Clauset, Shalizi, Newman *Power Law Distributions in Empirical Data*

Galen Wilkerson

March 19, 2013

The authors review problems that occur with 'conventional' fitting to power laws - linear fit to log-log data - and propose a new, more statistically founded method using Maximum Likelihood Estimators.

Definitions

- Maximum Likelihood Estimator:
- Kolmogorov-Smirnov Statistic:
- Goodness-of-fit:
- Likelihood ratio:
- CCDF (Complementary Cumulative Distribution):
- Hill Estimator:
- Hurwitz Zeta function:

Power Laws

Start out introducing power laws - differ from Gaussian, fat-tail, large/infinite variance, large mean, cannot be characterised by mean, variance.

General form:

$$p(x) \propto x^{-\alpha},$$

with $2 < \alpha < 3$ (roughly).

Discuss continuous versus discrete power laws.

Density:

$$p(x)dx = Cx^{-\alpha}dx,$$

with C a normalization constant.

(interesting that diverges as $x \to 0$, so there *must* be a lower bound to power-law behavior)

They find the continuous normalization constant:

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}$$

(I think they find this using the cumulative distribution, since $\int (\alpha - 1)x_{min}^{\alpha-1} = x^{\alpha}$)

In the discrete case,

$$p(x) = Cx^{-\alpha}.$$

They then find similarly that

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})},$$

with

$$\zeta(\alpha, x_{min}) = \sum_{n=0}^{\infty} (n + x_{min})^{-\alpha}.$$

The authors caution reader about using continuous distributions to approximate discrete distributions.

Observe that power law looks linear on log-log scale,

$$\ln p(x) = \alpha \ln x + const,$$

but that 'eyeballing' x_{min} and taking a log-log linear fit is erroneous.

Estimating α

Suggest MLE (maximum likelihood estimator) (Hill estimator) to find α:

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1},$$

with $x_i's$ as the observed n values of x s.t. $x_i \geq x_{min}$, and $\alpha > 1$.

[... skipping discrete case - can use continuous estimator for large n and x_{min} ...]

Discrete estimator very similar to continuous

Recommend visual form of CDF
The rest of the section (and much of the article) shows that their estimation method works very well using their synthetic data.

In particular, it shows that continuous MLE works best to get the correct α.

Estimating x_{min}

Two methods for finding x_{min}:

- (Handcock and Jones) model data above x_{min} by discrete power law; between 1 and x_{min}, model each by a separate probability p_k, with $1 \leq k \leq \hat{x}_{min}$.

  Cannot use MLE due to flexible number of parameters, instead "maximize marginal likelihood/'evidence' ... integrated over the parameters' possible values",

  use max of BIC (Bayes Information Criterion) (formula skipped) to find best $\hat{x}_{min}$ estimate.

  many parameters $\implies$ BIC may underestimate x_{min}.

- (Clauset)

  Choose the value of $\hat{x}_m in$ that makes the best model as 'similar as possible' above $\hat{x}_{min}$.

  Claim K-S statistic is best measure of distance for non-normal data:

  'The maximum distance between the CDFs of the data and fitted model.'

  $$D = \max_{x \geq x_{min}} |S(x) - P(x)|,$$

  with $S(x)$ the CDF for observations $> x_{min}$, and

  $P(x)$ is CDF for best-fit model where $x \geq x_{min}$.

  ?? Hmm, why the CDF ??

?? Also, does it make sense to do simple subtraction for distance between nonlinear models. ??

?? Why not distance in inverse-power-law space? or something else?? (i.e. transform power laws to linear, then take distance)

?? Why does max on a CDF capture the correct information ??

Their $\hat{x}_{min}$ minimizes D.

testing Clauset method

tested their method using a piecewise (first-derivative) continuous exponential, power law function:

tests using BIC and KS-MLE show better results for KS-MLE

Claim that for asymptotic power law functions,

$$p(x) = C(x+k)^{-\alpha},$$

$\hat{\alpha} \to \alpha$ as $n \to \infty$.
For related statistics Kuiper, get similar results as KS.
For Anderson-Darling, $\hat{x}_{min}$ gets overestimated.

Goodness of fit

Generate p-value to 'quantify plausibility':
'fraction of the synthetic distances (generated by the hypothesized best-fit model) (to the model) that are larger than the empirical distance (to the model)'

p close to 1: 'good fit' (i.e. empirical data is close to model)
p small (near 0) 'bad fit' (empirical data far from model)

Kolmogorov-Smirnov statistic to measure distance

procedure:

1. fit empirical data using MLE and KS statistic
2. generate many synthetic data sets according to best-fit model

3. fit each set to its power-law model and compute the K-S statistic
4. count how often the K-S statistic is greater than the value for empirical data → p-value

Care is needed with synthetic data generation and interpretation of p-values
Large p-value issues:

- other distributions may better match the data better
- if n is small, hard to make any definitive conclusion

Comparison to other distributions

Likelihood ratio test:
likelihood of data under two different distributions
higher likelihood wins

or use ratio or log of ratios

need sufficiently large log likelihood to be conclusive

depends on fluctuations σ

method to tell if R is statistically significant

handle nested distributions (families containing others - i.e. power law and truncated exponential, etc)

Rest of paper is testing, applications, demo.