Transcript of the video for

Information measures and geometry of the Zeta distributions and related distributions [1]*

https://www.youtube.com/watch?v=YvAbs0mjhN0

Frank Nielsen

Hi, I am Frank Nielsen and today I would like to talk about the Zeta distributions and their related distributions which are power laws distributions frequently met in applications. We shall consider information measures between zeta distributions and also their information geometry as discrete exponential families.

To start with, let us consider the Zipf's distributions which are discrete power law distributions with the following probability mass functions. The normalization constants to make those distributions probability measures are called the generalized harmonic numbers $H_{s,N}$. The support of Zipf's distributions are the integers ranging from 1 to N.

George Zipf was an American linguist who found that many empirical distributions of frequencies of words in human language texts follow such power law distribution. Here I plot two Zipf distributions: one in blue with discrete support ranging from 1 to 5, and one in red with discrete support ranging from 1 to 8. The corresponding power parameters are denoted by s

---

For example, in this plot we draw the log-rank versus the log-size for the 135 largest urban areas of the United States. In this log-log plot, we can see that the data are well fitted by a line on a range. This is explained by writing the log probability mass function and see that we get a linear equation in s. Thus the urban area size distributions follow a Zipf law.

Here we show yet another example of log-log plot which show the frequency distribution of US company sizes Again, we see that the data is well approximated by a line in this log-log plot. We see that we could estimate the power law parameters by line fitting method but this is not recommended in theory as power laws exhibit heavy tails. In this third and last example, the authors have considered the 100 translations of the Holy Bible and reported in a table the fitted Zipfs distributions including the power exponent and the support N which are the language vocabulary size used in the translation. For example, we show the log-log plot of the translation of the Holy Bible in Esperanto and see that the word frequencies follow nicely a Zipf power law distribution.

*Nielsen, Frank. "Comparing the Zeta Distributions with the Pareto Distributions from the Viewpoint of Information Theory and Information Geometry: Discrete versus Continuous Exponential Families of Power Laws." Physical Sciences Forum. Vol. 5. No. 1. MDPI, 2022 (arXiv:2104.10548). https://franknielsen.github.io/ZetaParetoExpFam/index.html

To analyze the specificities of various human languages and get an idea of their structure similarity and dissimilarities, we could use clustering techniques on their corresponding Zipf's distributions. Clustering requires a notion of proper distances between their features like Zipf distributions. For example, here we show a dendrogram obtained by hierarchical agglomerative clustering.

Now that I have introduced Zipf distributions and their use cases, let me define the Zeta distributions as limit of Zipf distributions when the support is the full range of integers. In that case, the former normalizing constants which were generalized harmonic numbers become the real Riemann zeta function. Because of the infinite summation, we now need the power exponent to be greater than 1. This contrasts with the Zipfs distributions which only required the power exponent to be positive. On the plot, I show the impact of the power exponent s on the probability mass functions of the zeta distributions.

In practice, we need to calculate efficiently an approximation of the Riemann zeta function. There are many algorithms for fast numerical approximations, and the zeta function is given as a primitive in many software packages like the free symbolic computing software Maxima which I use. We can also lower and upper bound the zeta distribution. Here I show the lower bound in red and the upper bound in blue on the zeta function. Notice that when s tends to infinity, zeta tends to one.

We shall now express the set of zeta distributions as a discrete exponential family. We can write the probability mass function of the zeta distributions as the following canonical expression of the densities of an exponential family. It comes that the natural parameter is theta equals s and the minimal sufficient statistic is $t(x) = -\log(x)$ The log-normalizer also called Cumulant function or log-partition function is the logarithm of the zeta function. Since the Fisher information of an exponential family is the negative expected Hessian of the log-density, we get the following expression for the Fisher information of the zeta distributions.

Here, for sanity check, we show the plot of the log zeta function and check that it is a strictly convex function. The log-normalizer are moreover always analytic.

A very interesting rewriting the Fisher information can be done using the so-called von Mangoldt function which is often used in number theory. The von Mangoldt function of I is defined as log p if I equals p to the power k for some prime p and integer k and 0 otherwise. The plot shows the von Mangoldt function for the first 100 integers. An interesting identity is to write the logarithm of any integer n as the sum of the von Mangoldt function for all dividers I of n.

Because the set of zeta distributions form a discrete exponential family, these zeta distributions can be interpreted as maximum entropy distributions. Thus if we maximize Shannon entropy for a distribution with the natural integer support with the constraint on the expectation on its sufficient statistic minus logarithm of x, we get the zeta distributions We can now introduce the dual parameterization of exponential families called the moment parameters or the mean parameters. Thus the dual parameter can be obtained as the derivative of the logarithm of the zeta function.

The Maximum Likelihood Estimator for the zeta distribution coincides with the method of moments for the sufficient statistics Therefore we get that the MLE in the eta parameter is the average minus logarithm of the identically and independently distributed data. To get the corresponding natural parameter, we need to inverse the log zeta derivative which is a complex task. In practice, several papers have studied different estimator strategies. The easiest approach is to do the line fitting in the log-log coordinate: This is called the quadratic distance estimator. It is shown efficient but suffers of the heavy tail property of the power law of zeta distributions.

The Fréchet-Darmois-Cramér-Rao lower bounds the variance of any unbiased estimator as the inverse of the Fisher information. Since the Fisher information of an identically and independently random vector is additive, we get that the Fisher information of a n-dimensional random vector is n times the Fisher information of the zeta distribution. Thus we get the following variance of the MLE expressed using the zeta and its first and second derivatives. Notice that this is CRLB bound is tight for exponential families only. We thus easily recover a result published in this paper by simply handling the zeta distributions as a discrete exp family.

We shall now compare the discrete power law with its continuous counterpart which is called the pareto distributions. The probability density function of a Pareto distribution on the support the interval [1,infinity] is given as follows where s is called the shape parameter of the Pareto distribution. The set of Pareto distributions form a continuous exponential family with the following canonical decomposition. The Pareto distributions was introduced by Vilfredo Pareto who was a sociologist. More general, the Pareto distributions can be defined by two parameters s and a and its Fisher-Rao information geometry has been characterized as a positive-curvature manifold.

In this synthetic table I compare the discrete versus the continuous power laws. That is the zeta distributions with the Pareto distributions.Observed that the sufficient statistic is the same but the range of the natural parameter is different. Also the cumulant function of the Pareto has a simple expression as minus the logarithm of its natural parameter minus one but the cumulant function of the zeta distributions is expressed with the Riemann zeta function. Similarly, the Fisher information for the Pareto distributions as a simplex expression but the Fisher information for the zeta distributions is much more complex. Both distributions are maximum entropy distributions. The Pareto with respect to the differential entropy.

We now come to define similarity coefficients and statistical distances between power law distributions. An important family of statistical divergences in information geometry are the alpha-divergences which can be expressed as a function of the Bhattaharryya similarity coefficient. When the densities ps1 and ps2 belong to the same exponential family, the Bhattacharrya coefficient can be expressed as the exponential of minus a Jensen divergence introduced by the convex function of the log-normalizer. Thus we obtain a simple expression of the alpha-divergences between two zeta distributions relying on the zeta function. For alpha=1/2 we get the squared Hellinger div, the only symmetric alpha-divergences.

A property of alpha-divergences is that they tend to the forward KLD and reverse KLD when alpha tends to 1 and -1, respectively. Thus by choosing in practice alpha values very close to these +1 ou − 1 values, we get an efficient way to approximate the Kullback-Leibler divergence between two zeta distributions.

We can also use the property that the KLD between two densities of an exponential family amounts to a reverse Bregman divergence for the log-normalizer of the zeta distributions to express the KLD between two zeta distributions. Furthermore, by using both the natural and dual parameters, we can express the Bregman divergene as a so-called Fenchel-Young divergence. Therefore we obtain the following expression of the KLD between two zeta distributions where I used the von Mangoldt function Lamda here.

Let us now cluster a set of n text documents by their corresponding Zipfs distributions relaxed to Zeta distributions. We can either use partition-based flat clustering like k-means or k-center clustering, or hierarchical clustering method. Since the densities are described by one parameter and since the KLD between zeta distributions amount to a Bregman divergence, a KLD kmeans amounts to a 1D Bregman k-means clustering which can be solved exactly in 1D using dynamic programming. This is because their Bregman Voronoi diagrams have convex cells and the clustering is called an interval clustering. If we want to cluster product of zeta distribution features, we end up with a generic Bregman k-means clustering and this problem is NP-hard.

Instead of relaxing the Zipf distributions to zeta distributions, we can also do k-means of Zipf distributions with the prototypes being zeta distributions. In that scenario, we need a distance between a Zipf distribution and a zeta distribution. The Zipf distribution can be interpreted as a right truncation of the zeta distribution and a formula for the Kullback-Leibler divergence between two truncated densities of an exponential family with nested support has been recently reported as a duo Bregman divergence. Here we show that formula of the KLD between a Zipf distribution and a zeta distribution

---

This figure visualizes the duo Bregman divergences induced by two strictly convex and smooth functions F1 and F2 with F1 dominating function F2 The equation of the duo Bregman divergence can be interpreted as the vertical gap in red. This is a pseudo-divergence when F1 is different than F2 because it cannot equal zero. When F1 equals F2, we recover the classic formula of the Bregman divergence.

Let me summarize our comparative study of discrete versus continuous power law distributions. I want to point out that the differential entropy of a Pareto distribution can be negative but the entropy of a zeta distributions is always positive. Also Observe that the MLE for the Pareto distributions admits a simple closed-form formula but this is not the case for the zeta distributions because we need to convert the moment parameter eta to the corresponding natural parameter.

To perform simulation and evaluate efficiency of various numerical schemes, we need to draw identically and independently variates of these power laws. Here I show some methods to sample a Zeta distribution using the acceptance/rejection method from the book of Devroye, and to sample a truncated Pareto distribution using the simple inverse transform technique.

To conclude, we have compared the Zeta and Pareto power distributions from the viewpoint of discrete and continuous exponential families. I would like to mention another recent related work which studies the discrete Gaussian distributions defined as maximum entropy discrete distributions on a lattice with first and second fixed raw moments.

Thank you very much for your attention.

References

[1] Frank Nielsen. Comparing the Zeta Distributions with the Pareto Distributions from the Viewpoint of Information Theory and Information Geometry: Discrete versus Continuous Exponential Families of Power Laws. In *Physical Sciences Forum*, volume 5, page 2. MDPI, 2022.