Computational Statistics and Data Analysis 54 (2010) 2187–2201

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

On the measure and the estimation of evenness and diversity

Josep Ginebra [*], Xavier Puig

Department of Statistics, Technical University of Catalonia, 08028, Barcelona, Spain

ARTICLE INFO

Article history:
Received 28 July 2008
Received in revised form 1 April 2010
Accepted 1 April 2010
Available online 7 April 2010

Keywords:
Overdispersion
Population size
Poisson mixture
Schur concavity
Sichel model
Species distribution
Stylometry
Uncertainty
Vocabulary distribution

ABSTRACT

Modelling word or species frequency count data through zero truncated Poisson mixture models allows one to interpret the model mixing distribution as the distribution of the word or species frequencies of the vocabulary or population. As a consequence, estimates of their mixing density can be used as a fingerprint of the style of the author in his texts or of the ecosystem in its samples. Definitions of measure of the evenness and of measure of the diversity within a vocabulary or population are given, and the novelty of these definitions is explained. It is then proposed that the measures of the evenness and of the diversity of a vocabulary or population be approximated through the expectation of these measures under the word or species frequency distribution. That leads to the assessment of the lack of diversity through measures of the variability of the mixing frequency distribution estimates described above.

1. Introduction

Some of the most useful tools in authorship attribution studies and in ecology rely on the analysis of word or species frequency count data. In the first case for example, texts are treated as samples from the vocabulary of their author and the word frequency counts in them are used to learn about his style and, in particular, about the size, evenness and diversity of his vocabulary, which might help distinguish his style from the style of other authors.

There has been a long lasting debate on which statistical models are most useful for word or species frequency count data, which has led experts to consider a large number of alternative models. Given that most words (species) appear very few times and very few words (species) are repeated many times, word and species frequency count data typically have reverse J-shaped distributions with long upper tails. Yule (1944) and Good (1953) conjectured that this skewness should be modelled through Poisson mixture models, which at the time was not well accepted by everyone. Ever since, the debate on which models best suit the analysis of that type of data has been posed mainly in terms of which models best fit them.

The *first goal* of the paper is to argue that what make Poisson mixture models truly special and useful is that they provide a simple mechanistic explanation that lets one interpret the model mixing distribution as the distribution of the word or species frequencies of the vocabulary or population from which the sample was created. That interpretation is lacking in all the many purely empirical motivated models considered for this kind of data. The word frequency distribution of the vocabulary of an author characterizes his style, and the species frequency distribution of an ecosystem characterizes its population and they determine the size, evenness and diversity of that vocabulary or population. Hence the importance of identifying Poisson mixture models that fit word or species frequency count data well, and yield estimates of the word or species frequency distribution that can be used as fingerprints of the style of an author in his texts and of the population of an ecosystem in its samples.

[*] Corresponding author. Tel.: +34 934011728; fax: +34 934016575.
E-mail address: josep.ginebra@upc.edu (J. Ginebra).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.04.001

To rank populations in terms of their evenness or diversity one needs real valued measures that capture the one aspect of evenness or diversity that one cares the most and induce a total ordering in the space of populations. Nevertheless, among practitioners there is a pervading feeling that the concepts of evenness and of diversity within a population cannot be precisely distinguished and characterized. The *second goal* of the paper is to propose definitions of measure of the evenness and of measure of the diversity within a population that can be used to compare populations with different total number of classes, which is the setting found most often when assessing evenness and diversity of vocabulary in stylometry, and of the population of an ecosystem in ecology.

An additional source of confusion is that the sample and the population versions of evenness and diversity measures are not always clearly distinguished. Typically the sample version is not a good estimate of the population version because it is biased with a bias that can be large. In stylometry and in ecology the estimation of diversity measures is made even more difficult due to the total number of words in the vocabulary or classes in the population being unknown. The *third goal* of the paper is to describe how one can estimate measures of the evenness and of the diversity of a population through the expectation of these measures under the frequency distribution estimates proposed early on in the paper. This leads to the assessment of the lack of diversity through measures of the variability of these model mixing distribution estimates.

The paper focuses on the analysis of word frequency counts even though it all extends to species frequency counts, and it is organized as follows. Section 2 describes word frequency count data, it motivates the use of zero truncated Poisson mixture models on them and it illustrates how both the zero truncated generalized inverse Gaussian–Poisson model as well as the zero truncated Tweedie–Poisson model provide excellent fits for the word frequency counts of very long texts. It also illustrates the use of the maximum likelihood estimate of the generalized inverse Gaussian and of the Tweedie model mixing densities as estimates of the density of the word frequency of the vocabulary of the author.

Section 3 proposes definitions of measure of evenness and of measure of diversity, and it explains the novelty of these two definitions when the total number of classes in the populations (words in the vocabularies) are unknown but known to be different among them. Readers not specially interested in foundational issues may want to skip the first part of this section.

Section 4 proposes approximating measures of the evenness and of the diversity through their expectation under the word or species frequency distribution of the population. This combined with the main argument in Section 2 links lack of diversity with the ratio between a measure of the variability of the model mixing distribution of truncated Poisson mixture models, and the expected value of that mixing distribution. As an illustration, the size and various measures of the diversity of the vocabulary behind seven texts are estimated twice, using both the truncated GIG-Poisson as well as the truncated Tweedie–Poisson mixture models and the estimates are compared with estimators in the literature; the proposed estimation approach seems to be very robust to the choice of model mixing distribution. The behavior of various diversity measures when word frequencies have generalized inverse Gaussian distributions is also explored.

2. Poisson mixture models and density of vocabulary

2.1. Vocabulary distribution and word frequency count data

To characterize the style of an author through his vocabulary, as in Holmes (1985), the basic assumption is that the author has available a list of all the words that he knows, and that the ith word in that list is characterized through the proportion of times that that word would be found in a text of infinite length by that author, which is denoted by π_i. The set of probabilities π_j when j ranges over all the v words known by an author, $(\pi_1, \dots, \pi_v)$, with $\sum_j^v \pi_j = 1$, constitute the probability function of the vocabulary of that author. By identifying v with the total number of words in the vocabulary one is assuming that $\pi_i > 0$ for $i = 1, \dots, v$.

For mathematical convenience one treats word frequencies, π_j, as realizations of a continuous random variable with a density function, $\psi(\pi)$. Note that the larger the number of words in a vocabulary, v, the smaller the π_j's, and the closer the probability mass of $\psi(\pi)$ is to 0, which links a small expectation of $\psi(\pi)$ with a *rich* vocabulary. Furthermore given v, the closer the vocabulary distribution $(\pi_1, \dots, \pi_v)$ is to the uniform distribution $(1/v, \dots, 1/v)$, the more peaked $\psi(\pi)$ is around $1/v$, and the more evenly represented are the words in texts from that vocabulary, which links variability of $\psi(\pi)$ with lack of *evenness* and of *diversity* of vocabulary.

Texts written by an author are treated as if they were random samples drawn from his vocabulary. If one denotes the total number of words (tokens) in a given text by n, the number of occurrences of the ith word by $n_{i:n}$, and the proportion of occurrences of that word in that text by $\hat{\pi}_{i:n} = n_{i:n}/n$, the expected value of $\hat{\pi}_{i:n}$ is π_i. Let v_n denote the number of different words (types) observed in that text, and let $v_{r:n}$ denote the number of different words appearing exactly r times in it. The proportion of different words appearing exactly r times in a text of size n is $\hat{p}_{r:n} = v_{r:n}/v_n$ and its expectation, which depends on n, will be denoted by $p_{r:n}$.

In a given text most words appear only a few times and only a few words are repeated many times, and the distribution of $(v_{1:n}, v_{2:n}, \dots, v_{n:n})$ is reverse J-shaped with an extraordinarily long upper tail. Table 1 presents the word frequency count of texts that will be used later on. Max Havelaar for example has a total of $n = 99\,767$ words out of which there are $v_n = 11\,161$ different words, 6004 words appear once, 1731 words appear twice and so on, with the most frequent word appearing a total of 4826 times.