat{G}(t)+z_{\alpha/2}\sqrt{\hat{V}(t)}\right), \qquad (16)$$

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

$$\left(1-\hat{G}(t)-z_{\alpha/2}\sqrt{\hat{V}(t)},1-\hat{G}(t)+z_{\alpha/2}\sqrt{\hat{V}(t)}\right). \qquad (17)$$

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

[4] 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.

[5] George, A.C. The history, development and the present status of the radon measurement programme in the United States of America. *Radiation Protection Dosimetry* 2015; **167**: 8-14.

[6] 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.

[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., Pan, L., Smith, P. and Wang, C. Estimation of Small Tail Probabilities by Repeated Fusion. *Mathematics and Statistics* 2019; **7**: 172 - 181.

[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] Owen, A. *Empirical Likelihood*, Chapman & Hall/CRC, Boca Raton, 2001.

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

[13] Zhang, X, Pyne, S, Kedem, B. Estimation of residential radon concentration in Pennsylvania counties by data fusion. *Appl Stochastic Models in Business and Industry*, 2020a; https://doi.org/10.1002/asmb.2546.

[14] Zhang X., Pyne S., Kedem B. Model selection in radon data fusion. *Statistics in Transition, new series*, Special Issue, August 2020b, 167-174.

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