Entropy **2013**, *15*, 4668-4699; doi:10.3390/e15114668

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Estimating Functions of Distributions Defined over Spaces of Unknown Size

David H. Wolpert [1,2,*] **and Simon DeDeo** [1,3]

[1] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA; E-Mail: simon@santafe.edu

[2] Information Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87501, USA

[3] School of Informatics and Computing, Indiana University, 901 E 10th St, Bloomington, IN 47408, USA

* Author to whom correspondence should be addressed; E-Mail: david.h.wolpert@gmail.com.

Received: 3 August 2013; in revised form: 11 September 2013 / Accepted: 17 October 2013 / Published: 31 October 2013

Abstract: We consider Bayesian estimation of information-theoretic quantities from data, using a Dirichlet prior. Acknowledging the uncertainty of the event space size m and the Dirichlet prior's concentration parameter c, we treat both as random variables set by a hyperprior. We show that the associated hyperprior, $P(c, m)$, obeys a simple "Irrelevance of Unseen Variables" (IUV) desideratum iff $P(c, m) = P(c)P(m)$. Thus, requiring IUV greatly reduces the number of degrees of freedom of the hyperprior. Some information-theoretic quantities can be expressed multiple ways, in terms of different event spaces, e.g., mutual information. With all hyperpriors (implicitly) used in earlier work, different choices of this event space lead to different posterior expected values of these information-theoretic quantities. We show that there is no such dependence on the choice of event space for a hyperprior that obeys IUV. We also derive a result that allows us to exploit IUV to greatly simplify calculations, like the posterior expected mutual information or posterior expected multi-information. We also use computer experiments to favorably compare an IUV-based estimator of entropy to three alternative methods in common use. We end by discussing how seemingly innocuous changes to the formalization of an estimation problem can substantially affect the resultant estimates of posterior expectations.

Keywords: Bayesian analysis; entropy; mutual information; variable number of bins; hidden variables; Dirichlet prior

---

1. Background

A central problem of statistics is estimating a functional of a probability distribution ρ from a dataset, $\vec{n}$, of N independent, identically distributed (IID) samples of ρ. A simple example of this problem is estimating the mean $Q(\rho)$ of a distribution, ρ, from a set of IID samples of that distribution. In this example, the functional $Q(.)$ depends linearly on ρ. More challenging versions of this problem arise when the functional $Q(.)$ is nonlinear. In particular, recent decades have seen a lot of work on estimating information-theoretic functionals [1,2] of a distribution from a set of IID samples of that distribution. Examples include estimating the Shannon entropy of the distribution, its mutual information, *etc*., from such a set of samples and, in particular, using the bootstrap to estimate associated error bars [3–5].

This work has concentrated on the case where the event space being sampled is countable. In addition, much of it has used non-Bayesian approaches. The first work that addressed the problem using Bayesian techniques was the sequence of papers [6–8] (hereafter abbreviated as WW), followed by similar work in [9] (see Appendix A for a list of corrections to some algebraic mistakes in [6]; a careful analysis of the numerical implementation of the formulas in WW can be found in [10]). This work showed how to calculate posterior moments of several nonlinear functionals of the distribution. Such moments provide both the Bayes-optimal estimate of the functional (assuming a quadratic loss function) and an error bar in that estimate. In particular, WW provided closed-form expression for the posterior first and second moments of entropy, mutual information and Kullback-Leibler distance, in addition to various non-information-theoretic quantities, like covariance.

Write the space of possible events as Z with elements written as z and distributions over Z as ρ. For tractability reasons, WW used a Dirichlet prior over the associated simplex of possible distributions, $P(\rho) \propto \prod_z \rho(z)^{cL(z)-1}$. (In the literature, the constant, c, is sometimes called a concentration parameter, and L is sometimes called a baseline distribution.)

In WW, Z was taken to be fixed (not a random variable), $L(z)$ was taken to be uniform over Z and c was taken to equal $|Z|$, the size of Z. This choice of a Dirichlet prior over ρ with uniform L has been the basis of all subsequent work on Bayesian estimates of information-theoretic functionals of distributions [11,12]. (Note though that recently, there has been some investigation of the extension of this work to mixture-of-Dirichlet distributions [12] and Dirichlet / Pittman-Yor processes [11]).) However, there has not been such consensus concerning c.

Although the choice of $c = |Z|$ was not explicitly advocated in WW, all the results in WW are for this special case. An important series of papers [12–14] (hereafter abbreviated as NSB) considered the generalization where $c = a|Z|$ for any positive constant, a. (WW is the special case where $a = 1$.) NSB considered the limit where we have such a c, and $|Z|$ is much larger than N, the number of samples of ρ. (Therefore, for non-infinitesimal a, $c \gg N$.) They showed that in this limit, the samples have little effect on the posterior moments of Q (e.g., for Q the Shannon entropy). Therefore, the data become irrelevant. In that large $|Z|$ limit, the posterior moments are dominated by the prior over ρ that is specified by c. This can be seen as a major shortcoming of WW.

To address this shortcoming, NSB noted that if c is a random variable with an associated prior, $P(c)$, it induces a prior distribution over the values of the Shannon entropy, $H(\rho) = -\sum_z \rho(z) \ln[\rho(z)]$. Specifically, for the case of a uniform L, for any potential entropy value h:

$$\begin{aligned} P(H = h) &= \int dc d\rho \; \delta(H(\rho) - h) P(\rho \mid c) P(c) \\ &= \int dc d\rho \; \delta(H(\rho) - h) \frac{\prod_z \rho(z)^{c/|Z|-1}}{\int d\rho' \; \prod_z \rho'(z)^{c/|Z|-1}} P(c) \end{aligned} \tag{1}$$

where the integrals over distributions are implicitly restricted to the associated simplex. NSB then decided that since their goal was to estimate entropy, they would set $P(c)$ so that the prior, $P(H)$, is flat. In essence, they formed a continuous mixture of Dirichlet distributions to (attempt to) obtain a flat prior over the ultimate quantity of interest, H. The $P(c)$ that results in flat $P(H)$ cannot be written down in closed form, but NSB showed how numerical computations can be used to approximate it.

As they are used in practice, both NSB and WW allow $|Z|$ to vary, giving it different values for different problems, values that are typically set in an *ad hoc* way. (Indeed, if they did not allow the number of bins to vary from one dataset to another, they could not be used on problems with datasets running over too large a number of bins.) They then set $P(c)$ based on that fixed value of $|Z|$.

One problem with this is that it means the posterior expected value of the functional, Q, can vary depending on how one expresses that functional. To illustrate this, say that Z is a Cartesian product, $X \times Y$, and let $H(X,Y)$, $H(X)$ and $H(Y)$ refer to the entropies of $\rho(x,y) \equiv \rho_{X,Y}(x,y)$, $\rho(x) \equiv \rho_X(x) \equiv \sum_y \rho(x,y)$ and $\rho(y) \equiv \rho_Y(y) \equiv \sum_x \rho(x,y)$, respectively. Recall that the mutual information between X and Y under any ρ can be written both as:

$$\begin{aligned} I_\rho(X;Y) &\equiv -\sum_x \rho(x) \ln[\rho(x)] - \sum_y \rho(y) \ln[\rho(y)] + \sum_{x,y} \rho(x,y) \ln[\rho(x,y)] \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{2}$$

and equivalently as:

$$I_\rho(X;Y) \equiv \sum_{x,y} \rho(x,y) \ln \left[\frac{\rho(x,y)}{\rho(x)\rho(y)} \right] \tag{3}$$

Now, let $\vec{n}$ be a set of IID of samples of $\rho_{X,Y}$, and let $\vec{n}_X$ and $\vec{n}_Y$ be the associated set of sample values of ρ_X and ρ_Y, respectively. Then, from Equations (2) and (3), the posterior expectation of the mutual information can be written as either:

$$\begin{aligned} \mathbb{E}(I(X;Y) \mid \vec{n}) &= -\mathbb{E}\left(\sum_x \rho_X(x) \ln[\rho_X(x)] \mid \vec{n}_X \right) - \mathbb{E}\left(\sum_y \rho_Y(y) \ln[\rho_Y(y)] \mid \vec{n}_Y \right) \\ &\qquad + \mathbb{E}\left(\sum_{x,y} \rho(x,y) \ln[\rho(x,y)] \mid \vec{n} \right) \\ &= \mathbb{E}(H(X) \mid \vec{n}_X) + \mathbb{E}(H(Y) \mid \vec{n}_Y) - \mathbb{E}(H(X,Y) \mid \vec{n}) \end{aligned} \tag{4}$$

or as:

$$\mathbb{E}(I(X;Y) \mid \vec{n}) = \mathbb{E}\left(\sum_{x,y} \rho(x,y) \ln \left[\frac{\rho(x,y)}{\rho(x)\rho(y)} \right] \;\middle|\; \vec{n} \right) \tag{5}$$

(See, for example, [11,15].)

If $P(c)$ is set in a way that depends on the size of the event space, then we would use a different $P(c)$ to evaluate each of the three terms in Equation (4), since the underlying event spaces (X, Y and $X \times Y$, respectively) have different sizes. However, there would only be a single $P(c)$ used to evaluate the expression in Equation (5), the same $P(c)$ as used for evaluating the third term in Equation (4). As a result, under either the NSB or WW approaches, the values given by Equation (4) and Equation (5) will differ in general; depending on which definition of mutual information we adopt, we would get a different estimate of posterior expected mutual information under those approaches. Indeed, to estimate mutual information in the NSB approach, one faces the choice of whether to set $P(c)$ to give a uniform prior distribution over values of the mutual information, $P(I(X;Y))$ (as it would appear, one must, since $I(X;Y)$ is what one wishes to estimate), or to set it to give a uniform $P(H(X,Y))$ (as in conventional NSB). It is not clear how to make this choice, in general.

2. Contribution of This Paper

In most of the earlier work on estimating an information-theoretic functional of ρ based on data, it is assumed that $|Z|$ is fixed, with a few exceptions (see, e.g., [16]). In many situations, the modeler is not completely certain *a priori* about the value of $|Z|$, and so should treat it as a random variable. For such scenarios, we need to specify a joint (hyper)prior, $P(c, |Z|)$, rather than just a prior, $P(c)$. In particular, if we set c from $|Z|$ as in either WW or NSB, then by specifying our uncertainty in Z, $P(|Z|)$, we set the joint prior $P(c, |Z|)$. Note that for both WW and NSB, this induced $P(c, |Z|)$ is not a product distribution $P(c)P(|Z|)$. (In particular, in NSB, the distribution, $P(c)$, is set independently of any data, in a way that varies with $|Z|$.)

Jaynes has argued convincingly for setting priors with invariance arguments concerning the fundamental nature of the problem domain [17]. In this paper, we show that the prior, $P(c, |Z|)$, obeys a simple "Irrelevance of Unseen Variables" (IUV) invariance, if and only if c and $|Z|$ are independent, *i.e.*, iff $P(c, |Z|) = P(c)P(|Z|)$. Therefore, if we require IUV, then rather than specify a full joint distribution over c and $|Z|$, we only need to specify a distribution over each of c and $|Z|$ separately. This greatly reduces the number of degrees of freedom in the prior that we need to specify (though not as much as would be the case if we used WW or NSB, were we only to specify $P(|Z|)$).

In this paper, we show that when IUV is obeyed, so that c and $|Z|$ are independent, the value for posterior expected mutual information does not change depending on whether we use Equation (2) or Equation (3) to define mutual information. In proving this, we derive an intermediate result that simplifies the calculation of some posterior moments. In particular, we show how to use this result to derive the formula for posterior expected mutual information given in WW in essentially a single line. We also show that both of these advantages extend to the calculation of multi-information, one of the ways proposed to generalize mutual information beyond two random variables [18]. Similarly, since Tsallis entropy with index q is just a weighted sum over i of the qth moments of the p_i, we can evaluate expected Tsallis entropy in closed form using our estimators. (However, higher-order moments of the Tsallis entropy do not simplify as easily.)

We then show that when c and $|Z|$ are independent under the prior, and $|Z|$ is averaged over according to a prior $P(|Z|)$ with some reasonable characteristics, the posterior expected value of information-theoretic quantities need not be dominated by the prior. In this sense, the problem that caused NSB to consider a non-conventional scheme for setting $P(c)$ does not exist if we allow $|Z|$ to be a random variable and require IUV.

We next discuss in detail various fully Bayesian schemes in which the random variables, c and/or $|Z|$, are integrated over to form estimates of posterior expectations. We also mention some schemes in which one or the other of those variables is given a single value (unlike in proper hierarchical Bayes). We run a few computer experiments as cursory "sanity checks". In these, we choose a naive IUV-based hyperprior and compare the associated estimators of posterior expected entropy and of posterior mutual information to the estimators considered in [12,16,19]. We find that the IUV-based estimator performs quite favorably.

There are several subtleties in how one models the statistical generation of Z, issues that do not arise if Z is fixed ahead of time. One of them involves the mapping of each newly sampled draw of ρ to an element of $\vec{n}$, *i.e.*, to a label for that draw. To formally justify the "intuitively obvious" model of how Z is generated that we have used up to now, we describe in detail a mapping of draws of ρ to elements of $\vec{n}$ that justifies that model.

After this, we describe a change one might make to the model of how Z is generated that would appear to be innocuous. We show that, in fact, this change can substantially affect the resultant estimations. Concretely, say there is a space, $\hat{Z}$, that is a grid of photoreceptors and that ρ is a distribution over $\hat{Z}$ that is IID sampled to generate counts of photons that are reflected from an object and focused onto elements of $\hat{Z}$. Say we know that the object being imaged may be occluded, so that, in fact, only a subset, $Z \subseteq \hat{Z}$, of the grid points can have a nonzero probability of a photon count. However, say we are uncertain of the size of Z and, therefore, of which precise pixels in $\hat{Z}$ it corresponds to. We show that the value of the posterior expected entropy in this scenario is different from its value in the conventional scenario, in which we are also uncertain of Z's size, but there is no encompassing set $\hat{Z}$ from which Z is formed.

This touches on the more general issue of the epistemological foundation of the probabilities (and probabilities of probabilities) considered in this paper. This issue, involving concepts like "degree of belief" and "objective probability", is deep and quite important, being fundamental to the differences between Bayesian and sampling theory statistics (see [20,21] for a discussion).

Here, we do not grapple with this issue. Rather, we adopt the "pragmatic Bayesian" perspective implicit in all earlier Bayesian work on the problem of estimating information-theoretic quantities from samples (including WW and NSB, in particular) and simply use probabilities as a part of a self-consistent calculus of uncertainty. Bayesian reasoning often relies on a choice of prior, and the work presented here is no exception. We emphasize that that is no "one true prior"; rather, the statistician must match the choice of model to their own prior knowledge about the system. While we have done our best to chose a set of priors with generality sufficient to avoid some of the common biases identified in the past, one of the main goals of this paper is to present our results, so that the readers may adapt our methods to the particular nature of their own research.

Some of the experiments reported here were run using the publicly available package, Thoth, available at http://thoth-python.org. The reader is directed to Appendix B for associated proofs. Note that as

discussed in WW, much of the analysis below can be modified for inference of arbitrary functionals of ρ from IID samples of ρ (e.g., estimation of covariances from IID samples rather than mutual information). The analysis is not limited to inferring information-theoretic functionals from IID samples.

3. Preliminaries

For any finite space, U, we write Δ_U to mean the simplex of possible distributions over U. We will also write T_k to mean the set of all k-dimensional vectors whose components are all integers greater than zero.

Throughout this paper, we restricted attention to Dirichlet distributions over distributions $\rho \in \Delta_Z$. For a fixed c and Z, we write such a distribution as:

$$\mathscr{D}_{c,Z}(\rho) = \frac{\prod_z \rho(z)^{[c/|Z|]-1}}{\int d\rho' \prod_z \rho'(z)^{[c/|Z|]-1}} \tag{6}$$

where we require c to be non-negative.

Say we are given a dataset, $\vec{n}$, of counts for each of the elements of Z, where $N \equiv \sum_z n_z$. For the Dirichlet prior, the posterior distribution is:

$$\begin{aligned} P(\rho \mid \vec{n}, c, |Z|) &= \frac{\mathscr{D}_{c,Z}(\rho) \prod_z \rho_z^{n_z}}{\int d\rho' \, \mathscr{D}_{c,Z}(\rho') \prod_z (\rho'_z)^{n_z}} \\ &\equiv \frac{\prod_z \rho_z^{n_z - 1 + c/|Z|}}{G(\vec{n}, c, |Z|)} \end{aligned} \tag{7}$$

Using [6], we can calculate:

$$G(\vec{n}, c, |Z|) = \frac{\prod_z \Gamma(n_z + c/|Z|)}{\Gamma(N + c)} \tag{8}$$

We will sometimes write the posterior given by Equation (7) as $\mathscr{D}_{c,Z}(\rho \mid \vec{n})$.

Note that Z is implicitly a random variable in Equation (7), since we condition on $|Z|$ there. This means that the event space over which ρ is defined is also a random variable, as is the event space over which $\vec{n}$ is defined. This means that when we average over Z's below, we must take a bit of care to define the space of all ρ's and the space of all $\vec{n}$'s, since ρ can be an element of any finite unit simplex and similarly for $\vec{n}$.

When this issue arises, we will define the set of all triples of Z, $\vec{n}$ and ρ as the infinite union of the triples, $(Z_1, \Delta_{Z_1}, T_{|Z_1|})$, $(Z_2, \Delta_{Z_2}, T_{|Z_2|})$, $(Z_3, \Delta_{Z_3}, T_{|Z_3|})$, *etc*., where each Z_i is defined as the set, $\{1, \ldots, i\}$. In such a fully formal approach, joint probability distributions over $Z, \vec{n}$ and ρ are defined over that infinite union by:

$$P(Z, \rho, \vec{n}) = 0 \text{ unless } \rho \in \Delta_Z, \vec{n} \in T_{|Z|} \tag{9}$$

and then using the multinomial and Dirichlet distributions to define the values of the conditional distributions, $P(\vec{n} \mid \rho, Z)$ and $P(\rho \mid Z)$, respectively, when the condition in Equation (9) is obeyed. For the simplicity of the exposition, we will minimize our use of this fully formal approach here.

We define $\mathscr{S}(\vec{n})$ to be the support of the dataset, $\vec{n}$, within Z. We write $I(.)$ to be 1/0, depending on whether its logical expression argument is true/false. For the case where $Z = X \times Y$, we define

$\vec{n}_X(x) \equiv \sum_y \vec{n}(x,y)$ and $\rho_X \equiv \sum_y \rho(x,y)$. For use below, as in WW, we define $\Delta\Phi^{(1)}(z_1, z_2) \equiv \Psi^{(0)}(z_1) - \Psi^{(0)}(z_2)$, where $\Psi^{(0)}$ is defined as $d[\ln \Gamma(z)]/dz$.

4. Irrelevance of Unseen Variables

4.1. The Problem of Unseen Variables

In general, there may be an "unrecorded" or "hidden" variable, $y \in Y$, whose values are not recorded in our dataset of values, $x \in X$. As an example, say our data are the set of all changes in the value of the US stock-market between opening and closing on all days it was open since 1970. A hidden variable is the age of the person recording each of the measurements when they made the recording. The change in stock market value is X, and the age of the person recording the value is Y.

The hidden variable in this example is chosen to be extreme, in that knowledge of its existence is clearly irrelevant to the statistical estimation problem. However it is hard to imagine an estimation problem where there are *not* in fact a potentially infinite number of such hidden variables. Some could perhaps be dismissed, as "clearly irrelevant, and therefore, not worthy of consideration". However, this approach is hard to justify axiomatically, since there are, of course, many instances when hidden variables *may* have some relevance to the estimation problem.

This issue is a bit of a philosophical hornet's nest. If at all possible, we would like to avoid having to consider it. In fact, we can do this by requiring that the existence of any hidden variables has no effect on our Bayesian estimate. More precisely, one can require that whether one does or does not have a single unseen random variable in one's model, and the (finite) size of its event space, if it does exist, must have no impact on posterior expected values of (functionals of the distribution over) the seen variables [22]. How to do this is the subject of this section. In the following section, we illustrate the practical benefits that result.

To analyze scenarios involving hidden variables, write $Z = X \times Y$, and say we have recorded the dataset, $\vec{n}_X(x) \equiv \sum_y \vec{n}_{X,Y}(x,y)$, not the full dataset, $\vec{n}_{X,Y}$. Next, write ρ_Z as a matrix of real numbers, $\{\rho_{X,Y}(x,y) : x \in X, y \in Y\}$. We can re-express any such $\rho_{X,Y} \in \Delta_{X\times Y}$ in an alternative coordinate system, as $(\rho_X, \rho_{Y|X})$, where $\rho_X(x) \equiv \sum_y \rho_{X,Y}(x,y)$, and $\rho_{Y|X}$ is the set of $|X||Y|$ real numbers given by $\rho(x,y)/\rho_X(x)$ for all $x \in X, y \in Y$. (Note that under a Dirichlet prior, no matter what our data are, there is both zero prior probability and zero posterior probability of a ρ_X, such that $P(\rho_X)(x) = 0$ for some x.) Therefore, for all x, y, $\rho_{X,Y}(x,y) = \rho_{Y|X}(y \mid x)\rho_X(x)$.

If we allow for such hidden variables, Y, then to do a proper Bayesian analysis, we must specify a prior, $P(\rho_{X,Y})$ (or just $P(\rho)$, for short) on the space, $\Delta_{X\times Y}$, *i.e.*, on the space of ρ's that run over $X \times Y$. It does not suffice to specify just a prior, $P(\rho_X)$, defined over Δ_X. Moreover, axiomatic derivations of Bayesian analysis counsel us to set this prior *without concern for what our likelihood function will be*. (The prior is our model of the underlying physical system. The likelihood instead has to do with the observation apparatus we happen to have handy to observe that system.) This implies that $P(\rho)$ should not reflect the fact that X is observed and Y is not, since what variable is observed is determined by the likelihood. Therefore, in particular, if we set $P(\rho)$ based on the size of the underlying event space, it should be based solely on the size of the space, $X \times Y$, with no consideration for just the size of X.

Now, in general, we do not even know how many values of a hidden variable Y there are, simply that there may (!) be some. Due to this uncertainty, we must let the cardinality of the set of hidden variables "float" as a random variable. Moreover, very often we are interested in a functional $Q(\rho^X(x)) = Q(\sum_y \rho_{X,Y}(x,y))$ of the observed variable, and this functional can be anything, depending on the statistical question we are interested in.

One might worry that for some such functional, Q, a Dirichlet prior over the the joint observed and hidden variables, $P(\rho_{X,Y})$, and some (visible) data vector, $\vec{n}_X$, the associated value of $\mathbb{E}(Q \mid \vec{n}_X)$ would vary depending on the number of degrees of freedom of the hidden variable, $|Y|$. If this were the case, how we set the prior over the number of degrees of freedom of the hidden variable would matter, which would confront us with the problem of how to set it. This would seem to be an intractable problem, since, in general, there are an infinite number of choices for what the hidden variable is.

The desideratum analyzed in this paper is that this problem does not arise. Formally, this is equivalent to requiring that no matter what Y is, the associated value of $\mathbb{E}(Q \mid \vec{n}_X)$ is exactly what it would be if there were no hidden variable at all:

Definition 1 *A distribution $\pi(c, |Z|)$ obeys **Irrelevance of Unseen Variables** (IUV) iff, for all finite spaces, X and Y, data vectors, $\vec{n}_X$, and functions, Q, defined over ρ_X:*

$$\int dc\, \pi(c \mid |X|) \int d\rho_X\, Q(\rho_X) \mathscr{D}_{c,X}(\rho_X \mid \vec{n}_X) = \int dc\, \pi(c \mid |X| \times |Y|) \int d\rho_{X,Y}\, Q(\rho_X) \mathscr{D}_{c,X\times Y}(\rho_{X,Y} \mid \vec{n}_X)$$

To help understand this desideratum, note that uncertainty about the size of a hidden variable space, $|Y|$, is different from uncertainty about the size of the observed variable space, $|X|$. Indeed, one could argue that $|Y|$ is always essentially infinite, up to any kinds of limits imposed by quantum mechanics. (As an illustration for the stock-market example, in addition to the age of the recorder of the stock-market's change in value, all other characteristics of the recorder are "hidden" and, therefore, arguably should be included in Y.) Moreover, in general, as the number of (IID) data grows, we will get more certain about $|X|$, or at least about the number of x for which $\rho(x)$ exceeds some preset threshold. In contrast, the size of the dataset has no effect on our uncertainty concerning $|Y|$; the latter is purely prior-dominated. Both of these properties mean that statistically estimating $|Y|$ is a more fraught exercise than estimating $|X|$.

Our desideratum says that π is arranged so that these difficulties are irrelevant. For a π obeying IUV, uncertainty in the value of $|Y|$ has no effect on our estimate of a functional $Q(\rho_X)$. In contrast, uncertainty in $|X|$ has a major effect on all estimators of functionals $Q(\rho_X)$ that we know of (including the ones we introduce below), regardless of π. Indeed, how to estimate the size of X from the observed data is so important that it has been analyzed for decades in statistics, under the name "coverage estimators" [16,23].

The π's used in both NSB and WW violate IUV. This is at the heart of their problems in estimating mutual information, which were discussed in Section 1.

4.2. Dirichlet-Independent (DI) Hyperpriors

Let T be a partition of Z. Then, there is a map, K, taking any distribution, ρ, over Z to a distribution, ρ_T, over the elements, $t \in T$. Using K, any distribution over ρ's induces a distribution over ρ_T's. In particular, Dirichlet distributions have the very nice property that the distribution $\mathscr{D}_{c,Z}(\rho)$ over ρ's induces the distribution, $\mathscr{D}_{c,T}(\rho_T)$, over ρ_T's, where the baseline distribution for T is given by applying K to the baseline distribution over Z. The crucial point about this property for us is that there is the same concentration parameter, c, in both Dirichlet distributions; Dirichlet distributions are consistent under marginalization. (Indeed, this property serves as a common definition of Dirichlet processes, the extension of Dirichlet priors to infinite spaces.)

As a special case, if $Z = X \times Y$, then X specifies a partition of Z in which each partition element is of the form $\{(x, y) : y \in Y\}$ for a different $x \in X$. Therefore, a Dirichlet distribution generating ρ's over Z induces a Dirichlet distribution generating ρ_X's over X that has the same concentration parameter.

Now note that because we are using a Dirichlet prior, the posterior, $P(\rho \mid \vec{n}_X)$, is a Dirichlet distribution. In light of the marginalization consistency of Dirichlet distributions just described, this means that the induced posterior, $P(\rho_X \mid \vec{n}_X)$, will also be a Dirichlet distribution, *with the same value of* c. This suggests that the posterior expectation of any functional of ρ_X will be the same, whether we evaluate it in X or $X \times Y$, so long as c is the same for both evaluations.

We can formalize this with the following lemma, proven in Appendix B:

Lemma 1 *Fix any finite spaces, X and Y, any set, $\vec{n}_{X,Y}$, of counts of each of the elements in $X \times Y$ and any two concentration parameters, c and c'. Then, $c = c'$ iff:*

$$\int d\rho_X \; Q(\rho_X) \mathscr{D}_{c,X}(\rho_X \mid \vec{n}_X) \;\; = \;\; \int d\rho_{X,Y} \; Q(\rho_X) \mathscr{D}_{c',X\times Y}(\rho_{X,Y} \mid \vec{n}_{X,Y})$$

for all functions, Q, defined over Δ_X.

There are several noteworthy implications of Lemma 1. To see the first one, consider the following modification of the definition of IUV, which involves $\vec{n}_{X,Y}$, the count vector of both seen and unseen bins:

Definition 2 *A distribution, $\pi(c, |Z|)$, obeys* ***strengthened IUV*** *iff for all finite spaces, X and Y, data vectors, $\vec{n}_{X,Y}$, and functions, Q, defined over ρ_X:*

$$\int dc \; \pi(c \mid |X|) \int d\rho_X \; Q(\rho_X) \mathscr{D}_{c,X}(\rho_X \mid \vec{n}_X) \;\; = \\ \int dc \; \pi(c \mid |X| \times |Y|) \int d\rho_{X,Y} \; Q(\rho_X) \mathscr{D}_{c,X\times Y}(\rho_{X,Y} \mid \vec{n}_{X,Y}).$$

An immediate corollary of Lemma 1 is the following:

Corollary 2 *IUV implies strengthened IUV.*

Proof: The integral on the RHSin the equation in Lemma 1 is the inner integral on the RHS of Definition 2. In addition, the integral on the LHSin the equation in Lemma 1 is the inner integral on the RHS of Definition 10. Therefore, applying Lemma 1 for $c = c'$ establishes the corollary. ■

Corollary 2 means that if IUV holds, then it does not matter how the counts, $\vec{n}_{X,Y}(x, y)$, are apportioned over Y, as far as calculating the associated posterior expected value of Q is concerned.

Assume there are no unobserved components of our data. Then, by Corollary 2, for any functional, Q, that depends purely on ρ_X, if IUV holds, we can evaluate expected $Q(\rho_X)$ conditioned on $\vec{n}_{X,Y}$ (given by an integral over $\Delta_{X\times Y}$) by calculating expected $Q(\rho_X)$ conditioned on $\vec{n}_X$ (given by an integral over Δ_X). In Section 5 below, we show that this property of IUV substantially simplifies calculations of expected moments of functionals defined over multi-dimensional spaces, e.g., the calculation of posterior expected mutual information.

To establish a second implication of Lemma 1, multiply both sides of the equality it establishes by $\sum_{\vec{n}_{X,Y}} P(\vec{n}_{X,Y} \mid \vec{n}_X)$. That leaves the LHS of that equality unchanged. However, it changes the RHS to $\int d\rho_{X,Y}\; Q(\rho_X)\mathscr{D}_{c',X\times Y}(\rho_{X,Y} \mid \vec{n}_X)$. This provides the following corollary:

Corollary 3 *Fix any finite spaces, X and Y, concentration parameter c, function Q defined over Δ_X and set $\vec{n}_{X,Y}$ of counts of each of the elements in $X \times Y$. Then:*

$$\int d\rho_X\; Q(\rho_X)\mathscr{D}_{c,X}(\rho_X \mid \vec{n}_X) \;\; = \;\; \int d\rho_{X,Y}\; Q(\rho_X)\mathscr{D}_{c,X\times Y}(\rho_{X,Y} \mid \vec{n}_X)$$

The integral on the RHS of Corollary 3 is the inner integral on the RHS of the definition of IUV, Definition 10. This establishes that if the conditional prior, $\pi(c \mid |X|)$, equals the conditional prior, $\pi(c \mid |X||Y|)$, then IUV holds.

We will use the term **Dirichlet-independent hyperprior** (DI hyperprior) to refer to any prior of the form $\pi(c, |Z|) = \pi(c)\pi(|Z|)$ over the hyperparameters of a Dirichlet distribution with uniform baseline distribution. We now present the main result of this section, which is that if we restrict attention to hyper priors meeting a particular technical condition [24], then IUV and a DI hyperprior are equivalent, as proven in Appendix B:

Proposition 4 *Assume that for any two finite spaces, $X and Y$, and associated count vector $\vec{n}_X$, there exists an $\epsilon > 0$, such that:*

$$((1+\epsilon)|X|)^c \frac{\pi(c \mid |X|) - \pi(c \mid |X||Y|)}{C(c, \vec{n}_X)}$$

is infinitely differentiable with respect to c at $c = 1$ and that its Fourier transform is analytic. Then, π is a DI hyperprior $\Leftrightarrow$ IUV holds.

We can combine these results to see that under the conditions in Proposition 4, we have a DI hyperprior iff strengthened IUV holds.

In some situations, the number of possible values of the "hidden variable" will vary with $x \in X$. This is a generalization of the currently considered scenario, where rather than $Z = \cup_{x\in X} Y_x$, where

$|Y_x|$ is the same for all x, we allow the sizes, $|Y_x|$, to vary with x. One can modify the definitions of IUV and strengthened IUV given above to address this situation, and all the results presented above (appropriately modified) still hold. This is due to the fact that Dirichlet distributions are "consistent under marginalization", as described at the beginning of this section.

5. Calculational Benefits

In addition to its "theoretical" advantage of being equivalent to a natural desideratum (IUV), DI hyperpriors also have practical advantages. In particular, recall that WW derived a complicated expression for posterior expected mutual information based on Equation (5). Part of what made that expression complicated was that it used a posterior over $\Delta_{X,Y}$ to evaluate $H(X)$ and $H(Y)$, as well as $H(X,Y)$.

However, when we have a DI hyperprior, the posterior expectation of $H(X)$ conditioned on $\vec{n}_X$ is the same, whether we evaluate it using a posterior over Δ_X or evaluate it using a posterior over $\Delta_{X,Y}$. This means we can evaluate that posterior expected mutual information by summing the posterior expected $H(X)$ under a posterior over Δ_X, posterior expected $H(Y)$ under a posterior over ρ_Y and posterior expected $H(X,Y)$ under a posterior over $\Delta_{X,Y}$. In turn, each of those three expectation values are given by a relatively simple formula from WW [Equation (29) in Appendix A].

This property concerns a scenario where we only have data $\vec{n}_X$. However, often when we want to estimate mutual information, we will have the full dataset with no unseen components, $\vec{n}_{X,Y}$. For that case, we can use Lemma 1 to justify decomposing the posterior expected mutual information into a sum of three posterior expected entropies and then evaluate those posterior expected entropies separately using Equation (29) in Appendix A. This directly gives us the following result, which does not rely on IUV:

Corollary 5 *Let X and Y be two spaces and $\vec{n}$, a sample generated by IID generating a distribution, ρ, across $X \times Y$, where ρ was generated by sampling a Dirichlet prior with concentration parameter c. Then:*

$$\begin{aligned}\mathbb{E}(I(X;Y) \mid \vec{n}, c) \;=\; & \sum_{x,y} \Delta\Phi^{(1)}(\vec{n}(x,y) + 1 + c/|X||Y|, N + |X||Y| + 1)\frac{\vec{n}(x,y) + c/|X||Y|}{N + c} \\ & - \sum_{x} \Delta\Phi^{(1)}(n_X(x) + 1 + c/|X|, N + |X| + 1)\frac{n_X(x) + c/|X|}{N + c} \\ & - \sum_{y} \Delta\Phi^{(1)}(n_Y(y) + 1 + c/|Y|, N + |Y| + 1)\frac{n_Y(y) + c/|Y|}{N + c}\end{aligned}$$

By Corollary 2, the analogous equality holds when we marginalize over c rather than condition on it, so long as we assume IUV.

Similarly to how the posterior first moment of mutual information can be defined using either Equation (2) or Equation (3), so can the posterior second moment:

$$\begin{aligned} E(I^2 \mid \vec{n}) &= \mathbb{E}(H(X)^2 \mid \vec{n}_X) + \mathbb{E}(H_Y^2 \mid \vec{n}_Y) + \mathbb{E}(H_{X,Y}^2 \mid \vec{n}) \\ &\quad - 2\mathbb{E}(H(X)H(Y) \mid \vec{n}) - 2\mathbb{E}(H(X)H(X,Y) \mid \vec{n}) - 2\mathbb{E}(H(Y)H(X,Y) \mid \vec{n}) \end{aligned} \tag{10}$$

or, alternatively:

$$\mathbb{E}(I^2 \mid \vec{n}) = \mathbb{E}\left(\left(\sum_{x,y} \rho(x,y) \ln \left[\frac{\rho(x,y)}{\rho(x)\rho(y)} \right] \right)^2 \,\middle|\, \vec{n} \right) \tag{11}$$

where the RHS of Equation (11) is evaluated using a Dirichlet prior over $\Delta_{X,Y}$, while the first two terms on the RHS in Equation (10) are instead evaluated using a Dirichlet prior over Δ_X and over Δ_Y, respectively.

Using the DI hyperprior, one gets the same answer whichever one of these expansions one uses. In addition, the first three terms in Equation (10) can be simplified under the DI hyperprior and evaluated using the formula in WW for posterior expected entropy. Unfortunately though, the remaining terms, e.g., $\mathbb{E}(H(X)H(X,Y)) \mid \vec{n})$, cannot be simplified the same way; evaluating them seems to require the kinds of techniques used in WW to evaluate the posterior variance of mutual information. This also applies to the use of our estimators for the computation of Tsallis entropy, since they also require use of higher-order moments..

There have been many ways proposed to generalize the idea of mutual information beyond two random variables. One of the most prominent is the **multi-information** of a set of random variable (see, e.g., Reference [18]). Just like mutual information among a pair of random variables can be defined either as a sum of the entropies of subsets of those random variables or as a function over the full event space, the same is true of multi-information. The definition of multi-information in terms of the entropies of subsets of the random variables is:

$$I(X_1, X_2, \ldots) \equiv \sum_i H(X_i) - H(X_1, X_2, \ldots) \tag{12}$$

Just as requiring IUV means that we do not have to worry about which of the ways to express the mutual information of two random variables (when calculating the posterior expectated mutual information based on data concerning only one of those random variables), the same is true of multi-information; requiring IUV means that we do not have to worry about how to express multi-information among a set of random variables when calculating its posterior expectation based on data concerning only a subset of those random variables. Moreover, just as IUV greatly simplifies the calculation of posterior expected information when our data concerns all the random variables, allowing us to just repeatedly apply Equation (29) in Appendix A, the same is true with multi-information.

6. Uncertainty in the Concentration Parameter and Event Space Size

Even when one adopts a DI hyperprior, so that prior c and prior $|Z|$ are statistically independent, there is still the issue of how to set $P(c)$ and $P(|Z|)$. In this section, we discuss some aspects of this issue.

6.1. Uncertainty in c

There are several natural choices of $P(c)$. For example, if we view c as a scale parameter, a logarithmic prior ($P(c) \propto 1/c$ up to a very large cut-off in c) would be reasonable.

Another approach is to set c to a single value that is "optimal" in some sense. Grassberger argued for setting $c = 1$ based on minimizing a rough approximation to the statistical bias. In addition, as subsequently pointed out by NSB, for a fixed Z, the choice of c equal to unity gives near-maximal prior variance of the entropy, *i.e.*:

$$\mathbb{E}([H - \mathbb{E}(H \mid c, Z)]^2 \mid c, Z) \;=\; \frac{c/|Z| + 1}{c + 1}\psi_1(c/|Z| + 1) - \psi_1(c + 1) \tag{13}$$

is near its maximum at $c = 1$. Confirming this nice property of $c = 1$, numerical calculation finds that, as $|Z|$ goes to infinity, the value of c maximizing that prior variance is $c_{\max} \approx 0.9222$. For small numbers of bins, $c_{\max}$ is smaller (e.g., for $|Z|$ equal to 5, $c_{\max} \approx 0.6997$).

Of course, one could also set c via a scheme other than hierarchical Bayes. In particular, setting it via maximum likelihood (*i.e.*, ML-II [25]) should often be reasonable.

6.2. Likelihood of the Event Space Size

Recalling the definition of G from Section 3, we can write:

$$P(\vec{n} \mid c, |Z|) \;=\; \frac{G(\vec{n}, c, |Z|)}{\sum_{\vec{n}} G(\vec{n}, c, |Z|)} \tag{14}$$

Using the results of WW to evaluate this, we get:

$$P(\vec{n} \mid c, |Z|) \;=\; \frac{\Gamma(c)}{\Gamma(c/|Z|)^{|\mathscr{S}(\vec{n})|}} \frac{\prod_{i=1}^{|\mathscr{S}(\vec{n})|} \Gamma(n_i + c/|Z|)}{\Gamma(N + c)} \tag{15}$$

Note that $\Gamma(x)$ diverges as $1/x$ as $x \to 0$. Thus, the factor of $\Gamma(c/|Z|)^{|\mathscr{S}(\vec{n})|}$ in the denominator means that Equation (15) is strongly weighted towards $|Z|$ small (but strictly greater than $|\mathscr{S}(\vec{n})|$, the number of observed bins.)

We have fixed the constant, c, in Equation (15). We could integrate over it instead, getting:

$$P(\vec{n} \mid |Z|) = \int \frac{\Gamma(c)}{\Gamma(c/|Z|)^M} \frac{\prod_{i=1}^{M} \Gamma(n_i + c/|Z|)}{\Gamma(N + c)} P(c)\, dc \tag{16}$$

where $P(c)$ must be independent of $|Z|$ in order to preserve IUV. Choosing different values of c shifts the posterior distribution, as can be seen in Figure 1.

Figure 1. Likelihoods for a dataset of one thousand samples drawn from a distribution ρ that was, in turn, drawn from a Dirichlet prior. The concentration parameter was $c = 1$, and $|Z|$ was 100. (The actual dataset was $\{691, 232, 24, 17, 14, 10, 6, 6\}$, with the remaining 92 entries equaling zero.) **Thick solid line**: $P(\vec{n} \mid |Z|)$ with a logarithmic prior for c (Equation (16)). Mean value: 8.2. **Thin solid line**: $P(\vec{n} \mid |Z|, c)$ with $c = 1$ (mean value: 8.4). **Dashed line**: $c = 0.01$ (mean value: 8.8). **Dotted line**: $c = 100$ (mean value 8.0.) Note that the maximum likelihood (ML) value is always $|\mathscr{S}(\vec{n})|$, the number of observed bins.

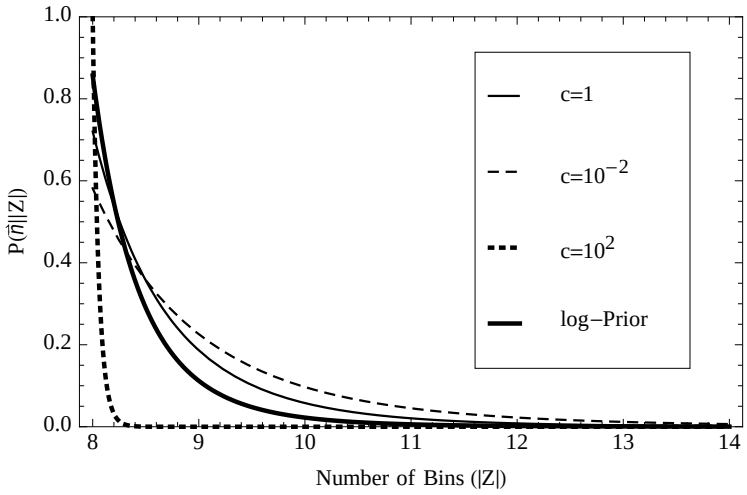
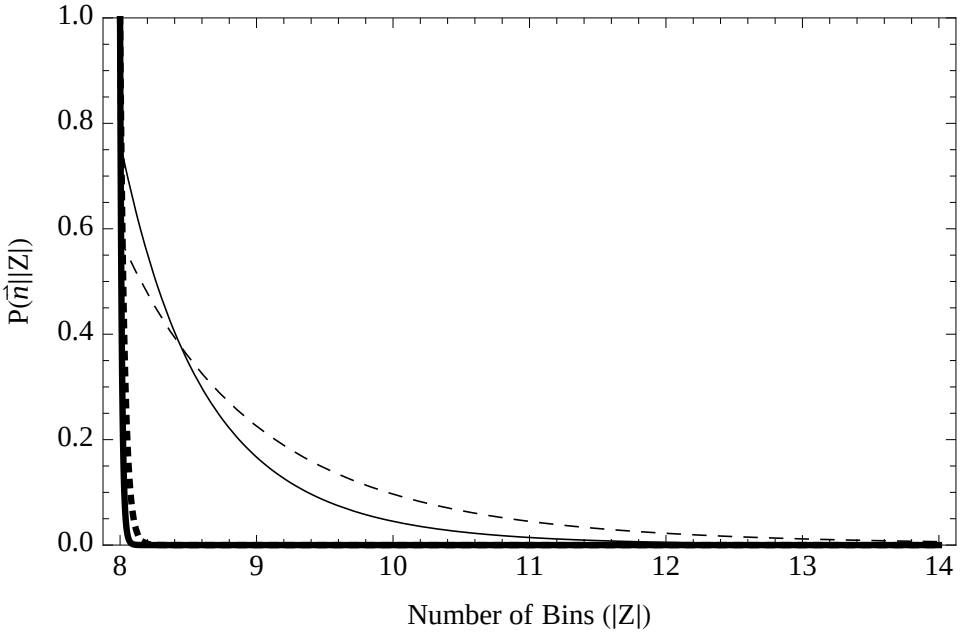


Figure 1 plots the likelihoods, $P(\vec{n} \mid |Z|)$ and $P(\vec{n} \mid c, |Z|)$, for various values of c. Note that $\mathscr{S}(\vec{n})$ is far smaller than $|Z|$. This reflects the fact that for $c \ll |Z|$, ρ is likely to be highly peaked about only a few bins and close to zero for all others. Note how strongly this likelihood prefers small values of $|Z|$. Even when the likelihood is evaluated with the value of c that was used to generate $\vec{n}$, it still strongly prefers $|Z|$ values that are far smaller than the one that was used to generate $\vec{n}$ [26].

In Figure 2, we again evaluate $P(\vec{n} \mid |Z|)$ and $P(\vec{n} \mid c, |Z|)$, where $\mathscr{S}(\vec{n})$ again equals eight, but now, $\vec{n}$ is uniform with the value 125 over the eight occupied bins. This dataset again implies with a high probability that ρ was highly peaked about the eight occupied bins and close to zero elsewhere. Therefore, again, the likelihood has a strong preference towards small values of $|Z|$.

Figure 2. Likelihoods for a dataset consisting of eight bins with 125 counts each, with the remaining 92 entries being zero. **Thick solid line**: $P(\vec{n} \mid |Z|)$ with logarithmic prior for c (Equation (16)). Mean value: 8.0. **Thin solid line**: $P(\vec{n} \mid |Z|, c)$ with $c = 1$ (mean value: 8.3). **Dashed line**: $c = 0.01$ (mean value: 8.8). **Dotted line**: $c = 100$ (mean value 8.0). Note that the ML value is always $|\mathscr{S}(\vec{n})|$, the number of observed bins.



To understand these results, recall "Bayes factors", which arise in Bayesian inference of the dimension of an underlying stochastic model based on samples of that model. These factors cause a strong *a priori* preference of the model dimension's likelihood towards small values. The preference of $P(\vec{n} \mid |Z|)$ for small $|Z|$ is a similar phenomenon, with the size of the space Z playing an analogous role here to the model dimension in Bayes factors. In both cases, it is the greatly increased likelihood of the data for a distribution from the smaller model that causes that model to have greater likelihood.

By comparing Figures 1 and 2, we see that for a logarithmic prior (one that is agnostic about c), changing the data without changing $|Z|$ or $\mathscr{S}(\vec{n})$ can have a marked effect on the likelihood, even when $|Z| \gg |\mathscr{S}(\vec{n})|$ [27]. Evidently, then, so long as we integrate over c's for each $|Z|$ (as in NSB) rather than fix a single value (as in WW), the dependence of the likelihood of $|Z|$ on the data is not overwhelmed by the precise choice of a prior. It does not seem that the particular $P(c \mid |Z|)$ adopted in NSB is necessary to have this data-sensitivity, at least as far as the likelihood of $|Z|$ is concerned and at least for the regime tested here.

6.3. Uncertainty in the Event Space Size

One of the core distinguishing features of the approach analyzed in this paper is to treat $|Z|$ as a random variable. In particular, the experimental comparisons with NSB discussed in Section 7.2 require specification of a prior, $P(|Z|)$.

There are several different ways of motivating such a prior over $|Z|$. Perhaps the simplest is to have it be uniform up to some cutoff, far larger than $\mathscr{S}(\vec{n})$. A somewhat more sophisticated approach would

be to assume that $P(|Z|)$ is set by a stochastic process that starts with $|Z| = 1$ and then iteratively adds a new "bin" to Z, the set of already selected bins, stopping after each iteration with a probability, $1 - \gamma$, ending at some upper cutoff, $\underline{m}$, if we get to $|Z| = \underline{m}$. Then, we can write:

$$P(|Z| \mid \underline{m}) \quad \propto \quad \gamma^{|Z|} \tag{17}$$

for all $|Z| \leq \underline{m}$, and $P(|Z| \mid \underline{m}) = 0$, otherwise. Alternatively, to avoid the need to specify an upper cutoff, one could use $P(|Z|) \propto \exp(-\alpha|Z|)$ for some hyperparameter, α.

Other models are also possible. For example, we might imagine that there is some upper bound value, $\underline{m}$, that is set *a priori* and $|Z|$ initially set to $\underline{m}$. Then, elements of Z are iteratively removed, at random, stopping after each iteration with a probability, $1 - \gamma$, ending at $|Z| = 1$, if we manage to get to $|Z| = 1$. Whereas the first $P(|Z|)$ is a decreasing function from $|Z| = 1$ up to $|Z| = \underline{m}$; this alternative $P(|Z|)$ is an increasing function over that range (see Section 8.2 for a discussion of these kinds of scenarios where Z is determined by randomly forming a subset of some original larger set).

A natural extension to both of these models is to allow $\underline{m}$ to vary and use an associated (hyper)prior. As always, the choice of random variables and associated (hyper)priors should match one's understanding of the underlying physical process by which the data is collected as accurately as possible.

Once we have a hyperprior, $P(|Z|)$, we can combine it with the likelihoods plotted in Figure 1 to get a posterior distribution over $|Z|$. For a $P(|Z|)$ that is nowhere-increasing, since the likelihoods are decreasing functions of $|Z|$, the posterior is also a decreasing function of $|Z|$. This would lead to an MAPestimate of $|Z| = |\mathscr{S}(\vec{n})|$, though, in general, $\mathbb{E}(|Z| \mid \vec{n})$ would be bigger than $|\mathscr{S}(\vec{n})|$. (In the scenario considered below in Section 8.2, both the MAP and posterior expected values of $|Z|$ can exceed $|\mathscr{S}(\vec{n})|$.)

6.4. Specifying a Single Event Space Size

Another way to address uncertainty in $|Z|$ is to set it to a single "optimal' value, rather than averaging over it. For example, we could use a "coverage estimator" [16,23] to set a single value of $|Z|$, m and, then, evaluate $\mathbb{E}(Q \mid \vec{n}, c, |Z| = m)$ using the formulas in WW [28]. Alternatively, if one has a distribution over c, then we would integrate over it while keeping $|Z|$ fixed. If we assume IUV, so $P(c \mid |Z|) = P(c)$, this would give:

$$\begin{aligned}
\mathbb{E}(Q \mid \vec{n}, |Z| = m) &= \int d\rho \; Q(\rho) P(\rho \mid \vec{n}, |Z| = m) \\
&= \int d\rho \; Q(\rho) \int dc \; P(\rho, c \mid \vec{n}, |Z| = m) \\
&= \int d\rho \; Q(\rho) \frac{\int dc \; P(\vec{n} \mid \rho, c, |Z| = m) P(\rho \mid c, |Z| = m) P(c)}{\int dc d\rho \; P(\vec{n} \mid \rho, c, |Z| = m) P(\rho \mid c, |Z| = m) P(c)} \\
&= \frac{\int dc d\rho \; P(c) Q(\rho) \left[\prod_z \rho_z^{n_z - 1 + c/m} / G(0, c, m) \right]}{\int dc d\rho \; P(c) \left[\prod_z \rho_z^{n_z - 1 + c/m} / G(0, c, m) \right]} \\
&= \frac{\int dc d\rho \; P(c) \Gamma(N + c) Q(\rho) \left[\prod_z \rho_z^{n_z - 1 + c/m)} / [\Gamma(c/m)]^m \right]}{\int dc d\rho \; P(c) \Gamma(N + c) \left[\prod_z \rho_z^{n_z - 1 + c/m} / [\Gamma(c/m)]^m \right]}
\end{aligned} \tag{18}$$

where we have used the results in Section 3 to derive the last line. Of course, we could also replace $P(c)$ in this formula with a distribution $P(c \mid |Z| = m)$ if we wish to violate IUV, e.g., as in the NSB estimator.

7. Posterior Expected Entropy When $|Z|$ is a Random Variable

7.1. Fixed c

How can we estimate the entropy of a system when the number of bins is unknown while c is fixed? Under IUV:

$$\begin{aligned} P(H = h \mid \vec{n}, c) &\propto \sum_{|Z|=1}^{\infty} \int d\rho \; P(\vec{n} \mid \rho, c, |Z|) P(\rho \mid Z, c) P(|Z| \mid c) \; \delta(H[\rho] - h) \\ &= \sum_{|Z|=1}^{\infty} \int d\rho \; P(\vec{n} \mid \rho, |Z|) \; \mathscr{D}_{c,Z}(\rho) P(|Z|) \; \delta(H[\rho] - h) P(|Z|) \\ &\equiv \sum_{|Z|=1}^{\infty} \int d\rho \; P(\vec{n} \mid \rho) \; \mathscr{D}_{c,Z}(\rho) P(|Z|) \; \delta(H[\rho] - h) \end{aligned} \tag{19}$$

where $P(|Z|)$ could be given by one of the priors discussed in Section 6.3.

Instead of trying to solve for this, as in WW, we can consider the posterior expected values of the moments of the entropy. Using IUV, the first moment is:

$$\begin{aligned} \mathbb{E}(H \mid \vec{n}, c) &= \sum_{|Z|=1}^{\infty} \int d\rho \; H(\rho) P(\rho, |Z| \mid \vec{n}, c) \\ &= \sum_{|Z|=1}^{\infty} \int d\rho \; H(\rho) \frac{P(\vec{n} \mid \rho) P(\rho \mid |Z|, c) P(|Z|)}{\sum_{|Z'|=1}^{\infty} \int d\rho' \; P(\vec{n} \mid \rho') P(\rho' \mid |Z'|, c) P(|Z'|)} \\ &= \frac{\sum_{|Z|=M}^{\infty} P(|Z|) \int d\rho \; H(\rho) \; P(\vec{n} \mid \rho) \; \mathscr{D}_{c,Z}(\rho)}{\sum_{|Z|=M}^{\infty} P(|Z|) \int d\rho \; P(\vec{n} \mid \rho) \; \mathscr{D}_{c,Z}(\rho)} \end{aligned} \tag{20}$$

The two integrals are straight-forward. We already know the denominator; the numerator is not that much harder. To simplify the expression of the result, define $M \equiv |\mathscr{S}(\vec{n})|$ and write $\{n_i : i \in \mathscr{S}(\vec{n})\}$ for the set, $\{n_Z(z) : z \in \mathscr{S}(\vec{n})\}$. Then, we get:

$$\begin{aligned} \mathbb{E}(H \mid \vec{n}, c) &= \frac{1}{\sum_{|Z|=M}^{\infty} P(|Z|) \frac{\prod_{i=1}^{M} \Gamma(n_i + c/|Z|)}{\Gamma(c/|Z|)^M}} \times \\ &\left\{ \sum_{|Z|=M}^{\infty} P(|Z|) \frac{\prod_{i=1}^{M} \Gamma(n_i + c/|Z|)}{\Gamma(c/|Z|)^M} \left(\left[\sum_{i=1}^{M} \frac{n_i + c/|Z|}{N + c} \Delta\Phi^{(1)}(n_i + c/|Z| + 1, c + N + 1) \right] \right. \right. \\ &\left. \left. + \; (|Z| - M) \frac{c/|Z|}{N + c} \Delta\Phi^{(1)}(c/|Z| + 1, c + N + 1) \right) \right\} \end{aligned} \tag{21}$$

where the term on the last line arises from empty bins and where $\Phi^{(n)}$ and $\Delta\Phi^{(n)}$ are defined in Section 3.

Recall from Section 6.2 that the likelihood, $P(\vec{n} \mid c, |Z|)$, is strongly weighted towards small $|Z|$. Therefore, if $P(|Z|)$ has an upper cutoff, $\mathscr{M}$, and is nowhere-increasing, then under a DI hyperprior,

the posterior, $P(|Z| \mid \vec{n}, c)$, must also be strongly weighted towards small $|Z|$. This will cause posterior moments of ρ to be dominated by values, $|Z|$, that are not much larger than M. In general, this will mean that these posterior moments will not be prior-dominated, in the sense that attributes of $\vec{n}$ will affect them significantly.

As an illustration, note that, in general, the innermost sum in Equation (21):

$$\left[\sum_{i=1}^{M} \frac{n_i + c/|Z|}{N+c} \Delta\Phi^{(1)}(n_i + c/|Z| + 1, c + N + 1)\right] \; + \; (|Z| - M)\frac{c/|Z|}{N+c}\Delta\Phi^{(1)}(c/|Z| + 1, c + N + 1)$$

goes to an $\vec{n}$-dependent constant as $|Z|$ becomes large. (The arguments of both $\Delta\Phi^{(1)}$'s become independent of $|Z|$, and the factors multiplying each of them go to a constant.) Furthermore, as discussed in Section 6.2, $\Gamma(c/|Z|)^{-M}$ goes to zero as $|Z|$ gets large. Accordingly, for our assumed form of $P(|Z|)$, once $\mathscr{M}$ is appreciably larger than M, posterior expected entropy does not change if $\mathscr{M}$ instead becomes hugely larger than M. In this sense, the posterior expected entropy does not become prior dominated as $\mathscr{M}$ becomes very large. (At a minimum, M—an attribute of the data, not the prior—is determining the range of relevant $|Z|$.) Therefore, to the degree that prior-dominance is avoided in Equation (21), treating $|Z|$ as a random variable and requiring IUV removes the phenomenon that caused NSB to adopt their scheme for setting $P(c)$.

A formula similar to Equation (21) gives the second moment of the posterior distribution over the entropy. (It is too long to write out here; we recommend that a package like Mathematica be used to evaluate it.) Combining that formula with Equation (21) provides the posterior variance of the entropy when the number of bins is a random variable.

7.2. Uncertain c

To allow for varying c, one can change the sum over $|Z|$ in Equation (21) for a fixed c to a new expression involving sums over $|Z|$ together with integrals over c with some prior for c. Care must be taken when doing this, since c is not independent of $\vec{n}$ (see Section 6.4 for another example of integrating over c when conditioning on $\vec{n}$). Using IUV, the result is:

$$\begin{aligned} \mathbb{E}(H \mid \vec{n}) &= \sum_{|Z|=1}^{\infty} \int d\rho dc \; H(\rho) P(\rho, |Z|, c \mid \vec{n}) \\ &= \frac{\sum_{|Z|=M}^{\infty} P(|Z|) \int d\rho dc \; H(\rho) \; P(\vec{n} \mid \rho) \; \mathscr{D}_{c,Z}(\rho) P(c)}{\sum_{|Z|=M}^{\infty} P(|Z|) P(c) \int d\rho dc \; P(\vec{n} \mid \rho) \; \mathscr{D}_{c,Z}(\rho) P(c)} \end{aligned} \tag{22}$$

To evaluate this, we need only separately apply $\int dc P(c)$ to both the numerator and denominator in Equation (21).

7.3. Experimental Tests

The primary focus of this paper is an analysis of the IUV desideratum and its implications for the hyperprior. However, as a sanity check, in this subsection, we compare the performance of posterior estimators of entropy and of mutual information that are based on a DI hyperprior to three estimators

of those quantities that were previously considered in the literature. These three alternative estimators are NSB, the estimator considered in [19] (which is an asymptotic version of NSB that allows for the estimation of entropy when the number of bins is unknown), and the "Coverage-Adjusted Estimator" (CAE) of [16]. To simplify the exposition, we will sometimes refer to any estimator based on a DI hyperprior as a "W&D" estimator.

¿From a decision-theoretic perspective, no estimator can do better under an assumed hyperprior than one that is Bayes-optimal for that hyperprior, *if* one quantifies performance with experiments that draw samples from that same hyperprior. Failures of a Bayes-optimal estimator always arise from a mismatch between the hyperprior used to construct the estimator and the hyperprior used to actually generate the data. Accordingly, to make meaningful comparisons between a W&D estimator and others, we must see how well Equation (22) performs "out of class", for data that is not generated from the DI hyperprior used to construct the W&D estimator.

To make these comparisons, we use a W&D estimator for a hyperprior that has logarithmic $P(c)$ and uniform $P(|Z|)$ and then consider two general types of data that are *not* drawn under this hyperprior. In particular, we consider: (1) distributions sampled from Dirichlet distributions with fixed c; and (2) power-law distributions of the form:

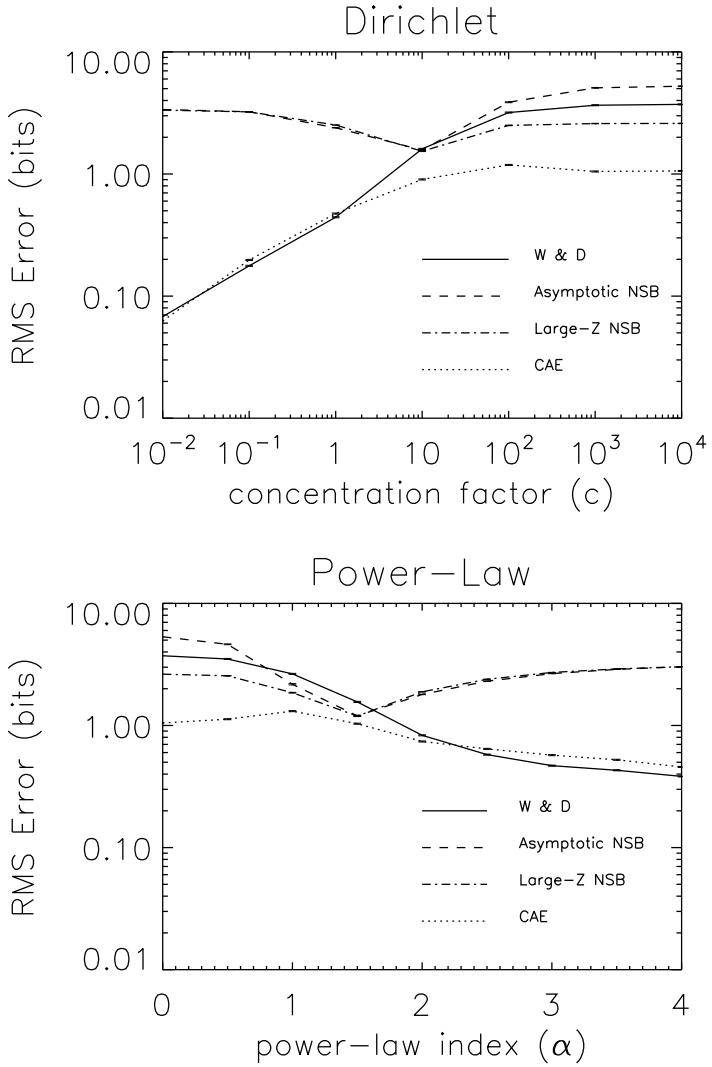
$$P(i) \propto \frac{1}{S[i]^{\alpha}}, \ [i = 1 \ldots m] \tag{23}$$

where $S[i]$ is a one-to-one map on the integers from one to $m = |Z|$. (Varying such $S[.]$ ensures that the order of the terms is not fixed, but can vary depending on the particular choice of $S[.]$; this is important when constructing joint probability distributions by re-interpreting a probability distribution over m categories as a joint distribution over $\sqrt{m} \times \sqrt{m}$ categories, as in estimates of mutual information between a pair of random variables.)

We estimated the entropy and mutual information based on datasets generated this way using Equation (22), the Coverage-Adjusted Estimator of [16] (Equation 18, with $n + 1$ correction to make it well-defined in the singleton case), the Asymptotic NSB estimator of [19] and a "large-Z" version of the standard NSB estimator of [12], where the bin size is set to a large value assumed to be larger than the possible number of bins. For both W&D and the large-Z NSB, we must include a maximum bin number. We take this to be $10,000$ (*i.e.*, a hundred times larger than the actual number), and in the case of mutual information estimation, we allow the large-Z NSB estimator to assume that the maximum bin number for each marginal is 100 (*i.e.*, ten times larger than the actual one), with the maximum bin number for the full distribution, as before, set to $10,000$.

We then computed the RMS error between these estimates and the truth for all three estimators. The results are shown in Figure 4, where we sample 100-bin processes in the "deeply undersampled regime", where N, the number of counts, is $\sqrt{m}$.

Figure 3. RMS error for the estimation of entropy with an unknown bin number. In both plots, the solid line shows the W&D estimator, Equation (22); the dashed line shows the asymptotic version of NSB (Equation (29) of [19]); the dotted line shows the Coverage-Adjusted Estimator of [16]; the dot-dashed line shows the "large-Z" version of NSB. The true bin number is 100, and the associated distribution, ρ, was sampled ten times (*i.e.*, it was radically under-sampled given the number of bins). (**Top**) Results when the distribution, ρ, used to generate the data is randomly generated under a Dirichlet distribution, whose concentration, c, is varied from 10^{-2} (highly non-uniform; many hits in a few bins) to 10^{4} (highly uniform), as indicated. (**Bottom**) Results when ρ is a a power-law distribution over the bin number with index α that is varied from zero (perfectly uniform) to four (highly non-uniform), as in Equation (23). (The Zipf distribution corresponds to $\alpha = 1$.)



The estimator of Equation (22) outperforms the asymptotic NSB estimator for a wide range of distributions, both Dirichlet and power-law. This is not entirely unexpected; the asymptotic estimator works only to zeroth order in $1/N$ and $1/m$. In the regimes where the true ρ is likely to have low-entropy, the large-Z estimator performs almost identically to the asymptotic NSB estimator, while it performs

somewhat better in the high-entropy regime, where it is competitive with Equation (22). For low-entropy samples (either low c or high α), Equation (22) is competitive with the Coverage-Adjusted Estimator. Inefficiencies in Equation (22) trace back to our prior over c; in cases where one has a strong belief that the data are drawn from high-entropy distributions, different c weights should be used.

Interestingly, at the α equal to the unity point (the Zipf distribution), all four methods are within a factor of two of each other RMS. The strongest differences between the methods emerge at low entropies.

We can use the same methods to compare the accuracies of the estimators of the mutual information. In particular, since Equation (22) respects IUV, we can decompose the mutual information into the sum and differences of entropies. In Figure 4, we plot RMS error for mutual information estimated in this fashion and compare it to the naive use of the asymptotic and large-Z NSB estimators and the Coverage-Adjusted Estimator of [16].

The differences between the estimators of mutual information are more extreme than the difference for the estimators of entropy; the W&D estimator based on Equation (22) and the Coverage-Adjusted Estimator perform comparably. The large-Z NSB Estimator performs comparably in the high-entropy regime; the asymptotic NSB estimator tends to perform poorly.

We emphasize that the choice of DI hyperprior used in these experiments was "naive", not based on any careful reasoning or desiderata. In particular, no consideration was given to what type of generative process plausibly underlies the construction of $|Z|$ (see Section 8). Potentially more compelling results would occur for a more careful choice of hyperprior.

Figure 4. RMSerror for the estimation of mutual information with unknown bin number. The true bin number is 100, interpreted as events over a 10×10 joint space, sampled ten times (*i.e.*, radically under-sampled compared to the number of bins). The solid line shows the estimate made using the W&D estimator, Equation (22); the dashed line, the asymptotic version of NSB (Equation (29) of [19]); the dotted line, the Coverage-Adjusted Estimator of [16]; the dot-dashed line, the "large-Z" version of NSB. (**Top**) Results when ρ is randomly generated under a Dirichlet distribution whose concentration, c, is varied from 10^{-2} (highly non-uniform; many hits in a few bins) to 10^4 (highly uniform), as indicated. (**Bottom**) Results when ρ is is a power-law distribution over the bin number with index α, which is varied from zero (perfectly uniform) to four (highly non-uniform). (The Zipf distribution corresponds to $\alpha = 1$.)

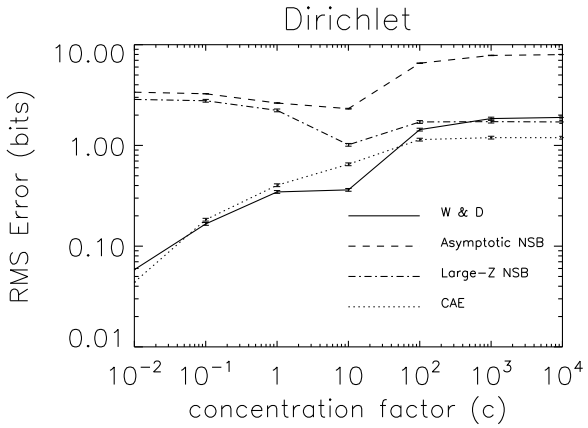
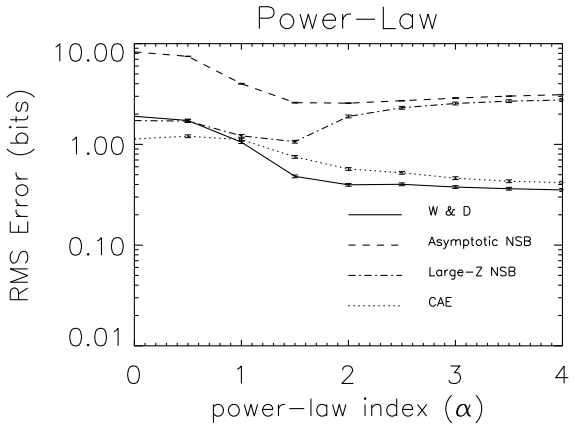


Figure 4. *Cont.*



8. Generative Models of Z

In this section, we discuss subtleties in how one models the statistical generation of Z. To ground the discussion, we will sometimes consider an example where a fishery's biologist randomly samples fish from a lake via a catch-and-release protocol, to try to ascertain quantities, like the number of fish species in the lake, the entropy of the distribution of the counts of members of those species, *etc*. In this example, Z is the set of all fish species in the lake (and is explicitly a random variable), and $\vec{n}$ is a set of the counts of the species of fish.

8.1. Mapping Physical Samples to Bin Labels

In the processes considered in the previous sections, $\pi(m)$ is sampled to produce m, and a set, Z, of m elements, labeled, for example, $\{1 \ldots, m\}$, is then created. Next, c is sampled from $\pi(c \mid m)$. After this, $P(\rho \mid c, m)$ is sampled to get a ρ, and finally, $\vec{n}$ is sampled from ρ.

Note that the $\vec{n}$ produced at the end of this process is a set of counts of the integers ranging from one to m. However, physically, $\vec{n}$ is not a set of counts of integers. This implies that we need a map from the physical characteristics of the samples in the real world into $\{1, \ldots, m\}$. As an illustration, in the fish-in-a-lake example, $\vec{n}$ is a set of the counts of species of fish that are distinguished by their physical characteristics (assuming no DNA sequencing or the like is used). Therefore, to apply the formulas derived in the previous sections, the biologist provided with a sample of the counts of fish species needs an invertible map sending each species of fish in their sample into $\{1 \ldots, m\}$.

How should the biologist create that map? One idea might be to randomly build an invertible map taking each of the distinct species of fish they have sampled to a different member of the set of integers from one to m. However, the biologist cannot do this, since they do not know m, and, so, cannot build such a map. (m is a random variable, whose value is not known with certainty to the biologist, even after the biologist gets $\vec{n}$). Another possibility would be to assign the species of the first fish the biologist samples to one, the second distinct species sampled to two, and so on. However, this would introduce major biases in the estimators (for example, before *any* data is generated, we would know that $n_1 \geq 1$, something we would not know for any n_i, where $i > 1$).

As an alternative, we can model the statistical process as one in which the biologist assigns a species "label" to each new fish as the biologist draws it from the lake, based on the physical characteristics of that fish. More precisely, assuming the biologist measures K real-valued physical characteristics of each fish that the biologist draws from the lake, we can model the sampling process as follows:

- $\pi(m)$ is sampled to get m, the number of fish species in the lake. At this point, nothing is specified about the physical characteristics of each of those m species.
- Next, a Dirichlet distribution is sampled that extends over distributions ρ that themselves are defined over m bins.
- Next, a vector, $v_j \in R^K$, is randomly assigned to each of $j = 1, \ldots, m$, e.g., where each of the m vectors is drawn from a Gaussian centered at zero (the precise distribution does not matter). v_j is the set of K real-valued physical characteristics that we will use to define an idealized canonical specimen of fish species, j. By identifying the subscript, j, on each v_j as the associated bin integer, we can view ρ as defined over m separate K-dimensional vectors of fish species characteristics.
- ρ is IID sampled to get a dataset of counts for each species, one through m. Physically, this means that the biologist draws a fish from bin j with probability ρ_j, *i.e.*, they draw a fish with characteristics v_j with probability ρ_j.

Note that we could interchange the order of steps 2 and 3. Note also that, in practice, since lakes do not contain "ideal fish", there will be some small noise added to v_j each time the biologist draws a member of species, j. We can assume that noise is small enough on the scale of the typical distance between the vectors of canonical fish species characteristics, so that the probability is infinitesimal, that the biologist misassigns what species a fish the biologist draws belongs to.

This generative model is more elaborate than the more informal one described at the beginning of this section. However, both models result in the same formulas, namely, those given in the previous sections.

8.2. Subset Selection Effects

There are other, simpler models that one might think solve the difficulty of how to map the physical members of $\vec{n}$ into integers $\{1, \ldots, |Z|\}$. However, many of these alternative models introduce subtle biases into the estimators. Seemingly trivial differences in the formulation of the problem of estimating functionals from data can have substantial effects on the resultant predictions.

To illustrate this, consider the common scenario where we are given a set, $\hat{Z}$, and know that $Z \subseteq \hat{Z}$, but do not know which precise subset of $\hat{Z}$ is Z. A simple example of such a scenario is a variant of the one where a field biologist wishes to estimate the entropy of the fish species in a particular lake by IID sampling fish in that lake. Say the biologist knows the set of all fish on Earth. However, assume they have no *a priori* knowledge that one fish is more likely than another to be a lake-dwelling fish. In this case, $\hat{Z}$ is the subset of all fish on Earth, and Z is the set of all lake-dwelling fish. While the biologist knows $\hat{Z}$, they are uncertain of Z.

We might presume that the estimation of quantities, like posterior expected entropy of the distribution of fish in the lake, would not depend on whether we calculate them for this scenario where we know

that Z is an (unknown) subset of $\hat{Z}$ or, instead, calculate them for the original scenario analyzed above, where we simply have uncertainty of Z, without any concern for an embedding set, $\hat{Z}$. However, it turns out that the estimation is quite different in these two scenarios. This illustrates how much care one must take in the statistical formulation of the estimation problem.

To see that estimation differs in this variant of the fish-in-a-lake example, first, as shorthand, define m to be the size of Z, $|Z|$. In the subset-of-$\hat{Z}$ scenario:

$$P(\vec{n} \mid \hat{Z}) = \sum_{m=|\mathscr{S}(\vec{n})|}^{|\hat{Z}|} P(m \mid \hat{Z}) P(\vec{n} \mid m, \hat{Z}) \tag{24}$$

(Note the implicit definition of $\vec{n}$ as the vector of counts for all elements in $\hat{Z}$.) $P(m \mid \hat{Z})$ plays the same role here as the prior, $P(|Z|)$, does in the analysis above of the original scenario where there is no $\hat{Z}$.

To evaluate the likelihood $P(\vec{n} \mid m, \hat{Z})$ in Equation (24), we need a stochastic model of how Z is formed from $\hat{Z}$. There are many such models possible. For simplicity, adopt the model that all Z's of a given size, m, are equally likely:

$$P(Z \mid m, \hat{Z}) = \frac{\delta_{|Z|,m}}{\binom{|\hat{Z}|}{m}} \tag{25}$$

Recalling the definitions of G and $\mathscr{S}$ in Section 3, we have the following (the proof is in Appendix B):

Proposition 6 *Under the conditional distribution in Equation* (25)*:*

- $P(\vec{n} \mid Z, m, \hat{Z}) \propto I(\mathscr{S}(\vec{n}) \subseteq Z) G(\vec{n}, c, m)$ *;*
- $P(\vec{n} \mid m, \hat{Z}) \propto \binom{|\hat{Z}| - |\mathscr{S}(\vec{n})|}{m - |\mathscr{S}(\vec{n})|} G(\vec{n}, c, m)$

In contrast to the likelihood $P(\vec{n} \mid m, \hat{Z})$ given by Proposition 6 for the subset-selection scenario, the likelihood for the original scenario analyzed above was:

$$\begin{aligned} P(\vec{n} \mid m) &= \int d\rho_Z \; P(\vec{n} \mid \rho_Z) P(\rho_Z \mid Z) \\ &\propto G(\vec{n}, c, m) \end{aligned} \tag{26}$$

where $|Z| = m$. Therefore, writing them out in full:

$$P(\vec{n} \mid m) = \frac{G(\vec{n}, c, m)}{\sum_{\vec{n}'} G(\vec{n}', c, m)} \tag{27}$$

$$P(\vec{n} \mid m, \hat{Z}) = \frac{\binom{|\hat{Z}| - |\mathscr{S}(\vec{n})|}{m - |\mathscr{S}(\vec{n})|} G(\vec{n}, c, m)}{\sum_{\vec{n}'} \binom{|\hat{Z}| - |\mathscr{S}(\vec{n}')|}{m - |\mathscr{S}(\vec{n}')|} G(\vec{n}', c, m)} \tag{28}$$

(The sum in the denominator in the second equation is implicitly restricted to those $\vec{n}' \in \hat{Z}$ whose support contains no more than m elements, and in both sums, we are implicitly restricting attention to those $\vec{n}'$ with the same total number of elements as $\vec{n}$.) Intuitively, the reason for the difference between

these two likelihoods is that in the subset-of-$\hat{Z}$ scenario, there are combinatorial effects reflecting the number of ways of assigning elements in $\hat{Z} \setminus \mathscr{S}(\vec{n})$ to the $|Z| - |\mathscr{S}(\vec{n})|$ bins in Z that are unoccupied in $\vec{n}$, whereas there are no such effects in the original scenario.

The extra combinatoric factor, $\binom{|\hat{Z}|-|\mathscr{S}(\vec{n})|}{m-|\mathscr{S}(\vec{n})|}$, in the subset-of-$\hat{Z}$ scenario's likelihood for m pushes that likelihood to prefer smaller values of $|\mathscr{S}(\vec{n})|$ compared to the original likelihood. It also distorts how the likelihood depends on m. To a degree, we can compensate for this second effect using the other term in the summand in Equation (24) besides $P(\vec{n} \mid m, \hat{Z})$, $P(m \mid \hat{Z})$. Even once we do this though, the precise estimates generated in the two scenarios will differ in general, since the prior, $P(m \mid \hat{Z})$, cannot fully compensate for an effect that depends on the data, $\vec{n}$.

More generally, one might even want to treat $|\hat{Z}|$ as a random variable. Returning to our concrete example involving fish, we would do this if we are uncertain about the total number of fish on Earth. This re-emphasizes the point that, as always, the choice of random variables and associated priors should match one's understanding of the underlying physical process by which the data is collected as accurately as possible.

9. Conclusions

The problem of estimating a functional of a distribution, ρ, based on samples of ρ is a core concern of statistics. In particular, in recent decades, there has been a great deal of work on estimating information-theoretic functionals of ρ based on samples of ρ.

Bayesian approaches to this problem began with WW, where the Dirichlet prior for ρ was adopted. This work concentrated on the case where the concentration parameter, c, of the Dirichlet prior equals the size of the underlying event space, $|Z|$. NSB pointed out that this special case has the problem that the resultant estimators are prior dominated whenever $|Z|$ is much larger than the support of the dataset. NSB realized that this problem could be addressed by using a hyperprior over c. They then advocated an approach to setting $P(c)$. Unfortunately, there is a substantial problem with the choice $c = |Z|$ analyzed in WW that is not fixed by the NSB choice of $P(c)$. In both approaches, the posterior expected value of a quantity, like mutual information, will depend on which of the many equivalent definitions of mutual information one adopts. In other words, that posterior expected value of such quantities is ill-defined.

In many situations, there will be uncertainty of $|Z|$, as well as c. Indeed, arguably, there is *always* an uncertain number of hidden degrees of freedom in the stochastic process that produced the data, degrees of freedom not recorded as components of that data. Since the stochastic process model must be set independently of the likelihood and it is the likelihood that determines what degrees of freedom are recorded, we must allow ρ to run over those hidden degrees of freedom, as well as the visible ones recorded in the data. Since we typically do not know how many such hidden degrees of freedom there are, this means we have an uncertain value of $|Z|$.

This reasoning argues that we should use a hyperprior, $P(c, |Z|)$. To do so, we must specify how the concentration parameter is statistically coupled with the size of the underlying event space. It is not at all clear how to do that in a hierarchical Bayesian way, where we cannot consider either the likelihood (which determines what variables are observed) or how the posterior estimate of ρ would be used (which

is what NSB uses to couple c and $|Z|$). It is also not at all clear how to specify a prior that extends over hidden degrees of freedom.

In this paper, we address the second concern by introducing the desideratum that *for any functional that only depends on those components of ρ corresponding to the recorded degrees of freedom*, the number of hidden degrees of freedom has no effect on our estimate of the functional. This desideratum says that our second problem is not a problem. We prove that this "Irrelevance of Unseen Variables" (IUV) desideratum can be satisfied, but only if c and $|Z|$ are independent. Therefore, IUV resolves both of our concerns.

In deriving this result, we prove an intermediate result that simplifies the calculation of some posterior moments. In particular, we show how to use it to derive the formula for posterior expected mutual information given in WW in essentially a single line.

We also show that by using a $P(c, |Z|)$ consistent with IUV rather than the one used in either WW or NSB, we resolve the problem shared by them that posterior expected mutual information is ill-defined. In addition, as we illustrate, using a hyperprior that respects IUV can also greatly simplify calculation of posterior moments of information-theoretic functionals. Another advantage of allowing $|Z|$ to vary and adopting IUV's hyperprior is that posterior expected values of information-theoretic quantities are no longer prior-dominated as they are under the hyperprior of WW. In this sense, there is no need for using approximations for setting $P(c)$, as, for example, the scheme in NSB.

After presenting these results, we discussed both hierarchical Bayesian approaches and other approaches for estimating information theoretic quantities when m and c are both random variables. We ended by discussing some changes to the statistical formulation of the estimation problem that would appear to be innocuous, but can actually substantially affect the resultant estimates.

Acknowledgments

We would like to thank John Young, Michael Hurley and Gordon Pusch for their assistance in compiling the errata. S.D. acknowledges the support of the Santa Fe Institute Omidyar Postdoctoral Fellowship, the National Science Foundation Grant EF-1137929, "The Small Number Limit of Biological Information Processing" and the Emergent Institutions Project. D.H.W. acknowledges the support of the Santa Fe Institute.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix A—Relevant Results and Errata from WW

In this appendix, we review relevant results from WW for the case of fixed Z and $c = |Z|$ and present errata for those results in WW. (A preliminary set of errata was reported in [8].)

It is straight-forward to generalize the reasoning in WW to derive:

$$\mathbb{E}(H \mid \vec{n}) = -\sum_z \frac{n_z + c/|Z|}{N + c} \Delta\Phi^{(1)}(n_z + 1 + c/|Z|, N + c + 1) \tag{29}$$

where care must be taken to replace the quantity "n_i" in WW with "$\vec{n}(x,y) - 1 + c/|Z|$", since WW considered the uniform (Laplace) prior, in which $c = |Z|$ and $L(.)$ is flat. Continuing to make the assumption of WW (and many others) that L is flat, we can evaluate the posterior variance for arbitrary c as:

$$\begin{aligned}\mathbb{E}([H - \mathbb{E}(H \mid \vec{n})]^2 \mid n) &= \sum_{z \neq z'} \frac{(n_z + 1)(n_{z'} + 1)}{(N + c)(N + c + 1)} A_{z,z'} \\ &\quad + \sum_{z} \frac{(n_z + 1)(n_z + 2)}{(N + c)(N + c + 1)} B_z\end{aligned}$$

where:

$$\begin{aligned}A_{z,z'} &= \Delta\Phi^{(1)}(n_z + 2, N + c + 2) \\ &\quad \times \Delta\Phi^{(1)}(n_{z'} + 2, N + c + 2) \\ &\quad - \Phi^{(2)}(N + c + 2)\end{aligned} \tag{30}$$

and:

$$\begin{aligned}B &= [\Delta\Phi^{(1)}(n_z + 3, N + c + 2)]^2 \\ &\quad + \Delta\Phi^{(2)}(n_z + 3, N + c + 2)\end{aligned}$$

where $\Phi^{(n)}$ and $\Delta\Phi^{(n)}$ are defined in Section 3.

The variance was incorrectly reported in WW: there was an error in its version of the second line of Equation (30). The mean expected entropy, in the absence of data and under the Laplace prior, is:

$$\mathbb{E}(H) = -\Delta\Phi^{(1)}(2, m + 1) = \sum_{q=2}^{m} \frac{1}{q} \tag{31}$$

(This was incorrectly reported in [29], but correct in WW.)

A complete list of errata in the published article follows. Errata unique to the arXiv versions are not shown.

- The Dirichlet prior equation in the continued paragraph on page 6843 should have the summation symbol replaced with the product symbol.
- Theorem 8 on page 6846—error as described above, corrected in Equation (30). The analogous equation, $\mathcal{E}_{\overline{IJMN}}$ (WW1, page 6852), does not contain the analogous error.
- Definitions necessary for various subsets, on page 6851, have errors. In particular, ν_i should be $n_i + 1$, and γ_n should be $\prod_{i=1}^{n} \Gamma(\nu_i)$.
- There is an error in the definition of $\mathcal{E}_{\overline{IN}}$ (page 6852.) In particular, the ν symbols in the denominators of the term:

$$\left(1 - \frac{\nu_{i\cdot} + \nu_{\cdot n} - 2\nu_{in}}{\nu} + \frac{(\nu_{i\cdot} - \nu_{in})(\nu_{\cdot n} - \nu_{in})}{\nu(\nu + 1)}\right)$$

should be replaced by $\bar{\nu}_{in}$, and the $1+$ symbols in the terms

$$\left(1+\frac{\nu_{i\cdot}-\nu_{in}}{\bar{\nu}_{in}+r}\right) \quad \text{and} \quad \left(1+\frac{\nu_{\cdot n}-\nu_{in}}{\bar{\nu}_{in}+r}\right)$$

should be replaced by:

$$\left(1-\frac{\nu_{i\cdot}-\nu_{in}}{\bar{\nu}_{in}+r}\right) \quad \text{and} \quad \left(1-\frac{\nu_{\cdot n}-\nu_{in}}{\bar{\nu}_{in}+r}\right)$$

Appendix B—Miscellaneous Proofs

Proof of Lemma 1: To begin, recall that since $Z = X \times Y$, we can write ρ_Z as a matrix of real numbers, $\{\rho(x,y) : x \in X, y \in Y\}$ or, alternatively, as $(\rho_X, \rho_{Y|X})$, where $\rho_{Y|X}$ is the set of $|X||Y|$ real numbers given by $\rho(x,y)/\rho_X(x)$ for all $x \in X, y \in Y$. In other words, the space of all pairs $(\rho_X, \rho_{Y|X})$ is a coordinate system of $\Delta_{X\times Y}$, given by the $|X|-1$ real numbers specifying ρ_X and the $|X|(|Y|-1)$ real numbers specifying $\rho_{Y|X}$, as is the matrix given by $|X||Y|-1$ real numbers. Writing the coordinate transformation between these coordinate systems as $\rho_{X,Y}(x,y) = \rho_X(x)\rho_{Y|X}(y \mid x)$, we see that there is an integrating factor, which we can write as:

$$\int d\rho_{X,Y} \ldots \;=\; \int d\rho_{X,Y} d\rho_X \prod_x \rho_X(x)^{|Y|-1} \ldots$$

where we subtract one from the exponent inside the product to reflect the normalization constraint on $\rho_{Y|X}$.

The proposition's hypothesized equality expands to:

$$\int d\rho_X \; Q(\rho_X) \frac{\prod_x \rho_X(x)^{\vec{n}_X(x)-1+c/|X|}}{C(c,\vec{n}_X)} \;=\; \int d\rho_{X,Y} \; Q(\rho_X) \frac{\prod_{x,y} \rho_{X,Y}(x,y)^{\vec{n}(x,y)-1+c'/|X||Y|}}{C(c',\vec{n}_{X,Y})}$$

Using the appropriate integrating factor, the RHS can be written as:

$$\int d\rho_X d\rho_{Y|X} \; Q(\rho_X) \prod_x \rho_X(x)^{|Y|-1} \frac{\prod_{x,y}[\rho_X(x)\rho_{Y|X}(y \mid x)]^{\vec{n}(x,y)-1+c'/|X||Y|}}{C(c',\vec{n}_{X,Y})}$$

$$= \int d\rho_X d\rho_{Y|X} \; Q(\rho_X) \prod_x \rho_X(x)^{|Y|-1} \frac{\prod_{x,y}[\rho_X(x)]^{\vec{n}(x,y)-1+c'/|X||Y|} \prod_{x,y}[\rho_{Y|X}(y \mid x)]^{\vec{n}(x,y)-1+c'/|X||Y|}}{C(c',\vec{n}_{X,Y})} \tag{32}$$

Rearranging the exponents in the products and then separately collecting all terms involving ρ_X and all terms involving $\rho_{Y|X}$, we can rewrite this as

$$= \int d\rho_X d\rho_{Y|X} \; Q(\rho_X) \frac{\prod_x [\rho_X(x)]^{n_X(x)-1+c'/|X|} \prod_{x,y}[\rho_{Y|X}(y \mid x)]^{\vec{n}(x,y)-1+c'/|X||Y|}}{C(c',\vec{n}_{X,Y})}$$

$$= \int d\rho_X \frac{Q(\rho_X)}{C(c',\vec{n}_{X,Y})} \prod_x [\rho_X(x)]^{n_X(x)-1+c'/|X|} \prod_x \int d\rho_{Y|X} \; \prod_y [\rho_{Y|X}(y \mid x)]^{\vec{n}(x,y)-1+c'/|X||Y|}$$

The inner integral is over the $|Y|$-dimensional simplex and evaluates to $\frac{\prod_y \Gamma(\vec{n}(x,y)+c'/|X|Y|)}{\Gamma(\vec{n}_X(x)+|Y|)}$. Therefore, rearranging terms, we get:

$$\frac{1}{C(c', \vec{n}_{X,Y})} \frac{\prod_{x,y} \Gamma(\vec{n}(x,y) + c'/|X|Y|)}{\prod_x \Gamma(\vec{n}_X(x) + |Y|)} \int d\rho_X \; Q(\rho_X) \prod_x [\rho_X(x)]^{n_X(x)-1+c'/|X|}$$

$$= \frac{\Gamma(N + c')}{\prod_x \Gamma(\vec{n}_X(x) + |Y|)} \int d\rho_X \; Q(\rho_X) \prod_x [\rho_X(x)]^{n_X(x)-1|+c'/|X|}$$

$$= \frac{1}{C(c', \vec{n}_X)} \int d\rho_X \; Q(\rho_X) \prod_x [\rho_X(x)]^{n_X(x)-1+c'/|X|}$$

By inspection, this expression equals the LHS of our hypothesized equality if $c = c'$. Going the other way, it is easy to see that if $Q(\rho_X) = \sum_x [\rho_X(x)]^2$, but $c \neq c'$, then our equality does not hold. ■

Proof of Proposition 4 We have just proven that a DI hyperprior implies that IUV holds. To go the other way, first, for any $\vec{k} \in \Delta_X$, define:

$$Q_{\vec{k}}(\rho_X) \equiv \frac{\delta(\rho_X - \vec{k})}{\prod_x k(x)^{n_X(x)-1}}$$

Next, plug Corollary 3 into the definition of IUV to show that if IUV holds, then for any Q:

$$\int dc \; F(c) \int d\rho_X \; Q(\rho_X) \mathscr{D}_{c,X}(\rho_X \mid \vec{n}_X) = 0$$

where $F(c)$ is defined as the analytic extension of $\pi(c \mid |X|) - \pi(c \mid |X||Y|)$. Therefore, in particular, for any $\vec{k}, c > 0$:

$$0 = \int dc \; F(c) \int d\rho_X \; Q_{c,\vec{k}}(\rho_X) \mathscr{D}_{c,X}(\rho_X \mid \vec{n}_X)$$

$$= \int dc \frac{F(c) \left[\prod_x k(x) \right]^{c/|X|}}{C(c, \vec{n}_X)}$$

Since $\prod_x k(x) \in [0, (1/|X|)^{|X|}]$, we see that for any $\alpha \in [0, 1/|X|]$:

$$\int dc \; B(c) \alpha^c = 0$$

where $B(c) \equiv \frac{F(c)}{C(c, \vec{n}_X)}$. Redefining α by dividing it by $(1 + \epsilon)|X|$ and redefining $B(c)$ by multiplying it by $((1 + \epsilon)|X|)^c$, we see that for any $\alpha \in [0, (1 + \epsilon)]$:

$$\int dc \; B(c) \alpha^c = 0$$

Since, by hypothesis, this rescaled $B(c)$ is analytic about $c = 1$, we can differentiate both sides of this equation with respect to α an arbitrary number of times and evaluate it at $\alpha = 1$. This establishes that all

moments of $B(c)$ must equal zero. Since the Fourier transform of B is assumed analytic, this means that the Fourier transform of $B(c)$ must equal zero identically. Therefore, $B(c)$ must equal zero identically, and therefore, $F(c)$ must.

This establishes that if IUV holds then for any spaces, $XandY$, it must be that $P(c \mid |X|) = P(c \mid |X||Y|)$. Relabeling X and Y then establishes that if IUV holds, for any spaces, $XandY$, $P(c \mid |X|) = P(c \mid |Y|)$. ■

Proof of Proposition 6: Recall that $I(\mathscr{S}(\vec{n}) \subseteq Z)$ equals one iff $\mathscr{S}(\vec{n})$, the support of $\vec{n}$ is a a subset of Z. Therefore,

$$\begin{aligned} P(\vec{n} \mid Z, m, \hat{Z}) &= I(\mathscr{S}(\vec{n}) \subseteq Z) \int d\rho_Z \; P(\vec{n} \mid \rho_Z, Z, m, \hat{Z}) P(\rho_Z \mid Z, m, \hat{Z}) \\ &\propto I(\mathscr{S}(\vec{n}) \subseteq Z) \int d\rho_Z \; \prod_{z \in Z} [\rho_Z(z)]^{n(z)} \mathscr{D}_{c,Z}(\rho_Z) \end{aligned}$$

Since we are using a Dirichlet prior with a uniform baseline distribution, by symmetry, the integral on the RHS must have the same value for all Z, such that $I(\mathscr{S}(\vec{n})) \subseteq Z$ and $|Z| = m$. That value is $G(\vec{n}, c, m)$. This establishes the first claim.

Next, write:

$$P(\vec{n} \mid m, \hat{Z}) = \sum_{Z \subseteq \hat{Z}} P(\vec{n} \mid Z, m, \hat{Z}) P(Z \mid m, \hat{Z}).$$

Combining this with our first result and with Equation (25):

$$P(\vec{n} \mid m, \hat{Z}) \propto \sum_{Z : |Z| = m} I(\mathscr{S}(\vec{n}) \subseteq Z) G(\vec{n}, c, m)$$

(The fact that $P(Z \mid m, \hat{Z})$ is uniform over all Z of size m means that it will cancel out once we divide by the appropriate sum to normalize $P(\vec{n} \mid m, \hat{Z})$.) For any set, $\mathscr{S}(\vec{n})$, and any integer, $m \in \{|\mathscr{S}(\vec{n})|, \ldots |\hat{Z}|\}$, there are a total of $\binom{|\hat{Z}| - |\mathscr{S}(\vec{n})|}{m - |\mathscr{S}(\vec{n})|}$ sets $Z \in \hat{Z}$, such that $I(\mathscr{S}(\vec{n})) \subseteq Z$ and $|Z| = m$. Combining establishes the second claim. ■

References and Notes

1. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 1991.
2. Mackay, D. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
3. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253.
4. Grassberger, P. Entropy estimates from insufficient samplings. **2003**, arXiv:physics/ 0307138
5. Korber, B.; Farber, R.M.; Wolpert, D.H.; Lapedes, A.S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 7176–7180.

6. Wolpert, D.; Wolf, D. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841–6854.
7. Wolf, D.R.; Wolf, D.R.; Wolpert, D.H. Estimating functions of probability distributions from a finite set of samples, Part II: Bayes Estimators for Mutual Information, Chi-Squared, Covariance, and other Statistics. **1994**, arXiv:comp-gas/9403002.
8. Wolpert, D.; Wolf, D. Erratum: Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1996**, *54*, 6973.
9. Hutter, M. Distribution of mutual information. *Adv. Neural Inform. Process. Syst.* **2002**, *1*, 399–406.
10. Hurley, M.; Kao, E. Numerical Estimation of Information Numerical Estimation of Information Theoretic Measures for Large datasets; MIT Lincoln Laboratory Technical Report 1169; Massachusetts Institute of Technology: Lexington, MA, USA, 2013; Available online: http://www.dtic.mil/dtic/tr/fulltext/u2/a580524.pdf (accessed on 28 October 2013).
11. Archer, E.; Park, I.; Pillow, J. Bayesian estimation of discrete entropy with mixtures of stick-breaking priors. In *Advances in Neural Information Processing Systems 25*, Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, 3–6 December 2012, Lake Tahoe, NV, USA; pp. 2024–2032.
12. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. In *Advances in Neural Information Processing System*; Dietterich, T., Ed.; MIT Press: Cambridge, MA, USA, 2003.
13. 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*, e056111.
14. Nemenman, I.; Lewen, G.D.; Bialek, W.; van Steveninck, R.R.D.R. Neural coding of natural stimuli: information at sub-millisecond resolution. *PLoS Comput. Biol.* **2008**, *4*, e1000025.
15. Archer, E.; Park, I.M.; Pillow, J.W. Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy* **2013**, *15*, 1738–1755.
16. Vu, V.; Yu, B.; Kass, R. Coverage-adjusted entropy estimation. *Stat. Med.* **2007**, *26*, 4039–4060.
17. Jaynes, E.T.; Bretthorst, G.L. *Probability Theory : The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
18. James, R.G.; Ellison, C.J.; Crutchfield, J.P. Anatomy of a bit: Information in a time series observation. **2011**, arXiv:1105.2988.
19. Nemenman, I. Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy* **2011**, *13*, 2013–2023.
20. Wolpert, D.H. Reconciling Bayesian and non-Bayesian Analysis. In *Maximum Entropy and Bayesian Methods*; Kluwer Academic Publishers: Dordrecht, Netherlands, 1996; pp. 79–86.
21. Wolpert, D.H. The Relationship between PAC, the Statistical Physics Framework, the Bayesian Framework, and the VC Framework. In *The Mathematics of Generalization*; Addison-Wesley: Indianapolis, IN, USA, 1995; pp. 117–215.
22. Note that no particular property is required of the relation among multiple unseen random variables that may (or may not) exist.
23. Bunge, J.; Fitzpatrick, M. Estimating the number of species: A review. *J. Am. Stat. Assoc.* **1993**, *88*, 364–373.

24. The proof of this proposition uses moment-generating functions with Fourier decompositions of the prior, π. To ensure we do not divide by zero, we have to introduce the constant $1 + \epsilon$ into that proof, and to ensure the convergence of our resultant Taylor decompositions, we have to assume infinite differentiability. This is the reason for the peculiar technical condition.
25. Berger, J.M. *Statistical Decision theory and Bayesian Analysis*; Springer-Verlag: Heidelberg, Germany, 1985.
26. This is an example of a more general phenomenon, that in some statistical scenarios there is a low probability that the likelihood of a randomly sampled dataset is concentrated near the truth (various "paradoxes" of statistics can be traced to this phenomenon). See also the discussion of Bayes factors below.
27. Intuitively, for that $P(c)$, the likelihood says that a dataset, $\{125, 125, \ldots\}$, would imply a relatively high value of c and, therefore, a high probability that ρ is close to uniform over all bins. Given this, it also implies a low value of $|Z|$, since if there were any more bins than the eight that have counts, we almost definitely would have seen them seen them for almost-uniform ρ. Conversely, for the dataset, $\{691, \ldots, 6\}$, the implication is that c must be low where some rare bins trail out, and as a result, there might be a few more bins who had zero counts.
28. Of course, since the formulas in WW implicitly assume $c = m$, care must be taken to insert appropriate pseudo-counts into those formulas if we want to use a value, c, that differs from m.
29. Wolpert, D.; Wolf, D. Estimating functions of probability distributions from a finite set of samples, Part 1: Bayes Estimators and the Shannon Entropy. **1994**, arXiv:comp-gas/9403001.

© 2013 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 license (http://creativecommons.org/licenses/by/3.0/).