A Fixed-Point Algorithm to Estimate the Yule-Simon Distribution Parameter

Juan Manuel Garcia Garcia

Computer Systems Department
Instituto Tecnológico de Morelia
Morelia, Mexico

Abstract

The Yule-Simon distribution is a discrete probability distribution related to preferential attachment processes such as the growth in the number of species per genus in some higher taxon of biotic organisms, the distribution of the sizes of cities, the wealth distribution among individuals, the number of links to pages in the World Wide Web, among others. In this paper we present an algorithm to, given a set of observations stemmed from a Yule process, obtain the parameter of the Yule-Simon distribution with maximum likelihood. In order to test our algorithm, we use a modified Polya urn process simulation to generate some data that was used as input to our algorithm. We make a comparison of our algorithm with another methods and also we show an application to some empirical data.

Keywords: Yule-Simon distribution; Yule process; preferential attachment

1. Introduction

The Yule-Simon distribution was formulated by Yule [1] to describe the distribution of species among genera in some higher taxon of biotic organisms. New genera are added to a taxon whenever a newly appearing species is considered sufficiently different from its predecessors that it does not belong in any of the current genera. New species are added as old ones speciate, that is, split into two species. The probability that a new species appears in a genus is proportional to the number of species the genus already has. Simon introduced essentially the same growth mechanism to explain the observed distribution of word frequencies in texts [2, 3, 4, 5]. This mechanism is also underlying to the preferential attachment model of complex networks proposed by Baràbasi and Albert [6, 7]. Then the Yule-Simon distribution is related not only to the distribution of species among genera, but also to many other stochastic data like distribution of words frequencies in a document, distribution of income, distribution of cities by number of habitants, among many others [8].

Email address: jgarciagarcia@acm.org (Juan Manuel Garcia Garcia)

The probability mass function of the Yule-Simon distribution depends on a single parameter. This papers presents an algorithm to estimate, given a set of observations stemmed from a Yule process, the Yule-Simon distribution parameter with the maximum likelihood.

2. Yule-Simon distribution

To explain how are distributed the different species among the genera, Yule [1] proposed the following model:

There are two types of mutations, ones that produce a new specie of the same genus, called *specific* mutations, and other that produce an entirely new genus, called *generic* mutations. Let g be the *generic mutation rate* and s the *specific mutation rate*. Assuming that populations of one specie grows exponentially, the probability that a genus of age t to be composed of a single specie is

$$p_1(t) = e^{-s \cdot t}. \tag{1}$$

The probability of a genus of age t to be composed by two species, $p_2(t)$ is equal to the probability of having a specific mutation at some time between time 0 and t and not having any further mutation, that is,

$$\begin{aligned} p_2(t) &= \int_0^t p_1(t') \cdot s \cdot e^{-2s(t-t')} dt' \\ &= e^{-st} \cdot \left(1 - e^{-st}\right). \end{aligned} \tag{2}$$

By induction, it can be showed that, for a genus of age t, the probability of having a k number of species is given by

$$p_k(t) = e^{-st} \cdot \left(1 - e^{-st}\right)^{k-1}. \tag{3}$$

And, since the distribution of ages of genera is given by $g \cdot e^{-gt}$, the probability that a given genus could have k species is given by

$$\begin{aligned} p_k &= \int_0^\infty p_k(t) \cdot g \cdot e^{-gt} dt \\ &= \int_0^\infty e^{-st} \cdot \left(1 - e^{-st}\right)^{k-1} \cdot g \cdot e^{-gt} dt \\ &= \frac{g}{s} \cdot \int_0^1 x^{g/s} (1-x)^{k-1} dx \\ &= \frac{g}{s} \cdot B(g/s + 1, k) \\ &= \frac{g}{s} \cdot \frac{\Gamma(k)\Gamma\left(\frac{g}{s} + 1\right)}{\Gamma\left(\frac{g}{s} + 1 + k\right)} \end{aligned} \tag{4}$$

where B is the Euler Beta function and Γ is the Gamma function [9]. If we make $\rho = g/s$ we have the Yule-Simon distribution [1, 2] probability mass function:

$$f(k;\rho) = \rho \cdot \frac{\Gamma(k) \cdot \Gamma(\rho+1)}{\Gamma(k+\rho+1)}. \tag{5}$$

where k is any positive integer and the $\rho > 0$ parameter, called the *shape*, is a real number.

This distribution was rediscovered by Simon [2] as that of words by the frequency of their use in a document. In the Simon model, when a text is being written each word is added in the following way: with probability α a new word is added and with probability $(1-\alpha)$ a previous word is selected. The probability that it will be a particular word already written is linearly proportional to the number of its previous occurrences. The probability that a particular word had been repeated k times along the document is given by (5), where $\rho = 1/(1-\alpha)$.

Simon observed that the same distribution can be applied to other cases like distributions of numbers of papers published by scientists, distribution of cities by population and distribution of incomes.

More recently, in order to explain the distribution of connectivity in the World Wide Web [10] and other networks Barabasi and Albert [6] proposed the preferential attachment model. This model reduces to Simon's model as was pointed out in [7]. For a review of the close relation of Yule's model with Polya urn models, branching processes, random graphs and coagulation models, among others, see [8].

3. Fixed Point Algorithm

In this section we present an algorithm to estimate the parameter ρ of the p.m.f. given by equation (5). Our algorithm is based on a maximum likelihood estimation of the Yule distribution parameter given a set of observations of a random variable.

Given $k_1, k_2, \ldots, k_N$ i.i.d. observations, the joint probability of these observations given a fixed shape ρ is

$$f(k_1, k_2, \ldots, k_N|\rho) = f(k_1;\rho) \cdot f(k_2;\rho) \cdots f(k_N;\rho) \tag{6}$$

Then the *likelihood* of the parameter ρ given the observations $k_1, k_2, \ldots, k_N$ is defined as

$$\mathcal{L}(\rho|k_1, k_2, \ldots, k_N) = \prod_{i=1}^{N} f(k_i;\rho) \tag{7}$$

Given the set of observations $k_1, k_2, \ldots, k_N$ we want to estimate the Yule distribution parameter ρ which maximizes the likelihood $\mathcal{L}$. For practical convenience we instead work with the *logarithmic average likelihood* defined as

$$\hat{l} = \frac{1}{N} \ln \mathcal{L} \tag{8}$$

and then

$$\hat{l} = \frac{1}{N} \ln \left[\prod_{i=1}^{N} f(k_i; \rho) \right] = \frac{1}{N} \sum_{i=1}^{N} \ln f(k_i; \rho) \tag{9}$$

Replacing (5) in (9) we get:

$$\hat{l} = \ln \rho + \ln \Gamma(\rho + 1) + \frac{1}{N} \sum_{i=1}^{N} \ln \Gamma(k_i) - \frac{1}{N} \sum_{i=1}^{N} \ln \Gamma(k_i + \rho + 1) \tag{10}$$

then

$$\frac{\partial \hat{l}}{\partial \rho} = \frac{1}{\rho} + \frac{\Gamma'(\rho + 1)}{\Gamma(\rho + 1)} - \frac{1}{N} \sum_{i=1}^{N} \frac{\Gamma'(k_i + \rho + 1)}{\Gamma(k_i + \rho + 1)} \tag{11}$$

that is

$$\frac{\partial \hat{l}}{\partial \rho} = \frac{1}{\rho} + \psi(\rho + 1) - \frac{1}{N} \sum_{i=1}^{N} \psi(k_i + \rho + 1) \tag{12}$$

where ψ is the *digamma* function [9].

In order to obtain the maximum loglikelihood we make (12) equal to zero to obtain

$$\frac{1}{N} \sum_{i=1}^{N} \psi(k_i + \rho + 1) = \frac{1}{\rho} + \psi(\rho + 1) \tag{13}$$

The digamma function satisfies the following recurrence relation [9]:

$$\psi(x + 1) = \psi(x) + \frac{1}{x} \tag{14}$$

which we can apply to reduce the terms inside the summation at the left side of (13) as follows:

$$\begin{aligned} \psi(k_i + \rho + 1) &= \psi(k_i + \rho) + \frac{1}{k_i + \rho} \\ &= \psi(k_i - 1 + \rho) + \frac{1}{k_i - 1 + \rho} + \frac{1}{k_i + \rho} \\ &\vdots \\ &= \psi(1 + \rho) + \frac{1}{1 + \rho} + \cdots + \frac{1}{k_i + \rho} \end{aligned}$$

that is

$$\psi(k_i + \rho + 1) = \psi(\rho + 1) + \sum_{j=1}^{k_i} \frac{1}{\rho + j} \tag{15}$$

and then

$$\frac{1}{N} \sum_{i=1}^{N} \psi(k_i + \rho + 1) = \psi(\rho + 1) + \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k_i} \frac{1}{\rho + j} \tag{16}$$

Replacing (16) in the left side of (13) we have

$$\psi(\rho+1)+\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k_i}\frac{1}{\rho+j}=\frac{1}{\rho}+\psi(\rho+1) \tag{17}$$

Then we obtain

$$\rho=\frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{k_i}1/(\rho+j)}. \tag{18}$$

The equation (18) suggests to us a fixed point algorithm to estimate the shape ρ with maximum likelihood. This fixed-point algorithm is showed as the algorithm 1.

Data: An array $k_1,\ldots,k_N$ of N observations of a Yule process, an initial approach ρ_0, and a tolerance ϵ
Result: Estimated parameter ρ of the corresponding Yule distribution
begin
  $f \leftarrow$ false;
  while $\neg f$ **do**
    $\rho_1 \leftarrow N/\left(\sum_{i=1}^{N}\sum_{j=1}^{k_i}1/(\rho_0+j)\right)$;
    $f \leftarrow (|\rho_1-\rho_0|<\epsilon)$;
    $\rho_0 \leftarrow \rho_1$;
  end
  return ρ_1;
end

Algorithm 1: Fixed Point Algorithm to estimate ρ

In order to test our algorithm we firstly use data generated by the simulation of a Yule process with some already known ρ parameter, so we can verify how accurate is the estimation made by the algorithm 1. For simplicity sake, we simulate a modified Polya urn process, as explained in the following section.

4. Modified Polya urn process simulation

Consider the following modified Polya urn process: Given finitely many bins each containing one ball, suppose that additional balls arrive one at a time. For each new ball, with probability α, a new bin is created and the ball is placed in that bin; and with probability $(1-\alpha)$, the ball is placed in an existing bin, such the probability the ball is placed in a particular bin is proportional to the number of balls in that bin [11, 12]. We can observe that this stochastic process is fully equivalent to the Simon's model. Then, the probability a bin has exactly k balls is given by equation (5) with $\rho=1/(1-\alpha)$.

The simulation algorithm of this modified Polya urn process is the algorithm 2. In algorithm 2, the probability α of a new bin creation in each time step, the amount b_M of balls to generate before stop and the initial number b_0 of balls are given as input and then it outputs the sequence of integers $k_1,k_2,\ldots,k_N$, where k_i is the number of balls that contains the i-th bin.

In order to test algorithm 1, we first use algorithm 2, giving a fixed α, to generate the sequence $k_1, \cdots, k_N$ and then we use this sequence as input of algorithm 1. The estimated parameter ρ computed by algorithm 1 must be close to $1/(1-\alpha)$.

```
Data: Parameter α of the modified Polya urn process, initial number of
      balls b_0, number of balls to generate b_M
Result: An array k_1, ..., k_N where k_i is the number of balls contained in
        the bin i
begin
    balls ← b_0;
    N ← b_0;
    for i ← 1 to N do
        k_i ← 1;
    end
    while balls ≤ b_M do
        Generate a pseudo random number 0 ≤ r_0 ≤ 1;
        if r_0 < α then // A new bin is generated
            N ← N + 1; k_N ← 1; balls←balls+1;
        else// Ball is placed in a preexisting bin
            Generate a pseudo random number 1 ≤ r_1 ≤ balls;
            b ← 0; s ← 1;
            while r_1 > b + k_s do // Preferential attachment
                b ← b + k_s; s ← s + 1;
            end
            k_s ← k_s + 1; balls ← balls + 1;
        end
    end
    return k_1, ..., k_N;
end
```

Algorithm 2: Polya urn process simulation

In table 1 we summarize some of the results we obtained from our experimental tests. In the first column we have the parameter α given to simulation algorithm 2. The second column is the value of $1/(1-\alpha)$, that is, the exact value of ρ which we want to estimate using algorithm 1. The following columns are the different estimations of ρ for various supplied data sets.

The algorithm 1 was always run with $\epsilon = 0.00001$ and initial approach $\rho_0 = 0$. For the estimate ρ_1 the data were generated with $b_M = 2 \times 10^5$. For ρ_2, data was generated with $b_M = 2.5 \times 10^5$. And b_M was 5×10^5 and 1×10^6, for estimates ρ_3 and ρ_4 respectively. As observed in table 1, good estimates of ρ were always obtained, and the accuracy mainly depends, as expected, on the number of observations used as input.

In all our experimental tests we always observe convergence of algorithm 1 and even for $N \approx 1 \times 10^6$ the running time was less than a second on a computer with a 1 GHz Pentium Dual Core processor. In most of our experimental tests convergence was reached in less than 10 iterations. In our future work the formal

α	$1/(1-\alpha)$	ρ_1	ρ_2	ρ_3	ρ_4
0.1	1.111111	1.108276	1.106701	1.109161	1.113740
0.2	1.250000	1.238459	1.244910	1.247314	1.248361
0.3	1.428571	1.446412	1.429618	1.426060	1.427149
0.4	1.666666	1.682525	1.671963	1.662993	1.664442
0.5	2.000000	2.035083	2.004017	1.993611	1.999833
0.6	2.500000	2.540717	2.505299	2.495214	2.499993
0.7	3.333333	3.357493	3.340851	3.325171	3.331717
0.8	5.000000	5.022355	5.019381	5.000228	4.995364
0.9	10.00000	10.09391	9.984415	9.994049	9.994425

Table 1: Experimental test results

proof on convergence properties of our algorithm remains.

5. Comparison with other methods

Given that the p.m.f. expressed by eq. (5) has the property that for sufficiently large k we have

$$f(k;\rho) \approx \frac{\rho \cdot \Gamma(\rho+1)}{k^{\rho+1}} \propto \frac{1}{k^{\rho+1}}, \tag{19}$$

then $f(k;\rho)$ can be estimated as an inverse power law with exponent $\rho+1$ [1, 2]. Two current methods to estimate, given a set of observations, the inverse power law exponent are the least-squares and maximum-likelihood fitting methods [13].

First of all, a power-law distribution is one described by a probability density $p(x)$ such that

$$p(x) = Cx^{-\beta} \tag{20}$$

where C is a normalization constant and the exponent β is called the *scaling parameter*.

In the well known least-squares fitting method, the scaling parameter of the power law is obtained by performing a least-squares linear regression on the histogram plotted on a log-log scale. The value of the parameter β is given by the absolute slope of the straight line. This procedure dates back to Pareto's work on the distribution of wealth at the end of the 19th century [14].

In the maximum-likelihood fitting method, the maximum likelihood estimator for the scaling parameter is given by:

$$\beta = 1 + N\left[\sum_{i=1}^{N} \ln(2x_i)\right]^{-1} \tag{21}$$

where x_i, $i = 1, \ldots, N$ are the observed values of x [13].

Then, another way to estimate the Yule-Simon distribution parameter ρ is applying least-squares or maximum-likelihood fitting methods to obtain the

scaling factor β and then we can estimate ρ as $\beta - 1$. We did this for the synthetic data generated by the modified Polya urn process simulator discussed on the previous section and the we compare the obtained estimations with those produced by our algorithm. Results of this tests are summarized on table 2.

α	$1/(1-\alpha)$	ρ_F	ρ_L	ρ_M
0.1	1.111111	1.109161	1.072947	0.715815
0.2	1.250000	1.247314	1.163400	0.761797
0.3	1.428571	1.426060	1.359513	0.812980
0.4	1.666666	1.662993	1.536816	0.870747
0.5	2.000000	1.993611	1.771245	0.935924
0.6	2.500000	2.495214	2.002797	1.010872
0.7	3.333333	3.325171	2.356327	1.097092
0.8	5.000000	5.000228	3.052245	1.196469
0.9	10.00000	9.994049	4.545002	1.309242

Table 2: Estimated parameters comparison (synthetic data)

In the first column of table 2 we have the parameter α of the modified Polya urn process and then, in the second column, the exact corresponding parameter $\rho = 1/(1-\alpha)$. The estimation ρ_F was obtained using our fixed-point algorithm with a tolerance $\epsilon = 0.00001$ and an initial approach $\rho_0 = 0$. The estimation ρ_L was obtained by the least-squares fitting method and ρ_M by the maximum-likelihood fitting method. The data was generated using algorithm 2 with $b_M = 2 \times 10^5$.

As can be observed, better results were obtained with our method. Matter of fact, $\rho_M < \rho_L < \rho_F$ and this divergence grows as ρ becomes greater. Again, on these tests we use synthetic data because simulation allows to fix an exact value of ρ and then we can measure how accurate are the obtained estimations. As a further test, we apply this methods to empirical data, as discussed below.

6. Application to empirical data

According to Simon [2], the Yule-Simon distribution applies to frequency of words occurrence in a text. In this second test, for a given text composed by N different words, k_i is the number of times that i-th word appears on this text. We use $k_1, \ldots, k_N$ as input data for algorithm 1 to estimate parameter ρ. For comparison, we also make estimations of ρ by the least-squares and maximum-likelihood fitting methods discussed on previous section.

As texts we choose the following novels: (A) *Ulysses* by James Joyce, (B) *Don Quixote* by Miguel de Cervantes, (C) *Moby Dick* by Herman Melville, (D) *War and Peace* by Leo Tolstoi and (E) *Les Miserables* by Victor Hugo. We use English translations of Cervantes, Tolstoi and Hugo novels. All texts were downloaded from the Gutenberg Project site.

Estimated parameter for each method are presented on table 3. As before, ρ_F was obtained with algorithm 1, ρ_L with least-squares and ρ_M with maximum-likelihood fitting methods.

Text	N	ρ_F	ρ_L	ρ_M
A	30938	1.108752	0.635097	0.721935
B	24329	0.905986	0.371927	0.640057
C	17752	0.886444	0.461332	0.624012
D	18505	0.629574	0.372494	0.493363
E	23975	0.700271	0.447398	0.534301

Table 3: Estimated parameters comparison (empirical data)

In order to measure the *goodness of fit* we follow the method exposed on [13] which generates a p value that quantifies the plausibility of a model. This procedure is as follows: Empirical data is fitted to a model using one of the reviewed methods (fixed point, least-squares or maximum-likelihood) and the corresponding Kolmogorov-Smirnov statistic is calculated for this fit. Next, a large number of synthetic data sets are generated with same parameters to those of the distribution that fits the observed data. Then the Kolmogorov-Smirnov statistic for each data set relative to its own model is calculated. Then is counted what fraction of the time the resulting statistic is larger than the value for the empirical data. This fraction is the p-value that measures the goodness of fit.

In table 4 we present p-values for each one of the estimated parameters for each text. p_F is the p-value for a Yule-Simon model with ρ obtained by algorithm 1, p_L is for the power-law model with parameters estimated by the least-squares fitting method and p_M for the power-law model obtained by the maximum likelihood method.

Text	p_F	p_L	p_M
A	0.627049	0.269625	0.358362
B	0.540541	0.278090	0.412921
C	0.313208	0.340580	0.387681
D	0.372494	0.456140	0.445175
E	0.381679	0.368298	0.487179

Table 4: Goodness of fit

From these results can be inferred that, in A and B texts, word frequencies are well fitted by a Yule-Simon distribution but in another cases a Yule-Simon model seems as good as a pure power-law. To find an explanation of why some texts seems to follow more closely a Yule model than others is beyond the aim of this paper.

7. Conclusions

In this paper we present an algorithm to, given a set of observations stemmed from a Yule process, obtain the parameter of the Yule-Simon distribution with maximum likelihood. In order to test our algorithm, we use a modified Polya urn process simulation to generate some data that was used as input to our algorithm. Then we compare the estimate of the parameter with the exact value of the corresponding parameter fixed on the simulator. Good properties of convergence were observed in all our experimental tests. We also made a comparison with another methods, obtaining better results for synthetic and for some empirical data.

The algorithm presented in this paper can be applied to easily adjust a Yule-Simon distribution to a set of observations coming from a preferential attachment process as those related to complex networks.

8. References

[1] G. U. Yule, A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S., Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213 (1925) 21–87.

[2] H. A. Simon, On A Class Of Skew Distribution Functions, Biometrika 42 (1955) 425–440.

[3] J. B. Estoup, Gammes Stenographique, Institute stenographique de France, Paris, 1916.

[4] E. U. Condon, Statistics of Vocabulary, Science 67 (1928) 300–+.

[5] G. K. Zipf, The Psycho-Biology of Language, MIT Press, Cambridge, Massachusetts, USA, 1935.

[6] A. L. Barabasi, R. Albert, Emergence of scaling in random networks, Science (New York, N.Y.) 286 (1999) 509–512.

[7] S. Bornholdt, H. Ebel, World wide web scaling exponent from simon's 1955 model, Phys. Rev. E 64 (2001) 035104–+.

[8] M. V. Simkin, V. P. Roychowdhury, Re-inventing willis, ArXiv Physics e-prints (2006).

[9] G. E. Andrews, R. A. Askey, R. Roy, Special functions; 2nd ed., Encyclopaedia of mathematics and its applications, Cambridge Univ. Press, Cambridge, 2001.

[10] R. Albert, H. Jeong, A. L. Barabasi, The diameter of the world wide web, Nature 401 (1999) 130–131.

[11] H. Mahmoud, Polya Urn Models, Chapman & Hall/CRC, 2008.

[12] F. Chung, S. Handjani, D. Jungreis, Generalizations of polya's urn problem, Annals of Combinatorics 7 (2003) 141–153. 10.1007/s00026-003-0178-y.

[13] A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-Law Distributions in Empirical Data, SIAM Review 51 (2009) 661–703.

[14] B. C. Arnold, Pareto Distribution, Encyclopedia of Statistical Sciences, John Wiley & Sons, Inc., 2004.