Estimating the Unseen: Improved Estimators for Entropy and other Properties

Yaswanth Duvvuru, SampathKumar Kilaparthi, Premsai Chinthamreddy
{112073592, 112079198, 1}

Abstract

This paper shows that a class of statistical properties of distributions can be used to estimate given a sub-linear sized sample, which includes practically relevant properties as entropy, the number of distinct elements, and distance metrics between pairs of distributions. The primary idea behind this is to estimate the empirical distribution of new elements, given a sample distribution. Estimating these properties has important implications in various fields. We also demonstrate that the estimators perform exceptionally well, in practice, for a variety of estimation tasks, on a variety of natural distributions, for a wide range of parameters. The key step in our approach is to first use the sample to characterize the "unseen" portion of the distribution effectively reconstructing this portion of the distribution as accurately as if one had a logarithmic factor larger sample.

1 Introduction

From the statistical perspective, estimating the entropy is by no means an unique problem among the problems of estimating functionals of parameters. And yet, not much is known and still is an area of interest because of its importance in practice. This is the reason why it has attracted so much attention and also we have a very poor understanding of the general problem of functional estimation. This work is an attempt to estimate the empirical distribution of new elements, given a sample distribution. Estimating these properties has important implications in various fields. The unseen distribution presented in this paper can be modelled to identify the distribution of species in the field of biology. These studies can be extended beyond and finds in application in identifying new species from the original species collection setting. For example, in genomics, the important question is: given the genetic variation already identified in the genomes of individuals from some population(say, East Asia), how many additional mutations do we expect to find by sequencing the genomes of additional individuals from East Asia. Likewise, this technique can be used to modeled to arrive at decisions with limited data. These challenges are encountered in various datasets including text corpus, customer data, genetic mutations etc.

1.1 Previous Work

Recently, a number of other estimators based on different arguments have been proposed. Among these, the most commonly used for discrete distributions are the Nemenman-Shafee-Bialek (NSB), the Centered Dirichlet Mixture, the Pitman-Yor mixture, and the Dirichlet process mixture. These are Bayesian estimators, which then hing on explicitly defined probabilistic assumptions. Similarly, non Bayesian measures have been suggested, such as the Coverage-Adjusted estimator, the Best Upper Bound, or the James Stein estimator. It should be highligted that there is no unbiased estimator in this context, and that the convergence rate of different estimators can vary in a considerable manner, in some cases being arbitrarily slow. However, for the specific question of the OP, which is based on a discrete distribution with finite range, a reasonable choice could be the NSB estimator, which uses an approximately flat prior distribution over the values of the entropy, built as a mixture of symmetric Dirichlet distributions. This estimator shows rapid convergence to the entropy and good performances in terms of robustness and bias.

2 Approach

In this paper, we show that a class of statistical properties of distributions can be used to estimate given a sub-linear sized sample, which includes

practically relevant properties as entropy, the number of distinct elements, and distance metrics between pairs of distributions. The solution to this interesting problem might be considerably easier if we could deal with a large number of samples. Regardless of the use of an empirical distribution function or of a distribution directly obtained by raw data, when we have a large number of samples. We can rapidly calculate entropy using the standard formula for Shannon entropy. Estimating entropy is not an easy problem and have been a subject of research for years. Studies like [Paninski 2003] revealed that there is no unbiased estimator for entropy. But there are plenty of good entropy estimators that have low bias and/or low variance.

However, there are issues to be considered in order to solve the unseen distribution. The first is that we are harvesting information regarding the analysis from a relatively small set of observations taken from a very larger set of possibilities. Estimating the unseen from such an under-sampled context is quite challenging. On the other hand, we should make clear assumptions of the underlying probability distributions to link our sample dataset to some entropy measure. Furthermore, estimating the entropy from a single set of under-sampled observations is not easy either. In practice, we have a set of unknowns to solve first.

This challenge of inferring properties of a distribution given a "too small" sample is encountered in a variety of settings, including text data (typically, no matter how large the corpus, around 30 of the observed vocabulary only occurs once), customer data (many customers or website users are only seen a small number of times), the analysis of neural spike trains, and the study of genetic mutations across a population. Additionally, many database management tasks employ sampling techniques to optimize query execution; improved estimators would allow for either smaller sample sizes or increased accuracy, leading to improved efficiency of the database system.

Specially in the field of machine learning, we face several challenges in analyzing raw data in real world; indeed there are several cases where large datasets represent only a tiny fraction of an underlying distribution we are trying of make sense of. Through this paper, the author introduced a general and robust approach for using a sample to characterize the unseen portion of the distribution. This begins with the unseen estimator, developed for sampling from a bare minimal sample and adjusts for the potential new elements that have not yet been observed in the samplethese become the new patterns or "words" in a spike train that have not yet been observed. Without a prior assumptions about the distribution, we try to estimate the symmetric properties like entropy, support size etc of the unseen portion of the distribution. Properties such as entropy, the number of distinct elements, and distance metrics between pairs of distributions, can be estimated given a sub-linear sized sample. To be precise , we estimate the number of unseen domains elements that occur in various probability ranges. And given such a reconstruction, we use it to estimate certain properties/ functional of the distribution which depends on shape or histogram of the given distribution.

The paper summarizes the performance guarantees of this approach in terms of the following three concrete and practically relevant questions, each defined with respect to an arbitrarily small constant error parameter $\epsilon > 0$.

Distinct Elements: Given n buckets, each of which contains one object that is not necessarily distinct from those in the other buckets, how many buckets must one inspect in order to estimate the total number of distinct objects to within $\pm\epsilon$k, with high probability ?

Entropy Estimation: Given a sample obtained by taking independent draws from a distribution, p, of support size at most k, how large does the sample need to be to estimate the entropy of the distribution, $H(p) := \Sigma_{x:p(x)>0} p(x) log(p(x))$, to within $\pm\epsilon$, with high probability ?

Distance: Given two samples obtained by taking independent draws from two distributions, p_1, p_2 of support size at most k, how large do the samples need to be to estimate the total variation distance between the distributions (also referred to as "statistical distance"), $D_{tv}(p_1, p_2) = |p_1(x)$

- $p_2(x)|$, to within $\pm\epsilon$, with high probability ?

The vast majority of estimators for entropy, for example, are linear functions of the summary statistics of the sample, F_1,F_2, ... where F_i denotes the number of domain elements that occur exactly i times in the sample. But however, in order to estimate the unseen, there exists a bound on the sample size with which we perform the empirical analysis. While the previous suggests that no explicit estimators were known to solve this problem using sample size of size o(k) even for e= 0.49. The sample mentioned in this paper is generally of size n = $\mathcal{O}(k/\log k)$ where the constant is dependent on the error parameter e. Given a sample consisting of independent draws from any distribution over at most k distinct elements, these properties can be estimated accurately using a sample of size $\mathcal{O}(k/\log k)$. This sample can be fruitfully used as a component within larger machine learning and statistical analysis systems.

The problem is posed as finding the simplest plausible histogram as a pair of linear programs. The approach is best described as the inverse problem given a sample, what is the histogram of the distribution from which it was drawn - an optimization problem. We then capture the essential features of this problem via a linear program. This makes it both computationally tractable as well as amenable to rich set of probabilistic and statistical analysis tools. Furthermore, this general linear program formulation allowed for a considerable flexibility in tailoring few constraints to suit our advantage.

3 Algorithm Description

This algorithm estimates any statistical property which is independent of the labeling of the elements (symmetric) and sufficiently smooth. Rather than directly trying to estimate a specific property of the distribution, we instead take the canonical approach and return to the original question what can we infer about the true distribution given a sublinear number of samples. This algorithm returns a distribution that is, with high probability, close in some sense to the true distribution. Specifically, we return a distribution D with the property that if we had taken our samples from the hypothetical D instead of from the unknown true distribution, then with high probability the number of support elements occurring once, twice, etc. in this sample will closely match the corresponding parameters of the actual sample. We find such a distribution via linear programming. Given the distribution D returned by our algorithm, to obtain an estimate for some property, we may simply evaluate the property on D. Unsurprisingly, this yields a very good estimate; surprisingly, one can actually prove this.

From a given sample of size n, drawn from a distribution we calculate the fingerprint F of it. Using the fingerprint drawn from out sample with a histogram h, we find the histogram h' that has the properties that if one were to take n independent draws from a distribution with histogram h', the fingerprint of the resulting sample would be similar to that of the observed fingerprint F. The idea behind that is that with similar h and h', we have similar entropies and support sizes.

We start by understanding how to obtain a plausible fingerprint from a histogram. Given a distribution D, and some domain element α occurring with probability

$$x = D(\alpha), \tag{1}$$

the probability that it will be drawn exactly i times in n independent draws from D is

$$Pr[Binomial(n, x) = i] \approx poi(nx, i). \tag{2}$$

By linearity of expectation, the expected i^{th} fingerprint entry will roughly satisfy

$$E[F_i] \approx \sum_{x:h_D(x)\neq 0} h(x)poi(nx,\ i) \tag{3}$$

This mapping between histograms and expected fingerprints is linear in the histogram, with coefficients given by the Poisson probabilities. Additionally, it is not hard to show that $Var(F_i) \leq E(F_i)$, and thus the fingerprint is tightly concentrated about its expected value. This motivates a "first moment" approach. We then invert the linear map from histograms to expected fingerprint entries, to yield a map from observed fingerprints, to plausible histograms h'. Another additional component is that, there will be a large space of equally possible histograms. For a given fingerprint, there

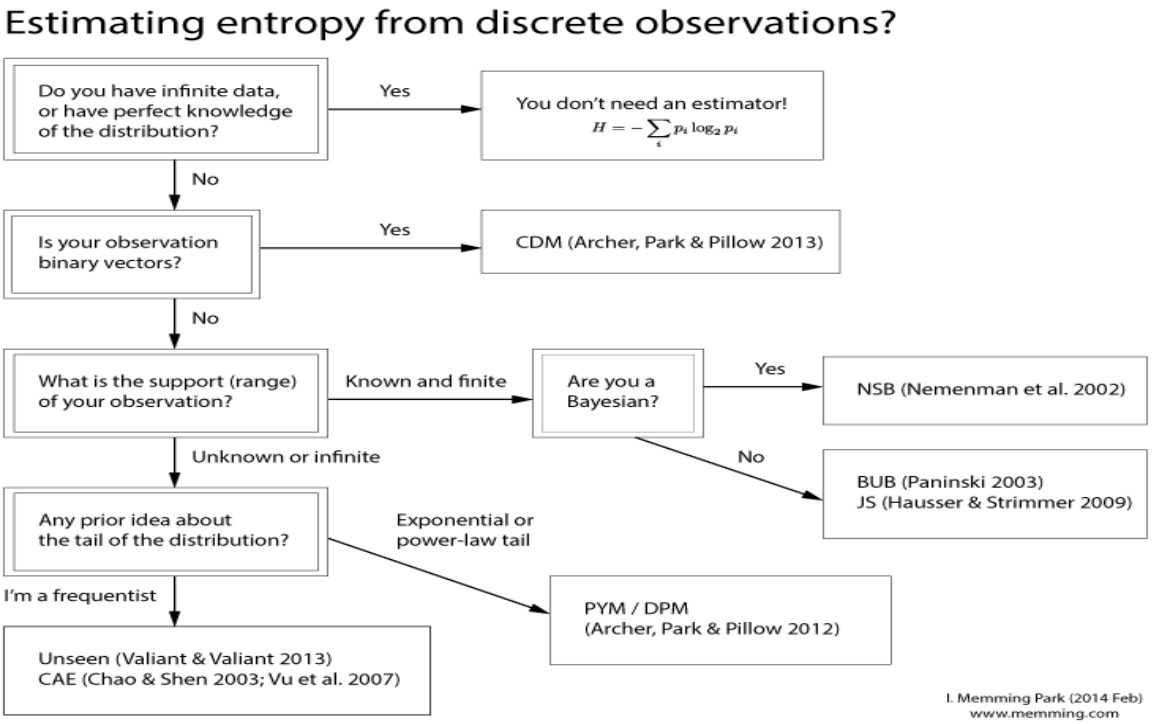

Figure 1: Flowchart for estimating entropy from unseen distributions[2]

could be several other distributions that can arrive at it. Given either distribution, the probability of obtaining the observed fingerprint from a set of samples is sizable, yet these distributions are quite different and have very different entropy values and support sizes. They are all very plausible. To resolve this issue in a principled fashion, we strengthen our initial goal of returning a histogram that could have plausibly generated the observed fingerprint: we instead return the simplest histogram that could have plausibly generated the observed fingerprint.

Example: Consider the sequence of trees as a sample from the distribution of trees on an island X = (Olive, Olive, Birch, Pine, Olive, Birch, Birch, Olive, Palm, Olive). We have F = (2,0,1,0,1) indicating that two species occurred exactly once(Pine and Palm),one species occurred exactly three times (Birch), and one species occurred exactly ve times (Olive). Consider the following distribution of trees: Pr(Olive) = 1/2, Pr(Birch) = 1/4, Pr(Pine) = Pr(Palm) = Pr(Willow) = Pr(Oak) = 1/16. The associated histogram of this distribution is h : (0,1] Z dened by h(1/16) = 4, h(1/4) = 1, h(1/2) = 1, and for all x not in 1/16,1/4,1/2, h(x) = 0.

3.1 Algorithm

We pose this problem of finding the simplest plausible histogram as a pair of linear programs. The first linear program will return a histogram h' that minimizes the distance between its expected fingerprint and the observed fingerprint, where we penalize the discrepancy between F_i and $E[F_i^{h'}]$ in proportion to the inverse of the standard deviation of F_i, which we estimate as $1/\sqrt{1+F_i}$, since Poisson distributions have variance equal to their expectation. The constraint that h' corresponds to a histogram simply means that the total probability mass is 1, and all probability values are non-negative. The second linear program will then find the histogram h" of minimal support size, subject to the constraint that the distance between its expected fingerprint, and the observed fingerprint, is not much worse than that of the histogram found by the first linear program.

To make the linear programs finite, we consider a fine mesh of values $x_1, \ldots, x_l \in (0, 1]$ that between them discretely approximate the potential support of the histogram. The variables of the linear program, $h_1^{'}, \ldots h_l^{'}$ will correspond to the histogram values at these mesh points, with variable $h_i^{'}$ representing the number of domain

elements that occur with probability x_i, namely $h'(x_i)$.

A minor complicating issue is that this approach is designed for the challenging "rare events" regime, where there are many domain elements each seen only a handful of times. By contrast if there is a domain element that occurs very frequently, say with probability 1/2, then the number of times it occurs will be concentrated about its expectation of n/2 (and the trivial empirical estimate will be accurate), though fingerprint $F_{n/2}$ will not be concentrated about its expectation, as it will take an integer value of either 0, 1 or 2. Hence we will split the fingerprint into the "easy" and "hard" portions, and use the empirical estimator for the easy portion, and our linear programming approach for the hard portion.

4 Implementation

The authors provided Matlab implementations of the unseen estimation. Here, we created an open source implementation of the algorithm as described in "Estimating the Unseen: Improved Estimators for Entropy and other Properties" using Python and other open source libraries. Our implementation has few components. First, we generate a sample of size k from the uniform distribution of support 100000. This sample is then passed to makefinger function to compute the corresponding fingerprint F. The makefinger function outputs a vector of fingerprints, given an input vector of integers, v . This fingerprint F is fed to unseen() function which outputs the approximation of histogram of the true distribution. The idea is that an approximation of the entropy of the true distribution can be computed using the output histogram from unseen().The unseen takes a fingerprint and returns the approximate histogram of the overall distribution using leveraging linear programming. The histogram is used to calculated the entropy of the original sample.

4.1 Code

The code is available here: https://github.com/sampathkumar81293/unseenEstimation

4.2 Results

We have tested our implementation with the given Matlab implementation with randomly generated datasets. We have observed minor deviations between the two results for some inputs. This can be explained by the implementation techniques used by Matlab and Python libraries for linear programming function. Also an important observation during the implementation is that the linear programming in Python will not terminate unless the values are rounded off to 4 digits.

5 Conclusion

We have implemented the given algorithm using open source libraries unlike the implementation provided by the authors which needs a third party software. From this vantage point, this is a leap from harnessing the power of linear algebra, to harnessing the power of linear programming. In addition to the more obvious directions for future investigation, an intriguing question is whether this additional power is necessary.

References

[1] [Unseen] Gregory Valiant, Paul Valiant. Estimating the Unseen:Improved Estimators for Entropy and other Properties. https://theory.stanford.edu/~valiant/papers/unseenJournal.pdf

[2] Il Memming Park. A guide to discrete entropy estimators. https://memming.wordpress.com/2014/02/09/a-guide-to-discrete-entropy-estimators/

[3] [CAE] A. Chao and T. Shen. Nonparametric estimation of Shannons index of diversity when there are unseen species in sample. Environmental and Ecological Statistics, 10(4):429443, 2003.

[4] [CDMentropy] E. Archer, I. M. Park, and J. Pillow. Bayesian entropy estimation for binary spike train data using parametric prior knowledge. In C.J.C. Burges and L. Bottou and M. Welling and Z. Ghahramani and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, 2013.

[5] [BUB] L. Paninski. Estimation of entropy and mutual information. Neural Computation, 15:11911253, 2003.

[6] [JS] J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. The Journal of Machine Learning Research, 10:14691484, 2009.

[7] [NSB] I. Nemenman, W. Bialek, and R. Van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. Physical Review E, 69(5):056111, 2004.