entropy

MDPI

Brief Report

On Generalized Schürmann Entropy Estimators

Peter Grassberger

Jülich Supercomputing Center, Jülich Research Center, D-52425 Jülich, Germany; p.grassberger@fz-juelich.de

Abstract: We present a new class of estimators of Shannon entropy for severely undersampled discrete distributions. It is based on a generalization of an estimator proposed by T. Schürmann, which itself is a generalization of an estimator proposed by myself.For a special set of parameters, they are completely free of bias and have a finite variance, something which is widely believed to be impossible. We present also detailed numerical tests, where we compare them with other recent estimators and with exact results, and point out a clash with Bayesian estimators for mutual information.

Keywords: entropy estimates; mutual information estimates; undersampling; Bayesian; bias; variance

Citation: Grassberger, P. On Generalized Schürmann Entropy Estimators. *Entropy* **2022**, *24*, 680. https://doi.org/10.3390/e24050680

Academic Editor: Leandro Pardo

Received: 8 January 2022
Accepted: 9 May 2022
Published: 11 May 2022

1. Introduction

It is well known that estimating (Shannon) entropies from finite samples is not trivial. If one naively replaces the probability p_i to be in "box" i by the observed frequency, $p_i \approx n_i/N$, statistical fluctuations tend to make the distribution look less uniform, which leads to an underestimation of the entropy. There have been numerous proposals on how to estimate and eliminate this bias [1–22]. Some make quite strong assumptions [5,7]; others use Bayesian methods [6,11,12,19,21,22]. As pointed out in [4,13,14,17], one can devise estimators with arbitrarily small bias (for sufficiently large N and fixed p_i), but these will then have very large statistical errors. As conjectured in [4,13–15,17], the variance of any estimator whose bias vanishes will have a diverging variance.

Another widespread belief is that Bayesian entropy estimators cannot be outperformed by non-Bayesian ones for severely undersampled cases. The problem with Bayesian estimators is of course that they depend on a good choice of prior distributions, which is not always easy, and they tend to be slow. One positive feature of a non-Bayesian estimator proposed in [14] is that it is extremely fast since it works precisely like the 'naive' (or maximum-likelihood) estimator, except that the logarithms used there are replaced by a function G_n defined on integers, which can be precomputed by means of a simple recursion. While the estimator of [14] seems in general to be a reasonable compromise between bias and variance, it was shown in [15] that it can be improved—as far as bias is concerned, at the cost of increased variance—by generalizing G_n to a one-parameter family of functions $G_n(a)$.

In the present paper, we show that the Grassberger–Schürmann approach [14,15] can be further improved by using different functions $G_n(a_i)$ for each different realization i of the random variable. Indeed, the a_i can be chosen such that the estimator is completely free of bias and yet has a finite variance—although, to be honest, the optimal parameters a_i can be found only if the exact distribution is known (in which case also the entropy can be computed exactly). We show that—even if the exact, optimal a_i is not known—the new estimator can reduce the bias very much, without inducing unduly large variances, provided the distribution is sufficiently much undersampled.

We test the proposed estimator numerically with simple examples, where we produce bias-free entropy estimates, e.g., from pairs of ternary variables, something which, to my knowledge, is not possible with any Bayesian method. We also use it for estimating mutual

information (MI) in cases where one of the two variables is binary, and the other one can take very many values. In the limit of severe undersampling and of no obvious regularity in the true probabilities, MI cannot be estimated unambiguously. In that limit, the present algorithm seems to choose systematically a different outcome from Bayesian methods for reasons that are not yet clear.

2. Basic Formalism

In the following, we use the notation of [14]. As in this reference, we consider $M > 1$ "boxes" (states, possible experimental outcomes, etc.) and $N > 1$ points (samples, events, and particles) distributed randomly and independently into the boxes. We assume that each box has weight p_i ($i = 1, \ldots M$) with $\sum_i p_i = 1$. Thus each box i will contain a random number n_i of points, with $\mathsf{E}[n_i] = p_i N$. Their joint distribution is multinomial,

$$P(n_1, n_2, \ldots n_M; N) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!}, \tag{1}$$

while the marginal distribution in box i is binomial,

$$P(n_i; p_i, N) = \binom{N}{n_i} p_i^{n_i} (1 - p_i)^{N - n_i}. \tag{2}$$

Our aim is to estimate the entropy,

$$H = -\sum_{i=1}^{M} p_i \ln p_i = \ln N - \frac{1}{N} \sum_{i=1}^{M} z_i \ln z_i, \tag{3}$$

with $z_i \equiv \mathsf{E}[n_i] = p_i N$, from an observation of the numbers $\{n_i\}$ (in the following, all entropies are measured in "natural units", not in bits). The estimator $\hat{H}(n_1, \ldots n_M)$ will of course have both statistical errors and a bias, i.e., if we repeat this experiment, the average of $\hat{H}$ will, in general, not be equal to H,

$$\Delta[\hat{H}] \equiv \mathsf{E}[\hat{H}] - H \neq 0, \tag{4}$$

as will also be its variance $\mathrm{Var}[\hat{H}]$. Notice that for computing $\mathsf{E}[\hat{H}]$, we need only Equation (2), not the full multinomial distribution of Equation (1). However, if we want to compute this variance, we additionally need the joint marginal distribution in two boxes,

$$\begin{aligned} P(n_i, n_j; p_i, p_j, N) \quad = \quad & \frac{N!}{n_i! n_j! (N - n_i - n_j)!} \times \\ & p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{N - n_i - n_j}, \end{aligned} \tag{5}$$

in order to compute the covariances between different boxes. Notice that these covariances were not taken into account in [13,17], whence the variance estimations in these papers are, at best, approximate.

In the following, we are mostly interested in the case where we are close to the limit $N \to \infty, M \to \infty$, with M/N (the average number of points per box) being finite and small. In this limit, the variance will go to zero (because essentially one averages over many boxes), but the bias will remain finite. The binomial distribution, Equation (2), can be replaced then by a Poisson distribution

$$P_{\mathrm{Poisson}}(n_i; z_i) = \frac{z_i^{n_i}}{n_i!} e^{-z_i}. \tag{6}$$

However, as we shall see, it is in general not good advice to jump right to this limit, even if we are close to it. More generally, we shall therefore also be interested in the case of large

but finite N, where also the variance is positive, and we will discuss the balance between demanding minimal bias versus minimal variance.

Indeed it is well known that the *naive* (or 'maximum-likelihood') estimator, obtained by assuming $z_i = n_i$ without fluctuations,

$$\hat{H}_{\text{naive}} = \ln N - \frac{1}{N}\sum_{i=1}^{M} n_i \ln n_i, \tag{7}$$

is negatively biased, $\Delta\hat{H}_{\text{naive}} < 0$.

In order to estimate H, we have to estimate $p_i \ln p_i$ or equivalently $z_i \ln z_i$ for each i. Since the distribution of n_i depends, according to Equation (2), on z_i only, we can make the rather general ansatz [4,14] for the estimator

$$\widehat{z_i \ln z_i} = n_i \phi(n_i) \tag{8}$$

with a yet unknown function $\phi(n)$. Notice that $\hat{H}$ becomes with this ansatz a sum over strictly positive values of n_i. Effectively this means that we have assumed that observing an outcome $n_i = 0$ does not give any information: if $n_i = 0$, we do not know whether this is because of statistical fluctuations or because $p_i = 0$ for that particular i.

The resulting entropy estimator is then [14]

$$\hat{H}_\phi = \ln N - \frac{M}{N}\overline{n\phi(n)} \tag{9}$$

with the overbar indicating an average over all boxes,

$$\overline{n\phi(n)} = \frac{1}{M}\sum_{i=1}^{M} n_i \phi(n_i). \tag{10}$$

Its bias is

$$\Delta H_\phi = \frac{M}{N}(\overline{z \ln z} - \overline{\mathsf{E}_{N,z}[n\phi(n)]}). \tag{11}$$

with

$$\mathsf{E}_{N,z}[f_n] = \sum_{n=1}^{\infty} f_n P_{\text{binom}}(n; p = z/N, N). \tag{12}$$

being the expectation value for a typical box (in the following we shall suppress the box index i to simplify notation, wherever this makes sense).

In the following, $\psi(x) = d \ln \Gamma(x)/dx$ is the digamma function, and

$$E_1(x) = \Gamma(0, x) = \int_1^\infty \frac{e^{-xt}}{t} dt \tag{13}$$

is an exponential integral (Ref. [23], paragraph 5.1.4). It was shown in [14] that

$$\mathsf{E}_{N,z}[n\psi(n)] = z \ln z + z[\psi(N) - \ln N] + z \int_0^{1-z/N} \frac{x^{N-1} dx}{1-x}\,, \tag{14}$$

which simplifies in the Poisson limit ($N \to \infty$, z fixed) to

$$\mathsf{E}_{N,z}[n\psi(n)] \to z \ln z + z E_1(z)\,. \tag{15}$$

Equations (14) and (15) are the starting points of all further analysis. In [14], it was proposed to re-write Equation (15) as

$$\mathsf{E}_{N,z}[nG_n] \to z \ln z + z E_1(2z)\,, \tag{16}$$

where

$$G_n = \psi(n) + (-1)^n \int_0^1 \frac{x^{n-1}}{x+1} dx. \tag{17}$$

The advantages are that G_n can be evaluated very easily by recursion (here $\gamma = 0.57721\ldots$ is the Euler–Mascheroni constant), $G_1 = G_2 = -\gamma - \ln 2$, $G_{2n+1} = G_{2n}$, and $G_{2n+2} = G_{2n} + \frac{2}{2n+1}$, and neglecting the second term, $zE_1(2z)$ gives an excellent approximation unless z is exceedingly small, i.e., unless the numbers of points per box are very small such that the distribution is very severely undersampled. Thus the entropy estimator proposed in [14] was simply

$$\hat{H}_G = \ln N - \frac{1}{N} \sum_{i=1}^{M} n_i G_{n_i}. \qquad \text{(Poisson)} \tag{18}$$

Furthermore, since $zE_1(2z)$ is strictly positive, neglecting it gives a negative bias in $\hat{H}_G$, and one can show rigorously that this bias is smaller than that of [1,3].

3. Schürmann and Generalized Schürmann Estimators

The easiest way to understand the Schürmann class of estimators [15] is to define, instead of G_n, a one-parameter family of functions

$$G_n(a) = \psi(n) + (-1)^n \int_0^a \frac{x^{n-1}}{x+1} dx. \tag{19}$$

Notice that $G_n(1) = G_n$ and $G_n(0) = \psi(n)$.

Let us first discuss the somewhat easier Poissonian limit, where

$$\begin{aligned} \mathsf{E}_{N,z} \quad & [n(G_n(a) - \psi(n))] = \\ & = \sum_{n=1}^{\infty} (-1)^n n P_{\text{Poisson}}(n,z) \int_0^a \frac{x^{n-1}}{x+1} dx \\ & = -ze^{-z} \int_0^a \frac{dx}{x+1} e^{-xz} \\ & = -z(E_1(z) - E_1((1+a)z)), \end{aligned} \tag{20}$$

which gives

$$\mathsf{E}_{N,z}[nG_n(a)] = z \ln z + zE_1((1+a)z). \tag{21}$$

Using—to achieve greater flexibility—different parameters a_i for different boxes, and neglecting the second term in the last line of Equation (20), we obtain finally by using Equation (3)

$$\hat{H}_{\text{Schuermann}} = \ln N - \frac{1}{N} \sum_{i=1}^{M} n_i G_{n_i}(a_i) \qquad \text{(Poisson)}. \tag{22}$$

Indeed, the last term in Equation (20) can always be neglected for sufficiently large a because $0 < E_1((1+a)z) < \exp(-(1+a)z)/(1+a)z$ for any real $a > -1$.

Equation (22) might suggest that using larger a_i would always give an improvement because bias is reduced, but this would not take into account that larger a_i might lead to larger variances. However, the optimal choices of the parameters a_i are not obvious. Indeed, in spite of the ease of derivations in the Poissonian limit, it is much better to avoid it and to use the exact binomial expression.

For the general binomial case, the algebra is a bit more involved. By somewhat tedious but straightforward algebra, one finds that

$$\begin{aligned}
\mathsf{E}_{N,z} \quad & [n(G_n(a) - \psi(n))] = \\
& = \sum_{n=1}^{\infty} (-1)^n n \binom{N}{n} p^n (1-p)^{N-n} \int_0^a \frac{x^{n-1}}{x+1} dx \\
& = -pN \int_0^a \frac{dx}{x+1} \sum_{n=1}^{\infty} \binom{N-1}{n-1} (-px)^{n-1} (1-p)^{N-n} \\
& = -pN \int_0^a \frac{dx}{x+1} (1-p-px)^{N-1} \\
& = -z \int_0^a \frac{dx}{x+1} [1 - \frac{(1+x)z}{N}]^{N-1}.
\end{aligned} \tag{23}$$

One immediately checks that this reduces, in the limit ($N \to \infty$, z fixed), to Equation (20). On the other hand, by substituting

$$x \to t = 1 - \frac{(1+x)z}{N} \tag{24}$$

in the integral, we obtain

$$\mathsf{E}_{N,z}[nG_n(a) - \psi(n))] = -z \int_{1-(1+a)z/N}^{1-z/N} \frac{t^{N-1} dt}{1-t}. \tag{25}$$

Finally, by combining with Equation (14), we find [15]

$$\begin{aligned}
\mathsf{E}_{N,z}[n(G_n(a)] \quad & = \quad z \ln z + z[\psi(N) - \ln N] + \\
& + \quad z \int_0^{1-(1+a)z/N} \frac{x^{N-1} dx}{1-x}
\end{aligned} \tag{26}$$

and, using again Equation (3),

$$\hat{H}_{\text{opt}} = \psi(N) - \frac{1}{N} \sum_{i=1}^{M} n_i G_{n_i}(a_i), \qquad \text{(binomial)} \tag{27}$$

with a correction term which is $1/N$ times a sum over the integrals in Equation (26). This correction term vanishes, if all integration ranges vanish. This happens when $1 - (1 + a_i)z_i/N = 0$ for all i, or

$$a_i = a_i^* \equiv \frac{1-p_i}{p_i} \qquad \forall i. \tag{28}$$

This is a remarkable result, as it shows that in principle, there exists always an estimator which has zero bias and yet finite variance. In [15], one single parameter a was used, which is why we call our method a generalized Schürmann estimator.

When all box weights are small, $p_i \ll 1$ for all i, then these bias-optimal values a_i^* are very large. However, for two boxes with $p_1 = p_2 = 1/2$, e.g., the bias vanishes already for $a_1 = a_2 = 1$, i.e., for the estimator of Grassberger [14]!

In order to test the latter, we drew 10^8 triplets of random bits (i.e., $N = 3$, $p_0 = p_1 = 1/2$), and estimated $\hat{H}_{\text{naive}}$ and $\hat{H}_G$ for each triplet. From these, we computed averages and variances, with the results $\hat{H}_{\text{naive}} = 0.68867(4)$ bits and $\hat{H}_G = 0.99995(4)$ bits. We should stress that the latter requires the precise form of Equation (27) to be used, with $\psi(N)$ neither replaced by $\ln N$ nor by G_N.

Since there is no free lunch, there must of course be some problems in the limit when parameters a_i are chosen to be nearly bias-optimal. One problem is that one cannot, in general, choose a_i according to Equation (28), because the p_i is unknown. In addition, it

is in this limit (and more generally when $a_i >> 1$) that variances blow up. In order to see this, we have to discuss in more detail the properties of the functions $G_n(a)$.

According to Equation (19), $G_n(a)$ is a sum of two terms, both of which can be computed, for all positive integer n, by recursion. The digamma function $\psi(n)$ satisfies

$$\psi(1) = -\gamma, \quad \psi(n+1) = \psi(n) + 1/n. \tag{29}$$

Let us denote the second term in Equation (19) as $g_n(a)$. It satisfies the recursion

$$g_1(a) = -\ln(1+a), \quad g_{n+1}(a) = g_n(a) - (-a)^n/n. \tag{30}$$

Thus, while $\psi(n)$ is monotonic and slowly increasing, $g_n(a)$ has alternating sign and increases, for $a > 1$, exponentially with n. As a consequence, also $G_n(a)$ is non-monotonic and diverges exponentially with n, whenever $a > 1$. Therefore an estimator such as $\hat{H}_{\text{opt}}$ gets, unless all n_i are very small, increasingly large contributions of alternating signs. As a result, the variances will blow up, unless one is very careful to keep a balance between bias and variance.

To illustrate this, we drew tuples of independent and identically distributed binary variables $\{s_1, \ldots s_N\}$ with $p_0 = 3/4$ and $p_1 = 1/4$. For a_0, we chose $a_0 = a_0^* = 1/3$ because this should minimize the bias and should not create problems with the variance. We should expect such problems, however, if we would take $a_1 = a_1^* = 3$, although this would reduce the bias to zero. Indeed we found for $N = 100$ that the variance of the estimator exploded for all practical purposes as soon as $a_1 > 1.4$, while the results were optimal for $0.5 < a_1 \leq 1$ (bias and statistical error were both $< 10^{-5}$ for 10^8 tuples). On the other hand, for pairs ($N = 2$), we had to use much larger values of a_1 for optimality, and $a_1 = 3$ gave indeed the best results (see Figure 1). A similar plot for ternary variables is shown in Figure 2, where we see again that a-values near the bias-optimal ones gave estimates with zero almost zero bias and acceptable variance for the most undersampled case $N = 2$. Again, using the the exact bias-optimal values would have given unacceptably large variances for large N.

The message to be learned from this is that we should always keep all a_i sufficiently small such that $a_i^{n_i}$ is not much larger than 1 for any of the observed values of n_i.

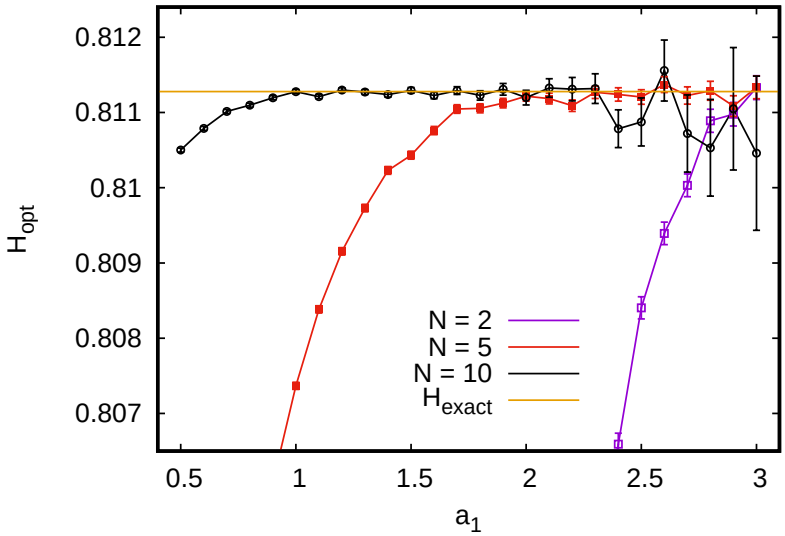

Figure 1. Estimated entropies (in bits) of N-tuples of independent and identically distributed random binary variables with $p_0 = 3/4$ and $p_1 = 1/4$, using the optimized estimator $\hat{H}_{\text{opt}}$ defined in Equation (27). The parameter a_0 was kept fixed at its optimal value $a_0 = 1/3$, while a_1 is varied in view of possible problems with the variances, and is plotted on the horizontal axis. For each N and each value of a_1, 10^8 tuples were drawn. The exact entropy for $p_0 = 3/4$ and $p_1 = 1/4$ is 0.811278... bits, and is indicated by the horizontal straight line.

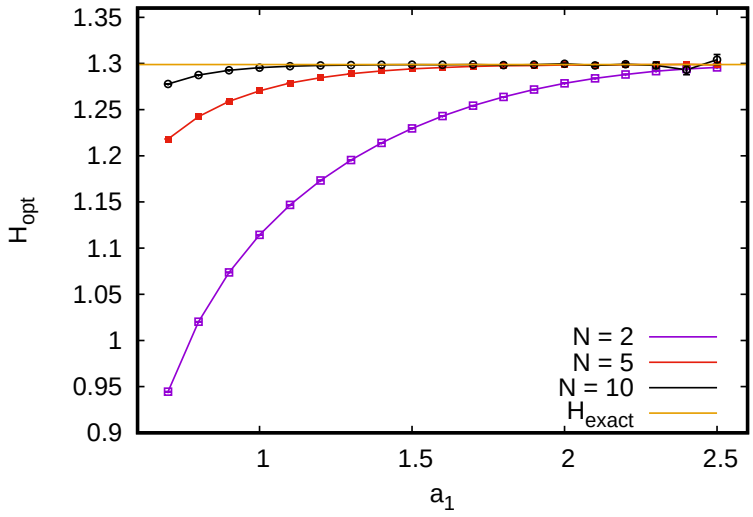

Figure 2. Estimated entropies (in bits) of N-tuples of independent and identically distributed random ternary variables with $p_0 = 0.625$, $p_1 = 0.25$, and $p_2 = 0.125$, using the optimized estimator $\hat{H}_{\mathrm{opt}}$ defined in Equation (27). The parameter a_0 was kept fixed at its optimal value $a_0^* = 0.6$, while a_1 and a_2 varied in view of possible problems with the variances. More precisely, we used $a_2 = 1 + 4(a_1 - 1)$, so that the plot ends at the bias-free value $a_2^* = 7.0$ and at a value of a_1 slightly smaller than $a_1^* = 2.5$. For each N and each value of a_1, 10^8 tuples were drawn. The exact entropy is 1.29879 . . . bits, and is indicated by the horizontal straight line.

4. Estimating Mutual and Conditional Information

Finally, we apply our estimator to two problems of mutual information (MI) estimation discussed in [22] (actually, the problems were originally proposed by previous authors, but we shall compare our results mainly to those in [22]). In each of these problems, there are two discrete random variables: X has many (several thousand) possible values, while Y is binary. Moreover, the marginal distribution of Y is uniform, $p(y = 0) = p(y = 1) = 1/2$, while the X distributions are highly non-uniform. Finally—and that is crucial—the joint distributions show no obvious regularities.

The MI is estimated as $I(X : Y) = H(Y) - H(Y|X)$. Since $H(Y) = 1$ bit, the problem essentially burns down to estimate the conditional probabilities $p(y|x)$. The data are given in terms of a large number of independent and identically distributed sampled pairs (x, y) (250,000 pairs for problem I, called 'PYM' in the following, and 50,000 pairs for problem II, called 'spherical' in the following). The task is to draw random subsamples of size N, to estimate the MI from each subsample, and to calculate averages and statistical widths from these estimates.

Results are shown in Figure 3. For large N, our data agree perfectly with those in [22] and in the previous papers cited in [22]. However, while the MI estimates in these previous papers all increase with decreasing N, and those in [22] stay essential constant (as we would expect, since a good entropy estimator should not depend on N, and conditional entropies should decrease with N for not so good estimators), our estimated MI decreases to zero for small N.

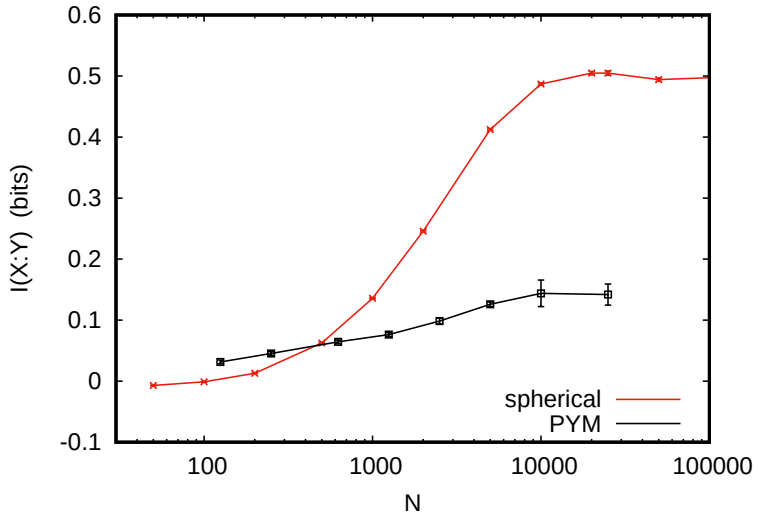

Figure 3. Estimated mutual information (in bits) of N-tuples of independent and identically distributed random subsamples from two distributions given in [22]. The data for "PYM", originally due to [24], consist of 250,000 pairs (x, y) with binary y with $p(y = 0) = p(y = 1) = 1/2$, and x being uniformly distributed over 4096 values. Thus each $x-$value is realized $\approx$60 times, and we classify them into 5 classes depending on the associated $y-$values: (i) very heavily biased toward $y = 1$, (ii) moderately biased toward $y = 1$, (iii) $y-$neutral, (iv) moderately biased toward $y = 0$, and (v) heavily biased toward $y = 0$. When we estimated conditional entropies $H(Y|X)$ for randomly drawn subsamples, we kept this classification and choose a_y accordingly: For class (iii) we used $a_0 = a_1 = 1$, for class (ii) we used $a_1 = 1, a_0 = 4$, for class (i) we used $a_1 = 1, a_0 = 7$, for class (iv) we used $a_1 = 4, a_0 = 1$, and finally for class (v) we used $a_1 = 7, a_0 = 1$. The data for "spherical", originally due to [21], consist of 50,000 (x, y) pairs. Here, Y is again binary with $p(y = 0) = p(y = 1) = 1/2$, but X is highly non-uniformly distributed over $\approx$4000 values. Again we classified these values as $y-$neutral or heavily/moderately biased toward or against $y = 0$ and used this classification to choose values of a_y accordingly.

This looks at first sight like a failure of our method, but it is not. As we said, the joint distributions show no regularities. For small N most values of X will show up at most once, and if we write the sequence of $y-$values in a typical tuple, it will look like a perfectly random binary string. The modeler knows that it actually is not random, because there are correlations between X and Y. However, no algorithm can know this, and any good algorithm should conclude that $H(Y|X) = H(Y) = 1$ bit. Why, then, was this not found in the previous analyses? In all these, Bayesian estimators were used. If the priors used in these estimators were chosen in view of the special structures in the data (which are, as we should stress again, not visible from the data, as long as these are severely undersampled), then the algorithms can, of course, make use of these structures and avoid the conclusion that $H(Y|X) = 1$ bit.

5. Conclusions

In conclusion, we gave an entropy estimator with zero bias and finite variance. It builds on an estimator by Schürmann [15], which itself is a generalization of [14]. It involves a real-valued parameter for each possible realization of the random variable, and bias is reduced to zero by choosing these parameters properly. However, this choice would require that we know already the distribution, which is of course not the case. Nevertheless we can reduce the bias very much for severely undersampled cases. In cases with moderate undersampling, choosing these zero-bias parameters would give very large variances and would thus be useless. Nevertheless, by judicious parameter choices, we can obtain extremely good entropy estimates. Finding good parameters is non-trivial, but is made less difficult by the fact that the method is very fast.

Finally, we pointed out that Bayesian methods, which have been very popular in this field, have the danger of choosing "too good" priors, i.e., choosing priors which are not justified by the data themselves and are thus misleading, although both the bias and the observed variances seem to be small.

I thank Thomas Schürmann for the numerous discussions, and Damián Hernández for both discussions and for sending me the data for Figure 3.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miller, G. Note on the bias of information estimates. In *Information Theory in Psychology II-B*; Quastler, H., Ed.; Free Press: Glencoe, IL, USA, 1955; pp. 95–100.
2. Harris, B. Colloquia Mathematica Societatis János Bolyai. *Infinite Finite Sets* **1973**, *175*, 323.
3. Herzel, H. Complexity of symbol sequences. *Syst. Anal. Mod. Sim.* **1988**, *5*, 435.
4. Grassberger, P. Entropy Estimates from Insufficient Samplings. *Phys. Lett.* **1988**, *128*, 369. [CrossRef]
5. Schmitt, A.O.; Herzel, H.; Ebeling, W. A new method to calculate higher-order entropies from finite samples. *Europhys. Lett.* **1993**, *23*, 303. [CrossRef]
6. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841. [CrossRef]
7. Poschel, T.; Ebeling, W.; Rose, H. Guessing probability distributions from small samples. *J. Stat. Phys.* **1995**, *80*, 1443. [CrossRef]
8. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. Network: Computation in neural systems. *Netw. Comput. Neural Syst.* **1996**, *7*, 87. [CrossRef] [PubMed]
9. Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*, 414. [CrossRef]
10. Strong, S.; Koberle, R.; van Steveninck, R.R.D.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200. [CrossRef]
11. Holste, D.; Grosse, I.; Herzel, H. Bayes' estimators of generalized entropies. *J. Phys. A* **1998**, *31*, 2551. [CrossRef]
12. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. In *Advances in Neural Information Processing 14*; Dietterich, T., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002.
13. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191. [CrossRef]
14. Grassberger, P. Entropy estimates from insufficient samplings. *arXiv* **2003**, arXiv:physics/0307138.
15. Schürmann, T. Bias analysis in entropy estimation. *J. Phys. A Math. Gen.* **2004**, *37*, L295. [CrossRef]
16. Vu, V.Q.; Yu, B.; Kass, R.E. Coverage-adjusted entropy estimation. *Stat. Med.* **2007**, *26*, 4039. [CrossRef]
17. Bonachela, J.A.; Hinrichsen, H.; Muñoz, M.A.M. Entropy estimates of small data sets. *J. Phys. A Math. Gen.* **2008**, *41*, 202001. [CrossRef]
18. Hausser, J.; Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.* **2009**, *10*, 1469.
19. Wolpert, D.H.; Deo, S.D. Estimating functions of distributions defined over spaces of unknown size. *Entropy* **2013**, *15*, 4668. [CrossRef]
20. Chao, A.; Wang, Y.; Jost, L. Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods Ecol. Evol.* **2013**, *2013*, 1091. [CrossRef]
21. Archer, E.; Park, I.M.; Pillow, J.W. Bayesian entropy estimation for countable discrete distributions. *J. Mach. Learn. Res.* **2013**, *2014*, 2833.
22. Hernández, D.G.; Samengo, I. Estimating the Mutual Information between Two Discrete, Asymmetric Variables with Limited Samples. *Entropy* **2019**, *21*, 623. [CrossRef] [PubMed]
23. Abramowitz, M.; Stegun, I. (Eds.) *Handbook of Mathematical Functions*; Dover: New York, NY, USA, 1965.
24. Schwartz-Ziv, R.; Tishby, N. Opening the black box of Deep Neural Networks via Information. *arXiv* **2017**, arXiv:1703.00810.