MDPI

Article

Inferring a Property of a Large System from a Small Number of Samples

Damián G. Hernández *[ORCID] and Inés Samengo *[ORCID]

Department of Medical Physics, Centro Atómico Bariloche and Instituto Balseiro,
San Carlos de Bariloche 8400, Argentina
* Correspondence: damian.hernandez@cab.cnea.gov.ar (D.G.H.); samengo@cab.cnea.gov.ar (I.S.)

Abstract: Inferring the value of a property of a large stochastic system is a difficult task when the number of samples is insufficient to reliably estimate the probability distribution. The Bayesian estimator of the property of interest requires the knowledge of the prior distribution, and in many situations, it is not clear which prior should be used. Several estimators have been developed so far in which the proposed prior us individually tailored for each property of interest; such is the case, for example, for the entropy, the amount of mutual information, or the correlation between pairs of variables. In this paper, we propose a general framework to select priors that is valid for arbitrary properties. We first demonstrate that only certain aspects of the prior distribution actually affect the inference process. We then expand the sought prior as a linear combination of a one-dimensional family of indexed priors, each of which is obtained through a maximum entropy approach with constrained mean values of the property under study. In many cases of interest, only one or very few components of the expansion turn out to contribute to the Bayesian estimator, so it is often valid to only keep a single component. The relevant component is selected by the data, so no handcrafted priors are required. We test the performance of this approximation with a few paradigmatic examples and show that it performs well in comparison to the ad-hoc methods previously proposed in the literature. Our method highlights the connection between Bayesian inference and equilibrium statistical mechanics, since the most relevant component of the expansion can be argued to be that with the right temperature.

Keywords: inference; bayesian; undersampled; entropy; mutual information

1. Introduction

Systems with a large number of states appear ubiquitously in the era of big data. Often, the number of available observations—which, in this paper are referred to as "samples"—is smaller than the number of states, sometimes considerably smaller. Although in these circumstances estimating the full probability distribution of the system is highly unreliable, here we show that even in the severe undersampled regime, fairly accurate inferences are often still possible. The clue is that, if only a single property of the system needs to be inferred, or maybe just a few properties, the complete description of the probability distribution is unnecessary. Examples of desired properties may be, for instance, the information processing capacity, the correlation between pairs of variables, or the diversity of reachable states in the system.

Let k be the number of available states of a discrete system, and $\boldsymbol{q} = (q_1, \ldots, q_k)$ the (unknown) probability distribution that governs the occupation of the states. The number of samples n is assumed to be small compared with the effective number of states $\exp[H(\boldsymbol{q})]$, where $H(\boldsymbol{q})$ is the entropy of the system. Still, n is assumed to be large enough so that, occasionally, some of the states are sampled more than once. In mathematical terms, this means that n is at least as large as the square root of the effective number of states; that is, $n \gtrsim \sqrt{\exp[H(\boldsymbol{q})]}$ [1]. For even smaller samples, most of the states remain unsampled,

and the few states that are sampled are likely to get no more than a single sample. When n reaches the order of $\sqrt{\exp[H(\boldsymbol{q})]}$, the probability that a small fraction of the states get more than a single sample becomes large [1]. We are interested in estimating a property $F(\boldsymbol{q}) = F(q_1, \ldots, q_k)$ whose functional form in terms of $\boldsymbol{q}$ is known. The naïve approach—which we here discourage—would be to first estimate $\boldsymbol{q}$ and then evaluate $F(\boldsymbol{q})$ in the estimated value. This strategy is called the "plug-in" estimator. Although it may work for properties that depend linearly on the probabilities, it is often appallingly biased in nonlinear cases—in particular, for information-related properties that typically bear a logarithmic dependence on q_i. To overcome this problem, some idiosyncratic solutions have emerged during the last decades that are valid for particular properties, such as entropy [2–6] and mutual information [7–9].

Here, we develop a principled way to approach this problem for properties $F(\boldsymbol{q})$ that may be arbitrary, as long as they vary smoothly with $\boldsymbol{q}$. The clue relies on the fact that the simplex $\mathcal{Q}$ containing the set of possible $\boldsymbol{q}$-vectors can be foliated by the level surfaces of $F(\boldsymbol{q})$. If only the value of F matters, estimating the full $\boldsymbol{q}$ is excessive; the only inference that needs to be made is about the level surface to which $\boldsymbol{q}$ belongs. This is a one-dimensional inference problem and hence is much simpler than the k-dimensional problem of estimating $\boldsymbol{q}$. We here estimate F within a Bayesian framework, using a prior distribution that weighs the different level surfaces according to a MaxEntropy principle.

The main contributions of the paper are as follows:

-   A new method to construct Bayesian estimators is proposed;
-   The method is applicable to arbitrary quantities $F(\boldsymbol{q})$;
-   A conceptual framework is provided, in which the one-dimensional nature of the problem is exploited;
-   The search of the prior is restricted to the level surfaces of $F(\boldsymbol{q})$;
-   The Maxentropy principle is employed to capitalize both what we know and what we do not know about the prior;
-   The performance of the method is tested in known examples;
-   The conceptual connection with equilibrium statistical mechanics is highlighted.

The paper is organized as follows. In the next section, the theoretical Bayesian framework is described in general: firstly, pointing out that only certain aspects of the prior matter in the inference of the property; secondly, using the symmetries of the property and our ignorance about the underlying distribution to constrain the search for the prior; thirdly, by finally constructing the prior with a MaxEntropy principle; and finally, showing how this approach relates to equilibrium statistical mechanics. In the third section, the method is applied to different example properties, including mutual information and entropy. Finally, in the last section, the implications and perspectives of the work are discussed and summarized.

2. Rationale of the Inference Process

Section 2 develops the theoretical underpinnings of the new estimator proposed in this paper.

2.1. A Bayesian Approach to the Inference Problem

In this subsection, we define the estimator to be used and formulate the problem within the Bayesian framework. The search for the prior is established as the main target of the paper.

Our goal is to estimate a property $F(\boldsymbol{q})$ of a stochastic system composed of a large number k of states, when we only have access to n samples, with $n \gtrsim \sqrt{\exp[H(\boldsymbol{q})]}$. The probabilities $\boldsymbol{q} = (q_1, \ldots, q_k)$ constitute a complete description of the system (see Figure 1).
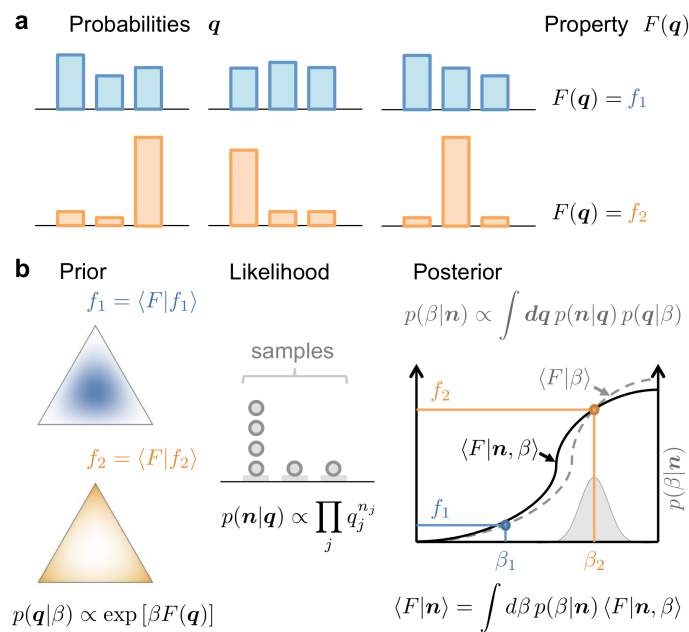
Figure 1. Conceptual framework used to estimate the property $F(q)$ of a stochastic system with probabilities q from a limited number of observations, or samples n. (**a**) Several probabilities q can produce the same value of the property. All such q-vectors belong to the same level surface of $F(q)$. (**b**) Left: Example case in which the property $F(q)$ has circular level surfaces on the simplex embedded in three-dimensional space. Two members of the family of base functions used to expand the prior are shown, containing the different q vectors shown in (a). These two members are solutions of the Maxentropy problem with two different expected values f of the property. Middle: The histogram n generated by the sampled data produces a likelihood $p(n|q)$ that selective favors a specific range of surface levels. Right: The most favored f value (or equivalently, β value) is the one for which the prior and posterior estimates of the property coincide.

We choose to work with the estimator $\hat{F} = \langle F(q|n) \rangle$ defined as the expected value of the property $F(q)$ when conditioned by the measured data n. The average $\langle \cdot \rangle$ is calculated with the posterior distribution $p(q|n)$, and the k-dimensional vector n has components $n_1, \dots, n_k$, with n_i equal to the number of samples in which the system was observed in the ith state. Therefore,

$$\hat{F} = \langle F|n \rangle = \int \mathrm{d}q\, F(q)\, p(q|n). \tag{1}$$

Throughout the paper, all integrals in q should be restricted to the space where q is defined: the simplex embedded in $\mathbb{R}^k$, or the cartesian product of several simplexes, in the multivariate case. The estimator of Equation (1) minimizes the mean square error of the estimation [10]. Using Bayes' rule, the posterior can be written in terms of the likelihood $p(n|q)$ and the prior $p(q)$, so that

$$\langle F|n \rangle = \int \mathrm{d}q\, F(q)\, \frac{p(n|q)p(q)}{p(n)}. \tag{2}$$

The likelihood $p(n|q)$ is the only factor that depends on the sampled data n, and for discrete states, it is a multinomial distribution, namely

$$p(n|q) \propto \prod_j q_j^{n_j}.$$

The only factor that still needs to be defined is the prior $p(q)$, and the strategy underlying this choice is the topic of this paper. We want the prior to produce an inductive bias that acknowledges our ignorance about $F(q)$. Such bias will be useful to overcome the

scarcity of samples, but it will only be suitable to estimate the specific $F(\boldsymbol{q})$ of the problem at hand and not others. In other words, the prior proposed here is individually tailored for $F(\boldsymbol{q})$.

2.2. Defining the Prior on the Level Surfaces of the Property

In this subsection, the apparent high-dimensionality of the problem is shown to be illusory. The prior we are searching for, although defined in a high-dimensional simplex, is here shown to affect the result of the inference process only through its integrated value inside each level surface of the quantity $F(\boldsymbol{q})$. In other words, the variations of $p(\boldsymbol{q})$ inside the level surfaces of $F(\boldsymbol{q})$ are irrelevant since they have no influence on the value of the estimator $\hat{F}$. These level surfaces can be indexed by a single-dimensional parameter, which will later allow us to drastically simplify the search for the prior.

To show this point, we start by noting that $p(\boldsymbol{q})$ is defined on the simplex embedded in $\mathbb{R}^k$. We introduce a change of variables $\boldsymbol{q} \rightarrow \boldsymbol{q}'$ to perform the integral of Equation (2). Specifically, the first two coordinates of the new variables are chosen to be

$$
\begin{aligned}
q_1'(\boldsymbol{q}) &= f = F(\boldsymbol{q}) = \text{property to be inferred} \\
q_2'(\boldsymbol{q}) &= \ell = q(\boldsymbol{n}|\boldsymbol{q}) = \text{likelihood} ,
\end{aligned}
\tag{3}
$$

and the remaining coordinates $q_3'(\boldsymbol{q}), \dots, q_k'(\boldsymbol{q})$ may be chosen arbitrarily, as long as the transformation $\boldsymbol{q} \rightarrow \boldsymbol{q}'$ be invertible. In the new variables,

$$
\begin{aligned}
\langle F|\boldsymbol{n}\rangle &= \int \mathrm{d}f\, f \int \mathrm{d}\ell\, \ell \int \mathrm{d}q_3' \dots \mathrm{d}q_k' \frac{p(\boldsymbol{q}')}{p(\boldsymbol{n})} \left| \frac{\partial q_1, \dots, q_k}{q_1', \dots, q_k'} \right| \\
&= \int \mathrm{d}f\, f \int \mathrm{d}\ell\, \ell\, p(f, \ell|\boldsymbol{n}), \quad \text{with} \\
p(f, \ell|\boldsymbol{n}) &= \int \mathrm{d}q_3' \dots \mathrm{d}q_k' \frac{p(\boldsymbol{q}')}{p(\boldsymbol{n})} \left| \frac{\partial q_1, \dots, q_k}{q_1', \dots, q_k'} \right| \\
&= \int \mathrm{d}\boldsymbol{q}\, \frac{p(\boldsymbol{q})}{p(\boldsymbol{n})}\, \delta[f(\boldsymbol{q}) - f]\, \delta[L(\boldsymbol{q}) - \ell].
\end{aligned}
\tag{4}
$$

Equation (4) shows that it does not matter how the prior $p(\boldsymbol{q})$ distributes the density inside the manifold $\mathcal{M}_{f\ell}$ obtained by intersecting the level surface f of the property F with a level surface ℓ of the likelihood L. Only the integral of $\boldsymbol{q}$ inside $\mathcal{M}_{f\ell}$ has an effect on $\hat{F}$. Therefore, we opt to consider only priors $p(\boldsymbol{q})$ that are constant inside $\mathcal{M}_{f\ell}$, since for any prior not constant in $\mathcal{M}_{f\ell}$, a prior that is constant in $\mathcal{M}_{f\ell}$ exists that produces the same estimation $\hat{F}$. Our task is therefore to design how the prior $p(\boldsymbol{q})$ changes with f and ℓ. A crucial point is that the level surfaces of $L(\boldsymbol{q})$ vary with the data $\boldsymbol{n}$. Yet, by definition, a prior cannot depend on the measured data. Therefore, we only work with priors that, when written in the $\boldsymbol{q}$-coordinates, only depend on $\boldsymbol{q}$ through $F(\boldsymbol{q})$. In other words, we assume that a function g exists such that $p(\boldsymbol{q}) = g[F(\boldsymbol{q})]$. By limiting the search to priors of this type, for each value of f, we impose maximal uncertainty about $\boldsymbol{q}$, as dictated by a maxentropy principle.

As a side remark, we point out that the assumption $p(\boldsymbol{q}) = g[F(\boldsymbol{q})]$ does not imply that, when written in the $\boldsymbol{q}'$-coordinates, the prior $p(\boldsymbol{q}')$ must be independent of $\ell = q_2'$, since the Jacobian of the transformation of variables may well depend on ℓ. Moreover, the marginal prior $p(f, \ell|\boldsymbol{n})$ may also depend on ℓ, since the level surface in which Equation (4) is calculated depends on $L(\boldsymbol{q})$. Therefore, $p(f, \ell|\boldsymbol{n})$ is allowed to depend on ℓ, but only due to the geometrical structure that the delta function in $L(\boldsymbol{q})$ restricts the integration region of Equation (4).

2.3. Decomposing the Prior as a Linear Combination of a Family of Base Priors

In this subsection, we exploit the fact that $p(\boldsymbol{q}) = g[F(\boldsymbol{q})]$ to decompose the prior as a linear combination of a collection of base priors indexed by a single real number

f. The decomposition allows us to reduce the search of a prior $p(q)$ defined in a high-dimensional simplex to the search of a density $g(f)$ defined as a function of f. This step brings about a dramatic simplification, but more importantly, it reduces the relevance of the subjective aspect involved in the selection of priors. We show here that the range of f-values that are relevant to the inference process (and thereby, the base priors that actually matter) is only partially defined by $g(f)$, which is the only element still containing subjective choices. In many applications, the sampled data n will turn out to be the truly defining factor in selecting the base priors that participate in the inference process.

We start to demonstrate these statements by noting that the set of priors $p(q)$ for which a function $g : \mathbb{R} \to \mathbb{R}$ exists such that $p(q) = g[F(q)]$ is equal to all the distributions of the form

$$p(q) = \int \mathrm{d}f \, g(f) \, \delta[F(q) - f] \tag{5}$$

that can be obtained by varying $g(f)$. Equation (5) implies that the problem of specifying $p(q)$ can be reduced to the problem of specifying $g(f)$, which is a drastically simpler object, given that q has $(k-1)$ dimensions, whereas f has only a single dimension.

We now show that, when the number of samples is sufficiently large, the determination of g becomes unnecessary, since for large k, the measured data n often enter into the likelihood in such a way that only a small range of f-values remains with significantly non-zero probability. Therefore, all the g-functions that do not vary drastically within the permitted range give rise to the same estimation. In such cases, the data alone dictate the value of the estimator through the likelihood, making the discussion about priors essentially inconsequential. We now prove these statements. Replacing Equation (5) in (2), the estimator becomes

$$
\begin{aligned}
\langle F|n \rangle &= \frac{1}{p(n)} \int \mathrm{d}q \, p(n|q) \, F(q) \, g[F(q)] \, \delta[F(q) - f] \\
&= \int \mathrm{d}f \, f \, p(f|n)
\end{aligned}
\tag{6}
$$

with

$$
\begin{aligned}
p(f|n) &= \frac{g(f)}{p(n)} \int \mathrm{d}q \, p(n|q) \, \delta[F(q) - f] \\
&\propto \frac{g(f)}{p(n)} \int \mathrm{d}q \, \mathrm{e}^{-n \, D_{\mathrm{KL}}(\frac{1}{n}n||q)} \, \delta[F(q) - f],
\end{aligned}
\tag{7}
$$

where $D_{\mathrm{KL}}(a||b)$ is the Kullback–Leibler divergence [11] between the distributions a and b, and the ith component of the vector $\frac{1}{n}n$ is n_i/n. The divergence is always non-negative, and it only vanishes when the two distributions coincide [12]. Therefore, if the number of samples n is sufficiently large, the result of the integral is significantly different from zero only for level surfaces that pass close enough to the sampled frequencies n/n and is maximal for the one that contains the sampled frequencies. The data n thereby select the range of f values that are compatible with the observations. The factor n in the exponent of Equation (7) implies that the allowed range of f-values becomes increasingly narrow as the number of samples grows. The crucial point in this reasoning is that for sufficiently narrow ranges, the shape of the prior $g(f)$ becomes irrelevant. If the range is much narrower than the typical scales in which $g(f)$ varies, for all practical matters, $g(f)$ is constant within this range, and has no bearing in the estimation. This is the situation in which we no longer need to bother about the prior.

When that limit is reached, the likelihood becomes a delta distribution, and the only q-vector that contributes to a given f-value is $q = n/n$. In this extreme case, when $g(f)$ is uniform, the estimator $\langle F|n \rangle$ converges to the plug-in estimator. Yet, the plug-in estimator is only justified when the likelihood indeed reaches the delta-like behavior, which only happens for $n \gg k$. Before this limit, the plug-in estimator completely neglects the width of the peak of the likelihood around the sampled frequencies n/n and thereby the possibility

that a whole collection of $\boldsymbol{q}$-vectors in the vicinity of $\boldsymbol{n}/n$ contribute to each f. Allowing for this possibility is important since, in the undersampled regime, there is a large degree of uncertainty about the true $\boldsymbol{q}$ that generated the data $\boldsymbol{n}$. By neglecting this uncertainty, the plug-in estimator produces the undesired biases mentioned before. The obvious solution is not to neglect the width of the likelihood. However, if no approximations are made, it is difficult to calculate or even to estimate $p(f|\boldsymbol{n})$ analytically, given that Equation (7) involves an integral over an arbitrarily shaped manifold.

To overcome this problem, we would like to design another estimator that retains insensitivity to the prior but still gives a chance for a collection of $\boldsymbol{q}$ vectors to contribute to each f. The size of this collection is determined by the width of the likelihood; that is, by the total number of samples n, and the obtained frequencies $\boldsymbol{n}/n$. The averaged contributions of all the $\boldsymbol{q}$ vectors that have non-vanishing likelihoods produces an estimated $\hat{F}$ that may be either smaller or larger than the plug-in value. In other words, all the $\boldsymbol{q}$-vectors different from $\boldsymbol{n}/n$ may tend to either increase or decrease F as compared with $F(\boldsymbol{n}/n)$, and the direction and size of this shift depends on the behavior of $F(\boldsymbol{q})$ around $\boldsymbol{q} = \boldsymbol{n}/n$. In what follows, we propose an alternative expansion of the prior different from Equation (5), which is formulated in terms of a parameter that controls the relative weight of the level surfaces that tend to increase vs. those that tend to decrease F.

Using the previous analysis as an inspiration, from now on, we consider a narrower set of priors that is a proper subset of the one defined by Equation (5). A restriction in the set of possible priors does not invalidate the analysis of how the width of the likelihood depends on the sampled data, it only reduces the set from which priors are selected. We begin by proposing the decomposition

$$p(\boldsymbol{q}) = \int \mathrm{d}f \, p(\boldsymbol{q}|f) \, g(f), \tag{8}$$

for a conveniently chosen family of priors $p(\boldsymbol{q}|f)$ that now need not coincide with $\delta[F(\boldsymbol{q}) - f]$ but are still assumed to only depend on $\boldsymbol{q}$ through the property $F(\boldsymbol{q})$. In other words, we assume that a family of normalizable functions $g_f : \mathbb{R} \to \mathbb{R}$ exists such that

$$p(\boldsymbol{q}|f) = g_f[F(\boldsymbol{q})].$$

Before continuing, it is important to note that in the new decomposition of Equation (8), the parameter f can no longer be identified as exactly the value of the property $F(\boldsymbol{q})$ in a given level surface, since the delta functions are no longer present to restrict $p(\boldsymbol{q}|f)$ to a single level surface of $F(\boldsymbol{q})$. Therefore, from now on, f should be regarded as a formal label that parametrizes the set of base functions used in the expansion. Yet, later on, we choose a family of base functions for which there is still a connection between f and $F(\boldsymbol{q})$, but the connection turns out to be probabilistic (see next subsection).

By inserting Equation (8) in Equation (2), the estimator becomes

$$
\begin{aligned}
\langle F|\boldsymbol{n} \rangle &= \int \mathrm{d}\boldsymbol{q} \, F(\boldsymbol{q}) \, \frac{p(\boldsymbol{n}|\boldsymbol{q})}{p(\boldsymbol{n})} \int \mathrm{d}f \, p(\boldsymbol{q}|f) \, g(f) \\
&= \int \mathrm{d}f \, \frac{g(f)}{p(\boldsymbol{n})} \, p(\boldsymbol{n}|f) \int \mathrm{d}\boldsymbol{q} F(\boldsymbol{q}) \, p(\boldsymbol{n}|\boldsymbol{q}) \, \frac{p(\boldsymbol{q}|f)}{p(\boldsymbol{n}|f)}
\end{aligned}
$$

where we have introduced the likelihood of the data for each value of the parameter,

$$p(\boldsymbol{n}|f) = \int \mathrm{d}\boldsymbol{q} \, p(\boldsymbol{q}|f) \, p(\boldsymbol{n}|\boldsymbol{q}). \tag{9}$$

With this definition, the estimator reads

$$\langle F|\boldsymbol{n} \rangle = \int \mathrm{d}f \, p(f|\boldsymbol{n}) \, \langle F|\boldsymbol{n}, f \rangle, \tag{10}$$

where

$$p(f|\boldsymbol{n}) = \frac{p(\boldsymbol{n}|f)\, g(f)}{p(\boldsymbol{n})} \tag{11}$$

represents the amount of evidence in favor of each f-value provided by the measured data $\boldsymbol{n}$, and

$$
\begin{aligned}
\langle F|\boldsymbol{n},f\rangle &= \int \mathrm{d}\boldsymbol{q}\, F(\boldsymbol{q})\; p(\boldsymbol{q}|\boldsymbol{n},f) \\
&= \int \mathrm{d}\boldsymbol{q}\, F(\boldsymbol{q})\, p(\boldsymbol{n}|\boldsymbol{q})\, \frac{p(\boldsymbol{q}|f)}{p(\boldsymbol{n}|f)}.
\end{aligned}
\tag{12}
$$

is the estimation of the property $F(\boldsymbol{q})$ conditional on the measured data $\boldsymbol{n}$ and the parameter f.

Equation (10) is homologous to Equation (6) for base priors that are not delta distributions. Using Bayes' rule (Equation (11)), the evidence $p(f|\boldsymbol{n})$ can be written in terms of the marginal likelihood $p(\boldsymbol{n}|f)$ of Equation (9). This evidence is defined by an integral in $\boldsymbol{q}$-space that contains the multinomial likelihood $p(\boldsymbol{n}|\boldsymbol{q})$ embodying the Kullback–Leibler divergence. Therefore, just as before, the data $\boldsymbol{n}$ still select a range of f-values; only now, we cannot identify f as an instantiation of F on one specific region. Still, the value of f that maximizes the posterior $p(f|\boldsymbol{n})$ is the one that makes the largest contribution to the integral in Equation (10). Before, only keeping the optimal f-value yielded to almost the plug-in estimator (except for the effect of $g(f)$). In the present case, only keeping the optimal f-value means to replace the integral in Equation (10) by the evaluation of $\langle F|\boldsymbol{n},f\rangle$ at the value of $f = f^*$ that maximizes $p(f|\boldsymbol{n})$. This procedure is henceforth denominated the *MAP estimator*, for *Maximum A Posteriori*. With the present decomposition, this procedure does not yield the plug-in estimator, because $\langle F|\boldsymbol{n},f^*\rangle$ contains an integral that sweeps through a whole range of level surfaces. For the family of base priors selected in the following subsection, the MAP estimator performs substantially better than the plug-in estimator and often also better than custom-made estimators designed for specific properties. In those cases in which the shape of $g(f)$ can be argued to play no relevant role, the MAP estimator can be replaced by an empirical Bayes estimator, in which the selected f^* value maximizes the marginal likelihood $p(\boldsymbol{n}|f)$ instead of the marginal posterior $p(f|\boldsymbol{n})$.

2.4. A Maxentropy Strategy to Select the Base of Functions to Expand the Prior

In this subsection, we select the family of distributions $p(\boldsymbol{q}|f)$ that serve as a base to expand the prior $p(\boldsymbol{q})$, as dictated by Equation (8). To do so, we employ a Maxentropy principle, maximizing the uncertainty about $\boldsymbol{q}$ under the constraint that the expected value of $p(\boldsymbol{q}|f)$ be f (Equation (13)).

Before, when the expansion in delta distributions was used, each element of the base was associated with a single f value, equal to the property F evaluated on the corresponding level surface. A natural relaxation of this condition, while still fulfilling the requirement that the elements $p(\boldsymbol{q}|f)$ have the same level surfaces as $F(\boldsymbol{q})$, is to demand that f be the *expected* value of $F(\boldsymbol{q})$, when the average is weighted by $p(\boldsymbol{q}|f)$. If, after imposing this requirement, we insist that the base functions have no additional structure, then the shape of $p(\boldsymbol{q}|f)$ can be derived from the Maxentropy principle [13,14], in which the differential entropy $\mathcal{H}[p(\boldsymbol{q}|f)]$ is maximized, conditioned to the restriction

$$\int \mathrm{d}\boldsymbol{q}\, F(\boldsymbol{q})\, p(\boldsymbol{q}|f) = f. \tag{13}$$

The solution of this maximization problem is

$$p(\boldsymbol{q}|f) = \frac{\mathrm{e}^{\beta F(\boldsymbol{q})}}{\int \mathrm{d}\boldsymbol{q}'\mathrm{e}^{\beta F(\boldsymbol{q}')}} = \frac{\mathrm{e}^{\beta F(\boldsymbol{q})}}{\mathcal{Z}_0(\beta)}, \tag{14}$$

where the hyperparameter β is a function of f; that is, of the expected value of the property. The correspondence between each f-value and each β-value implies that both these pa-

rameters constitute valid tags to designate a member of the base. For this reason, in what follows, we use the two parametrizations interchangeably, depending on what we want to stress. We warn the reader that we pass from one to the other, tagging the elements of the base equivalently as $p(q|f)$ or as $p(q|\beta)$, understanding that there is a one-to-one mapping between f and β. In the nomenclature of Amari [15,16], the parameter appearing linearly in the exponent of the distribution (for us, β) is referred to as an *exponential* system of coordinates of the space of parameters. When f is used, the coordinates are called *mixed*. In equilibrium statistical mechanics, for example, β is often proportional to the negative of the inverse of the temperature, whereas f is proportional to the mean energy of a state—and of course, the two are related.

In the exponential family of Equation (14), all level surfaces of $F(q)$ contribute to each $p(q|f)$, but some are more relevant than others. If β is large and positive, $p(q|f)$ is dominated by the level surface with maximal $F(q)$ and rapidly dies out when we depart from it. The mean value of $F(q)$ weighted by this $p(q|\beta \to +\infty)$ is equal to the maximum of F on the simplex. If $\beta = 0$, all level surfaces contribute uniformly. If β is large and negative, the level surface with minimal $F(q)$ has maximal relevance, and the mean value of $F(q)$ is equal to the minimum of F on the simplex. Therefore, β operates as a tuning knob that raises or lowers the relevance of different level surfaces, ranking them by the value of F. As happened when expanding the prior in a base of delta functions, the mean value of F shifts from its minimum to its maximum by adjusting the parameter. However, in contrast with the delta case, the expansion in exponential functions allows each element of the base to spread out to different level surfaces, allowing a whole diversity of surfaces to contribute to each $\langle F|\beta \rangle$.

2.5. The Estimation of the Property for Each Member of the Base

In this subsection, we show how the problem of estimating $F(q)$ for each fixed β value is formally equivalent to obtaining the mean value of a certain physical quantity in statistical mechanics. In this analogy, the data n of the inference problem produce exponential distortions of the partition function of the Statistical Mechanical formulation.

The posterior distribution conditioned on a specific member β of the base is

$$
\begin{aligned}
p(q|n,\beta) &= \frac{p(n|q)\,p(q|\beta)}{p(n|\beta)} \\
&= \frac{e^{\beta F(q)}\,\prod_j q_j^{n_j}}{\int dq'\,e^{\beta F(q')}\,\prod_j q'^{n_j}_j} \\
&= \frac{e^{\beta F(q)}\,\prod_j q_j^{n_j}}{\mathcal{Z}(n,\beta)}.
\end{aligned}
\tag{15}
$$

Replacing this expression in Equation (12)

$$
\langle F|n,\beta \rangle = \frac{1}{\mathcal{Z}(n,\beta)} \int dq\, F(q)\, e^{\beta F(q)} \prod_j q_j^{n_j},
$$

where the posterior partition function $\mathcal{Z}(n,\beta)$ is calculated with the product of the prior distribution $p(q|\beta)$ and the likelihood $p(n|q)$. Just as in statistical mechanics [13], this integral is equal to

$$
\langle F|n,\beta \rangle = \partial_\beta \log \mathcal{Z}(n,\beta).
\tag{16}
$$

Moreover, following the same reasoning used in statistical mechanics, the variance of the estimation reads

$$
\langle \Delta F^2|n,\beta \rangle = \partial_\beta^2 \log \mathcal{Z}(n,\beta).
$$

Importantly, in several applications, the partition function $\mathcal{Z}(n,\beta)$ can be calculated analytically. In those cases where an analytical treatment is impossible, the shape of $\mathcal{Z}(n,\beta)$

can often be well approximated by one-dimensional numerical integrals, which still enables the computation of $\langle F | \boldsymbol{n}, \beta \rangle$.

2.6. Considering Different β Values

In this subsection, we demonstrate that out of all the possible β-values, the one that has maximal influence on the estimation process (or equivalently, the one that tags the most relevant base prior) depends on two independent factors. The first factor depends on the prior density $g(\beta)$, comprising all the subjective elements of the inference process. The second factor is that for which the prior and posterior estimations of $F(\boldsymbol{q})$ coincide and is entirely determined by the sampled data $\boldsymbol{n}$. This result allows us to determine the role played by the different factors entering in the inference process.

Equation (10) states that the full estimation $\langle F | \boldsymbol{n} \rangle$ results from integrating $\langle F | \boldsymbol{n}, \beta \rangle$ in β, each contribution weighted by the evidence $p(\beta | \boldsymbol{n})$. Thus, we now analyze this evidence.

When $p(\boldsymbol{q} | \beta)$ is given by Equation (14), the evidence of Equation (11) becomes

$$\log p(\beta | \boldsymbol{n}) \propto \log p(\beta) + \log p(\boldsymbol{n} | \beta) = \log p(\beta) + \log \left[\frac{\mathcal{Z}(\boldsymbol{n}, \beta)}{\mathcal{Z}_0(\beta)} \right],\tag{17}$$

where the constant of proportionality does not depend on β.

The first term of Equation (17) contains the prior expectations on β. The second one is the marginal likelihood $p(\boldsymbol{n} | \beta)$ and describes the way in which the data $\boldsymbol{n}$ favor a specific range of level surfaces. It is equal to the ratio of the posterior and the prior partition functions. The β value that makes the largest contribution to this ratio, henceforth called β_0, can be found by taking the derivative of the logarithm of the ratio and setting it to zero,

$$\partial_\beta \log p(\boldsymbol{n} | \beta) \Big|_{\beta_0} = \langle F | \boldsymbol{n}, \beta_0 \rangle - \langle F | \beta_0 \rangle = 0.\tag{18}$$

Therefore, β_0 is the parameter for which the prior and posterior estimations of the property coincide. This idea is illustrated in the last panel in Figure 1.

3. Applications to Different Properties

In this section, we apply the method to estimate the marginal probability of a bivariate distribution (Section 3.1), the amount of mutual information (Section 3.2), and the entropy (Section 3.3). All these quantities are examples of properties $F(\boldsymbol{q})$ that have been extensively discussed in the literature [3,7,17–19]. Yet, as far as we know, all the methods developed so far—at least, those that perform decently in the severe undersampled regime—were idiosyncratically tailored to each specific property, whereas the method proposed here is general. In this subsection, we compare its performance to the more handcrafted approaches.

3.1. Marginal Probability

As a first example, we consider a system whose states are tagged by two labels: x, with many equiprobable states, and $y \in \{0, 1\}$, with only two. The property we are interested in estimating is the marginal probability of $y = 1$, denoted as q_1, defined as

$$q_1(\boldsymbol{q}) = \sum_x q_x \, q_{1|x} = \frac{1}{k_x} \sum_x q_{1|x} \,,$$

were k_x is the number of x-values. Since we assume all x-values have the same probability, here, $\boldsymbol{q}$ represents the collection of conditional probabilities $\{q_{1|x}\}_{x=1}^k$. This case constitutes a toy example, since the plug-in estimator $\hat{q}_1 = n_1 / n$ is unbiased and efficient, irrespective of the value of n_x. Still, we discuss this case to exemplify the procedure described in the previous section.

According to this procedure, the base functions used to expand the prior are solutions of a Maxentropy problem conditioned to have $q_1(\boldsymbol{q}) = f$, which in exponential coordinates reads (Equation (14)),

$$p(\boldsymbol{q}|\beta) \propto \exp\left(\frac{\beta}{k_x}\sum_x q_{1|x}\right) \propto \left(\prod_x e^{q_{1|x}}\right)^{\beta/k_x} \propto \prod_x p(q_{1|x}|\beta). \qquad (19)$$

For each β, this solution factorizes into independent terms. The exponential function favors either small or large values of $q_{1|x}$, depending on the sign of β. As the likelihood also factorizes $p(\boldsymbol{n}|\boldsymbol{q}) \propto \prod_x q_{1|x}^{n_{1x}}(1 - q_{1|x})^{n_{0x}}$, so does the posterior,

$$p(q_{1|x}|n_{1x}, n_{0x}, \beta) = \frac{q_{1|x}^{n_{1x}}(1 - q_{1|x})^{n_{0x}}\, e^{\beta q_{1|x}/k_x}}{M(n_{1x} + 1, n_x + 2, \beta/k_x)}, \qquad (20)$$

where $M(a, b, z)$ is the confluent hypergeometric function [20]. Then, the partition function corresponds to $\mathcal{Z}(\mathbf{n}, \beta) = \prod_x M(n_{1x} + 1, n_x + 2, \beta/k_x)$. The prior partition function $\mathcal{Z}_0(\beta)$ can always be obtained from $\mathcal{Z}(\mathbf{n}, \beta)$ by setting $\boldsymbol{n} = \mathbf{0}$.

In Equation (20), the exponential factor is inherited from the prior $p(q_{1|x}|\beta)$. In general, in the method proposed here, each β biases the estimation towards the corresponding f value in order to get an unbiased estimate of the property when all the β values are considered.

With the partition function, we derive the estimate for the property q_1 of each member β of the family as

$$\langle q_1|\boldsymbol{n}, \beta\rangle = \frac{1}{k_x}\sum_x \langle q_{1|x}|n_{1x}, n_{0x}, \beta\rangle = \frac{1}{k_x}\sum_x \frac{(n_{1x} + 1)}{(n_x + 2)}\frac{M(n_{1x} + 2, n_x + 3, \beta/k_x)}{M(n_{1x} + 1, n_x + 2, \beta/k_x)}. \qquad (21)$$

The parameter β_0 that makes a maximal contribution to the estimation is the one for which $\langle q_1|\boldsymbol{n}, \beta_0\rangle = \langle q_1|\beta_0\rangle$, as established in Equation (18). This equations is often difficult to solve analytically. However, in the present case, it can be easily solved in the extreme undersampled regime ($n \ll k_x$) with no coincidences ($n_x \leq 1$), where there is a fraction $f_1 = n_1/n$ of x states with $y = 1$ and a fraction $(1 - f_1)$ of states with $y = 0$. This condition translates into

$$(1 - f_1)\frac{M(2, 4, \gamma)}{3M(1, 3, \gamma)} + f_1\frac{2M(3, 4, \gamma)}{3M(2, 3, \gamma)} = \frac{M(2, 3, \gamma)}{2M(1, 2, \gamma)} = \langle q_1|\boldsymbol{n}, k_x\,\gamma\rangle, \qquad (22)$$

where $\gamma = \beta/k_x$. Taking into account the contiguous relations of the confluent hypergeometric function, this equation leads to $\langle q_1|\boldsymbol{n}, \beta_0\rangle = f_1 = n_1/n$, which is the naïve, plug-in, unbiased estimator. Usually for other properties, the plug-in estimator provides quite poor results in the severe undersampled regime, and there is no clear path to a low bias estimator. We believe that the proposed approach, focused on priors defined on the level surfaces, can be useful in these cases. In the two following subsections, we analyze other more challenging properties such as the mutual information and the entropy.

3.2. Mutual Information

The estimation of the amount of mutual information in a severely undersampled discrete system is a difficult task [2,21–25]. Several of the approaches developed previously [3,7] are based on separate estimators for the total and conditional entropies. Taking a different path, in a previous publication, we also developed a method that was focused on the dispersion of the conditional probabilities in relation to the marginal [8]. The performance was good, but still, the method was only applicable to mutual information.

In order to test the new method, we work with the same system introduced in the previous subsection. We assume $k_x \gg 1$, and all x-values have with equal probability $q_x = 1/k_x$, $\forall x$. The two states for the label $y \in \{0, 1\}$ are assumed to have equal marginal

probabilities, $q_y = {}^1\!/_2$, $\forall y$ (generalizations are explored in [8]). For this system, we want to find the mutual information $I = I(X;Y)$, namely

$$
\begin{aligned}
I &= \sum_{xy} q_x \, q_{y|x} \, \log\left(\frac{q_{y|x}}{q_y}\right) \\
&= \frac{1}{k_x} \sum_x \Big[\log 2 + q_{1|x} \log q_{1|x} + (1 - q_{1|x}) \log(1 - q_{1|x})\Big] = \frac{1}{k_x} \sum_x I_x(q_{1|x}).
\end{aligned}
\tag{23}
$$

As in the previous example, the property $I(\boldsymbol{q})$ depends on the set of conditional probabilities $\boldsymbol{q} = \{q_{1|x}\}_{x=1}^{k_x}$.

The prior $p(\boldsymbol{q}|\beta)$ of Equation (14) reads

$$
p(\boldsymbol{q}|\beta) \propto \exp[\beta I(\boldsymbol{q})] \propto \prod_x p(q_{1|x}|\beta).
\tag{24}
$$

Again, the prior factorizes. The individual factors are

$$
p(q_{1|x}|\beta) \propto \exp\left[\frac{\beta}{k_x} I_x(q_{1|x})\right] \propto \left[q_{1|x}^{q_{1|x}} \, (1 - q_{1|x})^{(1 - q_{1|x})}\right]^{\beta/k_x}.
\tag{25}
$$

This prior resembles the beta distribution $\propto q_{1|x}^{\nu} \, (1 - q_{1|x})^{\nu}$ proposed in [8] but has non-constant exponents. Each posterior reads

$$
p(q_{1|x}|n_{1x}, n_{0x}, \beta) = \frac{q_{1|x}^{n_{1x}} (1 - q_{1|x})^{n_{0x}} \, e^{\beta I_x(q_{1|x})/k_x}}{\mathcal{Z}(n_{1x}, n_{0x}, \beta)}.
\tag{26}
$$

This functional form does not allow us to continue the problem in an analytically tractable fashion. Yet, the factorization makes the problem uni-dimensional, so it is easy to calculate the partition function for different pairs of $\{n_{1x}, n_{0x}\}$ and the corresponding expectation values. Indeed, the partition function factorizes as $\mathcal{Z}(\boldsymbol{n}, \beta) = \prod_x \mathcal{Z}(n_{1x}, n_{0x}, \beta)$.

The estimated value of the amount of mutual information of each β-value—the member's guess— can be obtained with Equation (16), and the β-value that maximizes the marginal likelihood—the member's evidence— from Equation (18). Interestingly, not all x-states contribute to the marginal likelihood,

$$
\log p(\boldsymbol{n}|\beta) = \log\left[\frac{\mathcal{Z}(\boldsymbol{n}, \beta)}{\mathcal{Z}_0(\beta)}\right] = \sum_x \log\left[\frac{\mathcal{Z}(n_{1x}, n_{0x}, \beta)}{\mathcal{Z}(0, 0, \beta)}\right]
\tag{27}
$$

since the only non-vanishing terms are those in which the partition function associated to the x state is not proportional to the prior partition function; that is, $\mathcal{Z}(n_{1x}, n_{0x}, \beta)/\mathcal{Z}(0, 0, \beta)$ varies with β. For mutual information, this condition is only met by x states with coincidences ($n_x \geq 2$). In fact, only when coincidences begin to appear ($n \gtrsim \sqrt{k_x}$) can the sampled data start to provide evidence about a non-vanishing mutual information. No such condition was found in the previous section, since there, every sampled x-state contributed to the estimation of the marginal probability. This discrepancy stems from the fact that, while the marginal probability depends on the first moment of the set $\{q_{1|x}\}$ of conditional probabilities, the mutual information relates to the second moment, or the amount of dispersion, of this set [8]. We show a toy example that illustrates this approach in Figure 2. In this example, the proposed method performs as well as some of the best known estimators (Figure 2e) designed for such undersampled regime.
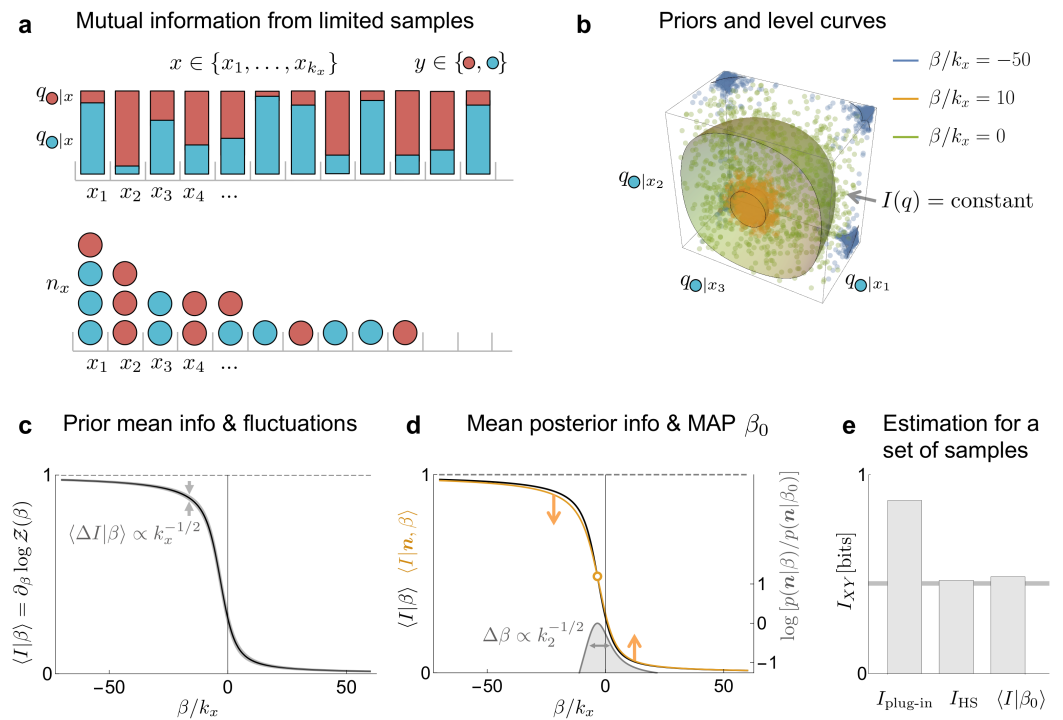
Figure 2. Toy example used to illustrate the estimation of the amount of mutual information in the same system as in Section 3.1. (**a**) The system is governed by a bivariate probability distribution, with states characterized by two labels: x and y. There are many x states, $k_x \gg 1$, and all are equally probable; that is, $q_x = 1/k_x$. The variable y is binary, and its two values are depicted as red and blue, corresponding to 1 and 0, respectively. The conditional probabilities were sampled from a symmetric beta distribution with parameter 0.5. The goal is to estimate the mutual information I_{XY} from n samples, with $n \gtrsim k_x$. (**b**) The level surfaces of the mutual information, as well as some sampled q-values, are displayed in a three-dimensional subspace of the full q-space, for different values of the hyperparameter β. (**c**) Prior mean mutual information as a function of the scaled hyperparameter β/k_x, and its fluctuations for $k_x = 100$. (**d**) Prior and posterior mean mutual information as a function of the scaled hyperparameter β/k_x for $n = 60$ samples. The intersection of these curves corresponds to the MAP estimation $\langle I | \beta_0 \rangle$. In gray, the posterior marginal evidence for the hyperparameter, whose width decreases as the square root of the number of states with at least two samples, k_2. (**e**) Comparison of the estimation of mutual information between different methods for the considered set of samples (I_{HS} is the estimator from [8]). The horizontal line corresponds to the true value of the mutual information.

3.3. Entropy

Finally, we consider the estimation of entropy, for which multiple estimators have been proposed in the past [1,3–6,26–28]. The entropy $H(q)$ is defined as

$$H(q) = -\sum_x q_x \log q_x. \tag{28}$$

Here, we use the natural logarithm for analytical convenience. We assume that the number of accessible states is unknown and that the number of effective states is quite large, so that $\exp[H(q)] \gg 1$.

Once more, we expand the prior in a family whose members are tagged with β and use Equation (14) to express each member of the family as

$$p(q|\beta) \propto \exp[\beta H(q)] = \prod_x q_x^{-\beta q_x}, \tag{29}$$

where q belongs to the simplex $\sum_x q_x = 1$. We can compare this prior to that used by NSB [3], a symmetric Dirichlet $p(q|\beta) \propto \prod_x q_x^{\beta-1}$. Although there are some similarities, the prior

in Equation (29) has non-constant exponents, which hamper the analytical tractability. The posterior reads

$$p(\boldsymbol{q}|\boldsymbol{n},\beta) = \frac{\delta\left(\sum_j q_j - 1\right)\exp[\beta H(\boldsymbol{q})]\prod_x q_x^{n_x}}{\mathcal{Z}(\boldsymbol{n},\beta)} = \frac{\delta\left(\sum_j q_j - 1\right)\prod_x q_x^{n_x - \beta q_x}}{\int \mathrm{d}\boldsymbol{q}'\,\delta\left(\sum_j q'_j - 1\right)\prod_x q'^{(n_x - \beta q'_x)}_x}. \quad (30)$$

To calculate the member's estimation $\langle H|\boldsymbol{n},\beta\rangle$ and the marginal likelihood $p(\boldsymbol{n}|\beta)$, the partition function $\mathcal{Z}(\boldsymbol{n},\beta)$ is needed.

We have found no evident way to solve the partition function analytically. Yet, its shape can be obtained by numerically integrating the partial convolutions that the delta over the simplex implies. With this result, we were able to obtain the prior and posterior estimates of the entropy, as well as the optimal β (Figure 3c). The MAP estimator of the entropy obtained with this procedure, even if numerical, was more accurate than the results derived from all previous methods we know of (Figure 3d).
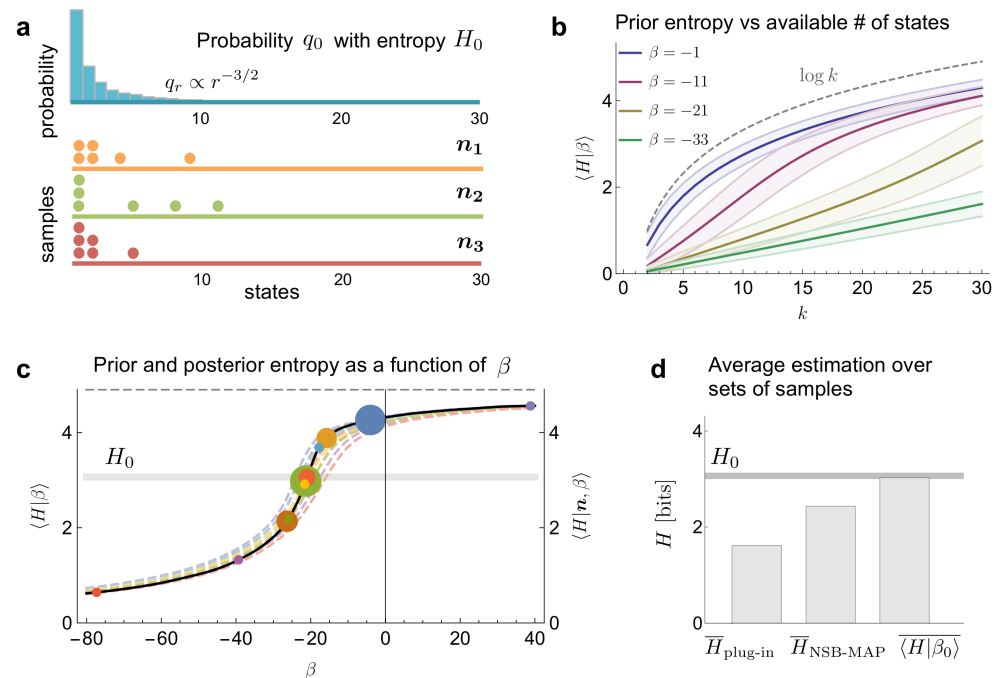

Figure 3. Entropy estimation in a toy example. (**a**) A distribution q_0 has ranked probabilities that decrease with a power law ($q_r \propto r^{-3/2}$) within a finite number of states ($k = 30$). Distributions such as these represent a challenge for the estimation of entropy. Three possible sets of $n = 6$ samples are displayed. (**b**) Prior mean entropy, plus/minus a standard deviation, as a function of the available number of states k for several values of the hyperparameter β. (**c**) Prior and posterior mean entropy as function of β for the different possible sets of multiplicities that can be obtained from $n = 6$ samples (three examples shown in panel a, with matching colors). The intersections between prior and posterior mean entropy (MAP estimation) are marked with circles. The size of the circles is proportional to the likelihood of the corresponding multiplicity (a minimum size is imposed, for visibility). (**d**) Comparison of the average estimation of entropy between different methods over all the multiplicity sets (discarding the set with no coincidences, and the set with all samples in one state), weighting according to their likelihood. The horizontal line corresponds to the true value of the entropy.

4. Discussion

In this paper, we have proposed a novel Bayesian estimator to make an inference about a one-dimensional property of a high-dimensional system. Our work focused on the selection of the prior. The method takes advantage of the drastic reduction of dimen-

sionality induced by the one-dimensional nature of the property of interest and employs a Maxentropy argument to eliminate the remaining indeterminate factors, thereby explicitly exploiting both what we know and what we do not know about the system. The approach is general in the sense that it can be applied to arbitrary properties and has excellent performance when compared to previous methods. Here, we discuss and summarize the main points of the paper.

Conditions of Validity

When inferring the value of properties $F(\boldsymbol{q})$, three sampling regimes are relevant. If the number of samples n is significantly smaller than $\sqrt{\exp[H(\boldsymbol{q})]}$, the sample is likely to contain no coincidences [1], and in this regime, no inference is possible. This study becomes useful when $\sqrt{\exp[H(\boldsymbol{q})]} \gtrsim n$, but still, the order of magnitude of n is not significantly larger than $\exp[H(\boldsymbol{q})]$. For a still larger n, the estimation of the probabilities q_i begins to be feasible, so it begins to make sense to consider the plug-in estimator.

Identifying the Factors That Matter When Searching for the Prior

The Bayesian approach is argued to be the only rational way to infer the value of an unknown property. Yet, the success of Bayesian inference relies on an adequate choice of the prior. If an unjustified prior is used, only because there are no evident reasons to choose a specific prior, the result of the inference is not guaranteed to make sense. It has been argued that more than a drawback of the Bayesian approach, this sensitivity to the prior stresses the fact that it is impossible to make inference without assumptions [29]. Yet, it is often the case that selecting a prior is difficult, and this is problematic when the result is too sensitive to the choice, which is often the case in the severe undersampled regime. Without intending to diminish the relevance of priors, in this paper, we have shown that not all aspects of the prior distribution are relevant, and thereby we hope to focus the efforts of the selection process on the relevant aspects. For example, we have proved that all the modulations of the prior within the intersection between a level surface of the property and a level surface of the likelihood bear no consequences on the result of the inference. This observation, and the understanding that the prior distribution cannot depend explicitly on the sampled data, reduce the search of a high-dimensional prior $p(\boldsymbol{q})$ to a one-dimensional prior $g(f)$. Moreover, in many circumstances, even the detailed definition of $g(f)$ may be more than required, although this statement only holds when the sampled data contain a non-negligible fraction of coincidences. When this condition holds, the data themselves select the range of f-values that actually contribute to the estimation. Therefore, only the variations of the prior distribution $g(f)$ within this range actually matter.

The Core of the Approach: Decomposing the Prior in a Conveniently Chosen Base of Priors

To make this selective role of the data explicit, and at the same time to be able to get analytical results, in this paper, we proposed the expansion of the prior distribution as a linear combination of base prior functions, each of which was required to have a different expected value of the property under study. By imposing no additional structure on the base priors (this requisite was instantiated by Maxentropy), the distributions of the base were derived to depend exponentially on the property under study. The most relevant element of the base (at least, inasmuchas the relevance was dictated by the data and not by $g(f)$) was shown to be the one for which the prior and posterior estimation of the property coincided (Equation (18)). This result is similar to the equality between the prior and posterior mean energy reported in the framework of statistical inference [30], and it relates to a scenario of parameter mismatch, as well as to the algorithm of expectation maximisation [31]. Even if the ratio of partition functions sometimes cannot be calculated analytically, the goal of this paper is to reveal the formal similarity between the two problems and thereby to prompt the reader to consider the battery of numerical algorithms that have been developed in expectation maximization to solve the type of one-dimensional problem posed here.

Assessing the Effectiveness of the Method

In Section 3, we tested the performance of the method for three well-known properties, for which many estimators exist. Even though our method is general, it accomplished good results when compared to those produced by other methods specifically designed for each individual property. Interestingly, the performance was good even using a single β-value; namely, the one for which the marginal likelihood $p(\beta|\boldsymbol{n})$ was maximal, without needing to integrate in β, and therefore, without needing to define the prior $g(\beta)$. One fortunate ingredient was given by the fact that in those examples, as well as in many cases of interest, the property to be estimated was additive and symmetric in the components q_i. Properties with these characteristics can be written as $F(\boldsymbol{q}) = \sum_j \varphi(q_j)$, for some function $\varphi : \mathbb{R} \to \mathbb{R}$. When the effective number of states $\exp[H(\boldsymbol{q})]$ grows, the constraint imposed by the normalization condition becomes increasingly irrelevant, so the sampled q_i can be assumed to be fairly independent of one another. In this case, the central limit theorem implies that the relative fluctuations of the property $\sigma_{F|\boldsymbol{n},\beta}/\langle F|\boldsymbol{n},\beta\rangle$ decay as the inverse of the square root of the effective number of states. Therefore, each member of the base makes a rather definite posterior guess $\langle F|\boldsymbol{n},\beta\rangle$ with minimal fluctuations, implying that even though the states $p(\boldsymbol{q}|\beta)$ spread all over the simplex, the posterior property is sharply defined within the state, as happened with the expansion in delta-like priors.

Future Perspectives

There are several directions in which this work can be continued. From the strictly theoretical point of view, in Section 2.3, we argued that the optimal decomposition makes use of prior base distributions, the density of which is compressed inside a level surface, and there, the density is constant (Equation (5)). This choice, however, makes the obtainment of the partition function in analytical form unfeasible, since the integrals that need to be calculated are typically intractable. Therefore, we have opted to expand the prior in a base composed of distributions of fixed mean value. Future work should either demonstrate the optimality of this choice or suggest other possibilities.

It is also important to make a more systematic characterization of the relevant statistics in the data. For example, here, we have shown that the estimation of mutual information is only sensitive to states that have been sampled more than once. What is the relative effect of states sampled $2, 3, \ldots, s$ times? Which are the relevant statistics determining the estimation of other properties different from mutual information? More importantly, are there thresholds in the amount of sampled data for the possibility to make inference at all [30]? How do these thresholds depend on the property to be estimated?

From a more practical perspective, the excellent performance obtained for estimations of entropy and mutual information should also be evaluated in broader conditions; for example, for non-uniform marginals, or for sets with arbitrary cardinality. Moreover, the method should also be applied to a broader collection of properties, such as the estimation of Pearson correlation coefficients, slopes in linear regression, skewness, kurtosis, conditional mutual information [12], transfer entropy [32], etc.

Conclusions

In summary, the method proposed here has the following attractive properties:

(a) It offers a general framework to select priors that is valid for arbitrary properties $F(\boldsymbol{q})$;
(b) It performs well in the estimation of well-known quantities, as marginal distributions, mutual information and entropy, with performance comparable or even better than previous estimators individually tailored for each property;
(c) It reveals the formal similarity between inference problems and equilibrium statistical mechanics. In particular, the parameter β is analogous to the inverse temperature of statistical physics, and the two alternative expansions of the prior (in a delta-like or an exponential-like base) correspond to the microcanonical and the canonical ensembles, respectively.

References

1. Nemenman, I. Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy* **2011**, *13*, 2013–2023. [CrossRef]
2. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
3. Nemenman, I.; Bialek, W.; van Steveninck, R.d.R. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E* **2004**, *69*, 056111. [CrossRef] [PubMed]
4. Archer, E.; Park, I.M.; Pillow, J.W. Bayesian entropy estimation for countable discrete distributions. *J. Mach. Learn. Res.* **2014**, *15*, 2833–2868.
5. 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]
6. Grassberger, P. Entropy estimates from insufficient samplings. *arXiv* **2003**, arXiv:physics/0307138.
7. 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]
8. Hernández, D.G.; Samengo, I. Estimating the Mutual Information between Two Discrete, Asymmetric Variables with Limited Samples. *Entropy* **2019**, *21*, 623. [CrossRef]
9. Contreras-Reyes, J.E. Mutual information matrix based on asymmetric Shannon entropy for nonlinear interactions of time series. *Nonlinear Dyn.* **2021**, *104*, 3913–3924. [CrossRef]
10. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer: Berlin/Heidelberg, Germany, 1998.
11. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
13. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *108*, 171–190. [CrossRef]
14. Jaynes, E.T. Information Theory and Statistical Mechanics II. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
15. Amari, S.I.; Hiroshi, N. *Methods of Information Geometry*; Oxford University Press: Oxford, UK, 2000.
16. Amari, S.I. Information Geometry on Hierarchy of Probability Distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711. [CrossRef]
17. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Netw. Comput. Neural Syst.* **1996**, *7*, 87. [CrossRef] [PubMed]
18. Samengo, I. Estimating probabilities from experimental frequencies. *Phys. Rev. E* **2002**, *65*, 046124. [CrossRef] [PubMed]
19. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
20. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Dover Publications: New York, NY, USA, 1965.
21. 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]
22. Kolchinsky, A.; Tracey, B.D. Estimating mixture entropy with pairwise distances. *Entropy* **2017**, *19*, 361. [CrossRef]
23. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual information neural estimation. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
24. Safaai, H.; Onken, A.; Harvey, C.D.; Panzeri, S. Information estimation using nonparametric copulas. *Phys. Rev. E* **2018**, *98*, 053302. [CrossRef]
25. Holmes, C.M.; Nemenman, I. Estimation of mutual information for real-valued data with error bars and controlled bias. *Phys. Rev. E* **2019**, *100*, 022404. [CrossRef]
26. Miller, G. Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*; Free Press: Glencoe, IL, USA, 1955.
27. 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*, 429–443. [CrossRef]

28. Berry II, M.J.; Tkačik, G.; Dubuis, J.; Marre, O.; da Silveira, R.A. A simple method for estimating the entropy of neural activity. *J. Stat. Mech. Theory Exp.* **2013**, *2013*, P03015. [CrossRef]
29. MacKay, D. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
30. Zdeborová, L.; Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Adv. Phys.* **2016**, *65*, 453–552. [CrossRef]
31. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22.
32. Staniek, M.; Lehnertz, K. Symbolic transfer entropy. *Phys. Rev. Lett.* **2008**, *100*, 158101. [CrossRef] [PubMed]