IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Entropy estimates of small data sets

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2008 J. Phys. A: Math. Theor. 41 202001

(http://iopscience.iop.org/1751-8121/41/20/202001)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.148
The article was downloaded on 03/06/2010 at 06:48

Please note that terms and conditions apply.

IOP FTC ▶▶▶
IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **41** (2008) 202001 (9pp) doi:10.1088/1751-8113/41/20/202001

FAST TRACK COMMUNICATION

Entropy estimates of small data sets

Juan A Bonachela[1,2], Haye Hinrichsen[2] and Miguel A Muñoz[1]

[1] Departamento de Electromagnetismo y Física de la Materia and Instituto de Física Teórica y Computacional Carlos I, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain
[2] Fakultät für Physik und Astronomie, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany

Received 21 December 2007, in final form 10 April 2008
Published 29 April 2008
Online at stacks.iop.org/JPhysA/41/202001

Abstract
Estimating entropies from limited data series is known to be a non-trivial task. Naïve estimations are plagued with both systematic (bias) and statistical errors. Here, we present a new 'balanced estimator' for entropy functionals (Shannon, Rényi and Tsallis) specially devised to provide a compromise between low bias and small statistical errors, for short data series. This new estimator outperforms other currently available ones when the data sets are small and the probabilities of the possible outputs of the random variable are not close to zero. Otherwise, other well-known estimators remain a better choice. The potential range of applicability of this estimator is quite broad specially for biological and digital data series.

PACS numbers: 89.75.Hc, 05.45.Xt, 87.18.Sn

1. Introduction

In statistical mechanics and information theory, entropy is a functional that measures the information content of a statistical ensemble or equivalently the uncertainty of a random variable. Its applications in physics, biology, computer science, linguistics, etc are countless. For example, it has become a key tool in data mining tasks arising from high-throughput biological analyses.

Historically, the most important example of such a functional is the Shannon (or *information*) entropy [1, 2]. For a discrete random variable x, which can take a finite number, M, of possible values $x_i \in \{x_1, \ldots, x_M\}$ with corresponding probabilities $p_i \in \{p_1, \ldots, p_M\}$, this entropy is defined by

$$H_S = -\sum_{i=1}^{M} p_i \ln(p_i). \tag{1}$$

Recently, various generalizations, inspired by the study of q-deformed algebras and special functions, have been investigated, most notably the Rényi entropy [3]

$$H_R(q) = \frac{1}{1-q}\ln\left(\sum_{i=1}^{M} p_i^q\right), \tag{2}$$

with $p \geqslant 0$, which, in particular, reduces to the Shannon entropy in the limit $q \to 1$. Also, the Tsallis entropy [4]

$$H_T(q) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{M} p_i^q\right), \tag{3}$$

although controversial, has generated a large burst of research activity.

In general, the full probability distribution for a given stochastic problem is not known and, in particular, in many situations only small data sets from which to infer entropies are available. For example, it could be of interest to determine the Shannon entropy of a given DNA sequence. In such a case, one could *estimate* the probability of each element *i* to occur, p_i, by making some assumption on the probability distribution, as for example (i) parametrizing it [5], (ii) dropping the most unlikely values [6] or (iii) assuming some *a priori* shape for the probability distribution [7, 8]. However, the easiest and most objective way to estimate them is just by counting how often the value x_i appears in the data set [9–15]. Denoting this number by n_i and dividing by the total size of the data set one obtains the relative frequency

$$\hat{p}_i = \frac{n_i}{N} \tag{4}$$

which *naïvely* approximates the probability p_i. Obviously, the entropy of the data set can be approximated by simply replacing the probabilities p_i by $\hat{p}_i$ in the entropy functional. For example, the Shannon entropy can be estimated by

$$H_S \approx \hat{H}_S^{\text{naive}} = -\sum_{i=1}^{M} \hat{p}_i \ln(\hat{p}_i) = -\sum_{i=1}^{M} \frac{n_i}{N}\ln\left(\frac{n_i}{N}\right). \tag{5}$$

The quantity $\hat{H}_S^{\text{naive}}$ is an example of an *estimator* of the entropy, in a very similar sense as $\hat{p}_i$ is an estimator of p_i. However, there is an important difference stemming from the nonlinear nature of the entropy functional. The frequencies $\hat{p}_i$ are *unbiased* estimators of the probabilities, i.e., their expectation value $\langle \hat{p}_i \rangle$ (where $\langle\cdot\rangle$ stands for ensemble averages) coincides with the true value of the estimated quantity

$$\langle \hat{p}_i \rangle = \frac{\langle n_i \rangle}{N} = p_i. \tag{6}$$

In other words, the frequencies $\hat{p}_i$ approximate the probabilities p_i with certain statistical error (*variance*) but without any systematic error (*bias*). Contrarily, *naïve* entropy estimators, such as $\hat{H}_S^{\text{naive}}$, in which p_i are simply replaced by n_i/N are always biased, i.e. they deviate from the true value of the entropy not only statistically but also systematically. Actually, defining an error variable $\epsilon_i = (\hat{p}_i - p_i)/p_i$, and replacing p_i in equation (1) by its value in terms of ϵ_i and $\hat{p}_i$, it is straightforward to verify that the bias, up to leading order, is $-\frac{M-1}{2N}$, which is a significant error for small N and vanishes only as $N \to \infty$ [12]. A similar bias, owing in general to the nonlinearity of the entropy functional, appears also for the Rényi and Tsallis entropies.

Therefore, the question arises whether it is possible to find improved estimators which reduce either the bias or the variance of the estimate. More generally, the problem can be formulated as follows. Given an arbitrary entropy functional of the form

$$H = \quad F\left[\sum_{i=1}^{M} h(p_i)\right] \tag{7}$$

(where F is a generic function) we want to find an estimator

$$\hat{H} = \quad F\left[\sum_{i=1}^{M} \chi_{n_i}\right] \tag{8}$$

such that the bias

$$\Delta = \langle \hat{H} \rangle - H \tag{9}$$

or the mean-squared deviation (the statistical error)

$$\sigma^2 = \langle (\hat{H} - \langle \hat{H} \rangle)^2 \rangle \tag{10}$$

or a combination of both are as small as possible. At the very end of such a calculation the estimator is defined by $N + 1$ real numbers χ_{n_i},[3] which depend on the sample size N. For example, the naïve estimator for the Shannon entropy would be given in terms of

$$\chi_{n_i}^{\text{naive}} = -\frac{n_i}{N} \ln\left(\frac{n_i}{N}\right). \tag{11}$$

The search for improved estimators has a long history. To the best of our knowledge, the first to address this question was Miller in 1955 [13], who suggested a correction to reduce the bias of the estimate of Shannon entropy, given by

$$\chi_{n_i}^{\text{Miller}} = -\frac{n_i}{N} \ln\left(\frac{n_i}{N}\right) + \frac{1}{2N}. \tag{12}$$

The correction exactly compensates the leading order of the bias, as reported above. In this case the remaining bias vanishes as $1/N^2$ as $N \to \infty$. This result was improved by Harris in 1975 [14], who calculated the next-leading order correction. However, his estimator depends explicitly on the (unknown) probabilities p_i, so that its practical importance is limited.

In another pioneering paper, Grassberger, elaborating upon previous work by Herzel [15], proposed an estimator which provides further improvement and gives a very good compromise between bias and statistical error [9]. For the Shannon entropy his estimator is given by

$$\chi_{n_i}^{\text{Grassberger}} = \frac{n_i}{N}\left(\ln N - \psi(n_i) - \frac{(-1)^{n_i}}{n_i(n_i+1)}\right), \tag{13}$$

where $\psi(x)$ is the derivative of the logarithm of the Gamma function, valid for all i with $n_i > 0$. According to [9], the function $\psi(x)$ can be approximated by

$$\psi(n_i) \approx \ln x - \frac{1}{2x} \tag{14}$$

for large x, giving

$$\chi_{n_i}^{\text{Grassberger}} \approx -\frac{n_i}{N} \ln\left(\frac{n_i}{N}\right) + \frac{1}{2N} - \frac{(-1)^{n_i}}{N(n_i+1)}. \tag{15}$$

This method can be generalized to q-deformed entropies.

More recently, a further improvement for the Shannon case has been suggested by Grassberger [10]

$$\chi_{n_i}^{\text{GS}} = \frac{n_i}{N}\left[\psi(N) - \psi(n_i) - (-1)^{n_i} \int_0^1 \frac{t^{n_i-1}}{1+t}\, \mathrm{d}t\right]. \tag{16}$$

[3] This is so because the estimator χ_{n_i} depends only on n_i, which can take $N + 1$ possible values.

This estimator can be recast (see equations (28), (29), (35) of [10]) as

$$\chi_{n_i}^{\mathrm{GS}} = \frac{n_i}{N}(\ln N - G_{n_i}), \tag{17}$$

where G_n satisfy the recurrence relation

$$G_1 = -\gamma - \ln 2 \tag{18}$$

$$G_2 = 2 - \gamma - \ln 2 \tag{19}$$

$$G_{2n+1} = G_{2n} \tag{20}$$

$$G_{2n+2} = G_{2n} + 2/(2n+1) \tag{21}$$

with $\gamma = -\psi(1)$. This estimator constitutes the state of the art for Shannon entropies, but unfortunately, it cannot be straightforwardly extended to more general q-deformed entropy functionals, for which [9] remains the best available option. These results were further generalized by Schürmann [11] with different balances between statistical and systematic errors.

It should be emphasized that an ideal estimator does not exist, instead the choice of the estimator depends on the structure of data to be analyzed [16]. For example, the above discussed estimators [9, 10] work satisfactorily if the probabilities p_i are sufficiently small. This is the case in many applications of statistical physics, where the number of possible states, M, in an ensemble is usually extremely large so that the probability p_i for an individual state i is very small. On the other hand, this assumption does not always hold for empirical data sets such as digital data streams and DNA sequences.

The performance of the estimators worsens as the values of p_i get larger. This is due to the following reason: the numbers n_i, which count how often the value x_i appears in the data set, are generically distributed as binomials, i.e. the probability P_{n_i} to find the value n_i is given by

$$P_{n_i}(p_i) = \binom{N}{n_i} p_i^{n_i}(1-p_i)^{N-n_i}, \tag{22}$$

where $\binom{N}{n_i} = \frac{N!}{n_i!(N-n_i)!}$ are binomial coefficients. For $p_i \ll 1$ this can be approximated by a Poisson distribution, which is the basis for the derivation of equation (13). For large values p_i, however, this assumption is no longer justified and this results in large fluctuations (even if the bias remains small).

It is important to note that it is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values. The existing studies have shown that there is always a delicate tradeoff between the two types of errors. For example, minimizing the bias usually comes at the expense of the variance, which increases significantly. Moreover, it can be proved that neither the variance nor the bias can be reduced to zero for finite N [17]. Therefore, it is necessary to study estimators with different balances between systematic and statistical errors, as it was done, e.g. in the work by Schürmann [11].

In the present work we introduce two estimators, which can be used to measure any of the entropy functionals discussed above. Both of them are specifically designed for short data series where the probabilities p_i take (in general) non-small values. The first one reduces the bias as much as possible at the expense of the variance, and is mostly of academic interest and discussed only for illustration purposes. The second one seeks for a robust compromise between minimizing bias and variance together, is very easy to implement numerically, and has a broad potential range of applicability. The estimator itself can be improved by adapting various of its elements to each specific problem.

2. The low-bias estimator

The starting point is the observation that the entropy H and its estimators $\hat{H}$ in equations (7) and (8) involve sums over all possible values of the data set. Therefore, as the bias can be minimized by minimizing the errors of each summand, the problem can be reduced to minimize

$$\delta(p_i) = \langle \chi_{n_i} \rangle - h(p_i) = \left(\sum_{n_i=0}^{N} P_{n_i}(p_i)\chi_{n_i} \right) - h(p_i) \tag{23}$$

over a broad range of p_i as much as possible.

A theorem by Paninski [17] states that it is impossible to reduce the bias to zero for all $p_i \in [0, 1]$ since an estimator is always a finite polynomial in p_i while the true entropy is usually not a polynomial. However, it is possible to let the bias vanish at $N + 1$ points p_i in the interval $[0, 1]$ because the determination of the different χ_{n_i} requires $N + 1$ independent equations.

For the sake of illustration, let us choose here equidistant points $p_j = j/N$, with $j = 0, 1, \ldots, N$. In general, other choices, more appropriate to each specific case, should be employed. The resulting set of linear equations reads

$$\delta(j/N) = 0 \Longrightarrow \sum_{n_i=0}^{N} P_{n_i}(j/N)\chi_{n_i} = h(j/N), \qquad j = 0, 1, \ldots, N. \tag{24}$$

Introducing the notation $h_j = h(j/N)$ and $P_{j,n_i} = P_{n_i}(j/N)$ this last expression takes the form

$$\sum_{n_i=0}^{N} P_{j,n_i}\chi_{n_i} = h_j, \qquad j = 0, 1, \ldots, N \tag{25}$$

or, in short, $\mathbf{P}\vec{\chi} = \vec{h}$, where $\mathbf{P}$ is the so-called *multinomial matrix* [18]. To find the solution $\vec{\chi} = \mathbf{P}^{-1}\vec{h}$, the matrix

$$P_{j,n_i} = \binom{N}{n_i} p_j^{n_i}(1 - p_j)^{N-n_i} = \binom{N}{n_i}\left(\frac{j}{N}\right)^{n_i}\left(1 - \frac{j}{N}\right)^{N-n_i}, \tag{26}$$

whose elements are binomial distributions, has to be inverted. For small N this inversion is most easily done numerically. However, we were also able to invert the matrix analytically, leading us to the closed form [19]

$$\hat{P}^{-1}_{i,j} = \sum_{k=0}^{N}\sum_{l=0}^{N} \binom{i}{k}\binom{l}{j} \frac{N^k k!(N-k)!}{N!} \frac{(-1)^{l+j}}{l!} s(l, k), \tag{27}$$

where $s(l, k)$ denotes the Stirling numbers of the first kind [20]. Having inverted the matrix, the numbers χ_{n_i} determining the estimators can be computed for any given entropy functional by a simple matrix multiplication.

Figure 1 illustrates a comparison for the Shannon case between the low-bias estimator and other well-known ones for the simple example of a binary sequence of $N = 20$ bits $x = 0, 1$ (i.e. $M = 2$), where the value 1 appears with probability p and 0 with probability $1 - p$. The bias of the low-bias estimator vanishes exactly only at values of p multiples of $1/20$, and takes small values in between (see inset of figure 1). On the other hand, the fluctuations for both the naïve estimator and the one in [10] remain bounded, while they diverge for the low-bias case (figure 1). This unbounded growing of statistical fluctuations makes the low-bias estimator useless for practical purposes.

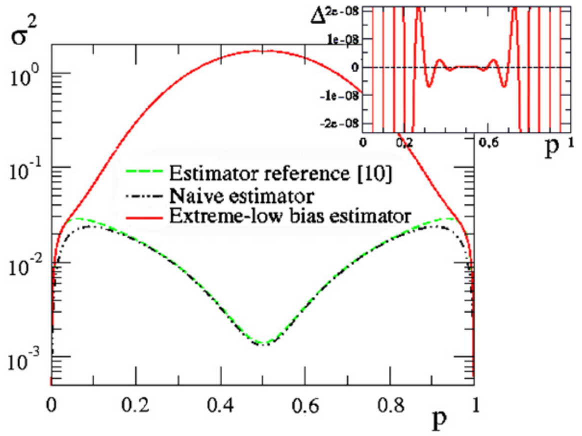

Figure 1. Fluctuations, σ^2, as defined by equation (10), for three different Shannon entropy estimates (the naïve one, the improved estimator introduced in [10] and the low-bias estimator defined in this paper) for a binary sequence ($M = 2$) of length $N = 20$. Inset: bias, Δ, as defined by equation (9), for the low-bias estimator showing the N+1 vanishing points with amplitude oscillations.

3. A balanced estimator

Aiming at solving the previously illustrated problem with uncontrolled statistical fluctuations, in this section we introduce a new *balanced estimator* designed to minimize simultaneously both the bias and the variance over a wide range of probabilities. This is of relevance for analyzing small data sets where statistical fluctuations are typically large and a compromise with minimizing the bias is required.

As before, ignoring correlations between the n_i both bias and statistical errors can be optimized by minimizing the errors of the summands in their corresponding expressions. Therefore, the problem can be reduced to minimize the bias for each state

$$\delta(p_i) = \langle \chi_{n_i} \rangle - h(p_i) \tag{28}$$

and the variance within such a state

$$\sigma^2(p_i) = \left\langle \left(\chi_{n_i} - \langle \chi_{n_i} \rangle\right)^2 \right\rangle \tag{29}$$

over a broad range of $p_i \in [0, 1]$, where $n_i \in 0, 1, \ldots, N$ is binomially distributed. Since we are interested in a balanced compromise error, it is natural to minimize the squared sum

$$\Phi^2(p_i) = \delta^2(p_i) + \sigma^2(p_i). \tag{30}$$

This quantity measures the total error for a particular value of p_i. Therefore, the average error over the whole range of $p_i \in [0, 1]$ is given by

$$\overline{\Phi_i^2} = \int_0^1 \mathrm{d}p_i w(p_i) \Phi^2(p_i), \tag{31}$$

where $w(p_i)$ is a suitable weight function that should be determined for each specific problem.

We discuss explicitly here the simplest case $w(p_i) \equiv 1$ (obviously, any extra knowledge of the probability values should lead to a non-trivial distribution of weights, resulting in improved results). Inserting equations (28) and (29) into equation (31), the average error is given by

$$\overline{\Phi_i^2} = \int_0^1 \mathrm{d}p_i \left[\left(\sum_{n_i=0}^{N} P_{n_i}(p_i) \chi_{n_i}^2 \right) + h^2(p_i) - 2h(p_i) \left(\sum_{n_i=0}^{N} P_{n_i}(p_i) \chi_{n_i} \right) \right]. \tag{32}$$

Now, we want to determine the numbers χ_{n_i} in such a way that the error given by equation (32) is minimized. Before proceeding, let us make it clear that instead of minimizing the mean-square error for each of the possible states $(i = 1, \ldots, M)$ one could also minimize the total mean-square error defined using equation (9) and (10) rather than equation (28) and (29) to take into account correlations between boxes which, in general, will improve the final result. For example, for binary sequences this can be easily done, and leads to the same result as reported on what follows [19].

As a necessary condition, the partial derivatives

$$\frac{\partial}{\partial \chi_{n_i}} \overline{\Phi_i^2} = 0 \tag{33}$$

have to vanish, i.e.

$$2\int_0^1 \mathrm{d}p_i\, P_{n_i}(p_i)[\chi_{n_i} - h(p_i)] = 0 \tag{34}$$

for all $n_i = 0, 1, \ldots, N$. Therefore, the balanced estimator is defined by the numbers

$$\chi_{n_i}^{\mathrm{bal}} = \frac{\int_0^1 \mathrm{d}p_i\, P_{n_i}(p_i)h(p_i)}{\int_0^1 \mathrm{d}p_i\, P_{n_i}(p_i)} = (N+1)\int_0^1 \mathrm{d}p_i\, P_{n_i}(p_i)h(p_i), \tag{35}$$

where we have explicitly integrated over p_i the binomial distribution.

In the Shannon case, where $h(p_i) = -p_i \ln(p_i)$, the integration can be explicitly carried out, leading to[4]

$$\chi_{n_i} = \frac{n_i+1}{N+2}\sum_{j=n_i+2}^{N+2}\frac{1}{j} \tag{36}$$

so that the final result for the balanced estimator of Shannon entropy is given by

$$\hat{H}_S^{\mathrm{bal}} = \frac{1}{N+2}\sum_{i=1}^{M}\left[(n_i+1)\sum_{j=n_i+2}^{N+2}\frac{1}{j}\right]. \tag{37}$$

Similarly, it is possible to compute χ_{ni} for a power $h(p_i) = p_i^q$, which is the basis for all q-deformed entropies

$$\chi_{n_i}(q) = \frac{\Gamma(N+2)\Gamma(n_i+1+q)}{\Gamma(N+2+q)\Gamma(n_i+1)}. \tag{38}$$

The balanced estimators for Rényi[5] and Tsallis entropy are then given respectively by

$$\hat{H}_R^{\mathrm{bal}}(q) = \frac{1}{1-q}\ln\left[\sum_{i=1}^{M}\chi_{n_i}(q)\right], \tag{39}$$

and

$$\hat{H}_T^{\mathrm{bal}}(q) = \frac{1}{q-1}\left[1 - \sum_{i=1}^{M}\chi_{n_i}(q)\right]. \tag{40}$$

To illustrate the performance of these estimators, let us consider again a binary sequence of N bits $x = 0, 1$ (i.e. $M = 2$) occurring with probabilities $1-p$ and p, respectively. In figure 2 we plot the mean-squared deviation $\Phi^2 = \langle(\hat{H} - H)^2\rangle$ of various estimators from the

[4] The calculation requires using the definition of Harmonic numbers of first-order and binomial coefficients.

[5] Here the influence of the nonlinearity of the logarithm on statistical averages is neglected.

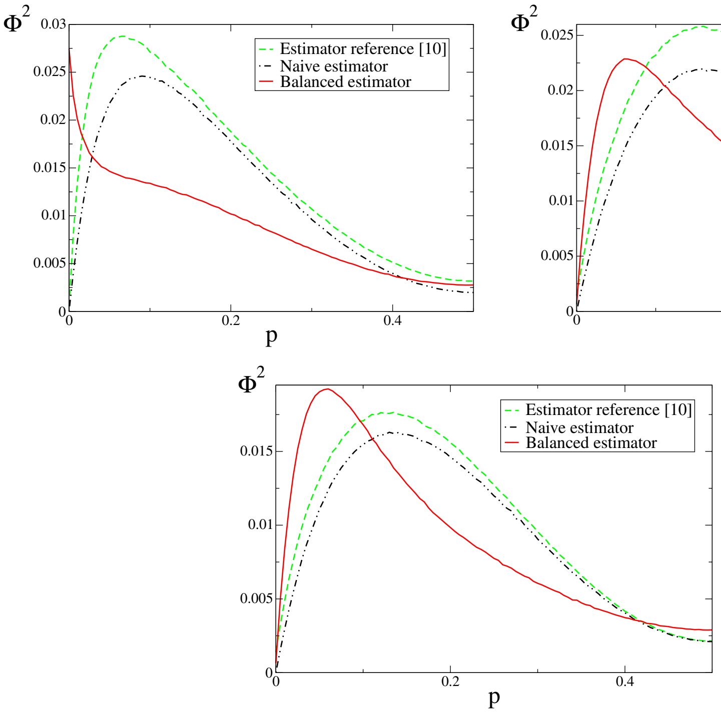

Figure 2. Mean-squared error $\Phi^2 = \langle(\hat{H} - H)^2\rangle$ of different entropy estimators (upper row: Shannon (left); Rényi with $q = 1.5$ (right); lower row: Tsallis with $q = 1.5$) for a binary sequence of $N = 20$, as a function of p. The set of possible values is $\{x_1 = 1, x_2 = 0\}$ and the probabilities, $\{p_1 = p, p_2 = 1 - p\}$, respectively.

true value of the Shannon as well as the Rényi entropy as a function of p. For such a short bit sequence, the performance of Grassberger's estimator *using the parameter* Φ^2, is even worse than the naïve one. This is not surprising since Grassberger's estimator is designed for small probabilities $p_i \ll 1$, while in the present example one of the probabilities p or $1 - p$ is always large and thus the estimator is affected by large fluctuations. The balanced estimator, however, reduces the mean-squared error considerably over an extended range of p while for small p and $0.4 < p < 0.6$ it fails. Similar plots can be obtained for the Tsallis entropy.

The advantage of the balanced estimator compared to standard ones decreases with increasing N. One of the reasons is the circumstance that the fluctuations of the estimator are basically determined by the randomness of n_i and, therefore, are difficult to reduce.

4. Conclusions

We have designed a new 'balanced estimator' for different entropy functionals (Shannon, Rényi and Tsallis) specially adequate for the analysis of small data sets where the possible states appear with not-too-small probabilities. To construct it, first we have illustrated a known result establishing that systematic errors (bias) and statistical errors cannot both be simultaneously reduced to arbitrarily small values when constructing an estimator for a limited data set. In particular, we have designed a low-bias estimator and highlighted that it leads to uncontrolled statistical fluctuations. This hinders the practical usefulness of such a low-bias estimator.

On the other hand, we have designed a new estimator that constitutes a good compromise between minimizing the bias and keeping controlled statistical fluctuations. We have illustrated how this balanced estimator outperforms (in reducing simultaneously bias and fluctuations) previously available ones in special situations the data sets are sufficiently small and the probabilities are not too small. Obviously situations such as in figure 2 are the 'worst case' for estimators like (13) and (16) which were designed to be efficient for large M. If any of these conditions is not fulfilled Grassberger's and Schürmann's estimator remains the best choice.

The balanced method fills a gap in the list of existing entropy estimators, is easy to implement for Shannon, Rényi and Tsallis entropy functional and therefore its potential range applicability is very large, specially in analyses of short biological (DNA, genes, etc) data series.

The balanced estimator proposed here is simple but by no means 'optimal' for two reasons. First, we made no effort to optimize the location of the mesh points p_j, which for simplicity are assumed to be equidistant. Moreover, we did not optimize the weights $w(p_j)$ toward a Bayesian estimate, as e.g. attempted by Wolpert and Wolf [8]. Further effort in this direction would be desirable.

Acknowledgments

We acknowledge financial support from the Spanish Ministerio de Educación y Ciencia (FIS2005-00791) and Junta de Andalucía (FQM-165).

References

[1] Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379

[2] Cover T M and Thomas J A 2006 *Elements of Information Theory* (Canada: Wiley)

[3] Rényi A 1960 *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability* 547
Rényi A 1970 *Probability Theory* (Amsterdam: North Holland)

[4] Tsallis C 1988 *J. Stat. Phys.* **52** 479 see also http://tsallis.cat.cbpf.br/ biblio.htm

[5] Pöschel T, Ebeling W and Rosé H 1995 *J. Stat. Phys.* **80** 1443

[6] Ebeling W and Nicolis G 1992 *Chaos Solitons Fractals* **2** 635
Kantz H and Schürmann T 1996 *Chaos* **6** 167
Dudok de Wit T 1999 *Eur. Phys. J.* B **11** 513

[7] Holste D, Herzel H and Grosse I 1998 *ICCS Proceedings* 589
Holste D, Grosse I and Herzel H 1998 *J. Phys. A: Math. Gen.* **31** 2551
Nemenman I, Bialek W and van Steveninck R R R 2004 *Phys. Rev.* E **69** 056111

[8] Wolpert D H and Wolf D R 1995 *Phys. Rev.* E **52** 6841

[9] Grassberger P 1988 *Phys. Lett.* A **128** 369 (see corrections in [10])

[10] Grassberger P 2003 *Preprint* condmat/0307138 unpublished

[11] Schürmann T 2004 *J. Phys. A: Math. Gen.* **37** L295

[12] Roulston M S 1999 *Physica* D **125** 285

[13] Miller G 1955 *Information Theory in Psychology II-B* ed H Quastler (Glencoe, Illinois: Free Press) p 95
See also Basharin G P 1959 *Theory Probab. Appl.* **4** 333

[14] Harris B 1975 *Topics on Information Theory* vol 323 ed I Csiszar (Amsterdam: North Holland)

[15] Herzel H 1988 *Syst. Anal. Model Sim.* **5** 435

[16] See, for instance, Schürmann T and Grassberger P 1996 *Chaos* **6** 414

[17] Paninski L 2003 *Neural Comput.* **15** 1191

[18] The matrix **P** is well known for biologists, as it is a common transition matrix in population genetics. See, Khazanie R G and McKean H E 1966 *Biometrika* **53** 37
Blythe R A and McKane A J 2007 *J. Stat. Mech.* P07018

[19] Bonachela J A, Hinrichsen H and Muñoz M A 2008 *Preprint*
See also Moak D S 1990 *Proc. Am. Math. Soc.* **108** 1

[20] Abramowitz M and Stegun I (ed) 1965 *Handbook of Mathematica Functions* (New York: Dover)