NOTE **Communicated by Peter Grassberger**

A Note on Entropy Estimation

Thomas Schürmann
Jülich Supercomputing Centre, Jülich Research Centre, 52425 Jülich, Germany

We compare an entropy estimator $\hat{H}_z$ recently discussed by Zhang (2012) with two estimators, $\hat{H}_1$ and $\hat{H}_2$, introduced by Grassberger (2003) and Schürmann (2004). We prove the identity $\hat{H}_z \equiv \hat{H}_1$, which has not been taken into account by Zhang (2012). Then we prove that the systematic error (bias) of $\hat{H}_1$ is less than or equal to the bias of the ordinary likelihood (or plug-in) estimator of entropy. Finally, by numerical simulation, we verify that for the most interesting regime of small sample estimation and large event spaces, the estimator $\hat{H}_2$ has a significantly smaller statistical error than $\hat{H}_z$.

1 Introduction

Symbolic sequences are typically characterized by an alphabet A of d different letters. We assume statistical stationarity: any letter-block (word or n-gram of constant length) w_i, $i = 1, \ldots, M$, can be expected at any chosen site to occur with a known probability $p_i = \text{prob}(w_i)$ and $\sum_{i=1}^{M} p_i = 1$.

In a classic paper, Shannon (1951) considered the problem of estimating the entropy,

$$H = -\sum_{i=1}^{M} p_i \log p_i, \tag{1.1}$$

of ordinary English. In principle, this might be done by dealing with longer and longer contexts until dependencies at the word level, phrase level, sentence level, paragraph level, chapter level, and so on have all been taken into account in the statistical analysis. In practice, however, this is quite impractical, for as the context grows, the number M of possible words explodes exponentially with n.

In the numerical estimation of the Shannon entropy, one can do frequency counting; hence, in the limit of large data sets, the relative frequency distribution yields an estimate of the underlying probability distribution. We consider samples of N independent observations and let k_i, $i = 1, \ldots, M$, be

Neural Computation **27, 2097–2106** (2015)
doi:10.1162/NECO_a_00775

the frequency of realization w_i in the ensemble. However, with the choice $\hat{p}_i = \frac{k_i}{N}$, the naive (or likelihood) estimate,

$$\hat{H}_0 = -\sum_{i=1}^{M} \hat{p}_i \log \hat{p}_i, \tag{1.2}$$

leads to a systematic underestimation of the Shannon entropy (Miller, 1955; Harris, 1975; Herzel, 1988; Schürmann & Grassberger, 1996; Grassberger, 2003; Schürmann, 2004). In particular, if M is on the order of the number of data points N, then fluctuations increase and estimates usually become significantly biased. By *bias*, we denote the deviation of the expectation value of an estimator from the true value. In general, the problem in estimating functions of probability distributions is to construct an estimator whose estimates both fluctuate with the smallest possible variance and are least biased.

On the other hand, there is the Bayesian approach to entropy estimation, building on an approach introduced by Nemenman, Shafee, and Bialek (2002), or a generalization recently proposed by Archer, Park, and Pillow (2014). There, the basic strategy is to place a prior over the space of probability distributions and then perform inference using the induced posterior distribution over entropy. Actually, a partial numerical comparison of the popular Bayesian entropy estimates and those discussed her can be found in the work of Archer et al. (2014). Unfortunately, these simulations consider only the bias of the entropy estimates, not their mean square error, which takes into account the important trade-off between bias and variance. However, in the considerations to be discussed below, for what we intend to demonstrate, no explicit prior information on distributions is assumed, and we will focus ourself on non-Bayes entropy estimates only.

To start, let us consider an estimator of the Shannon entropy that has recently been proposed and analyzed against the likelihood estimator by Zhang (2012). The development of this interesting estimator starts with a generalization of the diversity index proposed by Simpson (1949) and refers to the following representation of the Shannon entropy:[1]

$$H = \sum_{\nu=1}^{\infty} \frac{1}{\nu} \sum_{i=1}^{M} p_i \, (1 - p_i)^{\nu}. \tag{1.3}$$

Zhang (2012) has mentioned that there exists an interesting estimator of each term in equation 1.3, which is unbiased up to the order $\nu = N - 1$,

[1]For another interpretation of this representation, see Montgomery-Smith and Schürmann (2005).

namely, Z_ν/ν, where Z_ν is explicitly given by the expression

$$Z_\nu = \frac{N^{1+\nu}(N-\nu-1)!}{N!} \sum_{i=1}^{M} \frac{k_i}{N} \prod_{j=0}^{\nu-1} \left(1 - \frac{k_i}{N} - \frac{j}{N}\right), \tag{1.4}$$

such that

$$\hat{H}_z = \sum_{\nu=1}^{N-1} \frac{1}{\nu} Z_\nu \tag{1.5}$$

is a statistical consistent entropy estimator of H with (negative) bias

$$B_N = -\sum_{\nu=N}^{\infty} \frac{1}{\nu} \sum_{i=1}^{M} p_i \, (1-p_i)^\nu. \tag{1.6}$$

Indeed, the estimator is notable because a uniform variance upper bound has been proven by Zhang (2012) that decays at a rate of $\mathcal{O}(\log(N)/N)$ for all distributions with finite entropy, compared to $\mathcal{O}((\log(N))^2/N)$ of the ordinary likelihood estimator established by Antos and Kontoyiannis (2001). It should be mentioned here that the latter decay rate is an implication of the Efron-Stein inequality, whereas the former (faster) decay rate is derived within the completely different approach introduced by Zhang (2012). Actually, it seems hard to prove the same decay rate for the likelihood estimator.

In the following section, we show that $\hat{H}_z$ is algebraically equivalent to the estimator of Grassberger (2003) and Schürmann (2004),

$$\hat{H}_1 = \sum_{i=1}^{M} \frac{k_i}{N} \Big(\psi(N) - \psi(k_i)\Big), \tag{1.7}$$

while the summation is defined for all $k_i > 0$ and the digamma function $\psi(k)$ is the logarithmic derivative of the Gamma-function (Abramowitz & Stegun, 1965). Actually, the estimator equation, 1.7, is given for the choice $\xi = 1$ discussed by Schürmann (2004, equation 28). In the asymptotic regime $k_i \gg 1$, this estimator leads to the ordinary Miller correction $\hat{H}_1 \sim \hat{H}_0 + (M-1)/2N$. This can be seen by using the asymptotic relation $\psi(x) \sim \log(x) - 1/2x$.

The mathematical expression of the bias of $\hat{H}_1$ has also been derived by Schürmann (2004) and is explicitly given by

$$B_N^{(1)} = -\sum_{i=1}^{M} p_i \int_0^{1-p_i} \frac{t^{N-1}}{1-t}\, dt, \tag{1.8}$$

with a uniform upper bound:

$$|B_N^{(1)}| \le \frac{M}{N}. \tag{1.9}$$

The proof of the identity $B_N \equiv B_N^{(1)}$ will be suppressed here because it is sufficient to show the equivalence of the corresponding entropy estimators in the following section.

It should be mentioned that the numerical computation time of the estimator $\hat{H}_1$ is significantly faster than for $\hat{H}_z$. Actually, this improvement has not been taken into account by Archer et al. (2014; see Figure 11), where the authors still used expression 1.5 above.

In the third section, by numerical computation, we compare the mean square error of $\hat{H}_z$ with an entropy estimator corresponding to $\xi = 1/2$ of Schürmann (2004; equation 13) or Grassberger (2003; equation 35), which is explicitly given by the following representation:

$$\hat{H}_2 = \sum_{i=1}^{M} \frac{k_i}{N}\Big(\psi(N) - \psi(k_i) + \log(2) + \sum_{j=1}^{k_i-1} \frac{(-1)^j}{j}\Big). \tag{1.10}$$

This estimator is an extension of $\hat{H}_1$ by an oscillating term in the bracket on the right-hand side of equation 1.7. In both Grassberger (2003) and Schürmann (2004), this estimator has not been expressed in terms of a finite sum, but by integral expressions or infinite sum representations instead. However, it can be easily shown that the present form is equivalent to those discussed by Grassberger (2003) and Schürmann (2004), but the computation is less time-consuming. The bias of the estimator 1.10 is (Schürmann 2004)

$$B_N^{(2)} = -\sum_{i=1}^{M} p_i \int_0^{1-2p_i} \frac{t^{N-1}}{1-t}\, dt, \tag{1.11}$$

with uniform upper bound

$$|B_N^{(2)}| \le \frac{M+1}{2N}. \tag{1.12}$$

When we look at the right-hand side of equations 1.9 and 1.12, then we see that they mainly differ by a factor of 2 in the denominator. Thus, we can expect a faster convergence of $\hat{H}_2$ for sufficient large M and not very strongly peaked probability distributions. Actually, these are the distributions we are mainly interested in. The numerical comparison of the mean square error of $\hat{H}_z$ and $\hat{H}_2$ will be evaluated for the uniform probability distribution, the Zipf distribution, and the zero-entropy delta distribution.

2 Comparison of $\hat{H}_z$ and $\hat{H}_1$

In this section, we show the identity $\hat{H}_z \equiv \hat{H}_1$. Therefore, let $Z_{i,\nu}$ denote the ith term of equation 1.4:

$$Z_{i,\nu} = \frac{N^{1+\nu}(N-\nu-1)!}{N!}\,\frac{k_i}{N}\prod_{j=0}^{\nu-1}\left(1-\frac{k_i}{N}-\frac{j}{N}\right). \tag{2.1}$$

By extending with N in the product, this expression can be rewritten as

$$Z_{i,\nu} = \frac{(N-\nu-1)!}{N!}\,k_i\prod_{j=0}^{\nu-1}(N-k_i-j). \tag{2.2}$$

Next, the product is reformulated as a quotient of factorials,

$$\prod_{j=0}^{\nu-1}(N-k_i-j) = \frac{(N-k_i)!}{(N-k_i-\nu)!}, \tag{2.3}$$

and in terms of binomial coefficients, we get

$$Z_{i,\nu} = \frac{k_i}{N-\nu}\binom{N-\nu}{k_i}\Big/\binom{N}{k_i}. \tag{2.4}$$

Now, the ith term of the estimator, equation 1.5, is obtained by summation over ν,

$$\begin{aligned}\sum_{\nu=1}^{N-1}\frac{1}{\nu}Z_{i,\nu} &= \binom{N}{k_i}^{-1}k_i\sum_{\nu=1}^{N-1}\frac{1}{\nu(N-\nu)}\binom{N-\nu}{k_i}\\ &= \frac{k_i}{N}\left(\mathcal{H}_{N-1}-\mathcal{H}_{k_i-1}\right),\end{aligned} \tag{2.5}$$

while $\mathcal{H}_k = \sum_{n=1}^{k} 1/n$ is the kth harmonic number (Abramowitz & Stegun, 1965). Applying the identity $\mathcal{H}_{k-1} = \psi(k) + \gamma$ (with $\gamma = 0.5772\ldots$, the Euler-Mascheroni constant) and summation for $i = 1, 2, \ldots, M$, we obtain the estimator, equation 1.7, which proves the identity $\hat{H}_z \equiv \hat{H}_1$. In addition, we have the following proposition:

Proposition 1. *The estimator $\hat{H}_1$ is less biased than (or equally biased) the likelihood estimator $\hat{H}_0$, for all samples of size $N \geq 1$ and $M \geq 2$.*

Proof. Since we know from Schürmann (2014) that the bias of $\hat{H}_1$ is negative, it is sufficient to prove that $\psi(N) - \psi(k) > \log \frac{N}{k}$, for $0 < k < N$. Now, the following inequalities (Merkle, 2008),

$$\psi(N) \geq \log\left(N - \frac{1}{2}\right), \tag{2.6}$$

$$\psi(k) \leq \log(k) - \frac{1}{2k}, \tag{2.7}$$

can be applied such that we only have to check that

$$N > \frac{1/2}{1 - e^{-\frac{1}{2k}}}, \tag{2.8}$$

for all N with $0 < k < N$. For any finite $k > 0$, the inequality $1 + \frac{1}{2k} < \exp\left(\frac{1}{2k}\right)$ is satisfied. The proof is by Taylor series expansion of the exponential function. From this, by simple algebraic manipulations, it follows that the right-hand side of equation 2.8 is less than $k + \frac{1}{2}$ for any finite $k > 0$. It follows that equation 2.8 is satisfied for any k with $0 < k < N$. This proves that $\hat{H}_1$ is less biased than $\hat{H}_0$ for any $M \geq 2$.

3 Numerical Comparison of $\hat{H}_z$ and $\hat{H}_2$

In this section, we focus on the convergence rates of the root mean square error (RMSE) of $\hat{H}_z$ and $\hat{H}_2$. Here, the RMSE is defined by

$$\mathrm{RMSE} = \sqrt{\mathrm{E}[(\hat{H} - H)^2]}. \tag{3.1}$$

We choose this error measure because it takes into account the trade-off between bias and variance. Moreover, we want to mention that there is a slightly modified version $\hat{H}_z^*$ of the estimator $\hat{H}_z$, defined in equation 1.12 of Zhang (2012). Since the bias B_N of $\hat{H}_z$ is explicitly known, a correction is defined by subtraction of the bias term B_N with p_i replaced by its estimate

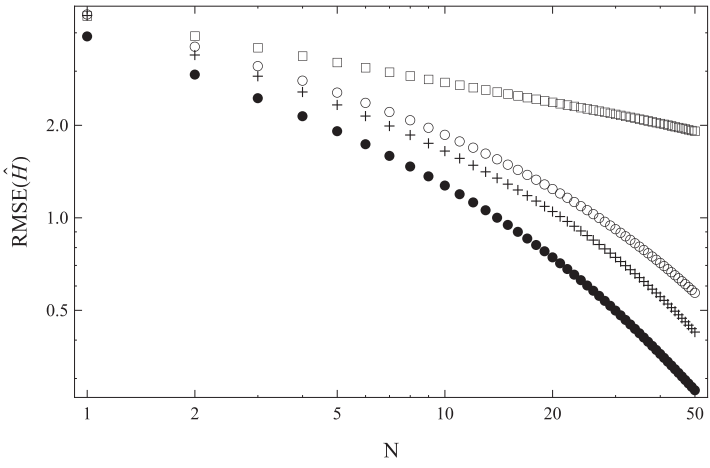

Figure 1: Statistical error of $\hat{H}_0$ (□), $\hat{H}_z$ (○), $\hat{H}_z^*$ (+), and $\hat{H}_2$ (•), for the uniform probability distribution with $M = 100$ (see text). The RMSE of $\hat{H}_2$ is significantly smaller than of $\hat{H}_z$ and $\hat{H}_z^*$. The exact value of the entropy is $H = 5.3$.

$\hat{p}_i$. The modified estimator is then given by $\hat{H}_z^* = \hat{H}_z - \hat{B}_N$, while $\hat{B}_N$ is the plug-in estimator of B_N. For simplicity, we deny applying the same procedure of bias correction for the estimator $\hat{H}_2$. Our first data sample is taken from the uniform probability distribution $p_i = 1/M$ for $i = 1, 2, \ldots, M$. In addition, we consider the (right-tailed) Zipf distribution with $p_i = c/i$, for $i = 1, 2, \ldots, M$ and normalization constant $c = 1/\mathcal{H}_M$ (reciprocal of the Mth harmonic number). The statistical error for increasing sample size N and given M is shown in Figures 1 and 2.

As we can see, the RMSE of all estimators is monotonic decreasing in N. The convergence of the naive estimator $\hat{H}_0$ is rather slow compared to the other estimators, while the performance of $\hat{H}_z^*$ is slightly better than for $\hat{H}_z$. On the other hand, the statistical error of $\hat{H}_2$ is significantly smaller than the statistical error of $\hat{H}_z$ and $\hat{H}_z^*$, and this behavior seems to be representative for large M. The statistical error for increasing M and fixed sample size N is shown in Figures 3 and 4. For $M \gg N$, the RMSE of $\hat{H}_z$ and $\hat{H}_z^*$ is greater than of $\hat{H}_2$. This phenomenon reflects the fact that the bias reduction becomes more and more relevant for increasing M compared to the contribution of the variance.

As we can see from both examples, the gap between $\hat{H}_z^*$ and $\hat{H}_2$ is slightly smaller for the peaked Zipf distribution compared to the uniform distribution. Thus, we ask for the performance in the extreme case of the delta distribution $p_i = \delta_{i,1}$, which has entropy zero. Indeed, in this

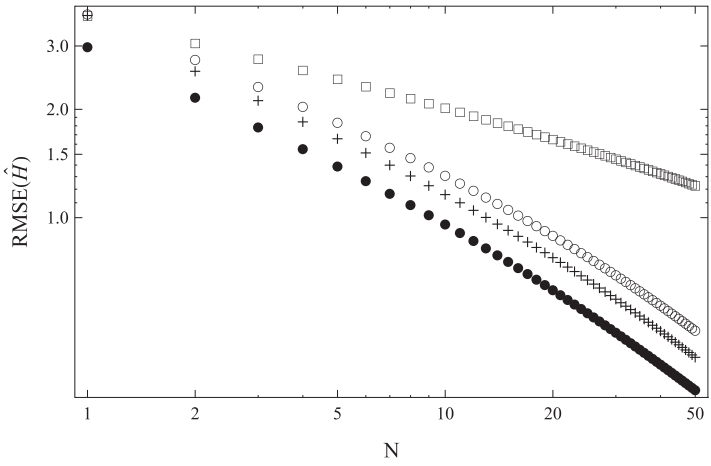

Figure 2: Same as in Figure 1 but for Zipf's probability distribution (see text). The exact value of the entropy is $H = 3.68$.

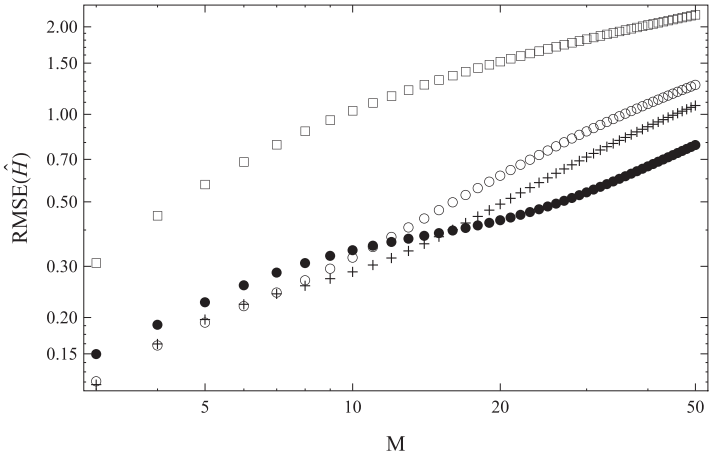

Figure 3: Statistical error of $\hat{H}_0$ (□), $\hat{H}_z$ (○), $\hat{H}_z^*$ (+), and $\hat{H}_2$ (•), for sample size $N = 10$ in the instance of the uniform probability distribution. Small sample estimation is expected when M is above the sample size N.

special case, we have $\hat{H}_0 = \hat{H}_1 = \hat{H}_z = \hat{H}_z^* = 0$ for any sample size N, but $\hat{H}_2 = \log(2) + \sum_{j=1}^{N-1} (-1)^j / j \to 0$ for $N \to \infty$. Actually, in this case, the statistical error of the latter scales like $\sim 1/2N$ for large N.

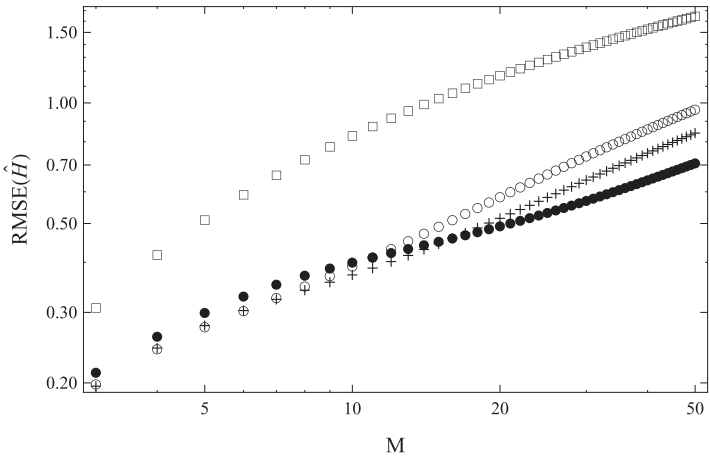

Figure 4: Same as in Figure 3 but for the Zipf distribution. There is a crossover for $M \approx N$.

4 Summary

In this note, we classified the entropy estimator $\hat{H}_z$ of Zhang (2012) within the family of entropy estimators originally introduced by Schürmann (2004). This reveals an interesting connection between two different approaches to entropy estimation: one coming from the generalization of the diversity index of Simpson and the other one from the estimation of p_i^q in the family of Renyi entropies. This connection is explicitly established by the identity $\hat{H}_z \equiv \hat{H}_1$. In addition, we proved that the statistical bias of $\hat{H}_1$ is smaller than the bias of the likelihood estimator $\hat{H}_0$. Furthermore, by numerical computation for various probability distributions, we found that $\hat{H}_z$ (or the heuristic estimator $\hat{H}_z^*$) can be improved by the estimator $\hat{H}_2$, which is an excellent member of the estimator family of Grassberger (2003) and Schürmann (2004). There is a uniform variance upper bound of $\hat{H}_z$ (and therefore of $\hat{H}_1$) that decays at a rate of $\mathcal{O}(\log(N)/N)$ for all distributions with finite entropy (Zhang, 2012). It would be interesting to know if this variance bound also holds for the estimator $\hat{H}_0$ or $\hat{H}_2$. The answer might be found in a forthcoming publication.

References

Abramowitz, M., & Stegun, I. (1965). *Handbook of mathematical functions*. New York: Dover.

Antos, A., & Kontoyiannis, I. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, *19*(3–4), 163–193.

Archer, E., Park, I. M., & Pillow, J. W. (2014). Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, *15*(1), 2833–2868.

Grassberger, P. (2003). Entropy estimates from insufficient samplings. arxiv:physics /0307138

Harris, B. (1975). *The statistical estimation of entropy in the non-parametric case*. In I. Csiszar (Ed.), *Topics in information theory* (pp. 323–355). Amsterdam: North-Holland.

Herzel, H. (1988). Complexity of symbol sequences. *Systems Analysis Modelling Simulation*, *5*(5), 435–444.

Merkle, M. (2008). Inequalities for the gamma function via convexity. In P. Cerone and S. S. Dragomir (Eds.), *Advances in inequalities for special function* (pp. 81–100). New York: Nova Science.

Miller, G. A. (1955). Note on the bias of information estimates. *Information theory in Psychology: Problems and Methods*, *2*, 95–100.

Montgomery-Smith, S., & Schürmann, T. (2005). *Unbiased estimators for entropy and class number*. arxiv:1410.5002, (2014).

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems*, *14*, (pp. 471–478). Cambridge, MA: MIT Press.

Schürmann, T. (2004). Bias analysis in entropy estimation. *Journal of Physics A: Mathematical and General*, *37*, L295–L301.

Schürmann, T., and Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *6*(3), 414–427.

Shannon, C. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*, 50–64.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*, 688–688.

Zhang, Z. (2012). Entropy estimation in turing's perspective. *Neural Computation*, *24*(5), 1368–1389.

Received April 8, 2015; accepted June 11, 2015.