Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species

Anne Chao\(^1\)*, Y. T. Wang\(^1\) and Lou Jost\(^2\)

\(^1\)Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan 30043; and \(^2\)EcoMinga Foundation, Via a Runtun, Banos, Tungurahua, Ecuador

Summary

1. Estimating Shannon entropy and its exponential from incomplete samples is a central objective of many research fields. However, empirical estimates of Shannon entropy and its exponential depend strongly on sample size and typically exhibit substantial bias. This work uses a novel method to obtain an accurate, low-bias analytic estimator of entropy, based on species frequency counts. Our estimator does not require prior knowledge of the number of species.

2. We show that there is a close relationship between Shannon entropy and the species accumulation curve, which depicts the cumulative number of observed species as a function of sample size. We reformulate entropy in terms of the expected discovery rates of new species with respect to sample size, that is, the successive slopes of the species accumulation curve. Our estimator is obtained by applying slope estimators derived from an improved Good-Turing frequency formula. Our method is also applied to estimate mutual information.

3. Extensive simulations from theoretical models and real surveys show that if sample size is not unreasonably small, the resulting entropy estimator is nearly unbiased. Our estimator generally outperforms previous methods in terms of bias and accuracy (low mean squared error) especially when species richness is large and there is a large fraction of undetected species in samples.

4. We discuss the extension of our approach to estimate Shannon entropy for multiple incidence data. The use of our estimator in constructing an integrated rarefaction and extrapolation curve of entropy (or mutual information) as a function of sample size or sample coverage (an aspect of sample completeness) is also discussed.

Key-words: diversity, Good-Turing frequency formula, mutual information, sample coverage, Shannon entropy, species accumulation curve, species discovery rate

Introduction

Shannon entropy and its exponential have been used to characterize class uncertainty and diversity in ecology, genetics, information theory, computer science, cryptography, physics, neuroscience, linguistics, bioinformatics and other fields (Shannon 1948; Lewontin 1972; Pielou 1977; Paninski 2003; Magurran 2004; Victor 2006; Sherwin 2010 among others). Shannon entropy of an assemblage is

\[
H = - \sum_{i=1}^{S} p_i \log p_i,
\]

where \(S\) is the number of species, and \(p_i\) is the relative frequency of the \(i\)-th species. While the ‘species’ may be biological species, words of a book, genes or alleles, or other discrete entities, in this paper, we concentrate on ecological applications to a community with \(S\) species. We assume that the true species richness \(S\) is unknown. Though we use biological terminology, our framework applies to all fields.

Shannon entropy was adopted by ecologists as a diversity measure in the 1950s (MacArthur 1955, 1965; Margalef 1958); see Ulanowicz (2001) for the early history of information theory in ecology. In most ecological applications, Shannon entropy can be interpreted as a measure of the uncertainty in the species identity of an individual randomly selected from a community. A higher degree of uncertainty implies greater diversity. Many other measures have also been proposed to characterize the diversity of a community (Magurran 2004). Some of them emphasize more abundant species (e.g. the Gini-Simpson index), while others ignore or under-emphasize abundance information (e.g. species richness). Shannon entropy and its monotonic transformations, such as \(\exp(H)\) (i.e. ‘the effective number of species’ or diversity of order one; see Jost 2007), are the only standard frequency-sensitive complexity measures that weigh species in proportion to their population abundances. This is not only the most natural weighing for many applications, but it facilitates hierarchical decomposition.

Shannon entropy and its transformations connect directly to the rich predictive mathematics of entropy and information theory (Shannon 1948 and thousands of articles since then). Shannon entropy is conveniently additive, which further facilitates hierarchical decomposition. The applications of Shannon entropy and related measures to ecology and evolution include the predictions for species-level diversity (Beck...
& Schwanghart 2010; Marcon et al. 2012), description of processes involving the information-carrying molecule in evolution (Sherwin et al. 2006; Sherwin 2010), development of the maximum entropy theory (Jaynes 2003), quantifying beta diversity and differentiation among communities (Jost 2007; Pfeiffer & Mezger 2012; Chiu, Jost & Chao 2013), and hierarchical diversity analysis between samples, plots, forests and regions (Marcon et al. 2012), enabling ecologists to understand the factors shaping the distribution of biodiversity.

The estimation of this seemingly simple function is surprisingly non-trivial when sample size is fixed and the number of species is unknown (May 1975). The observed entropy of the sample, which uses the sample fraction \( p_i = X_i/n \) in place of \( p_i \) in the formula for entropy (\( X_i \) denotes the sample species frequency of the \( i \)-th species in the sample of size \( n \)), is severely negatively biased (Basharin 1959) when there are undetected species in the sample. Blyth (1959) proved that an unbiased estimator does not exist for samples of fixed sizes. Therefore, the observed entropy of the sample, and its exponential, may severely underestimate the true entropy and diversity of the community (Peet 1974; Magurran 2004, p. 107). Also, when comparing multiple communities, the estimate of the differences in diversities of these communities may be seriously biased (Chao & Shen 2003).

A wide range of estimation methods have been proposed in various disciplines to obtain a reliable entropy estimator with less bias than that of the observed sample entropy. The diversity of the approaches reflects the wide range of applications and the importance of bias-reduction. We summarize them briefly here; more details and formulas along with more references are provided in Appendix S1 of Supporting Information.

(1) Traditional bias-reduction method: A direct approach to reduce bias is to evaluate the first-order bias of the observed entropy and remove it to obtain a bias-corrected estimator. However, the first-order bias is a function of species richness (see Appendix S1), which is more difficult to estimate than entropy itself (Chao 2005; O’Hara 2005; Gotelli & Chao 2013). Chao & Shen (2003) showed that using an estimated \( S \) in the bias-corrected formula results in increased variance and low accuracy.

(2) The jackknife estimator: The jackknife technique was developed as a statistical method to reduce the bias of a biased estimator. Here, the biased estimator is the observed entropy. The basic idea of the first-order jackknife method is to consider subdata by successively deleting one individual from the data. The resulting entropy estimator is commonly used (Zahl 1977).

(3) Coverage-adjusted method: In statistics, Chao & Shen (2003) combined the Horvitz & Thompson (1952) adjustment for undetected species and the concept of sample coverage (see below) to obtain a low-bias entropy estimator, which is referred to as the ‘coverage-adjusted estimator’ in Vu, Yu & Kass (2007); see Beck & Schwanghart (2010) for an updated comparison of this estimator with some other approaches.

(4) Estimator via Renyi’s entropies: In physics, there are several entropy estimators (see references in Appendix S1), but the most widely used is the Grassberger (2003) estimator. Since Shannon entropy is the limit of Renyi’s entropies when the order tends to unity, Grassberger derived an entropy estimator via estimation of Renyi’s entropies.

(5) Estimation via an expansion formula for entropy: Although Blyth (1959) in his pioneering work derived Taylor’s expansion of entropy and showed that the expansion could be used to construct a low-bias estimator, major contributions in this direction appeared only recently. Zhang (2012) and Vinck et al. (2012) each made advances. Their approach is to express entropy as a polynomial representation (Vinck et al. 2012) or a function of ‘generalized Simpson indices’ (Zhang 2012). Then, an entropy estimator is obtained by estimators of polynomials or generalized Simpson indices (details are provided in Appendix S1).

(6) Bayesian approach and others: In neuroscience, Nemenman, Shafee & Bialek (2002) proposed a Bayesian estimator, which has been referred to as the NSB estimator in the literature. Paninski (2003) proposed a ‘best upper bound’ estimator which minimizes an upper bound on the mean squared error. However, satisfactory performance of the two approaches requires knowledge of \( S \) and a prior distribution for species abundances (Zhang 2012). This information is usually available for neural data, but not in data from other sciences. Hauser & Strimmer (2009) derived a James-Stein type of entropy estimator, but they also require \( S \) to be known.

In this paper, we propose a novel estimation method that links entropy to the species accumulation curve. We first show that entropy is a simple function of the successive slopes of the species accumulation curve, which plots the cumulative number of species as a function of sample size. By applying slope estimators derived from an improved Good-Turing frequency formula (Chao & Jost 2012), we obtain a new accurate estimator of entropy. The performance of the proposed estimator is examined by extensive simulations from theoretical models and real surveys. We also compare the proposed estimator with some previous commonly used estimators in terms of bias and accuracy (low mean squared error). Extension to multiple incidence data and to estimate mutual information is discussed. The use of our proposed estimator to construct an ‘entropy (or mutual information) accumulation curve’ and other related issues are also discussed.

Materials and methods

ENTROPY AND THE SPECIES ACCUMULATION CURVE

We first link entropy to the species accumulation curve. Let \( S(k) \) be the expected number of species in a random sample of size \( k \) taken (with replacement) from a community with \( S \) species and relative abundances \( \{p_1, p_2, \ldots, p_S\} \). The theoretical expression for \( S(k) \) as a function of sample size \( k \) is (Good 1953):

\[
S(k) = \sum_{i=1}^{S} [1 - (1 - p_i)^k] = S - \sum_{i=1}^{S} (1 - p_i)^k. \tag{eqn 1}
\]

The species accumulation curve plots the expected sample richness \( S(k) \) versus sample size \( k \). The horizontal asymptote of this curve as \( k \) tends
to infinity is the species richness. The slope of the line connecting two adjacent points \((k, S(k))\) and \((k + 1, S(k + 1))\) is the expected change along the \(y\)-axis, \(\Delta(k) = S(k + 1) - S(k)\), divided by the corresponding change in the \(x\)-axis (which is an increment of one individual). From eqn 1, this slope is

\[
\Delta(k) = S(k + 1) - S(k) = \sum_{i=1}^{k} p_i (1 - p_i)^2. \tag{eqn 2}
\]

The slope \(\Delta(k)\) is the probability that the \((k + 1)\)-th individual represents a species that was missed in the previous sample of size \(k\) (Olszewski 2004). Thus, the curve’s successive slopes show the rates at which new species are detected in the sampling process. The slope \(\Delta(k)\) is a decreasing function of \(k\), implying that the rate declines as sample size is increased. This slope is also the expected ‘coverage deficit’, the complement of sample coverage (Chao & Jost 2012). Sample coverage, originally developed by Alan Turing (Good 1953, 2000), is the fraction of the population belonging to species represented in the sample; it is an objective measure of sample completeness. The coverage and its deficit are the aspects of undetected species that can be accurately estimated by sample data (Good 1953; Lande, DeVries & Walla 2000).

The link between the slopes \(\Delta(k)\) and the entropy is based on the following well-known expression for entropy using an expansion for the logarithm function:

\[
H = -\sum_{i=1}^{\infty} p_i \log p_i - \sum_{i=1}^{\infty} p_i \log[1 - (1 - p_i)] = \sum_{k=1}^{\infty} \frac{1}{k} \sum_{i=1}^{k} p_i (1 - p_i)^2. \tag{eqn 3}
\]

Our new perspective arises by noting that, by combining eqn 2 and eqn 3, the entropy can be expressed as the following infinite sum of functions of the successive slopes:

\[
H = \sum_{k=1}^{\infty} \frac{1}{k} \Delta(k). \tag{eqn 4a}
\]

NEW ENTROPY ESTIMATOR

The new formulation in eqn 4a opens a novel way to estimate entropy via discovery rates of new species. Since the expected slopes of a species accumulation curve can be accurately estimated when sample size is sufficiently large (Chao & Jost 2012), a very accurate entropy estimator can be obtained via estimation of these slopes. Based on a sample of fixed size \(n\), we separate the infinite sum into two parts: the first part with \(k < n\) and the second part with \(k \geq n\):

\[
H = \sum_{k=1}^{n-1} \frac{1}{k} \Delta(k) + \sum_{m=n}^{\infty} \frac{1}{m + m} \Delta(n + m). \tag{eqn 4b}
\]

The reason for this separation is because the first term can be estimated without bias, whereas there exists no unbiased estimator for the second term. Under the commonly used multinomial model, in which species frequencies \((X_1, X_2, \ldots, X_m)\) follow a multinomial distribution with cell total \(\sum_{i=1}^{m} x_i = n\) and cell probabilities \(\{p_1, p_2, \ldots, p_m\}\), we separately estimate each term in eqn 4b.

For the first sum, it is known from statistical theory for the multinomial distribution that the minimum variance unbiased estimator for the expected slope \(\Delta(k)\) exists for any size \(k < n\). This unbiased estimator for the first part is (Chao & Jost 2012)

\[
\hat{\Delta}(k) = \sum_{i=1}^{k} \frac{X_i}{n} \left( \frac{n - X_i}{k} \right), \quad k < n. \tag{eqn 5}
\]

The second sum in eqn 4b involves the expected slopes for sample sizes greater than \(n\) and no unbiased estimator exists. This part is usually dominated by rare undetected species whose effect on entropy cannot be ignored, especially when sample size is much smaller than the number of species. Thus, the burden of entropy estimation is shifted onto this second sum. Our new slope formulation provides a novel method for estimating this challenging sum. Chao & Jost (2012) derived slope estimators for successive slopes (see below) based on the wisdom of Turing and Good (Good 1953, 2000), who showed that the number of singletons and doubletons carries much information about the number of undetected rare species.

Let \(f_1\) denote the number of singletons and \(f_2\) denote the number of doubletons in the sample. Good-Turing’s original frequency formula implies that the estimated mean relative frequency of the singletons in the population is not \(1/n\), but \(2f_2/(n f_1)\), contrary to most people’s intuition. When \(n\) is sufficiently large, Chao & Jost (2012) derived a more accurate modification of Good-Turing’s frequency formula for singletons; we refer to this estimated mean relative frequency as \(A\) with the following expression:

\[
A = \begin{cases} 
\frac{2f_2/(n - 1)f_1 + 2f_2}{1} & \text{if } f_2 > 0 \\
\frac{2}{(n - 1)(f_1 - 1) + 2} & \text{if } f_2 = 0, \ f_1 \neq 0 \\
\frac{0}{1} & \text{if } f_2 = f_1 = 0.
\end{cases}
\]

Based on these formulas for \(A\), an estimator for the slope at sample size \(n + m\) is (Chao & Jost 2012)

\[
\hat{\Delta}(n + m) = \frac{f_1}{n} (1 - A)^{m+1}, \quad m \geq 0. \tag{eqn 6}
\]

Combining eqn 5 and eqn 6, we obtain an entropy estimator:

\[
\hat{H} = \sum_{k=1}^{n-1} \frac{1}{k} \sum_{1 \leq k \leq n - k} \frac{X_k}{n} \left( \frac{n - X_k}{k} \right) + \frac{f_1}{n} \sum_{m=0}^{\infty} \frac{1}{n + m} (1 - A)^{m+1}. \tag{eqn 7}
\]

Let \(\psi(x)\) be the digamma function. We can simplify each term to obtain our proposed entropy estimator (see Appendix S1):

\[
\hat{H} = \sum_{1 \leq k \leq n - 1} \frac{X_k}{n} \left[ \psi(n) - \psi(X_k) \right] - \frac{1}{n} \left[ (1 - A)^{-1 \psi} \left( -\log(A) - \sum_{k=1}^{n-1} \frac{1}{k} (1 - A)^{k-1} \right) \right]. \tag{eqn 7}
\]

The variance of the proposed estimator can be estimated by bootstrapping (details are provided in Appendix S2). The resulting variance estimate can then be used to construct a confidence interval for entropy to reflect sampling uncertainty.

In diversity analysis, we recommend using \(\exp(\hat{H})\) and the same bootstrapping method to make inferences for \(\exp(H)\), the exponential of entropy. Similar approach can be applied to estimate evenness measures (Ricotta 2003; Jost 2010), and other entropy-like measures.
NEW ESTIMATOR OF MUTUAL INFORMATION

The proposed method can also be applied to estimate the mutual information between any two discrete random variables Y and Z. This is an important issue especially in ecology, genetics, information science and neuroscience. In the context of ecology, mutual information is closely related to beta diversity and entropy-based measures of differentiation among communities, as will be explained after the definition of mutual information.

Let \( P_Y, P_Z(y, z) \) be the true relative frequencies for each possible outcome \((y, z)\) of \((Y, Z)\), where \((y, z) \in D\) and \(D\) is the set containing all possible discrete outcomes. The marginal frequencies of \(Y\) and \(Z\) are, respectively, \(P_Y(y) = \sum_{z \in Z} P_Y(y, z)\) and \(P_Z(z) = \sum_{y \in Y} P_Y(y, z)\). The mutual information between the two variables \(Y\) and \(Z\) is defined as

\[
I(Y; Z) = \sum_{y, z} P_Y(y, z) \log \frac{P_Y(y, z)}{P_Y(y) P_Z(z)}
\]

It measures the Kullback–Leibler distance from the joint probability density to the product of the two marginal probability densities. The mutual information can be expressed as a function of three entropies: \(H(Y) + H(Z) - H(Y; Z)\), eqn 8

where \(H(Y)\), \(H(Z)\) and \(H(Y; Z)\) are entropies with respect to the frequencies \(P_Y(y)\), \(P_Z(z)\) and \(P_Y(y, z)\).

Intuitively, mutual information quantifies how much information the value of one variable \(Y\) gives about the value of the other variable \(Z\) (e.g. see Sherwin 2010). Consider a region with \(S\) species and \(K\) communities. Let \(P_Y(y)\) denote the species relative abundance for species \(y\) in the combined community, \(P_Z(z)\) denotes the relative size of community \(z\), and \(P_Y(y, z)\) denotes the product of the relative size of community \(z\) and the relative abundance of species \(y\) within community \(z\). Then, \(H(Y)\) becomes gamma entropy, that is, entropy calculated from the combined community. Also, \(H(Y; Z) - H(Z)\) becomes alpha entropy, that is, the weighted (by community size) average of entropies of individual communities. It follows from eqn 8 that the mutual information \(I(Y; Z)\) becomes the difference of gamma and alpha entropies.

When communities are equally weighted, \(I(Y; Z)\) reduces to the beta diversity based on additive partitioning of gamma entropy (Jost 2007). In this case, the mutual information ranges from 0 to log \(K\), which varies with the number of communities \(K\). So the degree of differentiation between communities cannot be compared between two regions with different numbers of communities. To solve this and facilitate comparisons, beta diversity can be transformed to the interval \([0, 1]\) by normalization. The normalized mutual information \(I(Y; Z)/(\log K)\) measures the differentiation or dissimilarity among the \(K\) communities. For the special case of \(K = 2\), the normalized mutual information reduces to the Horn (1966) heterogeneity index. For a general value of \(K\), it reduces to the complement of the ‘true overlap’ measure developed in Chao et al. (2008). Because beta diversity, similarity and differentiation measures are important tools for comparing species composition along environmental gradients, the estimation of mutual information is essential in biodiversity studies. See Hauser & Strimmer (2009), Marcon et al. (2012) and Peiffer & Mezger (2012) for some biological applications.

Suppose \(N\) pairs \((y, z)\) are obtained by random sampling; we denote the set of sample frequencies as \(\{N_{yz}; \ (y, z) \in D\}\), where \(\sum_z N_{yz} = N\). Applying eqn 7, we can directly obtain the two marginal entropy estimators \(\hat{H}(Y)\) and \(\hat{H}(Z)\) based on \(\{N_{yz}; N_{y+} = \sum_z N_{yz}\}\) and \(\{N_{yz}; N_{z+} = \sum_y N_{yz}\}\), respectively.

For estimating \(\hat{H}(Y; Z)\), let \(F_1\) and \(F_2\) represent the number of singletons and doubletons in the set of frequencies \(\{N_{yz}; \ (y, z) \in D\}\). For \(F_2 > 0\), define \(B = 2F_2[(N - 1)F_1 + 2F_2]\) (with similar modifications as those for \(A\) in abundance data).

\[
\hat{H}(Y; Z) = \sum_{k=1}^{N-1} \sum_{k' = k}^{N-1} \frac{N_{yz}((N - N_{yz})}{N} \log \left( \frac{N_{yz}}{N_{yz} - N_{yz}} \right) + \frac{F_1}{N} \sum_{m=0}^{\infty} \frac{1}{N + m} (1 - B)^m.
\]

We can further simplify each term as follows:

\[
\hat{H}(Y; Z) = \sum_{k=1}^{N-1} \sum_{k' = k}^{N-1} \frac{N_{yz}((N - N_{yz})}{N} \log \left( \frac{N_{yz}}{N_{yz} - N_{yz}} \right) + \frac{F_1}{N} \left(1 - B\right)^{N-1} \left(-\log(B) - \sum_{m=1}^{\infty} \frac{1}{m} (1 - B)\right).
\]

Therefore, a very accurate mutual information estimator can be obtained by \(\hat{H}(Y) + \hat{H}(Z) - \hat{H}(Y; Z)\).

SIMULATION COMPARISONS OF ENTROPY ESTIMATORS

To examine the performance of the proposed entropy estimator, we carried out extensive simulations by generating data sets from several types of abundance models and comparing the proposed estimator with previous analytic estimators. We also did parallel simulations treating several large survey data sets as the ‘true’ entire assemblages. Simulations regarding the mutual information will be reported elsewhere.

Typically, in estimation, there is a trade-off between bias and variance. Any reduction in bias often leads to increase of variance. A measure of accuracy that takes into account bias and variance is mean square error (MSE, squared bias plus variance) or its root (RMSE). Our evaluation criteria are thus based on both bias and RMSE.

(A) DATA GENERATED FROM THEORETICAL ABUNDANCE MODELS

The number of species was fixed to be 100, 500 or 1000. Several species abundance models for \(p_1, p_2, \ldots, p_s\) were considered and are given below, where \(c\) in all cases is a normalizing constant such that \(\sum_s p_s = 1\). For each fixed model, we considered a range of sample sizes \((s = 25–500)\) in an increment of 25 if \(S < 500\), and \(n = 50–1000\) in an increment of 50 if \(S \geq 500\). For each combination of abundance model and sample size, 5000 simulated data sets were generated from the model. In each case, if \(p_1, p_2, \ldots, p_s\) are fixed parameters, then we give the true value of the entropy and the CV (coefficient of variation, which is the ratio of standard deviation and mean) of \(p_1, p_2, \ldots, p_s\).

The CV value quantifies the degree of heterogeneity of the probabilities \(p_1, p_2, \ldots, p_s\). When all probabilities are equal, CV of a normal distribution (e.g. log-normal or Dirichlet distribution as described below), then the average values of entropy and CV over 5000 simulated data sets are given to approximate the true theoretical values. Simulations are reported here from the following models:

A homogeneous model: \( p_i = 1/S, \ i = 1, 2, \ldots, S \). This is the model with no heterogeneity among species abundances. Three cases of \( S = (100, 500, 1000) \) corresponding to \( H = (4.605, 6.215, 6.908) \) are considered. CV = 0 for all three cases. The plot of \( S = 1000 \) is shown in Fig. 1(a) in order to examine the performance of our estimator in severely under-sampled cases; this is the situation where previous methods fail to estimate entropy well. Note in Fig. 1, sample sizes are less than the species richness so that there is a large fraction of undetected species in samples. The cases for \( S = 100 \) and 500 are shown in Figs. S3-1(a) and S3-1(b) of Appendix S3.

The log-normal model: \( p_i = c a_i, \ i = 1, 2, \ldots, S \), where \( (a_1, a_2, \ldots, a_S) \) are a random sample from a log-normal distribution with parameters \( \mu = 0, \sigma^2 = 1 \). The log-normal model is widely used for modelling species abundance distributions (Magurran 2004). For the three cases of \( S = (100, 500, 1000) \), the average values of \( H \) and CV over the 5000 generated data sets are \( H = (4.125, 5.722, 6.410) \) and \( CV = (1.215, 1.273, 1.296) \), respectively. The plot of \( S = 1000 \) is shown in Fig. 1(b). The cases for \( S = 100 \) and 500 are shown in Figs. S3-2(a) and S3-2(b) of Appendix S3.

The Zipf–Mandelbrot model (Zipf 1965; Mandelbrot 1977; Magurran 2004): \( p_i = c/i, \ i = 1, 2, \ldots, S \). This is a common model in

---

Fig. 1. Comparison of the average bias (left panels) and average root mean square error (RMSE, right panels) of entropy estimators over 5000 data sets for fixed sample size (lower horizontal axis) and the corresponding sample coverage (upper horizontal axis). A good estimator should closely approach the bias = 0 line and also closely approach the RMSE = 0 line. (a) Upper panels: data sets were generated from a homogeneous model with \( S = 1000 \) equally rare species (\( p_i = 0.001, \ i = 1, 2, \ldots, 1000 \)). The coefficient of variation (CV) of species abundances is CV = 0, and the true entropy is \( H = 6.908 \). (b) Lower panels: data sets were generated from a log-normal model with \( S = 1000 \) (\( p_i = c a_i, \ i = 1, 2, \ldots, 1000 \)), where \( (a_1, a_2, \ldots, a_{1000}) \) are a random sample from a log-normal distribution with parameters \( \mu = 0, \sigma^2 = 1 \). The average values of CV and H over 5000 simulated data sets are CV=6.410 and H = 1.296, respectively.
both ecology and linguistics. Three cases of \( S = (100, 500, 1000) \) corresponding to \( H = (3.681, 4.749, 5.191) \) and \( CV = (2.264, 4.103, 5.326) \) are considered. The plots for \( S = 100, 500 \) and 1000 are shown in Figs. S3-3(a-c) of Appendix S3.

4. The broken-stick model: \( p_i = c a_i \) where \( (a_1, a_2, ..., a_n) \) are a random sample from an exponential distribution. Or equivalently, \( (p_1, p_2, ..., p_n) \) is a Dirichlet distribution with parameter 1. This is MacArthur’s well-known broken-stick model (MacArthur 1957). For the three cases of \( S = (100, 500, 1000) \), the average values of \( H \) over the 5000 simulated data sets are \( H = (4.188, 5.791, 6.485) \). The average values of \( CV \) for the three cases are all very close to the theoretical value of 1.00. The plots are shown in Figs S3-4(a-c) of Appendix S3.

The following three models represent cases in which communities are dominated by a few extremely abundant species:

5. Power-decay model: \( p_i = c i^{\beta}, i = 1, 2, ..., n, CV = 6.316 \) and \( H = 1.570 \). The plot is given in Fig. S3-5 of Appendix S3.

6. Poisson model: \( p_i = c e^{-\lambda} / \lambda ^{i}, \lambda = c, i = 0, 1, ..., 99, CV = 4.089 \) and \( H = 1.877 \). The plot is given in Fig. S3-6 of Appendix S3.

7. Exponential-decay model: \( p_i = c e^{-\delta i}, i = 1, 2, ..., 100, CV = 6.758 \) and \( H = 1.041 \). The plot is given in Fig. S3-7 of Appendix S3.

For each generated data set, we compare the bias and RMSE of our entropy estimator with the observed entropy, the Grassberger (2003) estimator, the Zahl (1977) jackknife estimator, the Chao & Shen (2003) coverage-adjusted estimator and the Zhang (2012) \( H_Z \) estimator. See Appendix S1 for the formulas and details of these estimators. Our comparisons focus on estimators that do not require any specific prior information or hypotheses about the number of species or their relative abundances. Thus, Bayes estimators which generally require information of the number of species were not considered in our study.

In all figures of the main text and Appendix S3, the average bias (the left panel of each figure) and RMSE (the right panel of each figure) over the 5000 simulation trials are shown for the proposed estimator and the above five selected estimators as a function of sample size \( n \) in the lower horizontal axis and of sample coverage (i.e. \( 1 - \Delta(n) \) in eqn 6) in the upper horizontal axis. In each figure, we show the bias = 0 line and RMSE = 0 line. A good estimator should closely approach the bias = 0 line and also closely approach the RMSE = 0 line.

(B) USING LARGE SURVEYS AS THE TRUE COMMUNITIES

We used several large survey data sets as the ‘true’ communities and generated data sets from these communities. For example, in the simulation for Fig. 2(a), we treated the 150-year field observations of Miller & Wiegert (1989) for endangered and rare vascular plant species in the central portion of the southern Appalachian region as the true entire assemblage. The species abundance distribution for this survey is reproduced in Table S4-1 (Appendix S4); a total of 188 species were represented by 1008 individuals. The observed relative abundances are treated as the true relative abundances so that the entropy calculated from the frequencies is regarded as the true entropy. Let \( f_k \) be the number of species that have \( k \) representatives in the sample. The observed abundance frequency counts are as follows: \( f_1 = 61, f_2 = 35, f_3 = 18, ..., f_{60} = 1 \). The assemblage from which data were generated thus includes 188 species, 1008 individuals, and there are 61 species with a relative abundance of 1/1008, 35 species with a relative abundance of 2/1008, 18 species with a relative abundance of 3/1008, etc., and one species with a relative abundance of 67/1008. The data show a relatively high degree of heterogeneity among species abundances as reflected by \( CV = 1.563 \), and the true entropy is \( H = 4.574 \). For any fixed sample size \( n \) (between 25 and 500 in an increment of 25), we generated a sample of \( n \) individuals by sampling with replacement from this true assemblage. The average bias and RMSE over 5000 generated data sets as a function of sample size (from 25 to 500 in the lower \( x \)-axis) and of sample coverage (roughly from 40% to 90% in the upper \( x \)-axis) are shown in Fig. 2(a).

The simulations for Fig. 2(b) were based on a well-known Malayan butterfly survey (Fisher, Corbet & Williams 1943). There were S = 620 species represented by 9031 individuals. The species frequency distribution for this survey is given in Table S4-2 (Appendix S4). For these data, we have \( CV = 1.435 \) and \( H = 5.736 \). The patterns for average bias and average RMSE as a function of sample size (from 100 to 1000) and of sample coverage (roughly from 30% to 90%) are shown in Fig. 2(b). All the simulation procedures are the same as those above using abundance models. Three additional plots for other surveys are shown in Figs. S3-8–S3-10 in Appendix S3.

Results

Our discussion is based on the figures in the main text and all figures in Appendix S3. As indicated by many other researchers (e.g. Magurran 2004, p. 107), the observed entropy seriously underestimates in all cases and has the largest RMSE among all estimators. Thus, it is excluded in our following comparisons. We outline some specific results as follows:

1. When all species abundances are homogeneous (Fig. 1(a) and Fig. S3-1(a,b) in Appendix S3), the previous estimators either overestimate (for Chao and Shen estimator) or severely underestimate (for Grassberger, jackknife and Zhang’s \( H_Z \) estimators). Our estimator almost completely removes the bias. It quickly converges to the true entropy and has the smallest bias and RMSE.

2. In Figs 1 and 2, we specifically show the performance of various estimators for severely under-sampled cases. Such cases typically arise when sample sizes are small relative to species richness (Fig. 1). Under-sampling also arises when samples are taken from assemblages with a large fraction of rare species (Fig. 2). In severely under-sampled regimes, if the community is not dominated by a few extremely abundant species (such as in Figs S3-5–S3-7 in Appendix S3), our estimator is always superior to the existing methods and is the clear choice in terms of both bias and RMSE. Notice that even in the cases where a few species are extremely dominant, our estimator is inferior only for initial sample sizes, and the differences between our estimator and the best estimator are slight in bias and negligible in RMSE. In Appendix S3, the following cases with large species richness also represent under-sampling for the sample sizes considered in this study: Fig. S3-3(c) with \( S = 1000 \), Fig. S3-4(c) with \( S = 1000 \) and Fig. S3-9 with \( S = 1825 \). We have also tried many other under-sampling scenarios, and all show consistent results.

3. To compare the relative merits of all estimators, we first describe two general patterns: (3a) The Grassberger, jackknife and Zhang’s \( H_Z \) estimators are biased downwards and the three estimators have similar trends. The magnitudes of the biases show a consistent pattern: Grassberger is less biased...
than the jackknife, which is less biased than the Zhang’s estimator. The RMSE shows the same pattern. Our estimator in nearly all cases is better than the Grassberger, jackknife and Zhang’s $H_Z$ estimators based on the criteria of bias and RMSE. (3b) The Chao & Shen (2003) estimator generally has smaller bias and RMSE than the Grassberger estimator. Consequently, if our new estimator is not considered, Chao & Shen (2003) estimator in most cases is also preferable to the Grassberger, jackknife and Zhang’s $H_Z$ estimators.

(4) Based on the results described in (3a) and (3b), we only need to compare the proposed estimator and the Chao & Shen (2003) estimator for the cases not covered in (1) and (2). (4a) Our estimator exhibits a predictable behaviour: it always steadily converges to the true entropy when sample size is increased. In a wide variety of simulation scenarios, the relative bias for our estimator over the sample-size range considered in our simulations is much lower than 5%. It is nearly unbiased when sample size is not unreasonably small. (4b) The bias

Fig. 2. Comparison of the average bias (left panels) and average root mean square error (RMSE, right panels) of entropy estimators over 5000 data sets for fixed sample size (lower horizontal axis) and the corresponding sample coverage (upper horizontal axis). A good estimator should closely approach the bias = 0 line and also closely approach the RMSE = 0 line. (a) Upper panels: data sets were generated from the vascular plant community based on a survey in the southern Appalachians (Miller & Wiegert 1989). In the data set, there are 188 species among 1008 individuals. CV = 1.563, $H = 4.574$. See Table S4.1 (Appendix S4) for frequency count data. (b) Lower panels: data sets were generated from the Malayan butterfly survey (Fisher, Corbet & Williams 1943), $S = 620$ species, 9031 individuals, CV = 1.435, $H = 5.736$. See Table S4.2 (Appendix S4) for frequency count data.
direction and convergence patterns of the Chao & Shen (2003) estimator are less predictable. It is biased downwards in under-sampling cases; it is biased upward in the exponential-decay model (e.g. Fig. S3-7); in some other cases (e.g. Figs. S3-3(a), S3-3(b), S3-5), it underestimates entropy when sample size is small, but then goes above the true entropy and overestimates. In these cases, there is a limited range of sample sizes (near crossing points) where Chao and Shen’s estimator is less biased than our proposed estimator. However, this only occurs in a narrow range in each case and within the narrow range the difference of biases between the two estimators is little, leading to almost identical RMSE. When sample size is increased, the proposed estimator converges to the true value more quickly than the Chao and Shen estimator.

In summary, our extensive simulation studies generally reveal that the proposed estimator substantially reduces bias when compared with previous estimators and does so without a noticeable increase in variance, so that it generally still exhibits the lowest RMSE. Based on the overall assessment, we suggest using our proposed new estimator in all practical applications.

Discussion

In this paper, we have formulated Shannon entropy as an infinite sum of the slopes of species accumulation curve. Plugging in the slope estimators derived in Chao & Jost (2012), we have obtained a nearly unbiased estimator when sample size is not unreasonably small. Extensive simulations have shown that our proposed estimator generally outperforms the existing methods and is always superior when there are many undetected species in samples. Our approach is also extended to estimate mutual information and generalized to deal with multiple incidence data. The R script for computing all entropy estimators that are considered in our study (i.e. estimators in Figs 1 and 2) will be available in the R CRAN packages (http://cran.r-project.org/web/packages/) and the first author’s website (http://chao.stat.nthu.edu.tw/).

Note that from a statistical perspective, we could further apply the jackknife technique to our proposed entropy estimator (and any other estimators considered in this paper) to reduce a portion of the remaining bias for small sample sizes, but the variance of the resulting jackknifed estimator is increased, causing a larger RMSE; see Appendix S5 for details. If the goal is to minimize the magnitude of bias, then the jackknifed version of our proposed entropy estimator is recommended. If the goal is to balance bias and variance, then the proposed entropy estimator (without jackknifing) is nearly optimal.

Entropy (or Mutual Information) Accumulation Curve

Like observed species richness, the observed entropy in a sample strongly depends on sample size and thus also on sample coverage. Therefore, by analogy with the species accumulation curve, we can construct an ‘entropy accumulation curve’ by plotting the expected entropy \( H(m) \) as a function of sample size \( m \) or the expected coverage of a sample of size \( m \). Given the data for a sample of size \( n \) (called the ‘reference sample’), what are the analytic estimators for \( H(m) \)? Rarefaction (interpolation) refers to the case \( m < n \), whereas prediction (extrapolation) refers to the case \( m > n \). The integrated sample-size- or coverage-based rarefaction/extrapolation sampling curve represents the estimated entropy accumulation curve based on the reference sample. For species richness, the integrated sample-size- and coverage-based rarefaction and extrapolation were developed respectively by Colwell et al. (2012) and Chao & Jost (2012). The extension to entropy rarefaction and extrapolation requires a reliable asymptotic estimator for Shannon entropy. Our proposed estimator can be applied to develop the corresponding theory for entropy rarefaction and extrapolation. The results along this direction will be reported elsewhere. We are currently working on the rarefaction and extrapolation curve for mutual information.

Multiple Incidence Data

Our derivation in this paper is based on samples in which individuals are taken randomly from communities. In many ecological field surveys, the sampling unit is not an individual, but a trap, net, quadrat, plot or timed survey. It is these sampling units, not the individuals that are sampled randomly and independently. Often, it is not always possible to count individuals within a sampling unit (e.g. in plant communities or microbial assemblages), so estimation is usually based on a set of sampling units in which only the incidence (detection or non-detection) of each species is recorded. We refer to this type of data as ‘multiple incidence data’. Shannon entropy and its sampling framework as we discussed for abundance data needs proper modification to treat this kind of data. In Appendix S6, we provide a detailed model formulation and show that our derivation for individual-based data can be extended to obtain an accurate estimator of entropy for multiple incidence data.

Link to an Unbiased Estimator

In this paper, we only consider a fixed sample size. As discussed, no unbiased estimator of entropy exists under this condition. However, if sample size is not fixed in advance and can be allowed to be very large so that all species can be discovered in the sample, then unbiased estimators do exist. One of these unbiased estimators was proposed by Montgomery-Smith & Schürmann (2007). Let \( U_i \) be the observed minimum sample size needed to discover the \( i \)-th species, and suppose, all \( U_i \)s are observable when sample size can be arbitrarily large. Their unbiased estimator is expressed as \( \sum_{i=1}^{\delta} (U_i - 1)^{-1} I(U_i \geq 2) \) where \( I(U_i \geq 2) = 1 \) if \( U_i \geq 2 \), and 0 otherwise. Here, the species found in the first observation (i.e. \( U_1 = 1 \)) does not contribute to this estimator. In our formulation of entropy, we have \( H = \sum_{k=1}^{\delta} k^{-1} \Delta(k) \), where \( \Delta(k) \) denotes the slope of the line connecting the two points \((k, S(k))\) and \((k + 1, S(k + 1))\) in the species accumulation curve. When \( U_i \) individuals are
needed to discover the i-th species for the first time, the accumulated number of species is increased by one species when size is increased from $U_i-1$ to $U_i$. That is, for $U_i \geq 2$, the slope at size $U_i - 1$ is one, that is, $\Delta U_i - 1 = 1$, which gives $H = \sum_{k=0}^{U_i-1} k^{-1} \Delta(k) = \sum_{k=0}^{U_i-1} (U_i - 1)^{-1} I(U_i \geq 2)$. In other words, when sample size is not fixed and can be very large, substituting the observed slopes into our formulation, we obtain the unbiased estimator developed in Montgomery-Smith and Schürmann (2007). This is an elegant connection.

Acknowledgements
The authors thank the Executive Editor, the Associate Editor and two anonymous reviewers for thoughtful and helpful comments and suggestions, which substantially improved the paper. AC and YTW are supported by Taiwan National Science Council under Contract 100-2118-M-007-006-MY3. LJ acknowledges support from John V. Moore through a grant to the Population Biology Foundation. Both AC and LJ also thank the Centre de Recerca Matemàtica, Barcelona, for logistic support during their Exploratory Conference on the Mathematics of Biodiversity.

References


Received 14 February 2013; accepted 1 August 2013

Handling Editor: Derral Warin
Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. A brief review of entropy estimators.

Appendix S2. A bootstrap method to construct a variance estimator and confidence intervals.

Appendix S3. Supplementary plots (Fig. S3-1 to Fig. S3-10).

Appendix S4. Species frequency data used in Fig. 2 of the main text.

Appendix S5. Comparison of the proposed entropy estimator and its jackknifed version.

Appendix S6. Entropy estimation for multiple incidence data.