From the power law to extreme value mixture distributions Extremes Reading Group @ Lancaster

> Clement Lee (joint work with Emma Eastoe and Aiden Farrell)



Introduction

Discrete power law = Zipf distribution

$$p(x) = rac{x^{-lpha}}{\zeta(lpha, u+1)}, \qquad x = u+1, u+2, \ldots$$

 $\log p(x) = -\alpha \log x + \text{constant}$ 

The log-log plot

Frequency of occurrence of (sampled) unique words in the novel Moby Dick



## Another example

► The social network of Flickr users

http://konect.cc/networks/flickr-links/



### Particular interest in networks

Total degrees (undirected) or in-degrees (directed)

- Preferential attachment model
  - Barabási and Albert (1999, Science)
  - Generating networks using simple rules
  - The-rich-get-richer

Resulting degree distribution follows the power law

- Barabási, Albert and Jeong (1999, Physica A)
- Bollobás et al. (2001, Random Structures & Algorithms)

### Related works

- Wang and Resnick (2023, Extremes)
  - Reciprocity associated with extremal dependence between in-degrees & out-degrees
  - Original model underestimates reciprocity in real-life networks, hence unrealistic
- Here we focus on the degree distribution
  - Does the data really follow the power law?

# "Close to" power law?

Analysed by Valero et al. (2022, Physica A)

Facebook network at UCSC



# What about this

Partial power law?

#### Facebook network at Harvard



# Empirical frequencies mask the tail fit



- A distribution that fits the tail more adequately
- ▶ While retaining the (partial) power law
- And covering the possibility of curvature
- Simultaneously determine between the two

# Outline

- The Zipf-polylog distribution & its DoA
- Our mixture model
- Selection between power law or not
- Applications to real data

The Zipf-polylog distribution & its DoA

The Zipf / zeta distribution / discrete power law

$$p(x;\alpha) = \frac{x^{-\alpha}}{\sum_{k=w+1}^{\infty} k^{-\alpha}}, \qquad x = w+1, w+2, \dots$$

- Aka the zeta distribution
- ▶ The continuous counterpart the Pareto distribution
- No direct relationship between the two

# Relationships



The Zipf-polylog (ZP) distribution (Valero et al., 2022, Physica A)

$$p_{\mathsf{ZP}}(x;\alpha,\theta) = \frac{x^{-\alpha}\theta^x}{\sum_{k=w+1}^{\infty} k^{-\alpha}\theta^k}, \qquad x = w+1, w+2, \dots,$$

$$\blacktriangleright (\alpha, \theta) \in ((-\infty, \infty) \times (0, 1)) \cup ((1, \infty) \times \{1\})$$

Looks like a discrete version of Gamma, but not quite

- A disjoint union of Zipf  $(\theta = 1)$  and polylog  $(\theta \in (0, 1))$  distributions
- Accommodating curved data when  $\theta \in (0,1)$

## ZP inadequate for tails

- Going from  $\theta = 1$  to  $\theta \in (0, 1]$  is still insufficient for the right tail
- Consider the maximum domain of attraction (DoA) of ZP distribution

## Domain of attraction

A distribution F is in the DoA of an extreme value distribution H if there exists a<sub>n</sub> > 0, b<sub>n</sub> ∈ R such that

$$\lim_{n\to\infty}|F^n(a_nx+b_n)-H(x)|=0,$$

where H must be a negative Weibull, Gumbel, or Fréchet distribution.

- Applies to continuous & discrete distributions
- Poisson and geometric distributions do not belong to a DoA according to the definition

# Recovery to DoA for discrete distributions

- Shimura (2012, Extremes)
- If discrete F is the discretisation of continuous  $F_0$ , and  $F_0$  is in a DoA, then F is **recoverable** to the same DoA
- Geometric and Poisson are recoverable to the Gumbel DoA

## Key results for recovery (Shimura, 2012)

$$\Omega(F,x) := \left(\log rac{ar{F}(x+1)}{ar{F}(x+2)}
ight)^{-1} - \left(\log rac{ar{F}(x)}{ar{F}(x+1)}
ight)^{-1},$$

- ► If  $\lim_{x \to \infty} \Omega(F, x) = 0$ , then *F* is recoverable to the Gumbel DoA
- If  $\lim_{x\to\infty} \Omega(F,x) = \xi > 0$ , then F is in the Fréchet DoA with tail index  $\xi$

DoA of geometric( $\theta$ ) distribution

▶  $heta \in (0,1)$ 

$$\bar{F}(x) = \theta^{x}$$
$$\frac{\bar{F}(x+1)}{\bar{F}(x+2)} = \frac{\bar{F}(x)}{\bar{F}(x+1)} = \frac{1}{\theta}$$
$$\lim_{x \to \infty} \Omega(F, x) = \lim_{x \to \infty} \left[ \left( \log \frac{1}{\theta} \right)^{-1} - \left( \log \frac{1}{\theta} \right)^{-1} \right] = 0$$

Recoverable to the Gumbel DoA

# DoA of $ZP(\alpha, \theta)$ distribution

#### • When $\theta \in (0,1)$ i.e. the polylog distribution

 $\lim_{x\to\infty}\Omega(F,x)=0$ 

- Same limit i.e. also recoverable to the Gumbel DoA
- Proof similar to geometric case

# DoA of $ZP(\alpha, \theta)$ distribution

• When  $\theta = 1$  i.e. the Zipf distribution

$$\lim_{x\to\infty}\Omega(F,x)=1/(\alpha-1)$$

- ▶ In Fréchet DoA with tail index  $1/(\alpha 1)$
- Proof in the appendices of the paper

### Some remarks

- Voitalov et al. (2019, Physical Review Research) gave the result for the continuous version i.e. the Pareto distribution
- Regular variation arguments rather than GP distribution used
- ▶ Can't use result for our proof as Zipf  $\neq$  discretisation of Pareto

# Practically

 Approximate right tail of ZP(α, θ) by (discrete version of) GP distribution with shape parameter ξ

$$\xi = \mathbb{I}_{\{\theta=1\}}/(\alpha-1)$$

- Can't quite capture heavy tails of a different heaviness other than  $1/(\alpha 1)$  c.f.
  - For bivariate Gaussian(ρ),

$$\chi := \mathsf{Pr}(X > u | Y > u) = \mathbb{I}_{\{|
ho|=1\}}$$

Can't quite capture the spectrum of asymptotic independence

### Discrete version of GP distribution

- Integer generalised Pareto (IGP) distribution
- Prieto et al. (2014, Accident Analysis and Prevention)
- Rohrbeck et al. (2018, Annals of Applied Statistics)

Mixture model

### General framework

$$f(z) = \pi_1 f_1(z) + \pi_2 f_2(z) + \cdots + \pi_m f_m(z)$$

• Subject to 
$$\sum_{i=1}^{m} \pi_i = 1$$
, and  $0 < \pi_i < 1$ 

Usually same support for all components

#### In extremes

$$f(z) = \left\{egin{array}{ll} (1-\phi_u) imesrac{h(z)}{H(u)}, & z\leq u, \ \phi_u imes g_u(z), & z>u, \end{array}
ight.$$

- Disjoint support for the components
- Comprehensive review by Scarrott and MacDonald (2012, REVSTAT)
- R package evmix by Hu and Scarrot (2018, JSS)

## From continuous ...

$$f(z) = \left\{egin{array}{ll} (1-\phi_u) imesrac{h(z)}{H(u)}, & z\leq u, \ \phi_u imes g_u(z), & z>u, \end{array}
ight.$$

- h(z): bulk / body distribution
- ▶  $g_u(z)$ : GP density
- $\blacktriangleright \phi_u$ : exceedances rate

### ... to discrete

$$p(x) = \begin{cases} (1 - \phi_u) \times p_{\mathsf{TZP}}(x; \alpha_{\mathsf{mix}}, \theta_{\mathsf{mix}}, u, w), & x = w + 1, w + 2, \dots, u, \\ \phi_u \times [G_u(x; \sigma, \xi) - G_u(x - 1; \sigma, \xi)], & x = u + 1, u + 2, \dots \end{cases}$$

- $p_{\text{TZP}}(x)$ : density of **truncated** ZP distribution
- $G_u(x)$ : CDF of GP distribution
- ▶ *u*: a parameter, allowing threshold uncertainty
- ▶ w: fixed, as low as possible

### Schematic



Inference & selection between power law or not

### Bayesian inference

- ► To accommodate the threshold uncertainty
- Markov chain Monte Carlo (MCMC)
- ► Samples of  $\alpha_{mix}$ ,  $\theta_{mix}$ , u,  $\sigma$  and  $\xi$

▶ Interest in if 
$$\theta_{\mathsf{mix}} = 1$$
 or  $\theta_{\mathsf{mix}} \in (0,1)$ 

- ▶  $\theta_{mix}$  is continuous, never exactly 1 in the samples
- At the boundary makes it even more tricky
- ▶ Can't look at the proportion of  $\theta_{mix} = 1$  in the samples
- Proximity is insufficient as different tail behaviours implied

### Bayesian model selection

- 1. Define M which equals 0 if  $\theta_{\mathsf{mix}} \in (0,1)$ , and 1 if  $\theta_{\mathsf{mix}} = 1$
- 2. Assign Pr(M = 0) and Pr(M = 1)
- 3. Select between M = 0 and M = 1 in the MCMC
  - Gibbs variable selection (Carlin and Chib, 1995, JRSSB) or
  - Reversible jump MCMC (Green, 1995, Biometrika)
- 4. Calculate  $\hat{\Pr}(M = 0 | \text{data})$  and  $\hat{\Pr}(M = 1 | \text{data})$  from MCMC samples
- 5. Calculate the Bayes factor

Bayes factor

$$B_{10} = rac{\hat{\mathsf{Pr}}(M=1|\mathsf{data})}{\hat{\mathsf{Pr}}(M=0|\mathsf{data})} \left/ rac{\mathsf{Pr}(M=1)}{\mathsf{Pr}(M=0)} 
ight.$$

▶  $B_{10} > 1$ : evidence of "the **body** of the data follows the power law"

 $\triangleright$   $B_{10} < 1$ : evidence of "the **body** of the data does not follow the power law"

For  $ZP(\alpha, \theta)$  distribution as well

- $\blacktriangleright$  Can apply model selection to determine  $\theta=1$  or not
- Not ZP vs mixture though can determine visually
- ▶  $B_{10} > 1$ : evidence of "the **whole** of the data follows the power law"
- ▶  $B_{10} < 1$ : evidence of "the **whole** of the data does not follow the power law"

Applications

# Moby Dick (poweRlaw::moby)



- *u*: Moderate uncertainty but identified
- Not power law for body (left tail)

# Flickr users (Voitalov et al., 2019)



- 3-component mixture required
- Partial power law otherwise overlooked by ZP fit

# Facebook network at UCSC (Valero et al., 2022)



Mixture (IGP) better than ZP in the right tail

Could be power law or not for body

# Facebook network at Harvard (Valero et al., 2022)



Similar adequacy but strong evidence for partial power law

▶  $\alpha_{mix} < 1$ , would not be possible for Zipf fit over the whole of data

- ▶ For ZP,  $\theta \in (0,1)$  (polylog) almost always preferred to  $\theta = 1$  (Zipf)
- "Concavity" due to lighter right tail than implied had the power law in the body been extended
- Mixture resolves by replacing ZP by IGP for the tail

# Summary

- ZP distribution useful starting point for data that seems to follow the power law
- Generalises Zipf distribution, but inadequate for right tail
- Mixture model uses integer GP distribution instead
- Bayesian model selection decides if body follows the power law or not
- Applications show good fit and varying degrees of threshold uncertainty

Next steps

- More formal model comparison (ZP, 2-component mixture, 3-component mixture) via e.g. marginal likelihood
- Modified preferential attachment model that leads to such degree distributions

Thank you for listening!