Biometrika (2018), **xx**, x, pp. 1–8 *Printed in Great Britain*

Discrete Extremes

BY A. S. HITZ

Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK adrien.hitz@stats.ox.ac.uk

R. A. DAVIS

Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY rdavis@stat.columbia.edu

AND G. SAMORODNITSKY

School of Operations Research and Information Engineering, Cornell University, 220 Rhodes Hall, Ithaca, NY. gs18@cornell.edu

SUMMARY

We propose two peaks-over-threshold methods for discrete random variables, and show that ¹⁵ they can provide accurate tail probability estimates in simulated and real data.

Some key words: Extreme value theory; Tail approximation; Count data; Discrete distribution; Generalized Pareto distribution; Zipf distribution; Peaks over threshold.

1. INTRODUCTION

Extreme quantile estimation is an important but difficult problem in statistics, especially when the quantile is beyond the range of the data. In the univariate case, an approach that often works well in practice is to model observations above a large threshold with a parametric family of distributions which can be motivated as follows. Let X be a random variable taking values in $[0, x_F)$ for $x_F \in (0, \infty]$, and suppose that there exists a strictly positive sequence a_u such that

$$a_u^{-1}(X-u) \mid X \ge u \to Z,\tag{1}$$

5

10

30

in distribution as $u \to x_F$, for some Z following a non-degenerate probability distribution on $[0, \infty)$. Then, Z follows a generalized Pareto distribution, defined by its survival function

$$\bar{F}_{\text{GPD}}(x;\sigma,\xi) = \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi} \mathbb{1}_{\{x < \tau\}}, \quad x \ge 0,$$

with $\sigma > 0$, $\tau = \sigma/|\xi|$ if $\xi < 0$ and $\tau = \infty$ otherwise, $1_{\{x < \tau\}} = 1$ if $x < \tau$ and 0 otherwise, and $(1 + \xi x)^{1/\xi} = e^x$ if $\xi = 0$ (Pickands, 1975). Condition (1), written as $X \in \text{MDA}_{\xi}$, means that X is in the maximum domain of attraction of an extreme value distribution with shape parameter ξ (see e.g. Resnick (1987)). In this case, the sequence of cumulative distribution functions of $a_u^{-1}(X - u) \mid X \ge u$ converges uniformly to $1 - \bar{F}_{\text{GPD}}$ on $[0, \infty)$. Thus, the distribution of exceedances above a large threshold u (also called "peaks-over-threshold") can be approximated

c 2016 Biometrika Trust

in the following manner:

$$\operatorname{pr}(X - u > x \mid X \ge u) = \operatorname{pr}\left\{a_u^{-1}(X - u) > a_u^{-1}x \mid X \ge u \quad \approx \bar{F}_{\operatorname{GPD}}(x; \sigma a_u, \xi), \quad (2)\right\}$$

(Davison & Smith, 1990). This approximation, called the generalized Pareto approximation, is convenient in practice because it does not rely on a specific distributional assumption; X is only required to belong to some maximum domain of attraction, which holds for most common continuous distributions.

In the discrete case, by contrast, tail distributions are less understood. It is not clear how discrete exceedances over high threshold should be modeled and, consequently, the generalized Pareto approximation is often applied ignoring the discrete nature of the data. This poses two issues: first, a necessary condition for a discrete random variable X to be in some maximum domain of attraction in the case $x_F = \infty$ is that X is long-tailed,¹ i.e., $\overline{F}_X(u+1)/\overline{F}_X(u) \rightarrow 1$ as $u \rightarrow \infty$ (Shimura, 2012). However, many common discrete distributions, including geometric,

Poisson and negative binomial distributions, are not long-tailed.

Specific convergence results for maxima of discrete data have thus been derived (Anderson, 1970, 1980; Dkengne et al., 2016), but the limit proposed is always a continuous distribution, which leads to the second issue: treating discrete data as continuous introduces a bias in the likelihood function. Since the shape and location parameters ξ and σ of the generalized Pareto

approximation are unknown in practice, they must be estimated from the exceedance data. In this context, we will see that the bias renders the approximation inadequate if there are many tied observations — even when X is long-tailed, that is, when (2) is valid in theory.

- Our contribution is to overcome these limitations by proposing two peaks-over-threshold methods, each relying on a parametric family of discrete distributions: the discrete generalized Pareto and the generalized Zipf distribution. The latter distributions exist in the literature but have not been put forward by a limiting argument for modeling extremes. We will show that these new approximations can be theoretically motivated for X belonging to a broad class of
- ⁶⁰ discrete distributions, and that they outperform the generalized Pareto approximation when the data contain many tied observations. They perform very similarly to one another but it is still unclear if one of them should be preferred.

From now on, we assume that X is a discrete random variable with non-negative values, and $\xi \ge 0$. The first method is introduced by adapting condition $X \in \text{MDA}_{\xi}$ to the discrete case as follows. Suppose that there exists a random variable $Y \in \text{MDA}_{\xi}$ such that $\text{pr}(X \ge k) = \text{pr}(Y \ge k)$ for $k = 0, 1, 2, \ldots$, that is, the equality in distribution, $X = \lfloor Y \rfloor$, holds. In this case, we say that X is in the discrete maximum domain of attraction, which we write as $X \in \text{D-MDA}_{\xi}$. We call Y an extension of X and such an extension is not unique. Shimura (2012) proved that $X \in \text{MDA}_{\xi}$ if and only if $X \in \text{D-MDA}_{\xi}$ and X is long-tailed.² It was also shown by Shimura (2012)

that geometric, Poisson and negative binomial distributions belong to the discrete maximum domain of attraction. Therefore, $MDA_{\xi} \subseteq D-MDA_{\xi}$ for discrete distributions. If $X \in D-MDA_{\xi}$ and $Y \in MDA_{\xi}$ is a corresponding extension satisfying $X = \lfloor Y \rfloor$ in distribution, then, for large integers u, we use (2) to find

$$pr(X - u = k \mid X \ge u) = pr(Y - u \ge k \mid Y \ge u) - pr(Y - u \ge k + 1 \mid Y \ge u)$$

$$\approx p_{\text{D-GPD}}(k; \sigma a_u, \xi), \tag{3}$$

2

40

¹ All long-tailed distributions are heavy-tailed, but the converse is false.

² In this case, an extension of X can thus be X itself.

Discrete Extremes

where $p_{\text{D-GPD}}$ is the probability mass function of the discrete generalized Pareto distribution defined by

$$p_{\text{D-GPD}}(k;\sigma,\xi) = \bar{F}_{\text{GPD}}(k;\sigma,\xi) - \bar{F}_{\text{GPD}}(k+1;\sigma,\xi),$$

for k = 0, 1, 2, ... Equation (3) provides a method for modeling exceedances over threshold that we call the discrete generalized Pareto approximation. The latter distribution has been applied by Prieto et al. (2014) to model road accidents, while various aspects of discrete Pareto-type distributions were studied in Krishna & Pundir (2009), Buddana & Kozubowski (2014), and Kozubowski et al. (2015).

Whereas the first method is based on an extension of \overline{F}_X by a survival function in the maximum domain of attraction, the second method assumes instead an extension of p_X , the probability mass function of X. Suppose that there exists a non-negative random variable $Y \in \text{MDA}_{\xi/(1+\xi)}$ such that $p_X(k) = c \overline{F}_Y(k)$ for k = d, d + 1, d + 2, ..., for some c > 0 and $d \in \mathbb{N}_0 = \{0, 1, ...\}$. In this case, we say that p_X is in the discrete maximum domain of attraction which is denoted by $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$, and call \overline{F}_Y an extension of p_X . We will show that $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ implies $X \in \text{MDA}_{\xi}$ (under a mild condition in the case $\xi = 0$), and that geometric, Poisson and negative binomial satisfy $p_X \in \text{D-MDA}_0$. It follows from (2) that, for large integers u,

$$\operatorname{pr}(X - u = k \mid X \ge u) = \frac{\operatorname{pr}(Y > u + k)/\operatorname{pr}(Y > u)}{\sum_{i=0}^{\infty} \operatorname{pr}(Y > u + i)/\operatorname{pr}(Y > u)} \approx p_{\text{GZD}}\{k; (1 + \xi)\sigma a_u, \xi , (4)\}$$

where

$$p_{\text{GZD}}(k;\sigma,\xi) = \frac{\left(1+\xi\frac{k}{\sigma}\right)^{-1/\xi-1}}{\sum_{i=0}^{\infty} \left(1+\xi\frac{i}{\sigma}\right)^{-1/\xi-1}}, \quad k = 0, 1, 2, \dots$$
(5)

which is the probability function of a distribution that we call the generalized Zipf distribution. In the case $\xi = 0$, the latter is a geometric distribution (and so is the discrete generalized Pareto distribution), and in the case $\xi > 0$, it is a Zipf–Mandelbrot distribution (Mandelbrot, 1953). Zipf-type families have been fitted to various discrete datasets such as word frequencies (Booth, 1967), city sizes (Gabaix, 1999), company sizes (Axtell, 2001) and website visits (Clauset et al., 2009). The Zipf law, arising in the case $\xi = \sigma$, is sometimes presented as the discrete counterpart of the Pareto distribution (Arnold, 1983). We refer to the approximation procedure in (4) as the generalized Zipf approximation.

2. THEORETICAL RESULTS

We start by showing that the probability density and mass functions of the generalized Pareto, discrete generalized Pareto and Zipf distributions are asymptotically equivalent as σ tends to 105 infinity. Proofs are given in the Supplementary Material.

PROPOSITION 1. For $\sigma > 0, \xi \ge 0$ and $q, \tilde{q} \in \{f_{GPD}, p_{D-GPD}, p_{GZD}\}$, it holds

$$\lim_{\sigma \to \infty} \sup_{k=0,1,2,\dots} \frac{q(k;\sigma,\xi)}{\tilde{q}(k;\sigma,\xi)} - 1 = 0$$

(**-**))

This suggests that modeling a sample from $X - u \mid X \ge u$ by maximum likelihood using either f_{GPD} , $p_{\text{D-GPD}}$ or p_{GZD} should not differ too much if the estimated scale parameter $\hat{\sigma}$ is sufficiently large. When the sample size and u grow, $\hat{\sigma}$ only goes to infinity if the sequence a_u defined in

3

A. S. HITZ, R. DAVIS AND G. SAMORODNITSKY

(1) satisfies $a_u \to \infty$, which occurs if and only if X is long-tailed. Even in this case, a_u might converge too slowly for the three methods to be similar in practice, as we will see in Section 3.

We now attempt to put the approximation procedures on firmer theoretical grounds, starting with a convergence result for the discrete generalized Pareto approximation.

PROPOSITION 2. If $X \in D$ -MDA $_{\xi}$ for $\xi \ge 0$, then there exists a positive sequence $(a_u, u = 1, 2, ...)$ such that

$$\lim_{u \in \mathbb{N}_0, u \to \infty} \sup_{k=0,1,2,\dots} pr(X = u + k \mid X \ge u) - p_{D\text{-}GPD}(k; a_u, \xi) = 0.$$
(6)

120

We remark that (6) is not interesting if $a_u \to \infty$ because the two terms converge to 0. The next result sheds light on the approximation procedures when, this time, $p_X \in D$ -MDA. Recall that a distribution F is in MDA₀ if and only if the survival function has a representation

$$\bar{F}(x) = c(x) \exp\left\{-\int_0^x \frac{1}{a(y)} dy\right\}, \quad -\infty < x < x_F,$$
(7)

where $c(\cdot)$ is a positive function with $c(x) \to c > 0$ as $x \uparrow x_F$, and $a(\cdot)$ is a positive, differentiable function $a(\cdot)$ with $\lim_{x\uparrow x_F} a'(x) = 0$. If c(x) = c on (d, x_F) for some $d < x_F$, then we say that the distribution F satisfies the von Mises condition. The function $a(\cdot)$ in (7) is sometimes called the auxiliary function. Note that it is only uniquely defined on (d, x_F) under the von Mises condition; see Embrechts et al. (2013). We also remind that in the sequel we only consider the case of unbounded support, i.e. $x_F = \infty$.

THEOREM 1. If $p_X \in D$ -MDA_{$\xi/(1+\xi)$} and $\xi > 0$, then $X \in MDA_{\xi}$ and, for any sequence of nonnegative integers $(k_u)_{u \in \mathbb{N}_0}$ such that $\sup_u k_u/u < \infty$,

$$\lim_{u\in\mathbb{N},\,u\to\infty}\frac{pr(X=k_u+u\mid X\geq u)}{q(k_u;\xi u,\xi)}=1.$$
(8)

130

125

where
$$q \equiv f_{GPD}, p_{D-GPD}$$
 and p_{GZD} .

If $p_X \in D-MDA_0$ and if the auxiliary function of an extension \overline{F} of p_X satisfies $\lim_{x\to\infty} a(x) = \sigma > 0$, then $X \in D-MDA_0$ and

$$\lim_{u \in \mathbb{N}, u \to \infty} pr(X = k + u \mid X \ge u) = p_{D\text{-}GPD}(k; \sigma, 0) = p_{GZD}(k; \sigma, 0), \tag{9}$$

135 for $k = 0, 1, 2, \dots$

The condition $p_X \in D$ -MDA is satisfied, among others, by the Zipf–Mandelbrot, geometric, Poisson and negative binomial distributions as shown below and in the Supplementary Material.

Example 1. The probability mass function of a Zipf–Mandelbrot distribution is proportional to $(k+q)^{-1-1/\xi}$ for $k = 0, 1, 2, ..., q \ge 0, \xi > 0$, and satisfies $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ because it can be extended by $\bar{F}_Y(y) = c(y+q)^{-1-1/\xi}$ for $y \ge 0$ and some c > 0 with $Y \in \text{MDA}_{\xi/(1+\xi)}$. The probability mass function of a geometric distribution belongs to D-MDA₀ as it coincides up to a constant with the survival function of an exponential distribution. The latter distribution clearly satisfies the von Mises condition and thus is a member of MDA₀. The auxiliary function is, in fact, equal (eventually) to $1/\lambda$, where λ is the rate of the exponential distribution.

To summarize, for a discrete random variable X and $\xi \ge 0$, it holds $X \in \text{MDA}_{\xi}$ if and only if $X \in \text{D-MDA}_{\xi}$ and X is long-tailed. If $\xi > 0$, then $p_X \in \text{D-MDA}_{\xi/(1+\xi)}$ implies $X \in \text{D-MDA}_{\xi}$; the same implication holds in the case $\xi = 0$ if the auxiliary function of the extension of p_X satisfies $a(x) \to \sigma \in (0, \infty)$ as $x \to \infty$.



Fig. 1: On the left: frequency plot of a sample of $X - 2 \mid X \ge 2$ of size 739 simulated from (10). On the right: frequency plot of the length of the 2875 longest French words.

3. EMPIRICAL RESULTS

We assess the performance of the discrete generalized Pareto and the generalized Zipf approximations for estimating the probability of a rare event from discrete data, and illustrate why they should be preferred to the generalized Pareto approximation when there are many tied observations, whether X is long-tailed or not. Let $\alpha = 2$, $\beta = 1$ and

$$Y \sim \text{Inverse-gamma}(\alpha, \beta), \quad X = |Y|,$$
 (10)

where inverse-gamma distribution has density function $f(x) = \Gamma(\alpha)^{-1} \beta^{\alpha} x^{-\alpha-1} \exp(-\beta/x)$, ¹⁵⁵ x > 0. The experiment described below is repeated 500 times. An independent and identically distributed sample of size 8000 is drawn from the distribution of X. From these observations, the goal is to estimate the probability of the extreme region

$$p_e = \operatorname{pr}(X \ge |q_e|), \quad q_e = 70, \tag{11}$$

where q_e is the 99.99 percentile of Y, i.e., the value exceeded once every 10 000 times on average. The strategy pursued is to select an integer threshold u as the 95th empirical percentile of the sample,³ fit parametric distributions to the exceedances $X - u \mid X \ge u$, and use them to extrapolate the tail and estimate p_e . It holds $X \in D$ -MDA $_{\xi}$ and $X \in MDA_{\xi}$ for $\xi = 1/\alpha$, but it is not immediate if $p_X \in D$ -MDA $_{\xi/(1+\xi)}$ and we thus apply the approximations heuristically. The generalized Pareto distribution is fitted to the data either directly or after shifting them by a continuity correction $\delta = \frac{1}{2}$. As a benchmark, we will also estimate p_e from a sample of the continuous variable Y (as opposed to its discretization X). In this context, using a generalized Pareto distribution is motivated by the fact that $Y \in MDA_{1/2}$, and we fit it to $Y - u \mid Y \ge u$.

A frequency plot of a sample of $X - u \mid X \ge u$ is displayed on the left-hand side in Figure 1. For each model, we compute maximum likelihood estimators $\hat{\sigma}$ and $\hat{\xi}$ by performing a two

150

³ Selecting an appropriate threshold is crucial when estimating high quantiles and can be based on techniques such as mean residual plots (see e.g. Davison & Smith (1990)).

Table 1: Performance of several methods in estimating the probability p_e of a rare event defined in (11) from about 700 exceedances in each experiment. The table displays average maximum likelihood estimators for p_e , ξ and σ across 500 experiments. Coverage c, average length l and true length l^* of 90% confidence intervals are shown between brackets. The discrete generalized Pareto and Zipf approximations are superior in this case.

$\hat{p}_e \cdot 10^3 \ (c, l, l^{\star}) \\ 0.10$	$\hat{\xi}$ (c, l) 0.5	$\hat{\sigma}\left(l ight)$
0.10(87%, 0.16, 0.16)	0.49(95%, 0.22)	1.19(0.30)
0.10 (86%, 0.17, 0.16)	0.49(95%, 0.23)	1.19(0.33)
0.11 (88%, 0.17, 0.17)	0.50(95%, 0.24)	1.39(0.29)
0.04(20%, 0.07, 0.08)	0.37(22%, 0.18)	1.43(0.32)
7.93(83%, 23.97, 11.25)	$8.27 \ (0\%, 1.24)$	0.00(0.00)
	$\hat{p}_e \cdot 10^3 (c, l, l^*)$ 0.10 0.10 (87%, 0.16, 0.16) 0.10 (86%, 0.17, 0.16) 0.11 (88%, 0.17, 0.17) 0.04 (20%, 0.07, 0.08) 7.93 (83%, 23.97, 11.25)	$ \begin{array}{lll} \hat{p}_e \cdot 10^3 \ (c, l, l^{\star}) & \hat{\xi} \ (c, l) \\ 0.10 & 0.5 \\ \end{array} \\ \hline 0.10 \ (87\%, \ 0.16, \ 0.16) & 0.49 \ (95\%, \ 0.22) \\ \hline 0.10 \ (86\%, \ 0.17, \ 0.16) & 0.49 \ (95\%, \ 0.23) \\ 0.11 \ (88\%, \ 0.17, \ 0.17) & 0.50 \ (95\%, \ 0.24) \\ 0.04 \ (20\%, \ 0.07, \ 0.08) & 0.37 \ (22\%, \ 0.18) \\ 7.93 \ (83\%, \ 23.97, \ 11.25) & 8.27 \ (0\%, \ 1.24) \\ \end{array} $

Table 2: Fit of several distributions to the length of the 2875 longest French words, and to the number of tornadoes during the 435 most extreme outbreaks in the United States. The table displays p-value of discrete Kolmogorov–Smirnov tests (of the discretized model in the case of continuous models), negative log-likelihood $-\ell$ and maximum likelihood estimates with 90% confidence intervals and possible temporal trend $\hat{\sigma}_t$ in the scale parameter. Tied observations are much less frequent in the tornado data than in the word length data, explaining why the generalized Pareto approximation is not outperformed in the former case.

	p-val.	$-\ell$	ξ	$\hat{\sigma}_0$	$\hat{\sigma}_t$
Word length					
Discrete generalized Pareto dist.	0.85	3894.0	$0.02_{[-0.01,0.06]}$	$1.36_{[1.30,1.43]}$	
Generalized Zipf distribution	0.84	3894.0	$0.02_{[-0.01,0.06]}$	$1.37_{[1.32,1.43]}$	
Generalized Pareto dist., $\delta = \frac{1}{2}$	0.00		$-0.04_{[-0.06, -0.01]}$	$1.51_{[1.45, 1.57]}$	
Negative binomial	0.75	3893.9			
Tornado outbreak					
Discrete generalized Pareto dist.		1439.92	$0.27_{[0.16, 0.37]}$	$4.81_{[3.64,5.99]}$	$6.11_{[3.74,8.48]}$
Generalized Pareto dist., $\delta = \frac{1}{2}$		1439.93	$0.26_{[0.16, 0.37]}$	$4.86_{[3.68, 6.04]}$	$6.13_{[3.75, 8.50]}$

dimensional maximization using the function optim of R (R Core Team, 2015) with starting values (1, 1). We then compute \hat{p}_e and approximate 90% confidence intervals under asymptotic normality of the estimators. Table 1 displays: the average estimates \hat{p}_e , $\hat{\xi}$ and $\hat{\sigma}$ over the 500 experiments, the coverage⁴ of the confidence intervals, their average length and their true length.⁵ It

appears that the discrete generalized Pareto and Zipf approximations accurately estimate p_e from the discretized data with a coverage close to the correct one of 90%, and that their performance is good relative to the situation of full information where the continuous data are available — notice how they deliver very similar estimates to one another. On the other hand, the two variants of the generalized Pareto approximation perform poorly, the worst being the case $\delta = 0$. Decreasing α would increase $\hat{\sigma}$ and render the estimates from all these methods indistinguishable as expected

from Proposition 1. We point out that Poisson and negative binomial distributions would poorly estimate p_e in this example.

The ability of the discrete generalized Pareto and Zipf approximations to accurately estimate the probability of rare events is supported by complementary simulated cases covering $\xi = 0$ and $\xi < 0$ (Hitz, 2016, Chapter 2), and illustrated here on three real datasets. The first consists

⁴ Coverage indicates the proportion of time the truth lies in the confidence interval.

 $^{^{5}}$ True length is the length the intervals should have had to contain the estimates across the 500 experiments 90% of the time.

Table 3: On the left: frequency table of multiple births in the United States from 1995 to 2014. On the right: performance of several methods in estimating the probability p_e of an American women delivering quintuplets or more at birth using only one thousandth of the dataset. In each experiment, the threshold was u = 1 and there were about 2600 exceedances (see Table 1 for notation). The discrete generalized Pareto and Zipf provide useful techniques for such extrapolations.

		$\hat{p}_e \cdot 10^5 \ (c, l, l^\star)$
78178588	Truth	1.7
2500340	Discrete generalized Pareto distribution	1.4 (74%, 2.9, 2.9)
117603	Generalized Zipf distribution	$1.6 (87\%, 3.3, 2.8)^8$
8108	Negative Binomial	$1.2 \ (65\%, \ 2.3, \ 2.3)$
1353	Generalized Pareto distribution, $\delta = \frac{1}{2}$	n/a ⁹
	$78178588\\2500340\\117603\\8108\\1353$	78 178 588Truth2 500 340Discrete generalized Pareto distribution1 17 603Generalized Zipf distribution8 108Negative Binomial1 353Generalized Pareto distribution, $\delta = \frac{1}{2}$

of the frequency X of word length in the French lexicon (New et al., 2004); for instance, "anticonstitutionnellement" is the only word of 25 letters in French. We focus on describing the tail distribution and fit the usual models to $X - u \mid X \ge u$ with u = 15. A frequency plot of the 2875 exceedances is shown on the right-hand side of Figure 1. The discrete generalized Pareto and Zipf distributions deliver a good fit and similar estimations to one another, and clearly outperform the generalized Pareto approximation as shown by discrete Kolmogorov–Smirnov tests⁶ in Table 2 (Arnold & Emerson, 2011). Notice that the negative binomial also fits well in this case.

The second dataset studied in Tippett et al. (2016) reports the number X of tornadoes per extreme outbreak in the United States between 1965 and 2015, where an outbreak is a sequence 195 of twelve or more tornadoes occurring close to each other in time and that are high on Fujita scale. The authors found that the 435 observations from $X - u \mid X \ge u$ for u = 12 were well modeled by a generalized Pareto distribution with linear temporal trend in the scale parameter⁷ and continuity correction $\delta = \frac{1}{2}$. Table 2 shows that the discrete generalized Pareto distribution (and the generalized Zipf) yields almost the same estimates. Indeed, there are fewer ties in this 200 case: about 38% of the data consists of values shared with no more than 20 other observations, compared to 13% for the simulated data and 1% for the word length data. The location parameter is larger and, thanks to Proposition 1, the generalized Pareto approximation is appropriate.

The third dataset counts the number X of multiple births in the United States from 1995 to 2014 and is displayed on the left-hand side of Table 3 (Hamilton et al., 2015). The observations 205 only take 5 distinct values, and it is thus interesting to see if the discrete generalized Pareto and Zipf distributions can still describe the tail of the data in this non-standard estimation problem. We randomly select from the dataset a sample that contains thousand times fewer observations, and estimate from these the probability p_e that an American women delivers quintuplets or more. As the data are censored from above, we fit a right-censored version of the usual models to 210 $X^{C} - u \mid X^{C} \geq u$ for u = 1, where $X^{C} = \min(X, 5)$. The experiment was repeated 500 times, and each sample contained on average 9 quatruplets and 1 quintuplet or more. Table 3 shows that the discrete generalized Pareto and Zipf distributions outperform common alternatives, and seem to be useful techniques for inference from such limited data. The applicability of peaks-overthreshold methods when u is a particularly small integer should be more rigorously explored.

Future work could assess the validity of the approximations in the case $\xi < 0$. The discrete generalized Pareto distribution benefits from its closed-form survival and probability mass func-

190

p-values are here computed by Monte Carlo simulation.

⁷ More precisely, $\sigma_0 + \sigma_1 \mathbf{t}$, where \mathbf{t} is the covariate vector in years. In our case, we rescaled \mathbf{t} between [0, 1].

⁸ In 52 experiments out of 500, maximum likelihood estimates could not be computed numerically. In 70 experiments, the hessian matrix could not be computed numerically.

⁹ The log-likelihood function could not be maximized numerically.

A. S. HITZ, R. DAVIS AND G. SAMORODNITSKY

tion, allowing for exact likelihood based inference. It would be interesting to further understand how it relates to the generalized Zipf distributions which seems to deliver comparable performance in the data analysis carried out.

ACKNOWLEDGEMENT

The first author is grateful to the Berrow Foundation for financial support and would like to thank Robin Evans, Gesine Reinert, David Steinsaltz and Jonathan Tawn for valuable advice and encouragement. This research was partially supported by the ARO grant W911NF-12-10385.

220

REFERENCES

- ANDERSON, C. W. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability* **7**, 99–113.
- ANDERSON, C. W. (1980). Local limit theorems for the maxima of discrete random variables. *Mathematical Proceedings of the Cambridge Philosophical Society* **88**, 161–165.
- ARNOLD, B. C. (1983). Pareto distribution. International Cooperative Publishing House, Maryland.
- ARNOLD, T. B. & EMERSON, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal* **3**, 34–39.

AXTELL, R. L. (2001). Zipf distribution of US firm sizes. Science 293, 1818-1820.

- BOOTH, A. D. (1967). A law of occurrences for words of low frequency. *Information and Control* 10, 386–393.
- BUDDANA, A. & KOZUBOWSKI, T. J. (2014). Discrete Pareto distributions. *Economic Quality Control* 29, 143–156. CLAUSET, A., SHALIZI, C. R. & NEWMAN, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review* 51, 661–703.
 - DAVISON, A. C. & SMITH, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B. Methodological* 52, 393–442.
- ²⁴⁰ DKENGNE, P. S., ECKERT, N. & NAVEAU, P. (2016). A limiting distribution for maxima of discrete stationary triangular arrays with an application to risk due to avalanches. *Extremes* 19, 25–40.
 EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (2013). *Modelling extremal events: for insurance and finance*. Springer, New York.

GABAIX, X. (1999). Zipf's law and the growth of cities. The American Economic Review 89, 129-132.

245 HAMILTON, B. E., MARTIN, J. A., OSTERMAN, M. J., CURTIN, S. C. & TJ, M. (2015). Births: Final data for 2014. National Vital Statistics Reports 64.

HITZ, A. S. (2016). Modelling of Extremes. Ph.D. thesis, University of Oxford.

- KOZUBOWSKI, T. J., PANORSKA, A. K. & FORISTER, M. L. (2015). A discrete truncated Pareto distribution. *Statistical Methodology* **26**, 135–150.
- 250 KRISHNA, H. & PUNDIR, P. S. (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology* 6, 177–188.

MANDELBROT, B. (1953). Contribution à la théorie mathématique des jeux de communication. Publications de l'Institut de statistique de l'Université de Paris.

- NEW, B., PALLIER, C., BRYSBAERT, M. & FERRAND, L. (2004). Lexique 2: A new French lexical database. Behavior Research Methods, Instruments, & Computers **36**, 516–524.
 - PICKANDS, III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics* 3, 119–131.
 PRIETO, F., GÓMEZ-DÉNIZ, E. & SARABIA, J. M. (2014). Modelling road accident blackspots data with the discrete generalized Pareto distribution. *Accident Analysis & Prevention* 71, 38–49.
- R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
 - RESNICK, S. I. (1987). Extreme values, regular variation, and point processes. Springer, New York.

²²⁵

SHIMURA, T. (2012). Discretization of distributions in the maximum domain of attraction. *Extremes* 15, 299–317.
 TIPPETT, M. K., LEPORE, C. & COHEN, J. E. (2016). More tornadoes in the most extreme US tornado outbreaks. *Science* 354, 1419–1423.