Chaos, Solitons and Fractals 180 (2024) 114564

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

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

Bayesian estimation of information-theoretic metrics for sparsely sampled distributions

Angelo Piga [a,b,*], Lluc Font-Pomarol [a], Marta Sales-Pardo [a], Roger Guimerà [a,c]

[a] Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona, 43007, Catalonia, Spain
[b] Departament de Física de la Matèria Condensada, Universitat de Barcelona, Barcelona, 08028, Catalonia, Spain
[c] ICREA, Barcelona, 08010, Catalonia, Spain

ARTICLE INFO

Dataset link: https://zenodo.org/records/10592747

Keywords:
Shannon entropy
Kullback–Leibler divergence
Bayesian estimation
Sparse sampling
Information theory

ABSTRACT

Estimating the Shannon entropy of a discrete distribution from which we have only observed a small sample is challenging. Estimating other information-theoretic metrics, such as the Kullback–Leibler divergence between two sparsely sampled discrete distributions, is even harder. Here, we propose a fast, semi-analytical estimator for sparsely sampled distributions. Its derivation is grounded in probabilistic considerations and uses a hierarchical Bayesian approach to extract as much information as possible from the few observations available. Our approach provides estimates of the Shannon entropy with precision at least comparable to the benchmarks we consider, and most often higher; it does so across diverse distributions with very different properties. Our method can also be used to obtain accurate estimates of other information-theoretic metrics, including the notoriously challenging Kullback–Leibler divergence. Here, again, our approach has less bias, overall, than the benchmark estimators we consider.

1. Introduction

Information theory is gaining momentum as a methodological framework to study complex systems. In network science, information theory provides rigorous tools to predict unobserved links [1] and to infer community structure [2]. In neuroscience, Shannon entropy of spike train distributions characterizes brain activity from neural responses [3], while mutual information identifies correlations between brain stimuli and responses [4]. Recently, the Kullback–Leibler divergence [5] and its regularized version, the Jensen–Shannon distance, have also been successfully used in a wide variety of contexts: in neuroscience, to reconstruct structural connectivity from neuronal activity [6]; in cognitive science as a measure of “surprise”, to quantify and predict how human attention is oriented between changing screen images [7]; in quantitative social science, in combination with topic models, to track the propagation of political and social discourses [8,9] or to understand the emergence of social disruption from the analysis of judicial decisions [10]; and in machine learning, at the intersection between the statistical physics of diffusive processes, probabilistic models and deep neural networks [11].

Information theoretical metrics are measured on distributions. In practice, a distribution ρ over the possible states of a system, as well as functions $\mathcal{F}(\rho)$ of this distribution (such as Shannon entropy or other metrics), have to be inferred from experimental observations. However, this inference process is difficult for many real complex systems since we have no information about the underlying generative process of our data and, due to experimental limitations, the observations are often sparse [12], and statistical estimates of the distribution ρ and its functions can be severely biased.[1] Here, we focus on the particular yet important case of discrete (or categorical) distributions ρ_i, $i = 1, \dots, K$, where K is the number of possible states (or categories), which is known and fixed. Inferences about ρ and any function must be based on n_i, the number of observations in the ith state (with $N = \sum_i n_i$ the sample size). We assume that samplings from the underlying distribution are independent and identically distributed (but see, for example, Ref. [13]). In the undersampled regime we are interested in, $N \lesssim K$. The challenge is thus, from the sparse observations $\{n_i\}$, to infer the probability ρ_i of each category i and estimate metrics $\mathcal{F}(\rho)$.

A theoretically well-founded approach to tackle this problem is provided by the principles of conditional probability, encapsulated in Bayes’ theorem [14]. The probabilistic framework is in general preferable because of its transparency; it requires that all assumptions of the underlying generative model for the data are made explicit,

* Corresponding author at: Departament de Física de la Matèria Condensada, Universitat de Barcelona, Barcelona, 08028, Catalonia, Spain.
E-mail addresses: angelo.piga@ub.edu (A. Piga), lluc.font@urv.cat (L. Font-Pomarol), marta.sales@urv.cat (M. Sales-Pardo), roger.guimera@urv.cat (R. Guimerà).

[1] Given an estimator $\tilde{\mathcal{F}}$ for the quantity $\mathcal{F}$, its bias is defined as $\mathbb{E}[\tilde{\mathcal{F}}] - \mathcal{F}_{\text{true}}$, where $\mathbb{E}[\cdot]$ denotes the expected value.

https://doi.org/10.1016/j.chaos.2024.114564
Received 31 October 2023; Received in revised form 31 January 2024; Accepted 2 February 2024
Available online 7 February 2024

expressed via the choice of a likelihood function and a prior distribution that reflects the knowledge about the system before observing any data. In probabilistic reasoning, the combination of observations and prior distribution provides an updated (posterior) probability distribution of the quantity under study. Other estimation strategies make implicit assumptions and often provide only point estimates, as opposed to full distributions.

A class of expressive generative models for categorical distributions amenable to a Bayesian framework is the well-studied family of Dirichlet distributions. A specific infinite mixture of those prior was exploited by Nemenmann, Shafee, and Bialek (henceforth NSB), in their seminal work Ref. [15], to propose a very precise estimator of the Shannon entropy that works for a wide variety of distributions, even in the sparse sampling regime $N \lesssim K$.

An unbiased estimator for Shannon entropy does not exist [16]; however, NSB can be considered the state-of-the-art (among the Bayesian methods but also when compared to non-Bayesian approaches) for estimating entropy in the above-mentioned context [17]. Nevertheless, it is not simple to implement from scratch, and it does not provide estimates for the distribution $\boldsymbol{\rho}$, and in particular for the non-observed ρ_i. For this reason, its applicability is limited to estimating the Shannon entropy and related information theoretic quantities like mutual information and Jensen–Shannon distance, which can be expressed in terms of the former. By contrast, as we discuss below, it cannot be used to estimate the Kullback–Leibler divergence.

To cover this gap, Hausser and Strimmer (henceforth HS) derived a James–Stein-type shrinkage estimator for $\boldsymbol{\rho}$ [17], which has the advantage of being analytical and applicable to any information-theoretic metric, but at the price of making implicit *ad hoc* assumptions, of being more biased than NSB for the Shannon entropy, and of lacking error estimation. Besides NSB and HS, other methods have been proposed over the years to estimate information-theoretic metrics. Here, we limit ourselves to a comparison with these two methods because: (i) they are widely used; and (ii) previous benchmarking has shown that they are typically less biased than other approaches [17].

Here, we propose an alternative semi-analytical estimator that applies to generic discrete distributions and is simple to implement, and overall less biased than to NSB and HS for the diverse distributions we test. Its derivation is grounded in probabilistic considerations, without any *ad hoc* assumptions. We consider Dirichlet generative models and use a hierarchical Bayesian approach to extract as much information as possible from the few observations at hand. In the case of Shannon entropy, we can estimate the expected value and higher order moments with precision at least indistinguishable from the NSB estimator, and often better. Additionally, because our method provides estimates of the probability distribution, it can be used to obtain estimates of the Kullback–Leibler divergence. In this case, our approach also performs equally or better than the HS estimator.

2. Background

Let us consider a system with K possible output states whose observations follow an unknown discrete distribution $\boldsymbol{\rho} = \{\rho_i;\ i = 1, \dots, K\}$ with $\sum_i \rho_i = 1$. The vector $\mathbf{n} = \{n_i;\ i = 1, \dots, K\}$ represents the number of times each state was observed in a set of $\sum_i n_i = N$ independent observations of the system. We also consider a function $\mathcal{F}(\boldsymbol{\rho})$ of $\boldsymbol{\rho}$, such as, for example, the Shannon entropy

$$S(\boldsymbol{\rho}) = -\sum_{i=1}^{K} \rho_i \log \rho_i, \tag{1}$$

which we want to estimate from the set of observations. As noted above, here we assume that K is fixed and known, and that the samples from the distribution are independent and identically distributed.

The posterior distribution over the values of the function $\mathcal{F}$ given the observed counts $\mathbf{n}$ is

$$p(\mathcal{F}|\mathbf{n}) = \int d\boldsymbol{\rho}\ \delta(\mathcal{F} - \mathcal{F}(\boldsymbol{\rho}))\, p(\boldsymbol{\rho}|\mathbf{n}), \tag{2}$$

where $p(\boldsymbol{\rho}|\mathbf{n})$ is the posterior of the distribution $\boldsymbol{\rho}$ given the counts $\mathbf{n}$. We further assume that the prior over distributions depends on a parameter β, which becomes a hyperparameter of our generative model. Then, using the laws of conditional probability, we can write the posterior $p(\boldsymbol{\rho}|\mathbf{n}, \beta)$ as

$$p(\boldsymbol{\rho}|\mathbf{n}, \beta) = \frac{p(\mathbf{n}|\boldsymbol{\rho}, \beta)\, p(\boldsymbol{\rho}|\beta)}{p(\mathbf{n}|\beta)}, \tag{3}$$

where $p(\mathbf{n}|\boldsymbol{\rho}, \beta)$ is the likelihood, $p(\boldsymbol{\rho}|\beta)$ is the prior over distributions, and $p(\mathbf{n}|\beta) = \int d\boldsymbol{\rho}\, p(\mathbf{n}|\boldsymbol{\rho})\, p(\boldsymbol{\rho}|\beta)$ is the evidence and acts as normalization factor. The likelihood is the probability of the empirical observations $\mathbf{n}$ given $\boldsymbol{\rho}$; for independent multinomial samples, the probability of observing an event of type i is ρ_i, and the full likelihood is the product $p(\mathbf{n}|\boldsymbol{\rho}, \beta) = N! \prod_i^K \rho_i^{n_i}/n_i!$ and, given $\boldsymbol{\rho}$, it is independent of the hyperparameter β. The prior $p(\boldsymbol{\rho}|\beta)$ expresses the probability of each distribution $\boldsymbol{\rho}$ prior to observing any data, and plays a crucial role in the discussion below. Symmetric Dirichlet distributions are convenient priors because they are a generative model for a broad class of discrete distributions. Additionally, they have been widely used in this setting [18]; they are parametrized as follows

$$p(\boldsymbol{\rho}|\beta) = \frac{1}{\mathrm{B}_K(\beta)} \prod_{i=1}^{K} \rho_i^{\beta-1}, \quad \mathrm{B}_K(\beta) = \frac{\Gamma(\beta)^K}{\Gamma(\beta K)}, \tag{4}$$

where Γ is the gamma function, while the hyperparameter β is a real, positive number known as the concentration parameter (see Fig. 1, first row, for examples of categorical distributions sampled from symmetric Dirichlet priors).

Besides being very expressive, Dirichlet priors are conjugate distributions of categorical likelihoods, meaning that the posterior is still a Dirichlet distribution, a property that often makes the inference via Eqs. (2) and (3) analytically tractable. For example, when $\mathcal{F}(\boldsymbol{\rho}) = \boldsymbol{\rho}$, Dirichlet priors lead to expected posterior probabilities $\langle \rho_i \rangle$ given by the widely-used generalized Laplace's formula

$$\langle \rho_i \rangle = \frac{n_i + \beta}{N + K\beta}. \tag{5}$$

Note that by taking the limit $\beta \to 0$ (which is equivalent to assuming that the vast majority of observations fall in the same category), we recover the maximum likelihood (or frequency) estimator $\rho_i^{\mathrm{ML}} = n_i/N$. By contrast, for $\beta > 0$, Laplace's formula assigns non-zero probability to non-observed states, a desirable property when estimating Kullback–Leibler divergences for sparse observations, as it will become evident later. This result also illustrates how non-Bayesian approaches to inference make implicit and non-trivial assumptions; in this case, assuming $\beta \to 0$ amounts to assuming that infinitely concentrated distributions $\boldsymbol{\rho}$ are a priori much more plausible than more homogeneous ones.

Going back to the estimation of $\mathcal{F}$ from the observations $\mathbf{n}$, and given Eq. (5), one may be tempted to directly plug the value of $\langle \rho_i \rangle$ in the explicit expression of $\mathcal{F}(\boldsymbol{\rho})$ to get a point estimate. However, this is just an approximation; the exact procedure consists in finding and using the whole posterior $p(\mathcal{F}|\mathbf{n})$. Specifically, the expected value of this posterior $\langle \mathcal{F} \rangle = \int d\mathcal{F}\, \mathcal{F}\, p(\mathcal{F}|\mathbf{n})$ minimizes the mean-squared error [19], and its mode is a consistent estimator, meaning that it converges to the true value of $\mathcal{F}(\boldsymbol{\rho})$ when the number of observations increases, regardless of the prior and, in particular, regardless of the hyperparameter β. Wolpert and Wolf in Refs. [19,20] provided analytical formulas for all the moments of $p(\mathcal{F}|\mathbf{n})$ when $\mathcal{F}$ is the Shannon entropy considering Dirichlet priors (we report the formula for the mean in Eq. (17) and for the second moment in Appendix E).

However, even considering the whole posterior $p(\mathcal{F}|\mathbf{n})$, an unbiased estimation of $\mathcal{F}$ is not guaranteed for small samples. This is often the case for Dirichlet priors, especially when the parameter β is unknown. Several options for assigning a value of β have been proposed in literature, each one suitable to some specific case but deficient in others (for a discussion, refer to Refs. [15,17]). In [15], NSB suggested that, when samples are scarce, any attempt to find a single universal β is hopeless, the fundamental reason being that categorical distributions generated by a Dirichlet have a Shannon entropy that is narrowly determined by, and monotonically dependent on, β. In other words, for small samples, the posterior distribution (2) is dominated by the prior. To overcome this problem, Refs. [15,21] proposed, as the prior $p_{\mathrm{NSB}}(\boldsymbol{\rho})$, an infinite mixture of Dirichlet priors

$$p_{\mathrm{NSB}}(\boldsymbol{\rho}) \propto \int d\beta\, p_{\mathrm{NSB}}(\beta)\, p(\boldsymbol{\rho}|\beta), \tag{6}$$

where the weights $p_{\mathrm{NSB}}(\beta)$ were set so as to obtain a flat prior over entropies S, and have the functional form

$$p_{\mathrm{NSB}}(\beta) \propto \frac{d\,\mathbb{E}[S|n_i = 0, \beta]}{d\beta} = K\psi_1(K\beta + 1) - \psi_1(\beta + 1), \tag{7}$$

where $\mathbb{E}[S|\mathbf{n}, \beta]$ is the expected entropy given the observations $\mathbf{n}$, and then $\mathbb{E}[S|n_i = 0, \beta]$ is the expected entropy of the distributions $\boldsymbol{\rho}$ generated from a symmetric Dirichlet priors (that is if there are no observations), with fixed β and K, and $\psi_m(x) = \left(\frac{d}{dx}\right)^{m+1} \log \Gamma(x)$ are the polygamma functions. The NSB prior leads to very accurate estimates of the Shannon entropy. Even if best suited for situations in which the number of states K is known and fixed, it is quite versatile and has been later extended for countable infinite number of states [22] and further optimized for binary states [23] and long tail distributions [22]. Other estimators, for example the Chao-Shen estimator [24], perform at most as well as the NSB (or its derivatives), but never better (see [17] for a comprehensive review). Additionally, given an estimator of S, a number of other quantities can be indirectly estimated. For example, the mutual information M between two distributions $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ is $M(\boldsymbol{\rho};\boldsymbol{\sigma}) = S(\boldsymbol{\rho}) + S(\boldsymbol{\sigma}) - S(\boldsymbol{\pi})$, where $\boldsymbol{\pi}$ is the joint distribution of $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ [25]. Similar relations can be derived for Jensen–Shannon distance and other information-theoretic quantities [26].[2]

However, consider the estimation of the Kullback–Leibler divergence (D_{KL}) between two distributions $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ with the same dimension K

$$D_{\mathrm{KL}}(\boldsymbol{\rho}\|\boldsymbol{\sigma}) = \sum_{i=1}^{K} \rho_i \log \frac{\rho_i}{\sigma_i}. \tag{8}$$

To estimate D_{KL} from samples $\mathbf{n} = \{n_i;\, i = 1, \dots, K\}$ from $\boldsymbol{\rho}$, and $\mathbf{m} = \{m_i;\, i = 1, \dots, K\}$ from $\boldsymbol{\sigma}$, one cannot use the NSB approach. First, D_{KL} is not a combination of the Shannon entropies of the two underlying distributions $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$. Second, D_{KL} is unbounded, and any attempt to find a hyperprior in the spirit of Eq. (7) results in improper hyperpriors. Finally, with the NSB prior one renounces to any estimation of β and, therefore, to a good a point estimation of D_{KL} by means of Laplace's formula. In fact, lacking a way to directly estimate D_{KL}, we stress that a good estimation of the two involved probability distributions is necessary, and such estimation must also assign non-zero probabilities to non-observed states, to avoid potential singularities arising from the presence of $\boldsymbol{\sigma}$ in the denominator of Eq. (8). Our method, which we present in the next section addresses this issue by using Bayesian reasoning to correctly estimate β from observations.

Hausser and Strimmer proposed and alternative effective and widely used estimator [17]. Specifically, they extended the James–Stein shrinkage estimator [28] to the case of discrete probability distributions. In short, the method consists in finding the target probabilities ρ_i by using a convex combination of the estimates for ρ_i from two extreme models (the uniform and maximum likelihood estimators) as follows:

$$\rho_i^{\mathrm{HS}} = \lambda \frac{1}{K} + (1 - \lambda)\rho_i^{\mathrm{ML}}, \qquad \lambda \in [0, 1]. \tag{9}$$

The maximum likelihood estimator ρ_i^{ML} plays the role of a high-dimensional model that tends to overfit the observations, with a high variance and small bias. By contrast, the uniform distribution $\frac{1}{K}$ assigns equal probability to all states, and corresponds to a low-dimensional model with low variance and high bias. λ is the shrinkage intensity, whose optimum value $\lambda^\star$ balances the two extremes and, after a series of assumptions and simplifications [29] can be calculated as

$$\lambda^\star = \frac{1 - \sum_{i=1}^{K} (\rho_i^{\mathrm{ML}})^2}{(N - 1) \sum_{i=1}^{K} (\frac{1}{K} - \rho_i^{\mathrm{ML}})^2}. \tag{10}$$

Note that the maximum likelihood and the uniform distribution correspond to the two extreme choices of Dirichlet priors with $\beta \to 0$ and $\beta \to \infty$, respectively. The main advantage of the HS estimator lies in its analytical form and in its versatility, since it provides good results for different kinds of experimental data. However, this is at the price of making implicit ad hoc assumptions, of being less precise than NSB for the Shannon entropy, and of lacking error estimation.

3. Hierarchical Bayes point estimate for β

Here, we propose a new approach that addresses these limitations of the NSB and the HS estimators. We posit that the success of the NSB approach stems, not from mixing infinitely many values of the concentration parameter β, but rather from the flexibility to accommodate for *any particular value* of β. Indeed, we surmise that, in general, only a narrow interval of β values are compatible with a given observation $\mathbf{n}$ and therefore contribute to the mixture, whereas most others do not contribute. Motivated by this, we propose an approach that aims to directly estimate the value of β that most contributes to the posterior given the data $\mathbf{n}$.

First, we observe that the posterior $p(\boldsymbol{\rho}|\mathbf{n})$ can be written as

$$\begin{aligned} p(\boldsymbol{\rho}|\mathbf{n}) &= \int d\beta\, p(\boldsymbol{\rho}|\mathbf{n}, \beta)\, p(\beta|\mathbf{n}) \\ &= \int d\beta \frac{p(\mathbf{n}|\boldsymbol{\rho})\, p(\boldsymbol{\rho}|\beta)}{p(\mathbf{n}|\beta)}\, p(\beta|\mathbf{n}), \end{aligned} \tag{11}$$

where we have applied Bayes' rule, and the fact that, as mentioned above, $\mathbf{n}$ conditioned on $\boldsymbol{\rho}$ is independent of β, that is, $p(\mathbf{n}|\boldsymbol{\rho}, \beta) = p(\mathbf{n}|\boldsymbol{\rho})$.

Then, we assume that the conditional distribution $p(\beta|\mathbf{n})$ is very peaked around a given value $\beta^\star$, so that the posterior $p(\boldsymbol{\rho}|\mathbf{n})$ can be approximated as

$$p(\boldsymbol{\rho}|\mathbf{n}) \approx \frac{p(\mathbf{n}|\boldsymbol{\rho})\, p(\boldsymbol{\rho}|\beta^\star)}{p(\mathbf{n}|\beta^\star)}. \tag{12}$$

This approximation, sometimes referred to as *empirical Bayes*, is a point estimate for the fully hierarchical probabilistic model given by $p(\mathbf{n}|\boldsymbol{\rho})$ and $p(\boldsymbol{\rho}|\beta)$. Eq. (12) is identical to Eq. (3), with the difference that the concentration parameter is now the most likely value of β given the observed counts $\mathbf{n}$, that is,

$$\beta^\star = \underset{\beta}{\mathrm{argmax}}\; p(\beta|\mathbf{n}) = \underset{\beta}{\mathrm{argmax}}\; \frac{p(\mathbf{n}|\beta)\, p(\beta)}{p(\mathbf{n})}, \tag{13}$$

where $p(\mathbf{n}|\beta) = \int d\boldsymbol{\rho}\, p(\mathbf{n}|\beta, \boldsymbol{\rho}) p(\boldsymbol{\rho}|\beta)$. For Dirichlet priors as in Eq. (4), β^* satisfies

$$\sum_{i=1}^{K} \sum_{m=0}^{n_i-1} \frac{1}{m + \beta^\star} - \sum_{m=0}^{N-1} \frac{K}{m + K\beta^\star} + \frac{1}{p(\beta^\star)} \frac{d\, p(\beta)}{d\, \beta}\Big|_{\beta^\star} = 0, \tag{14}$$

which is the key analytical result of this paper (see Appendix A for a complete derivation of the equation and for an argument for the uniqueness of $\beta^\star$).

The hyperprior $p(\beta)$ reflects our prior knowledge about the shape of the distribution of the hyperparameter. To be completely agnostic in this regard, we can use a uniform hyperprior

$$p_{\mathrm{U}}(\beta) = \frac{1}{\Delta\beta} = \text{const.}, \quad \Delta\beta = \beta_{\max} - \beta_{\min}, \tag{15}$$

[2] As observed in Refs. [25,27], mutual information can be expressed in terms of different combinations of the Shannon entropy of the two distributions. But its estimations in general differ. The expression $M(\rho;\sigma) = S(\rho) + S(\sigma) - S(\pi)$ seems to be the less biased, however, in the absence of a unique consistent prior over the joint distribution, it is not guaranteed to minimize the mean-squared error.

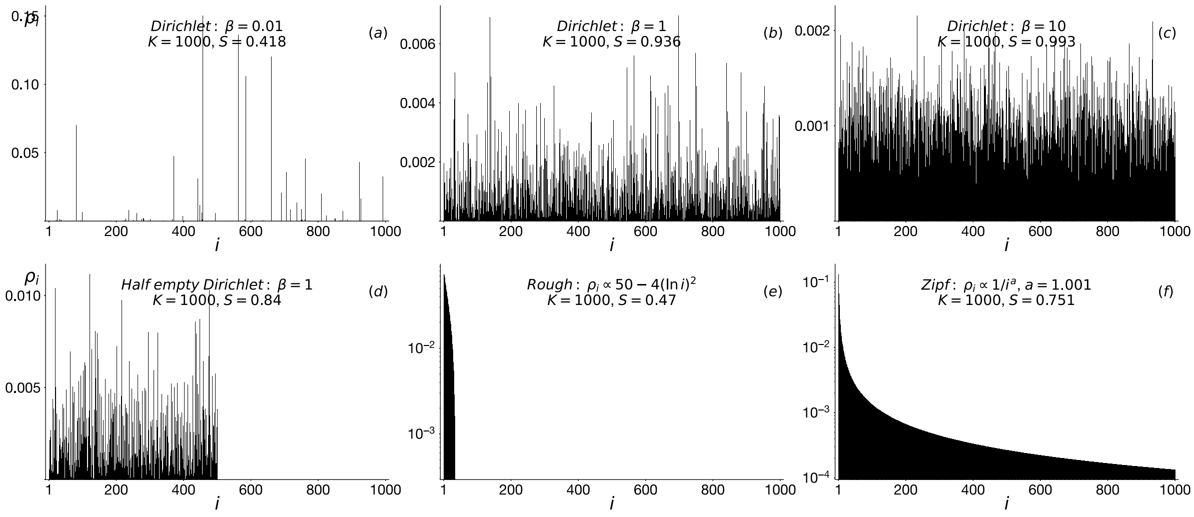

Fig. 1. Examples of synthetically-generated distributions used to test the estimators. First row: three categorical distributions sampled from symmetric Dirichlet with (a) $\beta = 0.01$, (b) $\beta = 1$,(c) $\beta = 10$. Second row: (d) categorical distribution sampled from a uniform Dirichlet with $\beta = 1$, but where half bins are set to zero; (d) fast decaying ("rough") distribution; (e) long tail Zipf's (power law) distribution with exponent $a = 1.001$.

with cut-offs $0 < \beta_{\min} < \beta_{\max} < \infty$. In this case, the derivative term in Eq. (14) disappears. The NSB hyperprior (7) is a valid alternative; in this case, the last term in Eq. (14) is (see Appendix A for details)

$$\frac{1}{p_{\rm NSB}(\beta^*)} \frac{d\, p_{\rm NSB}(\beta)}{d\,\beta}\bigg|_{\beta^*} = \frac{K^2 \psi_2(k\beta^\star + 1) - \psi_2(\beta^\star + 1)}{K\psi_1(k\beta^\star + 1) - \psi_1(\beta^\star + 1)}. \tag{16}$$

Despite the complex appearance of Eq. (14), β^* is not hard to obtain numerically,[3] giving a computational improvement with respect to the NSB estimator, whose algorithm has a higher computational cost. For a discussion about some algorithm details see Appendix B.

4. Results

We validate our method in a variety of scenarios and compare the results with the main alternative available estimators: the NSB [15,21] and the HS [17] for entropy, the HS and the Laplace's estimator (5) with $\beta = 1$ for $D_{\rm KL}$. Our validation includes both synthetically generated distributions and empirical data. Indeed, for the Shannon entropy, many other estimators have been proposed so far: by Miller-Madow [30], Schürman and Grassberger [31], and Grassberger [32], to name a few. However, these have been already carefully compared among them and against HS and SBN in Ref. [17], and showed lower performances. Finally, the Chao-Shen [24] and Valiant [33] estimators are designed for the cases where K is unknown, while for fixed K they do not provide better results than NSB (see Refs. [33,34] for a comparison).

4.1. Synthetic distributions

In our synthetic experiments, we generate target distributions and sample multinomial counts $\{n_i\}$ from those distributions. We fix $K = 1000$ and generate samples of increasing size $N = 20, \ldots, 10000$. After calculating $\beta^\star$ from (14), we estimate the Shannon entropy S and the Kullback–Leibler divergence $D_{\rm KL}$. For each case, we repeat this procedure 1000 times, generating a new distribution ρ each time. We always report averages over these repetitions and, in the case of the entropy, compare the averaged estimation with the average over the ground truth values (labeled $\hat{S}_{\rm true}$ in figures).[4]

As synthetic target distributions we consider both distributions that are *typical* in the Dirichlet prior (that is, they are generated by a symmetric Dirichlet prior; we use several values of concentration parameter $\beta = 0.01, 1, 10$; see first row in Fig. 1) and distributions that are *atypical* in the Dirichlet prior (that is, they have a negligible probability of being generated from a symmetric Dirichlet prior; second row in Fig. 1). Among the latter, we consider: (i) distributions with added structural zeroes: we sample from a symmetric Dirichlet prior with a given β, but half of the categories are then forced to have zero probability[5]; (ii) a distribution with fast-decaying probabilities, $\rho_i \propto 50 - 4(\log i)^2$;[6] (iii) a power-law (Zipf's) distribution: characterized by ranked probabilities $\rho_i \propto i^{-a}$, and a very slow decay with an exponent $a \gtrsim 1$ (long tail distribution; power laws serve as an approximate model in a plethora of contexts, in biological as well as social systems [35]).

To estimate the posterior $p(S|\mathbf{n})$ of the Shannon entropy we use the exact formulas of its moments (derived in Refs. [19,20] and later refined in Ref. [22]) with the estimated values of $\beta^\star$. The first moment is given by

$$\begin{aligned}\mathbb{E}[S|\mathbf{n}, \beta^\star] &= \int d\rho\; S(\rho|\beta^\star)\; p(\rho|\mathbf{n}) \\ &= \psi_0(N + K\beta^\star + 1) \\ &\quad - \sum_{i=1}^{K} \frac{n_i + \beta^\star}{N + K\beta^\star}\, \psi_0(n_i + \beta^\star + 1).\end{aligned} \tag{17}$$

In Appendix E, we also show the expression of the standard deviation.

In practice, given a dataset $\mathbf{n}$ we calculate the most probable $\beta^\star_{\rm f}$ from Eq. (14) by assuming a flat hyperprior, Eq. (15). Then, we compute the required moments of the Shannon entropy, indicated as $S(\beta^\star_{\rm f})$.

In Figs. 2, we show that our estimator is the most accurate estimator overall. In particular, $S(\beta^\star_{\rm f})$ is consistently more accurate than the

[3] The source code for the Python implementations can be accessed via GitHub at https://github.com/angelopiga/info-metric-estimation/ and is permanently archived on Zenodo: https://zenodo.org/records/10592747 (DOI 10.5281/zenodo.10592747).

[4] Averaging on multiple runs is preferable in order to highlight the scaling behaviors of the estimators while mitigating the effects of outliers (for example, very singular distributions or samples).

[5] This scenario corresponds to an experiment in which some states are not observable. Therefore, we are testing the robusteness of the method where K is unknown.

[6] In Refs. [15,21] a rigorous definition of atypicality is provided, related to the decaying rate of the probabilities ρ_i with respect to the rank.

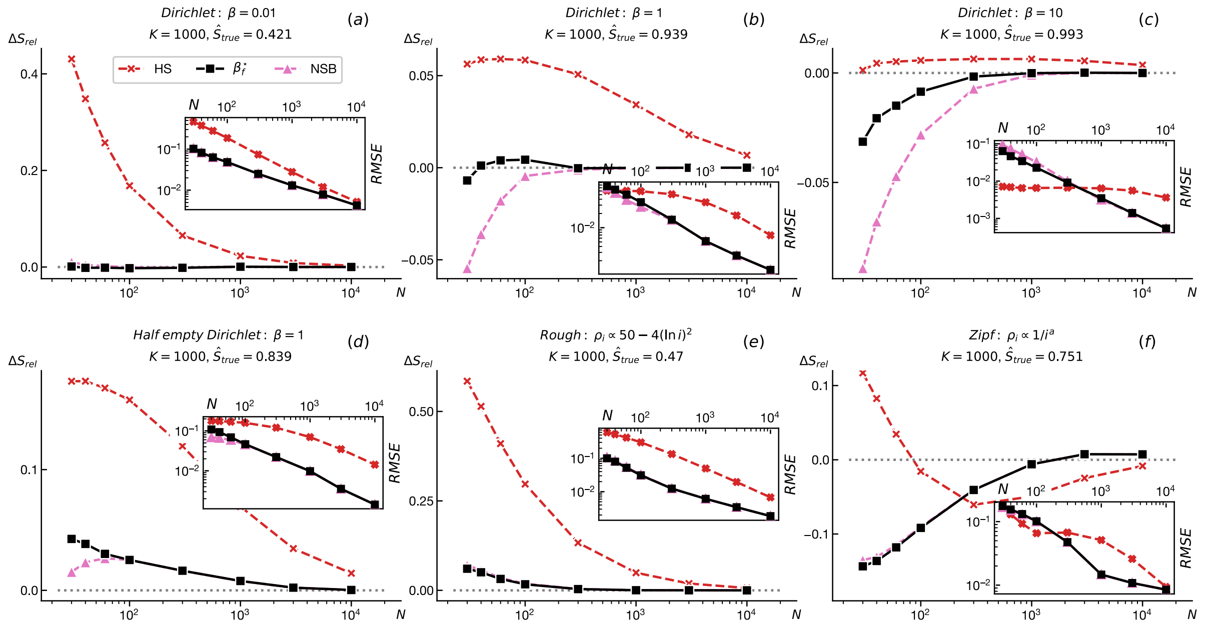

Fig. 2. Shannon entropy estimation for synthetic distributions ((a) to (f)) as in Fig. 1. Each point corresponds to an average over 1000 samples. Main plots: relative errors of entropies $\Delta S_{\mathrm{rel}} = (S_{\mathrm{est}} - \hat{S}_{\mathrm{true}})/\hat{S}_{\mathrm{true}}$, where S_{est} is the estimated entropy with different methods and $\hat{S}_{\mathrm{true}}$ is the average of the true entropies of the 1000 synthetically generated distributions, both measured in nats and divided by $\log(K)$ (so the maximum entropy is 1). Black squares: our estimator with $\beta^\star$ from a flat hyperprior. Pink upper triangle: NSB estimator. Red crosses: Hausser-Strimmer estimator. Insets: roots mean-squared errors (note the logarithmic scale in both axes). The value of $\hat{S}_{\mathrm{true}}$ in the titles serves as a reference and indicates the average over the entropies of the runs. Standard-errors bars of the main plots are smaller then symbols and are not shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

NSB estimator, except in the deep sparse regime $N < 30$ of one of the distributions atypical in the Dirichlet prior (the case with added structural zeroes), where it is comparable but slightly less accurate. The Bayesian estimators also behave better than the HS estimator S_{HS} except for very uniform distributions sampled from the Dirichlet prior with $\beta = 10$. Overall, the $S(\beta^\star_{\mathrm{f}})$ has little bias even in the very sparse regime and for distributions atypical in the Dirichlet prior.[7]

We also analyze the variability of the Shannon entropy estimates, as measured by the root mean squared error $\sqrt{\mathbb{E}[(S - S_{\mathrm{true}})^2]}$ (insets in Figs. 2). This analysis reveals that, besides having less bias, the $S(\beta^\star_{\mathrm{f}})$ estimator has a variability that is typically comparable to or smaller than the other estimators.[8]

Regarding the Kullback–Leibler divergence D_{KL}, there are no exact formulas for the moments of the posterior distribution $p(D_{\mathrm{KL}}|\mathbf{n})$. Therefore, we have to rely on a point estimate of the mean by first estimating the distributions via Laplace's formula Eq. (5) with the inferred $\beta^\star$, and then plugging these values into Ew. (8). The flat hyperprior in Eq. (15) is the only reasonable one to estimate $\beta^\star$ in this case, since the NSB prior can only be justified for the Shannon entropy.

We compare the results with Laplace's estimator Eq. (5) with $\beta = 1$ and with the HS estimator, since both have the same desirable property of assigning non-null probabilities to unobserved states ($n_i = 0$) and are suitable estimators for computing D_{KL}. Indeed, $\beta = 1$ in Laplace's formula is a common choice and amounts to assigning the same probability to all possible distributions. We test the estimators in a scenario typical in machine learning and variational inference, in which one wants to minimize the D_{KL} between a complex, target distribution and some model approximation. Here, after generating a synthetic discrete distribution ρ, we measure the $D_{\mathrm{KL}}(\rho;\hat{\rho})$, where $\hat{\rho}$ is the distribution estimated from counts; hence a good estimator should make D_{KL} as small as possible.

In Fig. 3, we show that our estimator and the HS estimator provide similar results, although $D_{\mathrm{KL}}(\beta^\star_{\mathrm{f}})$ is more accurate in the very sparse regime $N < 50$, and when the target distributions are atypical in the Dirichlet priors, especially in the important case of power-law distributions. The estimator based on Laplace's formula with $\beta = 1$ performs generally worse, unless in the case when the target distribution itself was also generated just from a Dirichlet with $\beta = 1$. Importantly, in this case, where $\beta = 1$ is optimal, our approach provides virtually identical results.

In Appendix D, the results are shown for simulations performed on the same synthetic distributions, but maintaining fixed the sample size N and increasing K. Also in these cases, our estimator outperforms the others.

4.2. Empirical networks

To assess our estimator on real data, we examine the degree distributions of empirical complex networks [38]. In complex network theory, a node's degree represents its number of links (edges), and the degree distribution describes the probabilities ρ_i of each degree i. This distribution offers some insights into the network structural properties. For example, power laws or mixtures of power laws (with distinct exponents for different ranks), sometimes truncated [39], are commonly observed. Particularly the entropy of the degree distribution is a preliminary measure of the complexity of the network [40].

We consider two empirical networks. The first is the network of citations between US patents from 1975 to 1999 [36], consisting of 3,774,768 nodes (patents) and 16,522,438 edges (citations). The second is the brain functional connectome [37], comprising 1,827,242 nodes (functional regions of the brain) and 143,158,339 edges (connections between regions).[9] Fig. 4 shows the degree counts and the

[7] Using our method but with $\beta^\star$ calculated from the NSB hyperprior (Eq. (14) and (16)) is generally worse than that with the flat hyperprior and it is discussed in Appendix C.

[8] It is worth noting that, unlike Bayesian estimators, for which all the moments can be estimated also from a single sample, the HS estimator is limited to a point estimate of the mean value of Shannon entropy.

[9] Both datasets are accessible at: Tiago P. Peixoto, "The Netzschleuder network catalogue and repository", https://networks.skewed.de/ (2020).

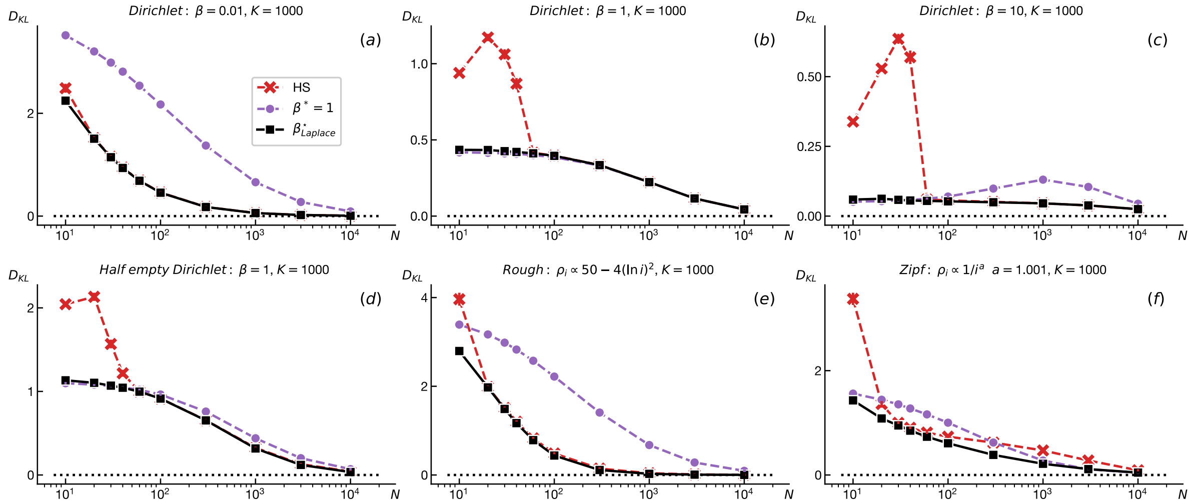

Fig. 3. Kullback–Leibler estimation for synthetic distributions ((a) to (f)) as in Fig. 1. Each point corresponds to an average over 1000 samples. Here, D_{KL} is taken between a given distribution and a second one estimated from a sampling of the former: the target value of D_{KL} is, therefore, $D_{KL}^{\text{true}} = 0$ and the analysis of the RMSE results unnecessary, since it equals the averaged value of D_{KL}. Black squares: our estimator, that is Laplace's formula with $\beta_{\text{f}}^{\star}$ estimated from a flat hyperprior. Red crosses: Hausser-Strimmer estimator. Purple circles: Laplace's estimator for uniform prior $\beta = 1$. Standard-errors bars are often smaller then symbols and are not visible. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

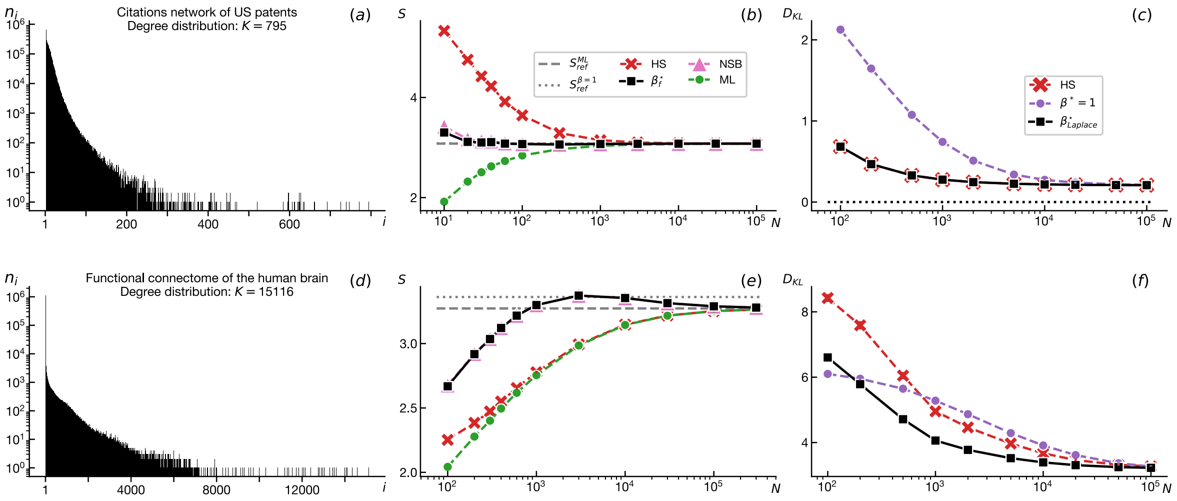

Fig. 4. Empirical networks' degree distributions. First row: US patents citation network from 1975 to 1999 [36] (3.774.768 vertices and 16.522.438 edges). Second row: brain functional connectome [37] (1.827.242 vertices and 143.158.339 edges). In both cases, for entropy as well as Kullback–Leibler divergence, each point corresponds to an average over 1000 samples of N nodes from the original network; the vector $\mathbf{n}$ of their degrees is used to estimate the target quantity (the Shannon entropy or Kullback–Leibler divergence) see also main text. (a), (d): the histograms of nodes' degrees. (b), (e): Shannon entropy. Black squares: our estimator, (17) with $\beta_{\text{f}}^{\star}$ estimated from a flat hyperprior. Red crosses: Hausser-Strimmer estimator. Pink upper triangle: NSB estimator. Green circles: maximum likelihood estimator. (c), (d): Kullbak–Leibler divergence. Black squares: our estimator, that is Laplace's formula with $\beta_{\text{f}}^{\star}$ estimated from a flat hyperprior. Red crosses: Hausser-Strimmer estimator. Purple circles: Laplace's estimator for uniform prior $\beta = 1$. Standard-errors bars are often smaller then symbols and are not visible. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results for the estimation of the Shannon entropy and Kullback–Leibler divergence.

Note that, as we have been discussing, typically only degree counts are available from empirical networks, and the true probability distribution and its metrics remain unknown. Therefore, strictly speaking, there is no "ground truth" in this case. In the absence of such ground truth, as a reference target entropy we use the entropies of probability distributions inferred using the Laplace formula Eq. (5) with $\beta = 1$ (denoted as $S_{\text{ref}}^{\beta=1}$) or with $\beta = 0$ (the frequentist estimation $S_{\text{ref}}^{\text{ML}}$, corresponding to the direct normalization of the histogram). When these two estimates are close, we assume that they are reasonable "ground truths".

Our experiment for entropy involves the following steps: (i) randomly sampling a subset of N nodes from the original networks (for US patents, N ranges from 10 to 10^5; for human brain connectome, N ranges from 100 to 3×10^5); (ii) measuring the degree of each node and constructing a new histogram that represents the vector $\mathbf{n}$ of observations; (iii) estimating the entropy from that histogram.

For the Kullback–Leibler divergence, a further step is necessary after step (ii) and before calculating D_{KL}. In fact, since D_{KL} is defined between two probability distributions (and not between a probability distribution and a set of counts), we first need to define a probability distribution from the original histogram of the degrees, which plays the role of ground truth. Since there is no unambiguous way to do that, for the sake of coherence, for each estimator we infer both distributions using the same estimator under inquiry. That is, in the case of our estimator, we first calculate (step (iii-A)) $\beta_{\text{f}}^{\star}$ from the original histogram to obtain $\rho(\beta_{\text{f}}^{\star})$ from Laplace's formula and hence (step (iii-B)) we do the same for the sample, obtaining $\beta_{\text{f}}^{\star,N}$ and $\rho(\beta_{\text{f}}^{\star,N})$. Finally (step (iii-C)), we can calculate $D_{KL}(\rho(\beta_{\text{f}}^{\star,N}); \rho(\beta_{\text{f}}^{\star}))$. Analogously, for HS estimator both distributions are estimated applying Eq. (9) and (10).

Our estimator performs similar to the NSB estimator and outperforms the HS and ML estimators for Shannon entropy. Regarding the Kullback–Leibler divergence, our estimator shows comparable performance to HS for the US patent network and surpasses HS for the human brain connectome, where the number of bins is larger. These results confirm the findings presented for synthetic networks above and in Appendix D, Fig. D.6, where we compare the different estimators for synthetic distributions with increasing K.

5. Conclusions

We have addressed the question of how to estimate categorical distributions and their information-theoretical metrics, such as the Shannon entropy or the Kullback–Leibler divergence, when only a small number of observations are available. The estimation problem in the sparse sampling regime is, theoretically and experimentally, unavoidable in complex systems [12]. Its rigorous study is then necessary, given the broad use of information-theoretical metrics in physics—especially Shannon entropy and related quantities such as the mutual information—and in data science and machine learning—Kullback–Leibler divergence is at the core of approaches as successful as variational autoencoders or diffusion models, to name just a couple of very prominent examples.

Very few estimators for Kullback–Leibler divergence have been proposed in the literature; they are more abundant for Shannon entropy. However, in both cases they suffer from limitations. First, most existing methods work well for specific contexts but fail in others, because of implicit *ad hoc* assumptions in their derivations. Second, with few exceptions, their application requires numerical algorithms that are difficult to implement from scratch. Third, they often only provide a point-wise estimation, without any estimation of the error. Finding methods that alleviate these drawbacks is crucial for putting analysis methods on solid grounds. Probabilistic approaches are particularly well suited for this purpose [14].

In such a framework, the NSB estimator is still perhaps the most widely used and often the most accurate [15], and has been of inspiration for many subsequent works extending the original method. Crucial in NSB (and in Bayesian analysis, in general) is the choice of the prior distribution, which explicitly expresses expectations about the generation of the data. The NSB prior is a clever mixture of Dirichlet distributions, which were broadly used well before NSB, as they are expressive generative models for discrete distributions; but they gave inconsistent estimations of the Shannon entropy. As pointed out by NSB, the explanation is to be sought in the properties of Dirichlet priors—they are defined by a set of hyperparameters β and, as NSB observed, when samples are scarce, the choice of β narrowly determines the estimation of the entropy. Since a prior-dependent inference is not useful, and in light of the difficulties in determining *a priori* the correct values of β, NSB circumvented the problem by integrating over all possible values of β, which ultimately results in the aforementioned mixture of priors. The impressive results of NSB at estimating Shannon entropy come with the shortcomings of not being able to estimate the probability distribution ρ itself, which might be necessary for other applications such as estimating the Kullback–Leibler divergence.

In this paper, we show that, whereas mixtures of priors are necessary to accommodate for any possible value of the hyperparameter β, in practice, considering a single value $\beta^{\star}$ leads to excellent estimation. As we have shown, this value of the hyperparameter can be found directly, given few observations. Far from being a mere technical point, knowing the hyperparameter $\beta^{\star}$ allows the full specification of the generative model and the estimation of the probability distribution. Importantly, our results still follow from a purely Bayesian framework; more precisely, from a hierarchical probabilistic model, where Bayes' rule is first applied at the higher level for the estimation of the prior parameters. This way, the overall simplicity of the assumptions and transparency of the derivation are preserved. The value of $\beta^{\star}$ is finally provided by a closed formula (Eq. (14)), which is easy to implement and depends only on the vector of observations.

Additionally, as shown by simulations over a variety of distributions, both synthetic and from real-world empirical data, our estimators provide results at least as accurate as, and most times more accurate than, the widely used NSB and HS estimators for Shannon entropy and Kullback–Leibler divergence.

Further study is still necessary to take into account priors other than the symmetric Dirichlet. Although Dirichlet priors are perhaps the most general generative models in the case of discrete distributions, we have observed that certain distributions, particularly in the regime of sparse sampling, are more challenging to estimate compared to “typical” ones under these priors. This is the case of distributions with long tails such as power laws. In specific scenarios where reliable theoretical models of the studied system exist, along with indications regarding the shape of the distributions, specific priors can be hypothesized, as in Refs. [22,24].

Some aspects that may deserve further study include the trade-off between bias and variance in the estimation of information-theoretic metrics, as recently discussed in Ref. [41]. Although a systematic comparison with their method would be illuminating, their estimator (a generalization of Ref. [32]) requires a parameter whose value is not determined without knowing the distribution. Further efforts should also be devoted to relaxing the hypothesis of having a fixed number K of categories, as studied in [24,27,33], or for systems with memory [13], where samples are not independent. Regarding the Kullback–Leibler divergence, further experiments must be conducted to test the estimator over samples from two different distributions, and efforts should be devoted to designing estimators that go beyond the point estimation.

CRediT authorship contribution statement

Angelo Piga: Conceptualization, Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review & editing. **Lluc Font-Pomarol:** Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Marta Sales-Pardo:** Conceptualization, Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review & editing. **Roger Guimerà:** Conceptualization, Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data to replicate simulations are publicy available at https://zenodo.org/records/10592747 (DOI 10.5281/zenodo.10592747).

Acknowledgments

This research was funded by the Social Observatory of the “la Caixa” Foundation as part of the project LCF/PR / SR19/52540009, by MCIN/AEI / 10.13039/501100011033 / FEDER, UE (Projects No. PID2019–106811GB-C31 and No. PID2022-142600NB-I00) and by the Government of Catalonia (Project No. 2017SGR-896).

Appendix A. Derivation of results (Eq. (14) in main text)

Let us suppose that we have K different categories (or types of random events) and that we observe N independent random events distributed in the K categories $\mathbf{n} = \{n_i;\ i = 1, \dots, K\}$, with $\sum_i n_i = N$. We also assume that the probabilities of observing counts in each category ρ_i are distributed according to a Dirichlet prior with the same hyper-parameters β for all $\boldsymbol{\rho} = \{\rho_i;\ i = 1, \dots, K\}$, so that

$$p(\boldsymbol{\rho}|\beta) = \frac{1}{B_K(\beta)} \prod_{i=1}^{K} \rho_i^{\beta-1}, \qquad B_K(\beta) = \frac{\Gamma(\beta)^K}{\Gamma(\beta K)}. \tag{A.1}$$

Our goal is to compute the most likely value of β given the observed counts $\{n_i\}$. To that end, we need to compute the conditional probability $p(\beta|\mathbf{n})$. We can do this by marginalizing over the possible combinations of $\boldsymbol{\rho} = \{\rho_i\}$ as follows:

$$p(\beta|\mathbf{n}) = \frac{p(\beta)}{p(\mathbf{n})} p(\mathbf{n}|\beta), \qquad p(\mathbf{n}|\beta) = \int d\boldsymbol{\rho}\, p(\mathbf{n}|\beta, \boldsymbol{\rho}) p(\boldsymbol{\rho}|\beta). \tag{A.2}$$

Since the probability of observing an event in category i is ρ_i, the probability of observing n_i events of type i is $\rho_i^{n_i}$. Therefore, for the integral in Eq. (A.2) we have that

$$p(\mathbf{n}|\beta, \boldsymbol{\rho}) = N! \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i}, \tag{A.3}$$

so that

$$p(\mathbf{n}|\beta) = \frac{1}{B_K(\beta)} \int d\boldsymbol{\rho} \prod_{i=1}^{K} \rho_i^{n_i+\beta-1}, \tag{A.4}$$

where we have used Eq. (A.1) for $p(\boldsymbol{\rho}|\beta)$ and the integral is over the simplex that satisfies the condition $\sum_i \rho_i = 1$.

To perform the integrals above we first evaluate the normalization condition for $\rho_k = 1 - R_{K-1}$ with $R_{K-1} = \sum_{i=1}^{K-1} \rho_i$ so that for ρ_{k-1} we have the following integral:

$$I_{K-1} = \int_0^{1-R_{K-2}} d\rho_{K-1}\, \rho_{K-1}^{n_{K-1}+\beta-1} \left(1 - \rho_{k-1} - R_{K-2}\right)^{n_K+\beta-1}. \tag{A.5}$$

To evaluate this integral we use the fact that

$$\int_0^{(1-R)} dx\, x^a (1-x-R)^b = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)} (1-R)^{a+b+1} \qquad \text{if} \quad \mathrm{Re}(R) < 1 \quad \text{and} \quad \mathrm{Im}(R) = 0 \tag{A.6}$$

so that

$$I_{K-1} = \frac{\Gamma(n_{K-1}+\beta)\Gamma(n_K+\beta)}{\Gamma(n_k+n_{K-1}+2\beta)} (1-R_{K-2})^{n_K+n_{K-1}+2\beta-1} \tag{A.7}$$

Which gives for ρ_{K-2} the following integral:

$$I_{K-2} = \int_0^{1-R_{K-3}} d\rho_{K-2}\, \rho_{K-2}^{n_{K-2}+\beta-1} \left(1 - \rho_{K-2} - R_{K-3}\right)^{n_K+N_{K-1}+2\beta-1} \tag{A.8}$$

$$= \frac{\Gamma(n_{K-2}+\beta)\Gamma(n_K+n_{K-1}+2\beta)}{\Gamma(n_k+n_{K-1}+n_{k-2}+3\beta)} (1-R_{K-3})^{n_K+n_{K-1}+n_{K-2}+3\beta-1} \tag{A.9}$$

which have evaluated using Eq. (A.6). If we do this for all $\boldsymbol{\rho}$ we end up having

$$\int d\boldsymbol{\rho} \prod_{i=1}^{K} \rho_i^{n_i+\beta-1} = \prod_{i=1}^{K} I_i = \frac{\prod_{i=1}^{K} \Gamma(n_i+\beta)}{\Gamma(N+K\beta)}. \tag{A.10}$$

Thus, we obtain the following expression for $p(\mathbf{n}|\beta)$

$$p(\mathbf{n}|\beta) = \frac{1}{B_K(\beta)} \frac{\prod_{i=1}^{K} \Gamma(n_i+\beta)}{\Gamma(N+K\beta)} = \frac{\Gamma(K\beta)}{\Gamma(\beta)^K} \frac{\prod_{i=1}^{K} \Gamma(n_i+\beta)}{\Gamma(N+K\beta)} \tag{A.11}$$

Our goal is to find $\beta^\star$ that maximizes $p(\beta|\mathbf{n}) = \frac{p(\beta)}{p(\mathbf{n})} p(\mathbf{n}|\beta)$. To that end we take the derivative of $\log p(\beta|\mathbf{n})$,

$$\log p(\beta|\mathbf{n}) = \log \Gamma(K\beta) - K \log \Gamma(\beta) + \sum_{i=1}^{K} \log \Gamma(n_i+\beta)$$

$$- \log \Gamma(N+K\beta) + \log p(\beta) - \log p(\mathbf{n}) \tag{A.12}$$

so that $\beta^\star$ is the one that satisfies the condition:

$$\left.\frac{d \log p(\beta|\mathbf{n})}{d\beta}\right|_{\beta=\beta^\star} = 0. \tag{A.13}$$

To evaluate this equation we use the following definitions and properties of the log Gamma function:

1. $$\left(\frac{d}{dx}\right)^{m+1} \log \Gamma(x) = \psi_m(x) \tag{A.14}$$

2. $$\psi_0(x+n) = \sum_{m=0}^{n-1} \frac{1}{x+m} + \psi_0(x). \tag{A.15}$$

Using the expressions above we obtain that:

$$\frac{d \log p(\beta|\mathbf{n})}{d\beta} = K\psi_0(K\beta) - K\psi_0(\beta) + \sum_{i=1}^{K} \psi_0(n_i+\beta)$$

$$- K\psi_0(N+K\beta) \tag{A.16}$$

$$= \sum_{i=1}^{K} \sum_{m=0}^{n_i-1} \frac{1}{m+\beta} - \sum_{m=0}^{N-1} \frac{K}{m+K\beta} + \frac{1}{p(\beta)} \frac{d\, p(\beta)}{d\, \beta}. \tag{A.17}$$

Therefore the condition that gives $\beta^\star$ is

$$\sum_{i=1}^{K} \sum_{m=0}^{n_i-1} \frac{1}{m+\beta^\star} - \sum_{m=0}^{N-1} \frac{K}{m+K\beta^\star} + \left.\frac{1}{p(\beta^\star)} \frac{d\, p(\beta)}{d\, \beta}\right|_{\beta=\beta^\star} = 0, \tag{A.18}$$

that is, the Eq. (14) in main text. For uniform hyperprior $p_{\mathrm{U}}(\beta) = \text{const.}$ the derivative term $\frac{1}{p(\beta)} \frac{d\, p(\beta)}{d\, \beta}$ disappears. If instead we consider a prior for β that results in a close-to-uniform distribution of Shannon entropy such as in Nemenman et al. [15,21] then

$$p_{\mathrm{NSB}}(\beta) = \frac{d\overline{S}}{d\beta}, \tag{A.19}$$

with $\overline{S} = \mathbb{E}[S|n_i = 0, \beta] = \psi_0(K\beta+1) - \psi_0(\beta+1)$, the average entropy of the distributions generated from a Dirichlet prior $p(\boldsymbol{\rho}|\beta)$. Note that this prior is already normalized since $\int_0^\infty d\overline{S}/d\beta d\beta = \overline{S(\infty; K)} - \overline{S(0; K)} = 1$. The derivative of the logarithm of this prior with respect to β is then

$$\frac{d \log p_{\mathrm{NSB}}(\beta)}{d\beta} = \frac{1}{p_{\mathrm{NSB}}(\beta)} \frac{d p_{\mathrm{NSB}}(\beta)}{d\beta} = \frac{1}{\frac{d\overline{S}}{d\beta}} \frac{d^2\overline{S}}{d\beta^2} = \frac{K^2\psi_2(k\beta+1) - \psi_2(\beta+1)}{K\psi_1(k\beta+1) - \psi_1(\beta+1)}. \tag{A.20}$$

The condition of the $\beta^\star$ that maximizes $p(\beta|n)$ is in this case:

$$\frac{d \log p(\beta|\mathbf{n})}{d\beta} = K\psi_0(K\beta) - K\psi_0(\beta) + \sum_i \psi_0(n_i+\beta)$$

$$- K\psi_0(N+K\beta) + \frac{1}{\frac{d\overline{S}}{d\beta}} \frac{d^2\overline{S}}{d\beta^2} =$$

$$= \sum_{i=1}^{K} \sum_{m=0}^{n_i-1} \frac{1}{m+\beta^\star} - \sum_{m=0}^{N-1} \frac{K}{m+K\beta^\star}$$

$$+ \frac{K^2\psi_2(k\beta^\star+1) - \psi_2(\beta^\star+1)}{K\psi_1(k\beta^\star+1) - \psi_1(\beta^\star+1)} = 0.$$

If the solution $\beta^\star$ exists, it is unique. Consider Eq. (A.18) (Eq. (14) in the manuscript) for flat hyperpriors, which indicates that the values of β that maximize (or minimize) $p(\beta|\mathbf{n})$ satisfy $A(\beta) - B(\beta) = 0$, with

$$A(\beta) = \sum_{i=1}^{K} \sum_{m=0}^{n_i-1} \frac{1}{m+\beta}, \qquad B(\beta) = \sum_{m=0}^{N-1} \frac{K}{m+K\beta}. \tag{A.21}$$

Both functions are non-negative, and decrease monotonically with β because their derivatives are always negative

$$\frac{dA}{d\beta} = -\sum_{i=1}^{K} \sum_{m=0}^{n_i-1} \frac{1}{(m+\beta)^2}, \qquad \frac{dB}{d\beta} = -\sum_{m=0}^{N-1} \frac{K^2}{(m+K\beta)^2}. \tag{A.22}$$

Table B.1
Estimated $\beta^\star$ values for Dirichlet distributions with $K = 1000$, and $\beta = 0.01$ (left) and $\beta = 1$ (right). For each sample size N, we show: (1) the mean $\langle\tilde{\beta^\star}\rangle$ of all the $\beta^\star$ when the algorithm succeeds (a minimum is found), along with the frequency (over 1000 repetitions) with which the algorithm converges to either of the cutoffs ($\%\beta_{\min}, \%\beta_{\max}$); (2) the median Median($\beta^\star$) of all $\beta^\star$s (including cases in which the algorithm converges to a cutoff value); and (3) the mean $\langle\beta^\star\rangle$ including the extreme cutoffs. Cutoffs are set as: $\beta_{\min} = 10^{-7}$, $\beta_{\max} = 10^7$.

N	Dirichlet $\beta = 0.01$			Dirichlet $\beta = 1$		
	$\langle\tilde{\beta^\star}\rangle, (\%\beta_{\min}, \%\beta_{\max})$	Median($\beta^\star$)	$\langle\beta^\star\rangle$	$\langle\tilde{\beta^\star}\rangle, (\%\beta_{\min}, \%\beta_{\max})$	Median($\beta^\star$)	$\langle\beta^\star\rangle$
30	0.010893 (0, 0)	0.009772	0.010893	0.552242 (0, 41.9)	0.735217	4190000.320853
40	0.010433 (0, 0)	0.009564	0.010433	1.734115 (0, 22.8)	3.425872	2280001.338737
60	0.010222 (0, 0)	0.010264	0.010222	2.268786 (0, 14.8)	1.342625	1480001.933006
100	0.010239 (0, 0)	0.009902	0.010239	5.868553 (0, 4.1)	1.075898	410 005.627942
300	0.009896 (0, 0)	0.009884	0.009896	1.068576 (0, 0)	0.996870	1.068576
1000	0.009952 (0, 0)	0.009839	0.009952	1.011258 (0, 0)	1.003212	1.011258
3000	0.009987 (0, 0)	0.009983	0.009987	1.002399 (0, 0)	0.998208	1.002399
10000	0.009982 (0, 0)	0.009935	0.009982	1.002354 (0, 0)	1.000571	1.002354

Additionally, the second derivatives are again non-negative

$$\frac{d^2A}{d\beta^2} = \sum_{i=1}^{K}\sum_{m=0}^{n_i-1}\frac{2}{(m+\beta)^3}, \qquad \frac{d^2B}{d\beta^2} = \sum_{m=0}^{N-1}\frac{2K^3}{(m+K\beta)^3}. \tag{A.23}$$

Therefore, $A(\beta)$ and $B(\beta)$ are both convex, monotonically decreasing functions, which means that they can only cross at one finite value of β, namely $\beta^\star$. It follows that $p(\beta|\mathbf{n})$ is unimodal. This argument is confirmed by numerical simulations.

Appendix B. Algorithm details

The zeros of Eq. (14) can be easily found with a Newton algorithm.[10] In general, both for the flat and the NSB hyperprior, the two extremes, $\beta_{\min}$ and $\beta_{\max}$, of the interval where β is searched must be defined, and can be made arbitrarily small and large, respectively. Importantly, they serve as regularizers when the sample is sparse and it is not possible to find the zero of the target function because of lack of convergence. Even though, as proved in Appendix A, a solution always exists and is unique and, in practice, in sparse cases, the optimal $\beta^\star$ may be arbitrarily large ($\beta^\star \to \infty$) or small ($\beta^\star \to 0$), depending on whether the original distribution was generated by large or small β, respectively. In particular, we found that if Newton algorithm does not converge (i.e., in our implementation, the algorithm gets stuck in the extreme $\beta_{\max}$), therefore one must choose $\beta^\star ! = \beta_{\min}$ (respectively $\beta_{\max}$) whenever the final value of the function is negative (respectively, positive).

In Table B.1, we summarize the estimated $\beta^\star$ for two paradigmatic cases in Fig. 2 of the main text, namely when the original distributions are Dirichlet with $\beta = 0.01$ and $\beta = 1$, respectively, and $\beta_{\min} = 10^{-7}$, $\beta_{\max} = 10^7$. For different sample sizes N, we show: (1) the mean $\langle\tilde{\beta^\star}\rangle$ of all the $\beta^\star$ when the algorithm succeeds (a minimum is found), along with the frequency (over 1000 repetitions) with which the algorithm converges to one of the two extreme cutoffs ($\%\beta_{\min}, \%\beta_{\max}$); (2) the median Median($\beta^\star$) of all $\beta^\star$s (including cases in which the algorithm converges to a cutoff value); and (3) the mean $\langle\beta^\star\rangle$ including the extreme cutoffs. The table shows that the algorithm is robust and $\beta^\star$ approaches the true value as N increases. Furthermore, in the Dirichlet case for $\beta = 0.01$ the algorithm always converges (at least for that particular set of 1000 distributions).

It is worth noting that when the algorithm converges to a cutoff it is typically because the true value of β is fully undetectable. In the case of large β (distributions close to uniform), this happens because each observed sample of the distribution falls in a different bin regardless of the precise value of β. Conversely, in the case of small β (most of the weight in a single bin), this happens because all observed samples of the distribution fall within a single bin, again independently of the precise β. Importantly, however, in either of these cases the entropy is still well approximated because, as shown in [15], high and low generative β values always lead to high (close to 1) and low (close to 0) values of the entropy, regardless of the precise β.

Appendix C. NSB hyperprior for β

In Fig. C.5, we consider the alternative NSB hyperprior (Eq. (7)), extracting $\beta^\star_{\mathrm{NSB}}$ from Eq. (16) and we compare the results, for entropy estimation, with the NSB method and with our method with the flat hyperprior. The data as the same as in Fig. 2 in the main text. Contrary to what one may expect, S_{NSB} differs from our estimate $S(\beta^\star_{\mathrm{NSB}})$ in that the latter generally underestimates entropy for small samples (unless in the case of faster decaying distribution, where it behaves slightly better than all other estimators). This happens because the NSB hyperprior (7) is a monotonically-decreasing distribution that assigns higher probabilities to smaller β's, while the Shannon entropy of distributions sampled from a symmetric Dirichlet is a monotonically-increasing function of β. However, it is not the same estimating $\beta^\star$ with the NSB hyperprior and then plugging it in (17) or directly estimating the Shannon entropy with the NSB prior (6); the latter in fact provides better results.

Appendix D. Results for increasing number of categories

In the main text, experiments on synthetic distributions were conducted with a fixed value of $K = 1000$. A pertinent question is to assess the efficiency of the estimator(s) for higher values of K, while keeping the sample size N constant. Figs. D.6 and D.7 illustrate the scaling of the entropy and Kullback–Leibler divergence estimators, respectively, for $K = 100, \dots, 300\,000$ and $N = 1000$, for the same synthetic distributions as in Fig. 1.

In terms of entropy estimation, our estimator performs similarly to NSB in all cases, except for typical distributions with $\beta = 1$ and $\beta = 10$, where it outperforms NSB by diverging from the target values at larger values of K. However, the HS estimator consistently performs the worst.

Regarding Kullback–Leibler divergence, our estimator and the HS estimator yield similar results in all cases, except for long- and short-tail distributions, where our estimator performs better.

Appendix E. Analytical moments of the Shannon entropy posterior

In the specific case of $S(\rho)$, instead of solving $p(\mathcal{F}|\mathbf{n}) = \int d\rho\, \delta(\mathcal{F} - \mathcal{F}(\rho))\, p(\rho|\mathbf{n})$ (Eq. (2) in main text) directly, it is possible to obtain closed-form expression for all the moments of the posterior [19,20,22]. Here we report the first two, the mean

$$\mathbb{E}[S|\mathbf{n},\beta] = \int d\rho\; S(\rho|\beta)\; p(\rho|\mathbf{n}) \;=\; \psi_0(N + K\beta + 1) - \sum_{i=1}^{K}\frac{n_i+\beta}{N+K\beta}\,\psi_0(n_i+\beta+1), \tag{E.1}$$

[10] The source code for the Python implementations can be accessed via GitHub at https://github.com/angelopiga/info-metric-estimation/ and is permanently archived on Zenodo: https://zenodo.org/records/10592747 (DOI 10.5281/zenodo.10592747).

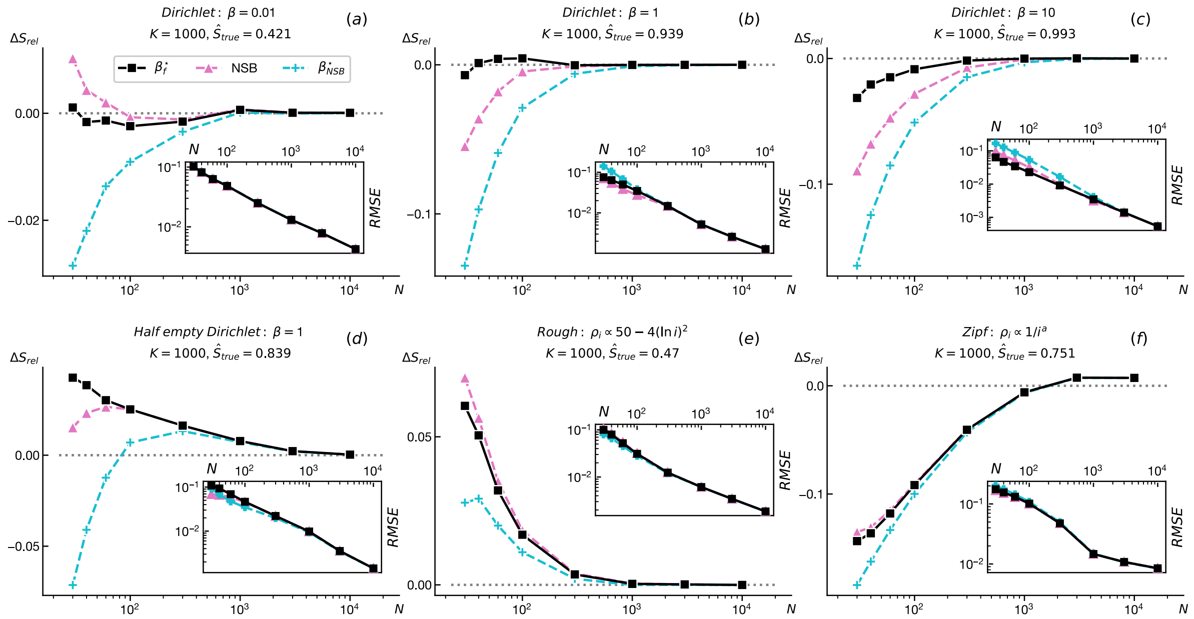

Fig. C.5. Comparison of Shannon entropy estimation with our method but different hyperpriors for β. Same data from synthetic distributions ((a) to (f)) as in Figs. 1 and 2. Each point corresponds to an average over 1000 samples. Main plots: relative errors of entropies $\Delta S_{\mathrm{rel}} = (S_{\mathrm{est}} - \hat{S}_{\mathrm{true}})/\hat{S}_{\mathrm{true}}$, where S_{est} is the estimated entropy with different methods and $\hat{S}_{\mathrm{true}}$ is the average of the true entropies of the 1000 synthetically generated distributions, measured in nats and divided by $\log(K)$ (so the maximum entropy is 1). Black squares: our estimator with $\beta^{\star}$ from a flat hyperprior. Cyan pluses: our estimator but with $\beta^{\star}$ from NSB hyperprior. Insets: roots mean-squared errors (note the logarithmic scale in both axes). The value of $\hat{S}_{\mathrm{true}}$ in the titles serves as a reference and indicates the average over the entropies of the runs. Standard-errors bars of the main plots are smaller then symbols and are not shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

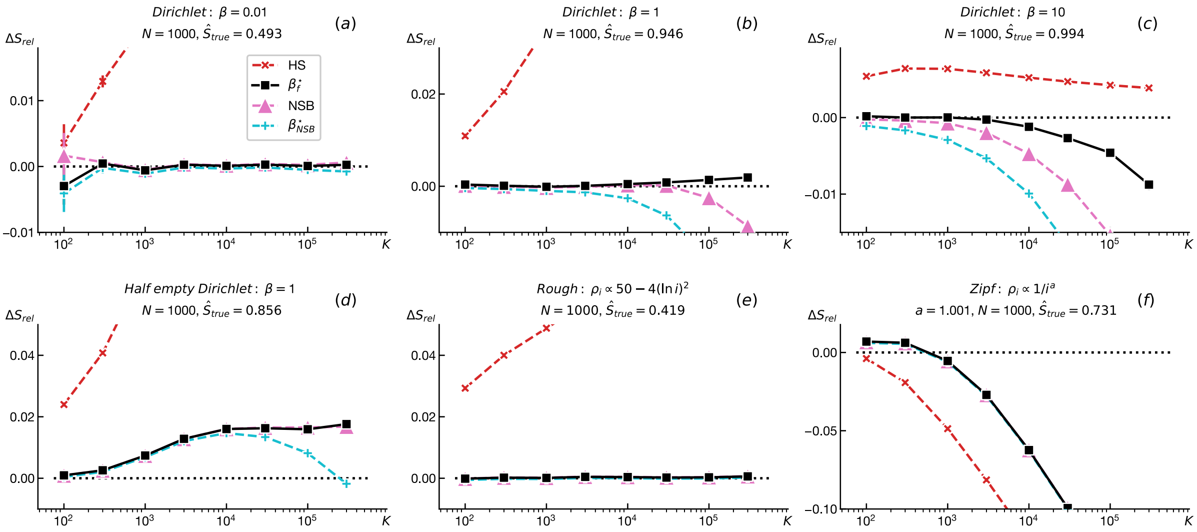

Fig. D.6. Estimation of Shannon entropy for different system sizes, $K = 100, \dots, 300\,000$, and fixed sample size $N = 1000$, for the synthetic distributions ((a) to (f)) as in Fig. 1. Each point corresponds to an average over 1000 samples. Main plots: relative errors of entropies $\Delta S_{\mathrm{rel}} = (S_{\mathrm{est}} - \hat{S}_{\mathrm{true}})/\hat{S}_{\mathrm{true}}$, where S_{est} is the estimated entropy with different methods and $\hat{S}_{\mathrm{true}}$ is the average of the true entropies over all 1000 target distributions, measured in nats and divided by $\log(K)$ (so the maximum entropy is 1). Black squares: our estimator with $\beta^{\star}$ from a flat hyperprior. Cyan pluses: our estimator but with $\beta^{\star}$ from NSB hyperprior. Pink upper triangle: NSB estimator. Red crosses: Hausser-Strimmer estimator. The value of S_{true} in the titles serves as a reference and indicates the average over the entropies of the runs. Standard-errors bars of the main plots are smaller than symbols and are not shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

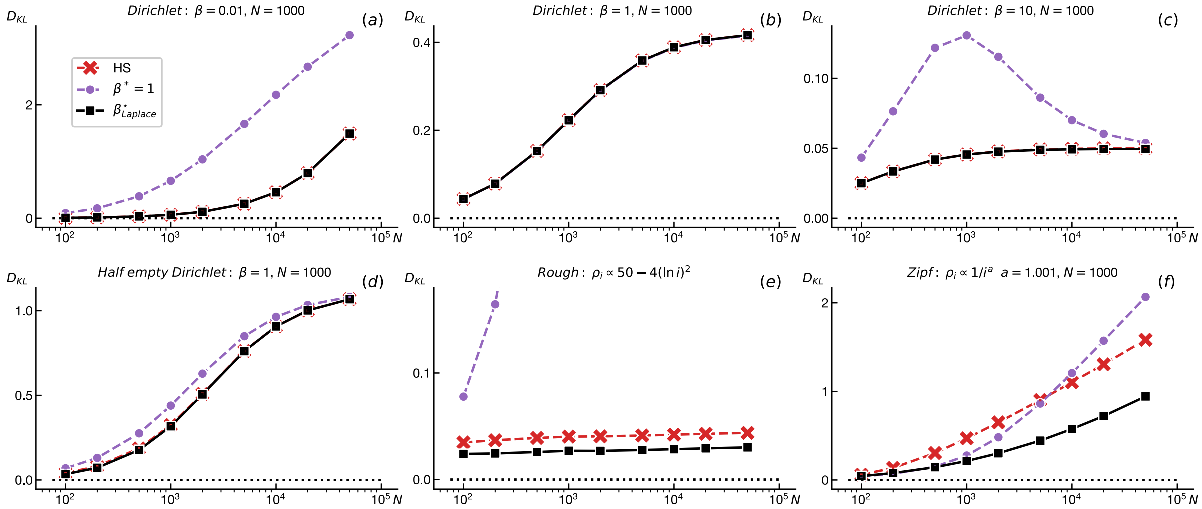

Fig. D.7. Estimation of Kullback–Leibler divergence for different system sizes, $K = 100, \dots, 300\,000$, and fixed sample size $N = 1000$, for synthetic distributions ((a) to (f)) as in Fig. 1. Each point corresponds to an average over 1000 samples. Here, D_{KL} is taken between a given distribution and a second one estimated from a sampling of the former: the target value of D_{KL} is, therefore, $D_{KL}^{\text{true}} = 0$. Black squares: our estimator, that is Laplace's formula with $\beta_f^\star$ estimated from a flat hyperprior. Red crosses: Hausser-Strimmer estimator. Purple circles: Laplace's estimator for a uniform prior with $\beta = 1$. Standard-errors bars are often smaller then symbols and are not visible. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and the second moment

$$\mathbb{E}[S^2|\mathbf{n},\beta] = \int d\rho\; S(\rho|\beta)^2\; p(\rho|\mathbf{n})$$
$$= \sum_{i\neq j}^{K} \frac{(n_i+\beta)(n_j+\beta)}{(N+K\beta+1)(N+K\beta)} I_{i,j}$$
$$+ \sum_{i=1}^{K} \frac{(n_i+\beta+1)(n_i+\beta)}{(N+K\beta+1)(N+K\beta)} J_i, \tag{E.2}$$

with

$$I_{i,j} = \Big(\psi_0(n_i+\beta+1) - \psi_0(N+K\beta+2)\Big)\cdot\Big(\psi_0(n_j+\beta+1) - \psi_0(N+K\beta+2)\Big) - \psi_1(N+K\beta+2);$$

$$J_i = \Big(\psi_0(n_i+\beta+2) - \psi_0(N+K\beta+2)\Big)^2 + \psi_1(n_i+\beta+2) - \psi_1(N+K\beta+2); \tag{E.3}$$

from which the standard deviation is in turn calculated as the square root of the variance $\mathrm{Var}(S|\mathbf{n},\beta) = \mathbb{E}[S^2|\mathbf{n},\beta] - \mathbb{E}[S|\mathbf{n},\beta]^2$.

References

[1] Guimerà R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. Proc Natl Acad Sci 2009;106(52):22073–8.

[2] Peixoto TP. Entropy of stochastic blockmodel ensembles. Phys Rev E 2012;85(5):056122.

[3] Rieke F, Warland D, Van Steveninck RdR, Bialek W. Spikes: exploring the neural code. MIT Press; 1999.

[4] Quian Quiroga R, Panzeri S. Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci 2009;10(3):173–85.

[5] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22(1):79–86.

[6] Orlandi JG, Stetter O, Soriano J, Geisel T, Battaglia D. Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging. PLoS One 2014;9(6):e98842.

[7] Itti L, Baldi P. Bayesian surprise attracts human attention. Vis Res 2009;49(10):1295–306.

[8] Barron AT, Huang J, Spang RL, DeDeo S. Individuals, institutions, and innovation in the debates of the French Revolution. Proc Natl Acad Sci 2018;115(18):4607–12.

[9] Gerlach M, Font-Clos F, Altmann EG. Similarity of symbol frequency distributions with heavy tails. Phys Rev X 2016;6(2):021009.

[10] Font-Pomarol L, Piga A, Teruel-Garcia RM, Nasarre-Aznar S, Sales-Pardo M, Guimerà R. Socially disruptive periods and topics from information-theoretical analysis of judicial decisions. EPJ Data Sci 2023;12. art. no. 2.

[11] Bahri Y, Kadmon J, Pennington J, Schoenholz SS, Sohl-Dickstein J, Ganguli S. Statistical mechanics of deep learning. Ann Rev Condens Matter Phys 2020;11(1).

[12] Levina A, Priesemann V, Zierenberg J. Tackling the subsampling problem to infer collective properties from limited data. Nat Rev Phys 2022;1–15.

[13] De Gregorio J, Sánchez D, Toral R. An improved estimator of Shannon entropy with applications to systems with memory. Chaos Solitons Fractals 2022;165:112797.

[14] Jaynes ET. Probability theory: the logic of science. Cambridge University Press; 2003.

[15] Nemenman I, Shafee F, Bialek W. Entropy and inference, revisited. Adv Neural Inf Process Syst 2001;14.

[16] Paninski L. Estimation of entropy and mutual information. Neural Comput 2003;15(6):1191–253.

[17] Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res 2009;10(7).

[18] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Chapman and Hall/CRC; 1995.

[19] Wolpert DH, Wolf DR. Estimating functions of probability distributions from a finite set of samples. Phys Rev E 1995;52(6):6841.

[20] Wolf DR, Wolpert DH. Estimating functions of distributions from a finite set of samples, part 2: Bayes estimators for mutual information, chi-squared, covariance and other statistics. 1994, arXiv preprint comp-gas/9403002.

[21] Nemenman I, Bialek W, Van Steveninck RDR. Entropy and information in neural spike trains: Progress on the sampling problem. Phys Rev E 2004;69(5):056111.

[22] Archer EW, Park IM, Pillow JW. Bayesian entropy estimation for countable discrete distributions. J Mach Learn Res 2014;15(1):2833–68.

[23] Archer EW, Park IM, Pillow JW. Bayesian entropy estimation for binary spike train data using parametric prior knowledge. Adv Neural Inf Process Syst 2013;26.

[24] Chao A, Shen T-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environ Ecol Stat 2003;10(4):429–43.

[25] Archer EW, Park IM, Pillow JW. Bayesian and quasi-Bayesian estimators for mutual information from discrete data. Entropy 2013;15(5):1738–55.

[26] DeDeo S, Hawkins RX, Klingenstein S, Hitchcock T. Bootstrap methods for the empirical study of decision-making and information flows in social systems. Entropy 2013;15(6):2246–76.

[27] Wolpert DH, DeDeo S. Estimating functions of distributions defined over spaces of unknown size. Entropy 2013;15(11):4668–99.

[28] James W, Stein C. Estimation with quadratic loss. In: Proc. fourth berkeley symp. math. statist. probab.. Vol. 1, California University Press; 1961, p. 361–79.

[29] Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 2005;4(1).

[30] Miller G. Note on the bias of information estimates. Inf Theory Psychol: Problems Methods 1955.

[31] Schürmann T, Grassberger P. Entropy estimation of symbol sequences. Chaos 1996;6(3):414–27.

[32] Grassberger P. Entropy estimates from insufficient samplings. 2003, arXiv preprint physics/0307138.

[33] Valiant G, Valiant P. Estimating the unseen: improved estimators for entropy and other properties. J ACM 2017;64(6):1–41.

[34] Contreras Rodríguez L, Madarro-Capó EJ, Legón-Pérez CM, Rojas O, Sosa-Gómez G. Selecting an effective entropy estimator for short sequences of bits and bytes with maximum entropy. Entropy 2021;23(5):561.

[35] Newman ME. Power laws, Pareto distributions and Zipf's law. Contemp Phys 2005;46(5):323–51.

[36] Hall BH, Jaffe AB, Trajtenberg M. The NBER patent citation data file: Lessons, insights and methodological tools. 2001.

[37] Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, et al. The WU-minn human connectome project: an overview. Neuroimage 2013;80:62–79.

[38] Newman M. Networks. Oxford University Press; 2018.

[39] Amaral LAN, Scala A, Barthélémy M, Stanley HE. Classes of small-world networks. Proc Natl Acad Sci USA 2000;97:11149–52.

[40] Anand K, Bianconi G. Entropy measures for networks: Toward an information theory of complex topologies. Phys Rev E 2009;80(4):045102.

[41] Grassberger P. On generalized Schürmann entropy estimators. Entropy 2022;24(5):680.