


Article

Estimating the Mutual Information between Two Discrete, Asymmetric Variables with Limited Samples

Damián G. Hernández * and Inés Samengo 

Department of Medical Physics, Centro Atómico Bariloche and Instituto Balseiro,
8400 San Carlos de Bariloche, Argentina; ines.samengo@gmail.com

* Correspondence: damian.hernandez@cab.cnea.gov.ar

Received: 3 May 2019; Accepted: 13 June 2019; Published: 25 June 2019



Abstract: Determining the strength of nonlinear, statistical dependencies between two variables is a crucial matter in many research fields. The established measure for quantifying such relations is the mutual information. However, estimating mutual information from limited samples is a challenging task. Since the mutual information is the difference of two entropies, the existing Bayesian estimators of entropy may be used to estimate information. This procedure, however, is still biased in the severely under-sampled regime. Here, we propose an alternative estimator that is applicable to those cases in which the marginal distribution of one of the two variables—the one with minimal entropy—is well sampled. The other variable, as well as the joint and conditional distributions, can be severely undersampled. We obtain a consistent estimator that presents very low bias, outperforming previous methods even when the sampled data contain few coincidences. As with other Bayesian estimators, our proposal focuses on the strength of the interaction between the two variables, without seeking to model the specific way in which they are related. A distinctive property of our method is that the main data statistics determining the amount of mutual information is the inhomogeneity of the conditional distribution of the low-entropy variable in those states in which the large-entropy variable registers coincidences.

Keywords: Bayesian estimation; mutual information; bias; sampling

1. Introduction

Inferring the statistical dependencies between two variables from a few measured samples is a ubiquitous task in many areas of study. Variables are often linked through nonlinear, relations, which contain stochastic components. The standard measure employed to quantify the amount of dependency is the mutual information, defined as the reduction in entropy of one of the variables when conditioning the other variable [1,2]. If the states of the joint distribution are well-sampled, the joint probabilities can be estimated by the observed frequencies. Replacing such estimates in the formula for the mutual information yields the so-called “plug-in” estimator of mutual information. However, unless all the states of the joint distribution are sampled extensively, this procedure typically over-estimates the mutual information [3–5]. In fact, when the number of samples is of the order of the effective number of joint states, even independent variables tend to appear as correlated.

The search for an estimator of mutual information that remains approximately unbiased even with small data samples is an open field of research [6–11]. Here, we focus on discrete variables, and assume it is not possible to overcome the scarceness of samples by grouping states that are close according to some metric. In addition to corrections that only work in the limit of large samples [12], the state of the art for this problem corresponds to quasi-Bayesian methods that estimate mutual information indirectly through measures of the entropies of the involved variables [8,13,14]. These approaches have the drawback of not being strictly Bayesian, since the linear combination of two or more Bayesian estimates

of entropies does not, in general, yield a Bayesian estimator of the combination of entropies [8]. The concern is not so much to remain within theoretical Bayesian purity, but rather to avoid frameworks that may be unnecessarily biased, or where negative estimates of information may arise.

Here, we propose a new method for estimating mutual information that is valid in the specific case in which there is an asymmetry between the two variables: One variable has a large number of effective states, and the other only a few. Examples of asymmetric problems can be found for instance in neuroscience, when assessing the statistical dependence between the activity of a population of neurons and a few stereotyped behaviors, as “lever to the left” or “lever to the right.” If the neural activity is represented as a collection of binary strings, with zeroes and ones tagging the absence or presence of spikes in small time bins, the set of possible responses is huge, and typically remains severely undersampled in electrophysiological experiments. If the behavioral paradigm is formulated in terms of a binary forced choice, the behavioral response only has two possible states.

In our approach, no hypotheses are made about the probability distribution of the large-entropy variable, but the marginal distribution of the low-entropy variable is assumed to be well sampled. The prior is chosen so as to accurately represent the amount of dispersion of the conditional distribution of the low-entropy variable around its marginal distribution. This prior is motivated in the framework of Bayesian estimation of information, and then tested with selected examples in which the hypotheses implied in the prior are fulfilled in varying degree, from high to low. The examples show that our estimator has very low bias, even in the severely under-sampled regime, where there are few coincidences that is, when any given state of the large-entropy variable has a low probability of being sampled more than once. The key data statistics that determine the estimated information is the inhomogeneity of the distribution of the low-entropy variable in those states of the high-entropy variable where two or more samples are observed. In addition to providing a practical algorithm to estimate mutual information, our approach sheds light on the way in which just a few samples reveal those specific properties of the underlying joint probability distribution that determine the amount of mutual information.

2. Bayesian Approaches to the Estimation of Entropies

We seek a low-bias estimate of the mutual information between two discrete variables. Let X be a random variable with a large number k_x of effective states $\{x_1, \dots, x_{k_x}\}$ with probabilities q_x , and Y be a variable that varies in a small set $y \in \{y_1, \dots, y_{k_y}\}$, with $k_y \ll k_x$. Given the conditional probabilities $q_{y|x}$, the marginal and joint probabilities are $q_y = \sum_x q_x q_{y|x}$ and $q_{xy} = q_x q_{y|x}$, respectively. The entropy $H(Y)$ is

$$H(Y) = - \sum_y q_y \log q_y, \quad (1)$$

and can be interpreted as the average number of well-chosen yes/no questions required to guess the sampled value of Y (when using a logarithm of base two). Since the entropy is a function of the probabilities, and not of the actual values taken by the random variable, here we follow the usual notation in which the expressions $H(Y)$ and $H(\mathbf{q}) = H(q_{y_1}, \dots, q_{y_{k_y}})$ are taken to represent the same concept, defined in Equation (1). The conditional entropy $H(Y|X)$ is the average uncertainty of the variable Y once X is known,

$$H(Y|X) = \sum_x q_x \left[- \sum_y q_{y|x} \log q_{y|x} \right] = \sum_x q_x H(Y|x). \quad (2)$$

The mutual information is the reduction in uncertainty of one variable once we know the other [2]

$$I(X, Y) = H(X) + H(Y) - H(X, Y) = H(Y) - H(Y|X). \quad (3)$$

Our aim is to estimate $I(X, Y)$ when Y is well sampled, but X is severely undersampled, in particular, when the sampled data contain few coincidences in X . Hence, for most values x ,

the number of samples n_x is too small to estimate the conditional probability $q_{y|x}$ from the frequencies n_{xy}/n_x . The plug-in estimators of entropy, conditional entropy and mutual information are defined as the ones in which all the probabilities appearing in Equations (1)–(3) are estimated naively by the frequencies obtained in a given sample. In fact, when $n_x \sim \mathcal{O}(1)$, the plug-in estimator typically underestimates $H(Y|x)$ severely [5], and often leads to an overestimation of $I(X, Y)$.

One possibility is to estimate $H(X)$, $H(Y)$ and $H(X, Y)$ using a Bayesian estimator, and then insert the obtained values in Equation (3) to estimate the mutual information. We now discuss previous approaches to Bayesian estimators for entropy, to later analyze the case of information. For definiteness, we focus on $H(X)$, but the same logic applies to $H(Y)$, or $H(X, Y)$.

We seek the Bayesian estimator of the entropy conditional to having sampled the state x_i a number n_i of times. We denote the collection of sampled data as a vector $\mathbf{n} = (n_1, \dots, n_{k_x})$, and the total number of samples $N = \sum_i n_i$. The Bayesian estimator of the entropy is a function that maps each vector \mathbf{n} on a non-negative real number $\hat{H}(\mathbf{n})$, in such a way that the posterior expected error is minimized [15,16]. In the Bayesian framework, the sampled data \mathbf{n} are known, whereas the underlying distribution \mathbf{q} originating the data is unknown, and must be inferred. In this framework, the estimator of the entropy is the function $\hat{H}(\mathbf{n})$ for which

$$\langle [\hat{H}(\mathbf{n}) - H]^2 \rangle = \int [\hat{H}(\mathbf{n}) - H(\mathbf{q})]^2 p(\mathbf{q}|\mathbf{n}) d\mathbf{q} \quad (4)$$

is minimized. The Bayesian nature of this framework is embodied in the fact that the integral weighs all candidate distributions \mathbf{q} that could have originated the data with the posterior distribution $p(\mathbf{q}|\mathbf{n})$, which can be related to the multinomial distribution $p(\mathbf{n}|\mathbf{q})$ by means of Bayes rule. The estimator $\hat{H}(\mathbf{n})$ can be found by taking the derivative of Equation (4) with respect to \hat{H} and equating to zero. The result is [17]

$$\begin{aligned} \hat{H}(\mathbf{n}) = \langle H|\mathbf{n} \rangle &= \int H(\mathbf{q}) p(\mathbf{q}|\mathbf{n}) d\mathbf{q} \\ &= [p(\mathbf{n})]^{-1} \int H(\mathbf{q}) p(\mathbf{n}|\mathbf{q}) p(\mathbf{q}) d\mathbf{q}, \end{aligned} \quad (5)$$

where the first equation results from the minimization procedure, and the second, from using Bayes rule. As a result, the Bayesian estimator is the expected value of $H(\mathbf{q}|\mathbf{n})$.

Since $p(\mathbf{n}|\mathbf{q})$ is the multinomial distribution

$$p(\mathbf{n}|\mathbf{q}) = N! \prod_x \frac{q_x^{n_x}}{n_x!}, \quad (6)$$

and since the normalization constant $p(\mathbf{n})$ can be calculated from the integral

$$p(\mathbf{n}) = \int d\mathbf{q} p(\mathbf{n}|\mathbf{q}) p(\mathbf{q}), \quad (7)$$

the entire gist of the Bayesian approach is to find an adequate prior $p(\mathbf{q})$ to plug into Equations (5) and (7). For the sake of analytical tractability, $p(\mathbf{q})$ is often decomposed into a weighted combination of distributions $p(\mathbf{q}|\beta)$ that can be easily integrated, each tagged by one or a few parameters, here generically called β that vary within a certain domain,

$$p(\mathbf{q}) = \int d\beta p(\beta) p(\mathbf{q}|\beta). \quad (8)$$

The decomposition requires to introduce a prior $p(\beta)$. Hence, the former search for an adequate prior $p(\mathbf{q})$ is now replaced by the search for an adequate prior $p(\beta)$. The replacement implies an assumption and also a simplification. The family of priors that can be generated by Equation (8) does not necessarily encompass the entire space of possible priors. The decomposition relies on the assumption that the remaining family is still rich enough to make good inference about the quantity of

interest, in this case, the entropy. The simplification stems from the fact that the search for $p(\beta)$ is more restricted than the search for $p(\mathbf{q})$ because the space of alternatives is smaller (the dimensionality of \mathbf{q} is typically high, whereas the one of β is low). Two popular proposals of Bayesian estimators for entropies are Nemenman–Shafee–Bialek (NSB) [13] and Pitman–Yor Mixture (PYM) [14]. In NSB, the functions $p(\mathbf{q}|\beta)$ are Dirichlet distributions, in which β takes the role of a concentration parameter. In PYM, these functions are Pitman–Yor processes, and β stands for two parameters: one accounting for the concentration, and the other for the so-called discount. In both cases, the Bayesian machinery implies

$$\langle H|\mathbf{n} \rangle = \frac{1}{p(\mathbf{n})} \int d\beta p(\beta) W(\beta|\mathbf{n}), \tag{9}$$

where $W(\beta|\mathbf{n})$ is the weight of each β in the estimation of the expected entropy

$$W(\beta|\mathbf{n}) = \int d\mathbf{q} H(\mathbf{q}) p(\mathbf{n}|\mathbf{q}) p(\mathbf{q}|\beta). \tag{10}$$

When choosing the family of functions $p(\mathbf{q}|\beta)$, it is convenient to select them in such a way that the weight $W(\beta|\mathbf{n})$ can be solved analytically. However, this is not the only requirement. In order to calculate the integral in β , the prior $p(\beta)$ also plays a role. The decomposition of Equation (8) becomes most useful when the arbitrariness in the choice of $p(\beta)$ is less serious than the arbitrariness in the choice of $p(\mathbf{q})$. This assumption is justified when $W(\beta|\mathbf{n})$ is peaked around a specific β -value, so that, in practice, the shape of $p(\beta)$ hardly has an effect. In these cases, a narrow range of relevant β -values is selected by the sampled data, and all assumptions about the prior probability outside this range play a minor role. For the choices of the families $p(\mathbf{q}|\beta)$ proposed by NSB and PYM, $W(\beta|\mathbf{n})$ can be calculated analytically, and one can verify that, indeed, a few coincidences in the data suffice for a peak to develop. In both cases, the selected β is one for which $p(\mathbf{q}|\beta)$ favours a range of \mathbf{q} values that are compatible with the measured data (as assessed by $p(\mathbf{n}|\mathbf{q})$), and also produce non-negligible entropies (Equation (10)).

When the chosen Bayesian estimates of the entropies are plugged into Equation (3) to obtain an estimate of the information, each term is dominated by its own preferred β . Since the different entropies are estimated independently, the β values selected by the data to dominate the priors $p(q_x)$ and $p(q_y)$ need not be compatible with the ones dominating the priors of the joint or the conditional distributions. As a consequence, the estimation of the mutual information is no longer Bayesian, and can suffer from theoretical issues, as, for example, yield negative estimates [8].

A first alternative would be to consider an integrable prior containing a single β for the joint probability distribution q_{xy} , and then replace H by I in the equations above, to calculate $\langle I \rangle$. This procedure was tested by Archer et al. [8], and the results were only good when the collection of q_{xy} values governing the data were well described by a distribution that was contained in the family of proposed priors $p(\mathbf{q}|\beta)$. The authors concluded that mixtures of Dirichlet priors do not provide a flexible enough family of priors for highly-structured joint distributions, at least for the purpose of estimating mutual information.

To make progress, we note that $I(X, Y)$ can be written as

$$I(X; Y) = \sum_x q_x \sum_y q_{y|x} \log \frac{q_{y|x}}{q_y} = \sum_x q_x D_{\text{KL}}(\mathbf{q}_{y|x} || \mathbf{q}_y), \tag{11}$$

where $\mathbf{q}_{y|x}$ and \mathbf{q}_y stand for the k_y -dimensional vectors $(q_{y_1|x}, \dots, q_{y_{k_y}|x})$ and $(q_{y_1}, \dots, q_{y_{k_y}})$, and D_{KL} represents the Kullback–Leibler divergence. The average divergence between $\mathbf{q}_{y|x}$ and \mathbf{q}_y captures a notion of spread. Therefore, the mutual information is sensitive not so much to the value of the probabilities $q_{y|x}$, but rather, to their degree of scatter around the marginal q_y . The parameters controlling the prior should hence be selected in order to match the width of the distribution of $q_{y|x}$ values, and not so much each probability. With this intuition in mind, in this paper, we put forward

a new prior for the whole ensemble of conditional probabilities $q_{y|x}$ obtained for different x values. In this prior, the parameter β controls the spread of the conditionals $q_{y|x}$ around the marginal q_y .

3. A Prior Distribution for the Conditional Entropies

Our approach is valid when the total number of samples N is at least of the order of magnitude of $\sqrt{e^{H(X)}}$, since in this regime, some of the x states are expected to be sampled more than once [18,19]. In addition, the marginal distribution q_y must be well sampled. This regime is typically achieved when X has a much larger set of available states than Y . In this case, the maximum likelihood estimators \hat{q}_y of the marginal probabilities q_y can be assumed to be accurate that is,

$$\hat{q}_y = \frac{n_y}{N} \approx q_y, \quad \forall y. \tag{12}$$

In this paper, we put forward a Dirichlet prior distribution centered at \hat{q}_y that is,

$$\begin{aligned} p(\{\mathbf{q}_{y|x}\}|\beta) &= \Gamma(\beta)^{k_x} \prod_{xy} \frac{q_{y|x}^{\beta \hat{q}_y - 1}}{\Gamma(\beta \hat{q}_y)} \\ &= \left[\frac{\Gamma(\beta)}{\prod_y \Gamma(\beta \hat{q}_y)} \right]^{k_x} e^{-\beta k_x H(\hat{\mathbf{q}}_y)} \frac{\exp[-\beta \sum_{xy} q_y (\log q_y - \log q_{y|x})]}{\prod_{xy} q_{y|x}} \\ &\propto \frac{\exp[-\beta \sum_x D_{KL}(\hat{\mathbf{q}}_y || \mathbf{q}_{y|x})]}{\prod_{xy} q_{y|x}}, \end{aligned} \tag{13}$$

where $\{\mathbf{q}_{y|x}\}$ contains the k_x conditional probabilities $\mathbf{q}_{y|x}$ corresponding to different x values. Large β values select conditional probabilities close to \hat{q}_y , while small values imply a large spread that pushes the selection towards the border of the k_y -simplex.

For the moment, for simplicity, we work with a prior $p(\{\mathbf{q}_{y|x}\})$ defined on the conditional probabilities $q_{y|x}$, and make no effort to model the prior probability of \mathbf{q}_x . In practice, we estimate the values of q_x with the maximum likelihood estimator $\hat{q}_x = n_x / N$. Since X is assumed to be severely undersampled, this is a poor procedure to estimate q_x . Still, the effect on the mutual information turns out to be negligible, since the only role of q_x in Equation (11) is to weigh each of the Kullback–Leibler divergences appearing in the average. If k_x is large, each D_{KL} -value appears in several terms of the sum, rendering the individual value of the accompanying q_x irrelevant, only the sum of all those with the same D_{KL} matters. In Section 6, we tackle the full problem of making Bayesian inference both in \mathbf{q}_x and $\{\mathbf{q}_{y|x}\}$.

The choice of prior of Equation (13) is inspired in three facts. First, β captures the spread of $q_{y|x}$ around q_y , as implied by the Kullback–Leibler divergence in Equation (13). Admittedly, this divergence is not exactly the one governing the mutual information (Equation (11)), since $\mathbf{q}_{y|x}$ and \mathbf{q}_y are swapped. Yet, it is still a measure of spread. The exchange, as well as the denominator in Equation (13), were introduced for the sake of the second fact, namely, analytical tractability. The third fact regards the emergence of a single relevant β when the sampled data begin to register coincidences. If we follow the Bayesian rationale of the previous section, now replacing the entropy by the mutual information, we can again define a weight $W(\beta|\mathbf{n})$ for the parameter β

$$\begin{aligned} W(\beta|\mathbf{n}) &= \int \{d\mathbf{q}_{y|x}\} I(\hat{\mathbf{q}}_x, \{\mathbf{q}_{y|x}\}) p(\mathbf{n}|\hat{\mathbf{q}}_x, \{\mathbf{q}_{y|x}\}) p(\{\mathbf{q}_{y|x}\}|\beta) \\ &= p(\beta|\mathbf{n}) F(\beta, \mathbf{n}), \end{aligned}$$

where $F(\beta, \mathbf{n})$ can be obtained analytically, and is a well behaved function of its arguments, whereas

$$\begin{aligned}
 p(\beta|\mathbf{n}) &= \frac{p(\beta)p(\mathbf{n}|\beta)}{p(\mathbf{n})} = \frac{p(\beta)}{p(\mathbf{n})} \int \{d\mathbf{q}_{y|x}\} p(\mathbf{n}|\hat{\mathbf{q}}_x, \{\mathbf{q}_{y|x}\}) p(\{\mathbf{q}_{y|x}\}|\beta) \\
 &= \frac{p(\beta)}{p(\mathbf{n})} \prod_x \frac{\Gamma(\beta)}{\Gamma(n_x+\beta)} \prod_{y=1}^{k_y} \frac{\Gamma(n_{xy}+\beta\hat{q}_y)}{\Gamma(\beta\hat{q}_y)}.
 \end{aligned}
 \tag{14}$$

For each x , the vector $\mathbf{q}_{y|x}$ varies in a k_y -dimensional simplex. For $p(\mathbf{n}|\hat{\mathbf{q}}_x, \{\mathbf{q}_{y|x}\})$ we take the multinomial

$$p(\mathbf{n}|\hat{\mathbf{q}}_x, \{\mathbf{q}_{y|x}\}) = N! \prod_{xy} \frac{[\hat{q}_x q_{y|x}]^{n_{xy}}}{n_{xy}!}.
 \tag{15}$$

The important point here is that the ratio of the Gamma functions of Equation (14) develops a peak in β as soon as the collected data register a few coincidences in x . Hence, with few samples, the prior proposed in Equation (13) renders the choice of $p(\beta)$ inconsequential.

Assuming that the marginal probability of Y is well-sampled, the entropy $H(Y)$ is well approximated by the plug-in estimator $\hat{H}(Y) = -\sum_y (n_y/N) \log(n_y/N)$. For each β , the expected posterior information can be calculated analytically,

$$\langle I|\mathbf{n}, \beta \rangle = \hat{H}(Y) - \sum_x \frac{n_x}{N} \left[\psi_0(\beta + n_x + 1) - \sum_y \frac{\beta\hat{q}_y + n_{xy}}{\beta + n_x} \psi_0(\beta\hat{q}_y + n_{xy} + 1) \right],
 \tag{16}$$

where ψ_0 is the digamma function. A code implementing the estimate of Equation (16) is publicly available in the site mentioned in the Supplementary Material below.

When the system is well sampled, $n_{xy} \gg 1$, so the effect of β becomes negligible, the digamma functions tend to logarithms, and the frequencies match the probabilities. In this limit, Equation (16) coincides with the plug-in estimator, which is consistent [20]. As a consequence, our estimator is also consistent. The rest of the paper focuses on the case in which the marginal probability of X is severely undersampled.

4. A Closer Look on the Case of a Symmetric and Binary Y -Variable

In this section, we analyze the case of a binary Y -variable, which, for simplicity, is assumed to be symmetric that is, $q_{y=0} = q_{y=1} = 1/2$, such that $H(Y) = \log 2$ nats. In this case, the Dirichlet prior in each factor of Equation (13) becomes a Beta distribution

$$p(q_{1|x}|\beta) = \frac{\Gamma(\beta)}{\Gamma(\beta/2)^2} [q_{1|x}(1 - q_{1|x})]^{\beta/2-1},
 \tag{17}$$

and $p(\{q_{1|x}\}|\beta) = \prod_x p(q_{1|x}|\beta)$. Large values of β mostly select conditional probabilities $q_{y|x}$ close to $1/2$. If all conditional probabilities are similar, and similar to the marginal, the mutual information is low, since the probability of sampling a specific y value hardly depends on x . Instead, small values of β produce conditional probabilities $q_{y|x}$ around the borders ($q_{y|x} \sim 0$ or $q_{y|x} \sim 1$). In this case, $q_{y|x}$ is strongly dependent on x (see Figure 1b), so the mutual information is large. The expected prior mutual information $\langle I|\mathbf{n} = \mathbf{0}, \beta \rangle$ can be calculated using the analytical approach developed by [14,17],

$$\langle I|\mathbf{n} = \mathbf{0}, \beta \rangle = \log 2 - \psi_0(\beta + 1) + \psi_0(\beta/2 + 1).
 \tag{18}$$

The prior information is a slowly-varying function of the order of magnitude of β , namely of $\log \beta$. Therefore, if a uniform prior in information is desired, it suffices to choose a prior on $\log \beta$ such that $p(\log \beta) \propto |\partial_{\log \beta} \langle I | \mathbf{n} = \mathbf{0}, \beta \rangle|$,

$$p(\log \beta) = \frac{\beta/2}{\log 2} |2\psi_1(\beta + 1) - \psi_1(\beta/2 + 1)|. \tag{19}$$

When $k_y = 2$, the expected posterior information (Equation (16)) becomes

$$\langle I | \mathbf{n}, \beta \rangle = \hat{H}(Y) - \sum_x \frac{n_x}{N} \left[\psi_0(\beta + n_x + 1) - \sum_{y \in \{0,1\}} \left(\frac{\beta/2 + n_{xy}}{\beta + n_x} \right) \psi_0(\beta/2 + n_{xy} + 1) \right]. \tag{20}$$

The marginal likelihood of the data given β is also analytically tractable. The likelihood is binomial for each x , so

$$\begin{aligned} p(\mathbf{n} | \beta) &= \prod_x \int_0^1 dq_{1|x} p(n_{x1}, n_{x0} | q_{1|x}) p(q_{1|x} | \beta) \\ &\propto \prod_x \frac{\Gamma(n_{x1} + \beta/2) \Gamma(n_{x0} + \beta/2) \Gamma(\beta)}{\Gamma(n_x + \beta) \Gamma(\beta/2)^2}. \end{aligned} \tag{21}$$

The posterior for β can be obtained by adding a prior $p(\beta)$, as $p(\beta | \mathbf{n}) \propto p(\mathbf{n} | \beta) p(\beta)$. The role of the prior becomes relevant when the number of coincidences is too low for the posterior to develop a peak (see below).

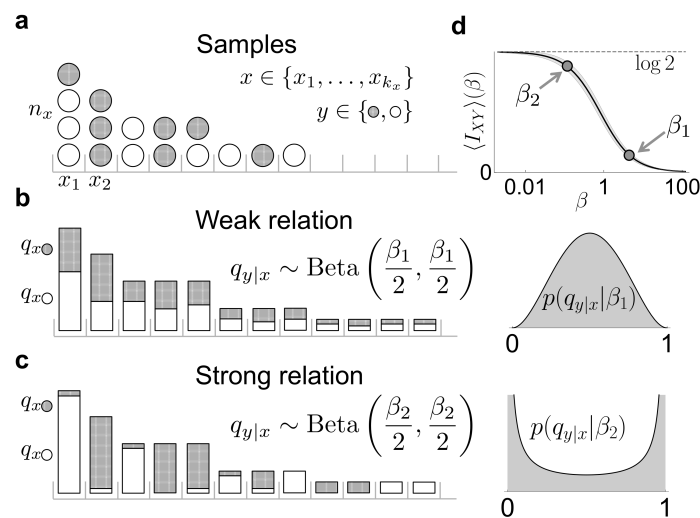


Figure 1. A scheme of our method to estimate the mutual information between two variables X and Y . (a) We collect a few samples of a variable x with a large number of effective states x_1, x_2, \dots , each sample characterized by a binary variable y (the two values represented in white and gray). We consider different hypotheses about the strength with which the probability of each y -value varies with x ; (b) one possibility is that the conditional probability of each of the two y -values hardly varies with x . This situation is modeled by assuming that the different $q_{y|x}$ are random variables governed by a Beta distribution with a large hyper-parameter β_1 ; (c) On the other hand, the conditional probability $q_{y|x}$ could vary strongly with x . This situation is modeled by a Beta distribution with a small hyper-parameter β_2 . (d) As β varies, so does the prior mutual information (Equation (18)). This prior is obtained by averaging all the $I(\{q_{1|x}\})$ values obtained from different possible sets of marginal distributions $\{q_{1|x}\}$ that can be generated when sampling the prior $p(\{q_{1|x}\} | \beta)$ of Equation (17). The shaded area around the solid line illustrates such fluctuations in $I(\{q_{1|x}\})$ when $k_x = 50$.

In order to gain intuition about the statistical dependence between variables with few samples, we here highlight the specific aspects of the data that influence the estimator of Equation (20). Grouping

together the terms of Equation (21) that are equal, the marginal likelihood can be rewritten in terms of the multiplicities $m_{nn'}$ that is, the number of states x with specific occurrences $\{n_{x1} = n, n_{x0} = n'\}$ or $\{n_{x1} = n', n_{x0} = n\}$,

$$\begin{aligned} \log p(\mathbf{n}|\beta) &= \sum_{n \geq n'} m_{nn'} \log \left[\frac{\Gamma(n + \beta/2)\Gamma(n' + \beta/2)\Gamma(\beta)}{\Gamma(n + n' + \beta)\Gamma(\beta/2)^2} \right] \\ &= \sum_{n \geq n'} m_{nn'} \log p_{nn'}(\beta), \end{aligned} \tag{22}$$

where

$$\begin{aligned} p_{10}(\beta) &= \frac{\beta/2}{\beta} = \frac{1}{2}, \\ p_{11}(\beta) &= \frac{(\beta/2)^2}{\beta(\beta + 1)} = \frac{\beta}{4(\beta + 1)}, \\ p_{20}(\beta) &= \frac{(\beta/2)(\beta/2 + 1)}{\beta(\beta + 1)} = \frac{(\beta/2 + 1)}{2(\beta + 1)}, \\ &\dots \\ p_{nn'}(\beta) &= \frac{(\beta/2)(\beta/2 + 1) \dots (\beta/2 + n - 1)(\beta/2)(\beta/2 + 1) \dots (\beta/2 + n' - 1)}{\beta(\beta + 1) \dots (\beta + n + n' - 1)}. \end{aligned} \tag{23}$$

The posterior for β is independent from states x with just a single count, as $p_{10}(\beta) = \text{constant}$. Only states x with coincidences matter. In order to see how the sampled data favor a particular β , we search for the β -value (denoted as β^*) that maximizes $\log p(\mathbf{n}|\beta)$ in the particular case where at most two samples coincide on the same x , obtaining

$$\frac{\partial}{\partial \beta} \log p(\mathbf{n}|\beta) = \frac{m_{11}}{\beta} + \frac{m_{20}}{\beta + 2} - \frac{m_{11} + m_{20}}{\beta + 1} \Big|_{\beta^*} = 0. \tag{24}$$

Denoting the fraction of 2-count states that have one count for each y -value as $f_{11} = m_{11} / (m_{11} + m_{20})$, Equation (24) implies that the likelihood $p(\mathbf{n}|\beta)$ is maximized at

$$\beta^* = \begin{cases} +\infty & \text{when } f_{11} \geq 1/2, \\ \frac{f_{11}}{1/2 - f_{11}} & \text{when } f_{11} < 1/2. \end{cases} \tag{25}$$

If the y -values are independent of x , we expect $f_{11} \sim 1/2$. This case corresponds to a large β^* and, consequently, to little information. On the other side, for small f_{11} , the parameter β^* is also small and the information grows. Moreover, the width of $p(\mathbf{n}|\beta)$ is also modulated by f_{11} . When the information is large, the peak around β^* is narrow. Low information values, instead, require more evidence, and they come about with more uncertainty around β^* .

In Equation (24), the data only intervene through m_{11} and m_{20} , which characterize the degree of asymmetry of the y -values throughout the different x -states. This asymmetry, hence, constitutes a sufficient statistics for β . If a prior $p(\beta)$ is included, the β that maximizes the posterior $p(\beta|\mathbf{n})$ may shift, but the effect becomes negligible as the number of coincidences grows.

We now discuss the role of the selected β in the estimation of information, Equation (20), focusing on the conditional entropy $\langle H_{Y|X} \rangle(\beta)$. First, in terms of the multiplicities, the conditional entropy can be rewritten as

$$\langle H_{Y|X} \rangle(\beta) = \sum_r f_r \sum_{n+n'=r} f_{nn'} H_{nn'}(\beta), \tag{26}$$

where f_r is the fraction of the N samples that fall in states x with r counts, and $f_{nm'}$ is the fraction of all states x with $n + n'$ counts, of which n correspond to one y -value (whichever) and n' for the other. Finally, $H_{nm'}(\beta)$ is the estimation of the entropy of a binary variable after $\{n, n'\}$ samples,

$$H_{nm'}(\beta) = \psi_0(n + n' + \beta + 1) - \frac{(n + \beta/2)\psi_0(n + \beta/2 + 1) + (n' + \beta/2)\psi_0(n' + \beta/2 + 1)}{n + n' + \beta}. \quad (27)$$

A priori, $\langle I | \mathbf{n} = 0, \beta \rangle = \log 2 - H_{00}(\beta)$, as in Figure 1d. Surprisingly, from the property $\psi_0(z + 1) = \psi_0(z) + 1/z$, it turns out that $H_{00} = H_{10}$ (in fact, $H_{nm} = H_{(n+1)n}$). Hence, if only a single count breaks the symmetry between the two y -values, there is no effect on the conditional entropy. This is a reasonable result, since a single extra count is no evidence of an imbalance between the underlying conditional probabilities, it is just the natural consequence of comparing the counts falling on an even number of states (2) when taking an odd number of samples. Expanding the first terms for the conditional entropy,

$$\langle H_{Y|X} \rangle = f_1 H_{00}(\beta) + f_2 f_{11} H_{11}(\beta) + f_2 f_{20} H_{20}(\beta) + \dots \quad (28)$$

In the severely under-sampled regime, these first terms are the most important ones. Moreover, when evaluating these terms in $\beta = \beta^*$ (Equation (25)), the conditional entropy simplifies into

$$\langle H_{Y|X} \rangle = \begin{cases} \log 2 & \text{when } f_{11} \geq 1/2, \\ (f_1 + f_2)H_{00}(\beta^*) + \mathcal{O}(f_3) & \text{when } f_{11} < 1/2. \end{cases} \quad (29)$$

Typically, $(f_1 + f_2)$ takes most of the weight, so the estimation is close to the prior H_{00} evaluated at the β -value that maximizes the marginal likelihood (or the posterior).

When dealing with few samples, it is important to have not just a good estimate of the mutual information, but also a confidence interval. Even a small information may be relevant, if the evidence attests that it is strictly above zero. The theory developed here also allows us to estimate the posterior variance of the mutual information, as shown in Appendix A. The variance (Equation (A4)) is shown to be inversely proportional to the number of states k_x , thereby implying that our method benefits from a large number of available states X , even if undersampled.

If an estimator, such as ours, is guaranteed to provide a non-negative estimate for all possible sets of sampled data, it cannot be free of bias, not at least if the samples are generated by an independent distribution ($q_{y|x} = q_y$), for which the true information vanishes. In Appendix B, we show that, in this specific case, the bias decreases with the square root of the number of coincidences. This number may be large, even in the severely undersampled regime, if $\exp[H(X)/2] < N \ll \exp[H(X)]$. If the distribution q_x is approximately uniform, the number of coincidences is proportional to the square of the total number of samples, so the bias is inversely proportional to N .

5. Testing the Estimator

We now analyze the performance of our estimator in three examples where the number of samples N is below or in the order of the effective size of the system $\exp(H_{XY})$. In this regime, most observed x -states have very few samples. In each example, we define the probabilities q_x and $q_{1|x}$ with three different criteria, giving rise to collections of probabilities that can be described with varying success by the prior proposed in this paper, Equation (17). Once the probabilities are defined, the true value I_{XY} of the mutual information can be calculated, and compared to the one estimated by our method in 50 different sets of samples \mathbf{n} of the measured data. We present the estimate obtained with $\langle I | \mathbf{n}, \beta^* \rangle$ from Equation (20) evaluated in the β^* that maximizes the marginal likelihood $p(\mathbf{n} | \beta)$. We did not observe any improvement when integrating over the whole posterior $p(\beta | \mathbf{n})$ with the prior $p(\beta)$ of Equation (19), except when m_{20} or m_{10} were of order 1. This fact implies the existence of a well-defined peak in the marginal likelihood.

In Figure 2, the performance of our estimator is compared with that of three other methods widely employed in the literature: Plug in, NSB and PYM. In addition, two other estimators were evaluated, but not shown in the figure to avoid cramming: the one of reference [21], which is a particularly convenient case of the Schuermann family of estimators [22], and the one of reference [23], extensively used in ecology. Their estimates fell between the plug-in estimator and NSB in the first case, and between NSB and PYM in the second case.

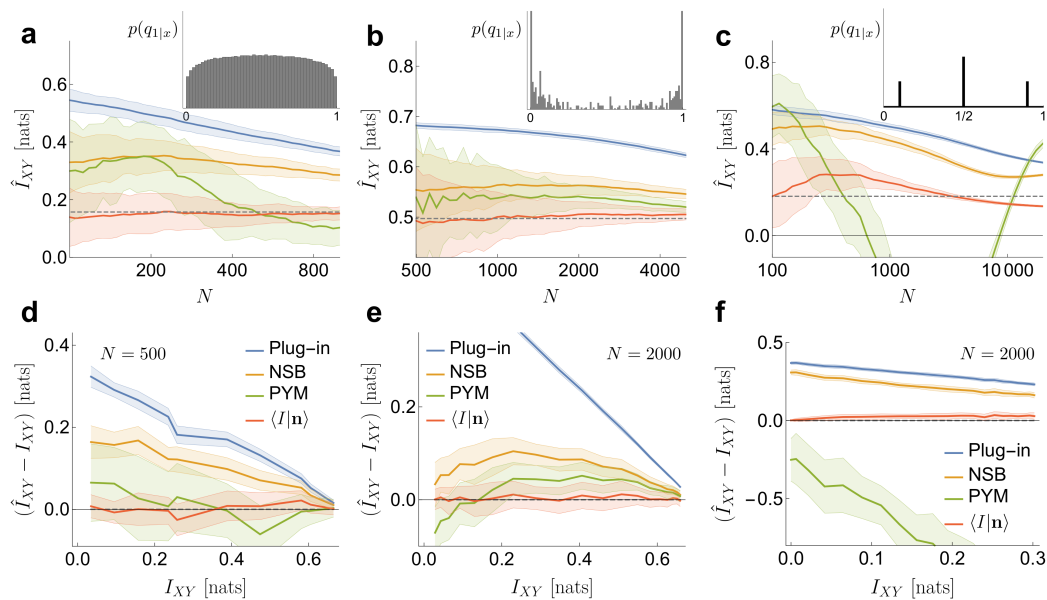


Figure 2. Comparison of the performance of four different estimators for I_{XY} : the plug-in estimator, NSB estimator used in the limit of infinite states, PYM estimator, and our estimator $\langle I|\mathbf{n} \rangle(\beta)$ (Equation (20)) calculated with the β that maximizes the marginal likelihood $p(\mathbf{n}|\beta)$ (Equation (21)). The curves represent the average over 50 different data sets \mathbf{n} , with the standard deviation displayed as a colored area around the mean. (a) estimates of mutual information as a function of the total number of samples N , when the values of $q_{1|x}$ are generated under the hypothesis of our method (Equation (17)). We sample once the marginal probabilities $q_x \sim \text{PY}(d = 0.55, \alpha = 50)$ (as described in [14]), as well as the conditionals $q_{y|x} \sim \text{Beta}(\beta/2, \beta/2)$ with $\beta = 2.3$. The effective size of the system is $\exp(H_{XY}) \simeq 800$. The exact value of I_{XY} is shown as a horizontal dashed line; (b) Estimates of mutual information, for data sets where the conditional probabilities have spherical symmetry. X , a binary variable of dimension 12, corresponds to the presence of 12 delta functions equally spaced in a sphere ($q_x = 2^{-12}$, for all x). We generate the conditional probabilities such that they are invariant under rotations of the sphere, namely $q_{y|x} = q_{y|R(x)}$, being R a rotation. To this aim, we set $q_{y|x}$ as a sigmoid function of a combination of frequency components ($\pi_0 - \pi_1 - \pi_2$) of the spherical spectrum [24]. The effective size of the system is $\exp(H_{XY}) \simeq 5000$; (c) estimates of mutual information, for a conditional distribution far away from our hypotheses. The x states are generated as Bernoulli ($p = 0.05$) binary vectors of dimension $D = 40$, while the conditional probabilities depend on the parity of the sum of the components of the vector. When the sum is even we set $q_{y|x} = 1/2$, and when is odd, $q_{y|x}$ is generated by sampling a mixture of two deltas of equal weight $q_{y|x} \sim [\delta(q - q_0) + \delta(q - 1 + q_0)]/2$ with $q_0 = 0.1$. The resulting distribution of $q_{y|x}$ -values contains three peaks, and therefore, cannot be described with a Dirichlet distribution. The effective size of the system is $\exp(H_{XY}) \simeq 4000$; (d) bias in the estimation as a function of the value of mutual information. Settings remain the same as in (a), but fixing $N = 500$ and changing $\beta \in (0.04, 14)$ in the conditional; (e) bias in the estimation as a function of the value of mutual information. Settings as in (b), but fixing $N = 2000$ and changing the gain of the sigmoid in the conditional; (f) bias in the estimation as a function of the value of mutual information. Settings as in (c), but fixing $N = 2000$ and changing $q_0 \in (0.01, 0.4)$ in the conditional.

In the first example (Figure 2a,d), the probabilities q_x are obtained by sampling a Pitman–Yor distribution with concentration parameter $\alpha = 50$ and tail parameter $d = 0.55$ (as described in [14]). These values correspond to a Pitman–Yor prior with a heavy tail. The conditional probabilities $q_{y|x}$ are defined by sampling a symmetric Beta distribution $q_{y|x} \sim \text{Beta}(\beta/2, \beta/2)$, as in Equation (17). In Figure 2a, we use $\beta = 2.3$. Once the joint probability q_{xy} is defined, 50 sets of samples \mathbf{n} are generated. The effective size of the system is $\exp(H_{XY}) \simeq 800$. We compare our estimator to the plug-in estimator (Plug-in), NSB and PYM when applied to H_X and H_{XY} (all methods coincide in the estimation of H_Y). Our estimator has a low bias, even when the number of samples per effective state is as low as $N/e^{H_{XY}} = 0.15$. The variance is larger than in the Plug-in estimator, comparable to NSB and smaller than PYM. All the other methods (Plug-in, NSB and to a lesser extend PYM) overestimate the mutual information. In Figure 2d, the performance of the estimators is also tested for different values of the exact mutual information I_{XY} , which we explore by varying $\beta \in (0.04, 14)$. For each β , the conditional probabilities $q_{1|x}$ are sampled once. Each vector \mathbf{n} contains $N = 500$ samples, and \mathbf{n} is sampled 50 times. Our estimates have very low bias, even as the mutual information goes to zero—namely, for independent variables.

Secondly, we analyze an example where the statistical relation between X and Y is remarkably intricate (example inspired by [25]), which underscores the fact that making inference about the mutual information does not require inferences on the joint probability distribution. The variable x is a binary vector of dimension 12. Each component represents the presence or absence of one of a maximum of 12 delta functions equally spaced on the surface of a sphere. There are 2^{12} possible x vectors, and they are governed by a uniform prior probability: $q_x = 2^{-12}$. The conditional probabilities are generated in such a way that they be invariant under rotations of the sphere that is, $q_{y|x} = q_{y|R(x)}$, where R is a rotation. Using a spherical harmonic representation [24], the frequency components $\pi_\ell(f(x))$ of the spherical spectrum are obtained, where $f(x)$ is the combination of deltas. The conditional probabilities $q_{y|x}$ are defined as a sigmoid function of $(\pi_0 - \pi_1 - \pi_2)$. The offset of the sigmoid is chosen such that $q_{y=1} \simeq 0.5$, and the gain such that $I_{XY} \simeq 0.5$ nats. In this example, and, unlike the Dirichlet prior implied by our estimator, $p(q_{y|x})$ has some level of roughness (inset in Figure 2b), due to peaks coming from the invariant classes in $\{x_1, \dots, x_{2^{12}}\}$. Hence, the example does not truly fit into the hypothesis of our method. With these settings, the effective size of the system is $\exp(H_{XY}) \simeq 5000$. Our estimator has little bias (Figure 2b,e), even with $N/e^{H_{XY}} = 0.2$ samples per effective state. In this regime, around $\sim 80\%$ of the samples fall on x states that occur only once ($f_1 \simeq 0.8$), $\sim 19\%$ on states that occur twice and $\sim 1\%$ on states with three counts, or maybe four. As mentioned above, in such cases, the value of I_{XY} is very similar to the one that would be obtained by evaluating the prior information $\langle I | \mathbf{n} = \mathbf{0}, \beta \rangle$ of (Equation (18)) at the β^* that maximizes the marginal likelihood $p(\mathbf{n}|\beta)$, which in turn is mainly determined by f_{11} . In Figure 2e, the estimator is tested with a fixed number of samples $N = 2000$ for different values of the mutual information, which we explore by varying the gain of the sigmoid. The bias of the estimate is small in the entire range of mutual informations.

In the third place, we consider an example where the conditional probabilities are generated from a distribution that is poorly approximated by a Dirichlet prior. The conditional probabilities are sampled from three Dirac deltas, as $q_{y|x} \sim [0.5\delta(q - 1/2) + 0.25\delta(q - q_0) + 0.25\delta(q - 1 + q_0)]$, with $q_0 = 0.1$. The delta placed in $q = 1/2$ could be approximated by a Dirichlet prior with a large β , while the other two deltas could be approximated by a small β , but there is no single value of β that can approximate all three deltas at the same time. The x states are generated as Bernoulli ($p = 0.05$) binary vectors of dimension $D = 40$, while the conditional probabilities $q_{1|x}$ depend on the parity of the sum of the components of the vector x . When the sum is even, we assign $q_{y|x} = 1/2$, and when it is odd, we assign $q_{y|x} = q_0$ or $q_{y|x} = 1 - q_0$, both options with equal probability. Although in this case our method has some degree of bias, it still preserves a good performance in relation to the other approaches (see Figure 2c,f). The marginal likelihood $p(\mathbf{n}|\beta)$ contains a single peak in an intermediate value of β , coinciding with none of the deltas in $p(q_{1|x})$, but still capturing the right value of the mutual information. As in the previous examples, we also test the performance of the estimator for different

values of the mutual information, varying in this case the value of q_0 (with $N = 2000$). Our method performs acceptably for all values of mutual information. The other methods, instead, are challenged more severely, probably because a large fraction of the x states have a very low probability, and are therefore difficult to sample. Those states, however, provide a crucial contribution to the relative weight of each of the three values of $q_{1|x}$. PYM, in particular, sometimes produces a negative estimate for I_{XY} , even on average.

Finally, we check numerically the accuracy of the analytically predicted mean posterior information (Equation (20)) and variance (Equation (A4)) in the severely under-sampled regime. The test is performed in a different spirit than the numerical evaluations of Figure 2. There, averages were taken for multiple samples of the vector \mathbf{n} , from a fixed choice of the probabilities \mathbf{q}_x and $\mathbf{q}_{y|x}$. The averages of Equations (20) and (A4), however, must be interpreted in the Bayesian sense. The square brackets in $\langle I|\mathbf{n} \rangle$ and $\langle H_{Y|X}^2 \rangle$ represent averages taken for a fixed data sample \mathbf{n} , and unknown underlying probability distributions \mathbf{q}_x and $\mathbf{q}_{y|x}$. We generate many such distributions with $q_x \sim \text{DP}(\alpha)$ (a Dirichlet Process with concentration parameter α) and $q_{y|x} \sim \text{Beta}(\beta/2, \beta/2)$. A total of 13,500 distributions q_{xy} are produced, with $\log \beta$ sampled from Equation (19), and three equiprobable values of $\alpha = \{e^4, e^5, e^6\}$. For each of these distributions, we generate five (5) sets of just $N = 40$ samples, thereby constructing a list of $5 \times 13,500$ cases, each case characterized by specific values of $\alpha, \beta, \mathbf{q}_x, \{\mathbf{q}_{y|x}\}, I(\mathbf{q}_x, \{\mathbf{q}_{y|x}\}), \mathbf{n}, \langle I|\mathbf{n} \rangle$ and $\sigma^2(I|\mathbf{n})$. Following the Bayesian rationale, we partition this list in classes, each class containing all the cases that end up in the same set of multiplicities $\{m_{nn'}\}$ —for example, $\{m_{10} = 36, m_{20} = 2\}$. For each of the 100 most occurring sets of multiplicities (which together cover 70% of all the cases), we calculate the mean and the standard deviation of the mutual information $I(\mathbf{q}, \{\mathbf{q}_{y|x}\})$ of the corresponding class, and compare them with our predicted estimates $\langle I|\{m_{nn'}\} \rangle$ and $\langle \sigma_I^2|\{m_{nn'}\} \rangle$, using the prior $p(\log \beta)$ from Equation (19). Figure 3 shows a good match between the numerical (y -axis) and analytical (x -axis) averages that define the mean information (panel a) and the standard deviation (b). The small departures from the diagonal stem from the fact that the analytical average contains all the possible \mathbf{q}_x and $\{\mathbf{q}_{y|x}\}$, even if some of them are highly improbable for one given set of multiplicities. The numerical average, instead, includes the subset of the 13,500 explored cases that produced the tested multiplicity. All the depicted subsets contained many cases, but, still, they remained unavoidably below the infinity covered by the theoretical result.

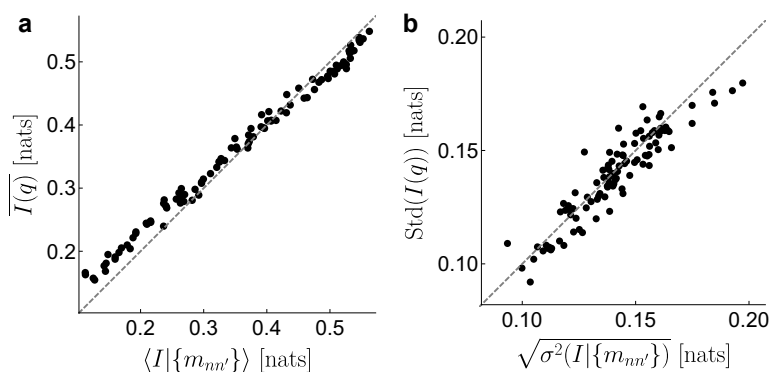


Figure 3. Verification of the accuracy of the analytically predicted mean posterior information (Equation (20)) and variance (Equation (A4)) in the severely under-sampled regime. A collection of 13,500 distributions q_{xy} are constructed by sampling $q_x \sim \text{DP}(\alpha)$ and $q_{y|x} \sim \text{Beta}(\beta/2, \beta/2)$, with α varying in the set $\{e^4, e^5, e^6\}$ and $\log \beta$ from Equation (19). Each distribution q_{xy} has an associated $I_{XY}(q_{xy})$. From each q_{xy} , we take five (5) sets of just $N = 40$ samples. (a) the values of $I(q_{xy})$ are grouped according to the multiplicities $\{m_{nn'}\}$ produced by the samples, averaged together, and depicted as the y component of each data point. The x component is the analytical result of Equation (20), based on the sampled multiplicities; (b) same analysis for the standard deviation of the information (the square root of the variance calculated in Equation (A4)).

We have also tested cases where Y takes more than two values, and where the marginal distribution q_y is not uniform, observing similar performance of our estimator.

6. A Prior Distribution for the Large Entropy Variable

The prior considered so far did not model the probability \mathbf{q}_x of the large-entropy variable X . Throughout the calculation, the probabilities \mathbf{q}_x were approximated by the maximum likelihood estimator $\hat{q}_X = n_x/N$. Here, we justify such procedure by demonstrating that proper Bayesian inference on \mathbf{q}_x hardly modifies the estimation of the mutual information. To that end, we replace the prior of Equation (13) by another prior that depends on both \mathbf{q}_x and $\{\mathbf{q}_{y|x}\}$.

The simplest hypothesis is to assume that the prior $p(\mathbf{q}_x, \{\mathbf{q}_{y|x}\})$ factorizes as $p(\mathbf{q}_x) p(\{\mathbf{q}_{y|x}\})$, implying that the marginal probabilities q_x are independent of the conditional probabilities $q_{y|x}$. We propose $q_x \sim \text{DP}(\alpha)$, so that the marginal probabilities q_x are drawn from a Dirichlet Process with concentration parameter α , associated with the total number of pseudo-counts. After integrating in \mathbf{q}_x and in $\mathbf{q}_{y|x}$, the mean posterior mutual information for fixed hyper-parameters β and α is

$$\begin{aligned} \langle I | \mathbf{n}, \alpha, \beta \rangle &= \frac{N}{N+\alpha} \left\{ \hat{H}(Y) - \sum_{x, n_x > 0} \frac{n_x}{N} \left[\psi_0(\beta + n_x + 1) - \sum_y \frac{\beta \hat{q}_{y|x} + n_{xy}}{\beta + n_x} \psi_0(\beta \hat{q}_y + n_{xy} + 1) \right] \right\} \\ &+ \frac{\alpha}{N+\alpha} \left[\hat{H}(Y) - \psi_0(\beta + 1) + \sum_y \hat{q}_y \psi_0(\beta \hat{q}_y + 1) \right]. \end{aligned} \quad (30)$$

Before including the prior $p(\mathbf{q}_x)$, in the severely undersampled regime, the mean posterior information was approximately equal to the prior information evaluated in the best β (Equation (16)). The new calculation (Equation (30)) contains the prior information explicitly, weighted by $\alpha/(N+\alpha)$, that is, the ratio between the number of pseudo-counts from the prior and the total number of counts. Thereby, the role of the non-observed (but still inferred) states is established.

The independence assumed between \mathbf{q}_x and $\{\mathbf{q}_{y|x}\}$ implies that

$$p(\mathbf{n} | \alpha, \beta) = p(\mathbf{n}_x | \alpha) p(\mathbf{n} | \beta). \quad (31)$$

The inference over α coincides with the one of PYM with the tail parameter as $d = 0$ [14], since

$$p(\mathbf{n}_x | \alpha) \propto \frac{\Gamma(1 + \alpha)}{\Gamma(N + \alpha)} \alpha^{k_1 - 1}, \quad (32)$$

where $k_1 = \sum_{x, n_x > 0} 1$ is the number of states x with at least one sample. With few coincidences in x , $p(\mathbf{n}_x | \alpha)$ develops a peak around a single α -value that represents the number of effective states. Compared to the present Bayesian approach, maximum likelihood underestimates the number of effective states (or entropy) in x . Since the expected variance of the mutual information decreases with the square root of the number of effective states, the Bayesian variance is reduced with respect to the one of the Plug-in estimator.

7. Discussion

In this work, we propose a novel estimator for mutual information of discrete variables X and Y , which is adequate when X has a much larger number of effective states than Y . If this condition does not hold, the performance of the estimator breaks down. We inspire our proposal in the Bayesian framework, in which the core issue can be boiled down to finding an adequate prior. The more the prior is dictated by the data, the less we need to assume from outside. Equation (11) implies that the mutual information $I(X, Y)$ is the spread of the conditional probabilities of one of the variables (for example, $q_{y|x}$, but the same holds for $q_{x|y}$) around the corresponding marginal (q_y or q_x , respectively). This observation inspires the choice of our prior (Equation (13)), which is designed to capture the same idea, and, in addition, to be analytically tractable. We choose to work with a hyper-parameter β that regulates the scatter of $q_{y|x}$ around q_y , and not the scatter of $q_{x|y}$ around q_x because the asymmetry in

the number of available states of the two variables makes the β of the first option (and not the second) strongly modulated by the data, by the emergence of a peak in $p(\mathbf{n}|\beta)$.

Although our proposal is inspired in previous Bayesian studies, the procedure described here is not strictly Bayesian, since our prior (Equation (13)) requires the knowledge of \hat{q}_y , which depends on the sampled data. However, in the limit in which q_y is well sampled, this is a pardonable crime, since \hat{q}_y is defined by a negligible fraction of the measured data. Still, Bayesian purists should employ a two-step procedure to define their priors. First, they should perform Bayesian inference on the center of the Dirichlet distribution of Equation (13) by maximizing $p(q_y|\mathbf{n})$, and then replace \hat{q}_y in Equation (13) by the inferred q_y . For all practical purposes, however, if the conditions of validity of our method hold, both procedures lead to the same result.

By confining the set of possible priors $p(\{\mathbf{q}_{y|x}\})$ to those generated by Equation (13), we relinquish all aspiration to model the prior of, say, $q_{y|x=3}$, in terms of the observed frequencies at $x = 3$. In fact, the preferred β -value is totally blind to the specific x -value of each sampled datum. Only the number of x -values containing different counts of each y -value matters. Hence, the estimation of mutual information is performed without attempting to infer the specific way the variables X and Y are related, a property named *equitability* [26], and that is shared also by other methods [8,13,14]. Although this fact may be seen as a disadvantage, deriving a functional relation between the variables can actually bias the inference on mutual information [26]. Moreover, fitting a relation is unreasonable in the severe under-sampled regime, in which not all x -states are observed, most sampled x -states contain a single count, and few x -states contain more than two counts—at least without a strong assumption about the probability space. In fact, if the space of probabilities of the involved variables has some known structure or smoothness condition, other approaches that estimate information by fitting the relation first may perform well [9–11]. Part of the approach developed here could be extended to continuous variables or spaces with a determined metric. This extension is left for future work.

In the explored examples, our estimator had a small bias, even in the severely under-sampled regime, and it outperformed other estimators discussed in the literature. More importantly, the second and third examples of Section 5 showed that it even worked when the collection of true conditional probabilities $q_{y|x}$ was not contained in the family of priors generated by $p(q_{y|x}|\beta)$. In these cases, the success of the method relies on the peaked nature of the posterior distribution $p(\beta|\mathbf{n})$. Even if the selected $p(q_{y|x}|\beta)$ provides a poor description of the actual collection of probabilities, the dominant β captures the right value of mutual information. This is the sheer instantiation of the equitability property discussed above.

When the number of samples N is much larger than the total number of available states $k_x \times k_y$, our estimator of mutual information tends to the plug-in estimator, which is known to be consistent [20]. Consequently, our estimator is also consistent. By construction, $\langle I|\mathbf{n} \rangle$ is equal to the mutual information averaged over all distributions \mathbf{q}_x and $\{\mathbf{q}_{y|x}\}$ compatible with the measured data, each weighted by its posterior probability. As such, it can never produce a negative result, which is a desirable property. The down side is that the estimator must be positively biased, at least, when the true information vanishes. The derivation in Appendix B shows that, when the number of samples $N \in [e^{H(X)/2}, e^{H(X)}]$, this bias is inversely proportional to the square root of the number of pairwise coincidences or, when q_x is fairly uniform, to the inverse of the total number of samples. Moreover, the factor of proportionality is significantly smaller than the one obtained in the bound of the bias of other frequentist estimators [3,27]. If the number of samples N grows even further, the bias tends to zero, since the bias of all consistent estimators vanishes asymptotically [28].

Our method provides also a transparent way to identify the statistics that matter, out of all the measured data. Quite naturally, the x -states that have not been sampled provide no evidence in shaping $p(\beta|\mathbf{n})$, as indicated by Equation (14), and only shift the posterior information towards the prior (Equation (30)). More interestingly, the x -states with just a single count are also irrelevant, both in shaping $p(\beta|\mathbf{n})$ and in modifying the posterior information away from the prior. These states are unable to provide evidence about the existence of either flat or skewed conditional probabilities $q_{y|x}$.

Only the states x that have been sampled at least twice contribute to the formation of a peak in $p(\beta|\mathbf{n})$, and in deviating the posterior information away from the prior.

Our method can also be extended to generalizations of mutual information designed to characterize the degree of interdependence of more than two variables [29–41], as long as all but one of the variables are extensively sampled. The applicability of the method to these measures will be the object of future work, once a certain degree of consensus has been reached regarding their meaning and range of applicability.

Several fields can benefit from the application of our estimator of mutual information. Examples can be found in neuroscience, when studying whether neural activity (a variable with many possible states) correlates with a few selected stimuli or behavioral responses [12,42,43], or in genomics, to understand associations between genes (large-entropy variable) and a few specific phenotypes [44]. The method can also shed light on the development of rate-distortion methods to be employed in situations in which only a few samples are available. In particular, it can be applied within the information bottleneck framework [45,46], aimed at extracting a maximally compressed representation of an input variable, but still preserving those features that are relevant for the prediction of an output variable. The possibility of detecting statistical dependencies with only few samples is of key importance, not just for analyzing data sets, but also to understand how living organisms quickly infer dependencies in their environments and adapt accordingly [47].

8. Conclusions

We have proposed a novel estimator for the mutual information $I(X; Y)$ between two variables, applicable to those cases in which the marginal distribution of one of the variables—the one with minimal entropy—is well sampled. The other variable, as well as the joint and conditional distributions, can be severely undersampled. We obtain a consistent estimator that presents very low bias, outperforming previous methods discussed in the literature. The main data statistics determining the estimated value is the inhomogeneity of the conditional distribution of the low-entropy variable in those states in which the large-entropy variable registers coincidences.

Supplementary Materials: The code with our estimator is available online at <https://github.com/dghernandez/info-estimation>.

Author Contributions: Conceptualization, D.G.H.; methodology, D.G.H.; software, D.G.H.; validation, D.G.H.; formal analysis, D.G.H. and I.S.; resources, I.S.; data curation, D.G.H.; writing—original draft preparation, D.G.H.; writing—review and editing, D.G.H. and I.S.; visualization, D.G.H.

Funding: This research was funded by CONICET, CNEA, ANPCyT Raíces 2016 grant number 1004.

Acknowledgments: We thank Ilya Nemenman for his fruitful comments and discussions.

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

Appendix A. Expected Variance for a Symmetric, Binary Y-Variable

The posterior variance of the mutual information is

$$\sigma^2(I|\mathbf{n}) = \langle (I|\mathbf{n})^2 \rangle - \langle I|\mathbf{n} \rangle^2. \quad (\text{A1})$$

In the first place, we demonstrate that this quantity is proportional to k_x^{-1} , implying that our estimator becomes increasingly accurate as the number of states of the X -variable increases. Given that

$$\langle I|\mathbf{n} \rangle \approx \hat{H}(Y) - \langle H_{Y|X} \rangle(\mathbf{n}), \quad (\text{A2})$$

with

$$\begin{aligned} H_{Y|X}(\{q_{1|x}\}) &= -\sum_x \hat{q}_x H_{Y|x}(q_{1|x}), \\ H_{Y|x}(q_{1|x}) &= -\left[q_{1|x} \log(q_{1|x}) + (1 - q_{1|x}) \log(1 - q_{1|x}) \right], \end{aligned} \quad (\text{A3})$$

it is easy to show that

$$\sigma^2(I|\mathbf{n}) \approx \sigma^2(H_{Y|X}|\mathbf{n}) = \langle H_{Y|X}^2(\{q_{1|x}\}) \rangle - \langle H_{Y|X}(\{q_{1|x}\}) \rangle^2. \tag{A4}$$

In other words, if the marginal entropy is well sampled, the variance in the information is mainly due to the variance in the conditional entropy. In turn, $H_{Y|X}$ is defined as an average of k_x terms (Equation (A3)). The independence hypothesis implied in the prior of Equation (13), and in the way different $q_{1|x}$ and q_x factor out in $q(\mathbf{n}|\mathbf{q}_x, \{\mathbf{q}_{y|x}\})$ (Equation (6)), imply that the different terms of (Equation (A3)) are all independent from each other. The average of k_x independent terms has a variance proportional to $1/k_x$, so the estimator proposed here becomes increasingly accurate as k_x grows.

We now derive the detailed dependence of $\langle H_{Y|X}^2 \rangle - \langle H_{Y|X} \rangle^2$ on the sampled data \mathbf{n} . The mean conditional entropy $\langle H_{Y|X} \rangle$ can be written in terms of $\langle H_{Y|x} \rangle(n_{x1}, n_{x0}, \beta)$, that is, of the entropy of the variable Y for a particular state x with $n_x = n_{x0} + n_{x1}$ counts at fixed β

$$\begin{aligned} \langle H_{Y|X} \rangle(\mathbf{n}) &= \int p(\beta|\mathbf{n}) d\beta \sum_x \frac{n_x}{N} \langle H_{Y|x} \rangle(n_{x0}, n_{x1}, \beta), \\ \langle H_{Y|x} \rangle(n_{x0}, n_{x1}, \beta) &= - \left\{ \psi_0(\beta + n_x + 1) - \sum_{y \in \{0,1\}} \left(\frac{\beta/2 + n_{xy}}{\beta + n_x} \right) \psi_0(\beta/2 + n_{xy} + 1) \right\}. \end{aligned} \tag{A5}$$

Similarly, for the second moment,

$$\begin{aligned} \langle H_{Y|X}^2 \rangle(\mathbf{n}) &= \int d\beta p(\beta|\mathbf{n}) \prod_x \int dq_{1|x} p(q_{1|x}|n_{x0}, n_{x1}, \beta) \left[\sum_{x'} \hat{q}_{x'} H_{Y|x'}(q_{1|x'}) \right]^2 \\ &= \int d\beta p(\beta|\mathbf{n}) \left[\sum_{x \neq x'} \hat{q}_x \hat{q}_{x'} \langle H_{Y|x} \rangle(n_{x0}, n_{x1}, \beta) \langle H_{Y|x'} \rangle(n_{x0}, n_{x1}, \beta) + \sum_x \hat{q}_x^2 \langle H_{Y|x}^2 \rangle(n_{x0}, n_{x1}, \beta) \right] \\ &= \int d\beta p(\beta|\mathbf{n}) \left[\langle H_{Y|X} \rangle^2(\mathbf{n}, \beta) + \sum_x \hat{q}_x^2 \text{Var}[H_{Y|x}](n_{x0}, n_{x1}, \beta) \right], \end{aligned} \tag{A6}$$

In turn, $\text{Var}[H_{Y|x}](n_{x0}, n_{x1}, \beta) = \langle H_{Y|x}^2 \rangle(n_{x0}, n_{x1}, \beta) - \langle H_{Y|x} \rangle(n_{x0}, n_{x1}, \beta)^2$, where the first moment $\langle H_{Y|x} \rangle(n_{x0}, n_{x1}, \beta)$ is given in Equation (A5). The second moment is [14],

$$\begin{aligned} \langle H_{Y|x}^2 \rangle(n_{x1}, n_{x0}, \beta) &= \int_0^1 dq_{1|x} p(q_{1|x}|n_{x0}, n_{x1}, \beta) H_{Y|x}^2(q_{1|x}) \\ &= \int_0^1 dq_{1|x} p(q_{1|x}|n_{x0}, n_{x1}, \beta) \left[q_{1|x} \log(q_{1|x}) + (1 - q_{1|x}) \log(1 - q_{1|x}) \right]^2 \\ &= \frac{2 \binom{\beta/2 + n_{x0}}{\beta/2 + n_x} \binom{\beta/2 + n_{x1}}{\beta/2 + n_x}}{(\beta + n_x + 1) (\beta + n_x)} F \left(\frac{\beta}{2} + n_{x0}, \frac{\beta}{2} + n_{x1} \right) + \\ &\quad + \sum_{y \in \{0,1\}} \frac{\binom{\beta/2 + n_{xy}}{\beta/2 + n_x} \binom{\beta/2 + n_{xy} + 1}}{(\beta + n_x + 1) (\beta + n_x)} G \left(\frac{\beta}{2} + n_{xy}, \beta + n_x \right). \end{aligned} \tag{A7}$$

In this equation,

$$\begin{aligned} F(z_0, z_1) &= -\psi_1(z_0 + z_1 + 2) + \prod_{i \in \{0,1\}} [\psi_0(z_i + 1) - \psi_0(z_0 + z_1 + 2)], \\ G(z_i, z) &= [\psi_0(z_i + 2) - \psi_0(z + 2)]^2 + \psi_1(z_i + 2) - \psi_1(z + 2), \end{aligned} \tag{A8}$$

and $\psi_1(z)$ is the first polygamma function.

Replacing the obtained expressions in Equation (A4), the variance of the estimated information is obtained. The two terms of Equation (A6) represent the two sources of uncertainty of the conditional entropy: the uncertainty of β (first term), manifest in the width of $p(\beta|\mathbf{n})$, and the uncertainty of the conditional entropies for a fixed β (second term), manifest in the width of $p(\{q_{1|x}\}|\beta)$. As the number of samples decreases, the uncertainty in β becomes the dominant term.

The approximate symbol in Equation (A2) stems from the fact that $H(Y)$ is assumed to be well approximated by the plug-in estimator. The error in the marginal entropy H_Y is thereby neglected, and the error in the mutual information is assumed to derive from the uncertainty in the conditional entropy $H_{Y|X}$ (Equation (A4)) alone. The assumption is well justified in the context explored in this paper, that is, when $\exp[H(X)] \gg \exp[H(Y)]$.

Appendix B. Bounding the Bias for Independent Variables

The core idea of a Bayesian approach is that the measured datum \mathbf{n} is given, and the underlying probabilities are unknown. Bayesian estimators are hence constructed by averaging over all possible inferred probabilities. In this context, the calculation of the bias of an estimator is somehow at odds with the core idea, since the bias is obtained by fixing the underlying probability, and averaging on the possible outcomes of the sampled data. Perhaps for this reason, the bias of Bayesian estimators is hardly ever calculated. Yet, if a Bayesian estimator is regarded as a function of \mathbf{n} , it is always possible to average the estimate over all the possible sampled data for a fixed underlying distribution. This average is performed in this appendix for the estimator $\langle I|\mathbf{n}, \beta^* \rangle$, evaluated at the β -value that maximizes $p(\mathbf{n}|\beta)$, for a binary and symmetric Y variable, in the severely undersampled regime, and when the variables X and Y are independent (that is, $q_{y|x} = q_y = 1/2, \forall x$). In this regime, the true information vanishes, so the calculated average equals the bias.

We work with a number of samples N in which the previous estimators begin to fail, but ours is still useful that is, $\exp[H(X)/2] < N \ll \exp[H(X)]$. In this regime, the majority of the x -states are sampled at most once, and most of the x -states with coincidences are sampled only twice. As an approximation, we assume that no x -states are sampled more than twice. Since the β -value that maximizes $p(\mathbf{n}|\beta)$ cannot be calculated in the absence of coincidences (see Equation (25)), we assume that there is at least one coincidence. The total number m of x -states with two-fold coincidences is decomposed into $m = m_{20} + m_{11}$, where m_{20} is the number of states sampled with two equal y -values (whichever), and $m_{11} = m - m_{20}$ is the number of states with two different y -values.

We define a variable $z = m_{20}/m - 1/2$. Equation (25) implies that the optimal β is $\beta^*(z) = (1/2 - z)/z$, if $z > 0$ and $\beta^* \rightarrow \infty$ if $z \leq 0$. In turn, assuming no more than two-fold coincidences occur (see Equations (24) and (29)), $\langle I(\mathbf{n}, \beta^*) \rangle$ can be expressed as

$$\langle I(\mathbf{n}, \beta^*) \rangle \simeq \log 2 - H_{00}(\beta^*(z)) = \begin{cases} \log 2 - \psi\left(\frac{1}{2z}\right) + \psi\left(\frac{1/2+z}{2z}\right), & \text{if } z > 0, \\ 0, & \text{if } z \leq 0. \end{cases} \tag{A9}$$

In order to calculate the bias in the non-informative case, we take $\mathbf{q}_{y|x} = 1/2, \forall x$ (a distribution with $I(X; Y) = 0$). The bias B is obtained by averaging $\langle I(\mathbf{n}, \beta^*) \rangle$ over $p(\mathbf{n}|\mathbf{q}_x, \{\mathbf{q}_{y|x}\})$,

$$\begin{aligned} B &= \sum_{\mathbf{n}} \langle I(\mathbf{n}, \beta^*) \rangle p(\mathbf{n}|\mathbf{q}_x, \{\mathbf{q}_{y|x}\}) \\ &= \sum_{m>0} p(m|\mathbf{q}_x) \sum_{m_{20}}^m p(m_{20}|m, \{\mathbf{q}_{y|x}\}) \langle I(\mathbf{n}, \beta^*) \rangle. \end{aligned} \tag{A10}$$

The distribution $p(m_{20}|m, \{\mathbf{q}_{y|x}\})$ is a Binomial($1/2, m$). For non-negligible values of m , the Binomial can be approximated by a Gaussian that we write in terms of the variable z , such that $p(z|m, \{\mathbf{q}_{y|x}\}) \sim \mathcal{N}(0, 1/4m)$. Then, Equation (A10) becomes

$$B \simeq \sum_{m>0} p(m|\mathbf{q}_x) \int_0^\infty dz \frac{e^{-2mz^2}}{\sqrt{\pi/2m}} \langle I(\mathbf{n}, \beta^*) \rangle. \tag{A11}$$

The function $\langle I(\mathbf{n}, \beta^*) \rangle$ of Equation (A9) is continuous in z , and starts growing from zero when $z = 0$ ($m_{20} = m/2$). As the main contributions to the integral of Equation (A11) come from small values of z , we expand $\langle I(\mathbf{n}, \beta^*) \rangle$ in a Taylor series, obtaining $\langle I(\mathbf{n}, \beta^*) \rangle \simeq z + z^2 - 2z^4 + \mathcal{O}(z^5)$. The integral of Equation (A11) can be calculated using the moments of the Half Gaussian, yielding

$$B \simeq \frac{1}{2\sqrt{2\pi}} \overline{m^{-1/2}} + \frac{1}{8} \overline{m^{-1}} - \frac{3}{16} \overline{m^{-2}}, \quad (\text{A12})$$

where the averages indicated with a horizontal line are weighted by $p(m|\mathbf{q}_x)$. We have checked in simulations that in the regime $\exp[H(X)/2] < N \ll \exp[H(X)]$, the dependence of the bias on m scales as predicted by Equation (A12). It is worth noticing that, when m is as small as 20 coincidences, the bias drops to $B \simeq 0.05$ nats. If the probabilities q_x are approximately homogeneous, then the number coincidences scales as $m \simeq 0.5N^2 / \exp[H(X)]$ [18]. In that case, the leading term in the bias is

$$B \sim \frac{1}{2\sqrt{\pi}} \frac{\sqrt{e^{H(X)}}}{N}. \quad (\text{A13})$$

For comparison, the bias of the entropy estimator H_G proposed by Grassberger [21] is bounded by $-\Delta H_G < 0.14 \dots e^{H(X)}/N$. Although the scaling in N is the same, the accompanying factor is much larger than in our estimator, especially in the undersampled regime considered here, where $\exp[H(X)/2] < N \ll \exp[H(X)]$.

References

- Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
- Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
- Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Network Comput. Neural Syst.* **1996**, *7*, 87–107. [[CrossRef](#)] [[PubMed](#)]
- Samengo, I. Estimating probabilities from experimental frequencies. *Phys. Rev. E* **2002**, *65*, 046124. [[CrossRef](#)] [[PubMed](#)]
- Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [[CrossRef](#)]
- Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [[CrossRef](#)]
- Montemurro, M.A.; Senatore, R.; Panzeri, S. Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput.* **2007**, *19*, 2913–2957. [[CrossRef](#)] [[PubMed](#)]
- Archer, E.; Park, I.M.; Pillow, J.W. Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy* **2013**, *15*, 1738–1755. [[CrossRef](#)]
- Kolchinsky, A.; Tracey, B.D. Estimating mixture entropy with pairwise distances. *Entropy* **2017**, *19*, 361. [[CrossRef](#)]
- Belghazi, I.; Rajeswar, S.; Baratin, A.; Hjelm, R.D.; Courville, A. MINE: Mutual information neural estimation. *arXiv* **2018**, arXiv:1801.04062.
- Safaai, H.; Onken, A.; Harvey, C.D.; Panzeri, S. Information estimation using nonparametric copulas. *Phys. Rev. E* **2018**, *98*, 053302. [[CrossRef](#)]
- Strong, S.P.; Koberle, R.; van Steveninck, R.R.D.R.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *80*, 197. [[CrossRef](#)]
- Nemenman, I.; Bialek, W.; van Steveninck, R.R.D.R. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E* **2004**, *69*, 056111. [[CrossRef](#)] [[PubMed](#)]
- Archer, E.; Park, I.M.; Pillow, J.W. Bayesian entropy estimation for countable discrete distributions. *J. Mach. Learn. Res.* **2014**, *15*, 2833–2868.
- Wolpert, D.H.; DeDeo, S. Estimating functions of distributions defined over spaces of unknown size. *Entropy* **2013**, *15*, 4668–4699. [[CrossRef](#)]

16. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2007.
17. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841. [[CrossRef](#)]
18. Ma, S.k. Calculation of entropy from data of motion. *J. Stat. Phys.* **1981**, *26*, 221–240. [[CrossRef](#)]
19. Nemenman, I. Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy* **2011**, *13*, 2013–2023. [[CrossRef](#)]
20. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms* **2001**, *19*, 163–193. [[CrossRef](#)]
21. Grassberger, P. Entropy estimates from insufficient samplings. *arXiv* **2003**, arXiv:0307138.
22. Schürmann, T. Estimating probabilities from experimental frequencies. *J. Phys. Math. Gen.* **2004**, *37*, L295–L300. [[CrossRef](#)]
23. Chao, A.; Wang, Y.; Jost, L. Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods Ecol. Evol.* **2013**, *4*, 1091–1100. [[CrossRef](#)]
24. Kazhdan, M.; Funkhouser, T.; Rusinkiewicz, S. Rotation invariant spherical harmonic representation of 3 d shape descriptors. *Symp. Geom. Process.* **2003**, *6*, 156–164.
25. Shwartz-Ziv, R.; Tishby, N. Opening the black box of deep neural networks via information. *arXiv* **2017**, arXiv:1703.00810.
26. Kinney, J.B.; Atwal, G.S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3354–3359. [[CrossRef](#)] [[PubMed](#)]
27. Grassberger, P. Entropy estimates from insufficient samples. *Archive* **2001**, *412*, 787.
28. Barlow, R.J. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 1993.
29. Amari, S.i. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711. [[CrossRef](#)]
30. Panzeri, S.; Schultz, S.R.; Treves, A.; Rolls, E.T. Correlations and the encoding of information in the nervous system. *Proc. R. Soc. B Biol. Sci.* **1999**, *226*, 1001–1012. [[CrossRef](#)]
31. Panzeri, S.; Schultz, S.R. Temporal Correlations and Neural Spike Train Entropy. *Phys. Rev. Lett.* **2001**, *86*, 5823–5826.
32. Panzeri, S.; Schultz, S.R. A Unified Approach to the Study of Temporal, Correlational, and Rate Coding. *Neural Comput.* **2001**, *13*, 1311–1349. [[CrossRef](#)]
33. Pola, G.; Hoffmann, K.P.; Panzeri, S. An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network* **2003**, *14*, 35–60. [[CrossRef](#)]
34. Hernández, D.G.; Zanette, D.H.; Samengo, I. Information-theoretical analysis of the statistical dependencies between three variables: Applications to written language. *Phys. Rev. E* **2015**, *92*, 022813. [[CrossRef](#)] [[PubMed](#)]
35. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
36. Harder, M.; Salge, C.; Polani, D. Bivariate Measure of Redundant Information. *Phys. Rev. E* **2015**, *87*, 012130. [[CrossRef](#)] [[PubMed](#)]
37. Timme, N.; Alford, W.; Flecker, B.; Beggs, J.M. Synergy, redundancy, and multivariate information measures: An experimentalist's perspective. *J. Comput. Neurosci.* **2013**, *36*, 119–140. [[CrossRef](#)] [[PubMed](#)]
38. Griffiths, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 159–190.
39. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [[CrossRef](#)]
40. Ince, R.A.A. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy* **2017**, *19*, 318. [[CrossRef](#)]
41. Yu, S.; Giraldo, S.; Gonzalo, L.; Jenssen, R.; Príncipe, J.C. Multivariate Extension of Matrix-based Rényi's α -order Entropy Functional. *arXiv* **2018**, arXiv:1808.07912.
42. Tang, C.; Chehayeb, D.; Srivastava, K.; Nemenman, I.; Sober, S.J. Millisecond-scale motor encoding in a cortical vocal area. *PLoS Biol.* **2014**, *12*, e1002018. [[CrossRef](#)]
43. Maidana Capitán, M.; Kropff, E.; Samengo, I. Information-Theoretical Analysis of the Neural Code in the Rodent Temporal Lobe. *Entropy* **2018**, *20*, 571. [[CrossRef](#)]

44. Butte, A.J.; Tamayo, P.; Slonim, D.; Golub, T.R.; Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12182–12186. [[CrossRef](#)]
45. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:0004057.
46. Still, S.; Bialek, W. How many clusters? An information-theoretic perspective. *Neural Comput.* **2004**, *16*, 2483–2506. [[CrossRef](#)] [[PubMed](#)]
47. Fairhall, A.L.; Lewen, G.D.; Bialek, W.; van Steveninck, R.R.D.R. Efficiency and ambiguity in an adaptive neural code. *Nature* **2001**, *412*, 787. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).