Entropy **2011**, *13*, 2013-2023; doi:10.3390/e13122013

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Coincidences and Estimation of Entropies of Random Variables with Large Cardinalities

Ilya Nemenman

Departments of Physics and Biology and Computational and Life Sciences Initiative, Emory University, 400 Dowman Dr., Atlanta, GA 30322, USA; E-Mail: ilya.nemenman@emory.edu

Received: 1 November 2011; in revised form: 8 December 2011 / Accepted: 15 December 2011 / Published: 19 December 2011

Abstract: We perform an asymptotic analysis of the NSB estimator of entropy of a discrete random variable. The analysis illuminates the dependence of the estimates on the number of coincidences in the sample and shows that the estimator has a well defined limit for a large cardinality of the studied variable. This allows estimation of entropy with no a priori assumptions about the cardinality. Software implementation of the algorithm is available.

Keywords: entropy estimation; coincidences; bias-variance tradeoff; model selection

Classification: MSC 94A17, 62F12, 62F15

1. Introduction

Estimation of functions of a discrete random variable with an unknown probability distribution is one of the simplest problems in statistics. However, the simplicity vanishes in an extremely undersampled regime, where K, the cardinality or the alphabet size of the variable, is much larger than N, the number of the samples. In this case, the average number of samples per possible outcome, or *bin*, is less than one, and the relative uncertainty about the underlying probability distribution and its various statistics is large. To decrease the posterior error, one may turn to Bayesian statistics and bias the set of a priori admissible distributions. However, finding an optimal bias-variance tradeoff is not easy. For severely undersampled cases, controlling the variance often make an estimator a function of the prior, rather than of the measured data.

This is often the case for inference of the Boltzmann-Shannon entropy, $H = -\sum_{i=1}^{K} q_i \ln q_i$ (here q_i is probability of an event i), an important characteristics of a discrete variable. In this paper, all logarithms are natural, and the unit of entropy is *nat*. Simple estimators of entropy have low variances but high biases that are difficult to calculate due to the divergence of the logarithm near zero [1]. Developments driven in part by computational biology applications have solved this problem in the moderately undersampled regime, $N \sim K$ and $N \sim e^H$ [1–9]. Interestingly, they also resulted in the understanding that it is impossible to estimate entropy with zero bias uniformly over all distributions for a smaller N. However, Ma has argued [10] that, since coincidences in data start to occur at $N \sim \sqrt{e^H}$, it is possible to estimate entropies even in the deeply undersampled regime, at least for some classes of probability distributions, such as uniform ones. Similar arguments are well-known in the literature on estimation of population sizes from capture-recapture data (see, e.g., [11] for recent developments). There it has been recognized that the population size (and the population entropy) can be estimated long before every possible individual outcome has been sampled with a high probability [12].

In 2002, Nemenman, Shafee and Bialek introduced a method for entropy estimation, hereafter called NSB [13]. While the estimator has proven successful in the Ma square-root regime [14,15], a theoretical basis for the success has not been presented in the literature. Here we review the method and perform its asymptotic analysis. We verify the intuition that the estimator works in the Ma regime by counting coincidences. We point out that the method can be viewed as finding the number of yet unseen bins with nonzero probability given K, the maximum cardinality of the variable. While estimation of K by model selection techniques cannot work (see below), we show that the method has a non-trivial limit as $K \to \infty$. Thus one should be able to calculate entropies of discrete random variables even without knowing their cardinalities. Our analysis allows for an efficient numerical implementation of the NSB estimator, which we have made available from [16].

2. Summary of the NSB Method

We use Bayes rule to expresses the posterior probability of a probability distribution $\mathbf{q} \equiv \{q_i\}$, $i = 1 \dots K$, of a discrete random variable with a help of its a priori probability, $\mathcal{P}(\mathbf{q})$. Thus if n_i i. i. d. samples from $\mathbf{q}$ are observed in bin i, such that $\sum_{i=1}^{K} n_i = N$, then the posterior, $P(\mathbf{q}|\mathbf{n})$, is

$$P(\mathbf{q}|\mathbf{n}) = \frac{P(\mathbf{n}|\mathbf{q})\mathcal{P}(\mathbf{q})}{P(\mathbf{n})} = \frac{\prod_{i=1}^{K} q_i^{n_i} \mathcal{P}(\mathbf{q})}{\int_0^1 d^K q \prod_{i=1}^{K} q_i^{n_i} \mathcal{P}(\mathbf{q})} \tag{1}$$

Following [13], we focus on the popular Dirichlet family of priors, indexed by a hyperparameter β:

$$\mathcal{P}_\beta(\mathbf{q}) = \frac{1}{Z_\beta} \delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1}, \quad Z_\beta = \frac{\Gamma^K(\beta)}{\Gamma(K\beta)} \tag{2}$$

Here the δ-function and Z_β enforce normalizations of $\mathbf{q}$ and $\mathcal{P}_\beta(\mathbf{q})$, respectively; and Γ stands for Euler's Γ-function. These priors are common in statistics since they result in an analytically tractable, multinomial posteriors. For example, Wolpert and Wolf [17] calculated posterior averages, here denoted as $\langle \dots \rangle_\beta$, of many interesting quantities, including the distribution itself,

$$\langle q_i \rangle_\beta = \frac{n_i + \beta}{N + \kappa}, \quad \kappa \equiv K\beta \tag{3}$$

and the moments of its entropy, which we will not reprint here.

According to Equation (3), Dirichlet priors add extra β samples (pseudocounts) to each bin. Thus for $\beta \gg N/K$, the data are unimportant, on average, and $P(\mathbf{q}|\mathbf{n})$ is dominated by almost uniform distributions, $\mathbf{q} \approx 1/K$. Then the posterior mean of the entropy is strongly biased upward to its maximum possible value of $H_{\max} = \ln K$. Similarly, for $\beta \ll N/K$, distributions in the vicinity of the frequentist's maximum likelihood estimate, $\mathbf{q} = \mathbf{n}/N$, are important, and $\langle H \rangle_\beta$ is biased downward [1].

[13] traced this problem to properties of the Dirichlet family. Its members encode reasonable a priori assumptions about $\mathbf{q}$, but not about $H(\mathbf{q})$. Instead, a priori assumptions about the entropy are strongly biased, as seen from the a priori moments:

$$\xi(\beta) \equiv \langle H|_{N=0} \rangle_\beta = \psi_0(\kappa + 1) - \psi_0(\beta + 1) \tag{4}$$

$$\sigma^2(\beta) \equiv \langle (\delta H)^2|_{N=0} \rangle_\beta = \frac{\beta + 1}{\kappa + 1}\,\psi_1(\beta + 1) - \psi_1(\kappa + 1) \tag{5}$$

Here $\psi_m(x) = (d/dx)^{m+1} \ln \Gamma(x)$ are the polygamma functions. $\xi(\beta)$ varies smoothly from 0 for $\beta = 0$, through 1 for $\beta \approx 1/K$, and to $\ln K$ for $\beta \to \infty$. $\sigma(\beta) \sim 1/\sqrt{K}$ for almost all β [13], which is negligibly small for large K. Thus $\mathbf{q}$ that is typical in $\mathcal{P}_\beta(\mathbf{q})$ has the entropy extremely close to some predetermined β-dependent value. This bias persists even when $N \sim K$ data are collected.

One should strive for the a priori distribution of entropy, $\mathcal{P}(H(\mathbf{q}))$, to be approximately uniform to have a chance for an unbiased estimator. NSB achieves the uniformity (but not necessarily zero bias) by noting that, following Equations (4) and (5), for large K, $\mathcal{P}_\beta(H)$ is almost a δ-function. Thus a prior that averages over all non-negative values of β (and, correspondingly, over all a $\xi \in [0; \ln K]$) may reduce the bias in the entropy estimation even for $N \ll K$. [13] proposed the following infinite mixture of Dirichlet priors [18] for the averaging:

$$\mathcal{P}(\mathbf{q}; \beta) = \frac{1}{Z}\,\delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1} \frac{d\xi(\beta)}{d\beta}\,\mathcal{P}(\beta) \tag{6}$$

Here Z is again a normalizing coefficient, and $d\xi/d\beta$ ensures uniformity for ξ, rather than for β. A non-constant prior on β, $\mathcal{P}(\beta)$, may be used if needed, but we will not focus on this term from now on. Such Dirichlet mixture results in $\mathcal{P}(\mathbf{q}) \neq \text{const}$, introducing biases in estimation of $\mathbf{q}$ as a tradeoff for a possibly accurate estimation of H.

Inference with the prior, Equation (6), involves additional averaging over β (or, equivalently, ξ). The a posteriori moments of the entropy are

$$\widehat{H^m} = \frac{\int_0^{\ln K} d\xi\, \rho(\xi, \mathbf{n}) \langle\, H^m \rangle_{\beta(\xi)}}{\int_0^{\ln K} d\xi\, \rho(\xi|\mathbf{n})} \tag{7}$$

where the unnormalized posterior density is

$$\rho(\xi|\mathbf{n}) = \mathcal{P}\left(\beta\left(\xi\right)\right) \frac{\Gamma(\kappa(\xi))}{\Gamma(N + \kappa(\xi))} \prod_{i=1}^{K} \frac{\Gamma(n_i + \beta(\xi))}{\Gamma(\beta(\xi))} \tag{8}$$

Note that, for $N = 0$, $\rho(\xi|0) = P(\beta(\xi))$. Thus if we choose $P(\beta(\xi)) = \text{const}$, then the a priori assumptions about ξ are exactly uniform, as we had hoped to achieve. We note again that the uniformity of the prior is not equivalent to zero posterior bias.

An additional reason for the choice of averaging over the model families, as in Equation (6), is provided by the theory of Bayesian model selection [13,19–23]. Specifically, families of probabilistic models of data that incorporate more models (have larger volumes in the model space) usually have high explanatory powers and include some models that are very likely a posteriori. However, they also include many extremely unlikely models, and the posterior probability averaged over the entire family is low. Thus the competition between the "goodness of fit" and the volume of the model space (the *Occam factor*) often attributes much of the posterior weight to model families that are relatively simple, but explain the data well. In the case of the NSB prior, different values of β index different model families. For small β, the estimates in Equation (3) are closer to the frequentist's maximum likelihood, explaining the data better. However, there is less smoothing, and the space of models is larger. Thus as argued in [13], one expects that the integrals in Equation (7) are dominated by some β^* with a small posterior variance, and then $\widehat{\cdots} \approx \langle \cdots \rangle_{\beta^*}$.

In this work, we start with investigating whether a maximum of the integrand in Equation (7), indeed, exists. We then study its properties. The results of the analysis leads to a deeper understanding of the NSB method.

3. Saddle Point Analysis

We calculate integrals in Equation (7) using the saddle point (a. k. a. Laplace) approximation. Since $\langle H^m \rangle_\beta$ does not depend on N, for $N \to \infty$, only the Γ-terms in ρ define the saddle. We write

$$\rho(\xi|\mathbf{n}) = \mathcal{P}(\beta(\xi)) \exp\left[-\mathcal{L}_K(\mathbf{n}, \beta)\right] \tag{9}$$

$$\mathcal{L}_K(\mathbf{n}, \beta) = -\sum_i \ln\Gamma(\beta + n_i) + K \ln\Gamma(\beta) - \ln\Gamma(\kappa) + \ln\Gamma(\kappa + N) \tag{10}$$

Differentiating, we obtain the following equation for the saddle point (or the maximum likelihood) value, $\kappa^* = K\beta^*$:

$$\frac{1}{K}\sum_i^{n_i>0} \psi_0(n_i + \beta^*) - \frac{K_1}{K}\psi_0(\beta^*) + \psi_0(\kappa^*) - \psi_0(\kappa^* + N) = 0 \tag{11}$$

where K_m denotes the number of bins that have, at least, m counts. Note that $N > K_1 > K_2 > \cdots$.

If $K \gg N$, and if there are many bins with multiple counts, *i.e.*, $N - K_1 \gg 1$, then the (unknown) $\mathbf{q}$ is likely non-uniform. Thus the entropy is significantly smaller than its maximum possible value $H_{\max}$. Since for any $\beta = O(1)$, $\langle H \rangle_\beta \approx H_{\max}$ [13], small entropy estimate is achievable only if $\beta^* \to 0$ as $K \to \infty$. Thus we will look for

$$\kappa^* = \kappa_0 + \frac{1}{K}\kappa_1 + \frac{1}{K^2}\kappa_2 + \ldots \tag{12}$$

where none of κ_j depends on K. Plugging Equation (12) into Equation (11), we get an equation for κ_0:

$$\frac{K_1}{\kappa_0} = \psi_0(\kappa_0 + N) - \psi_0(\kappa_0) \tag{13}$$

The leading terms in the expansion of κ^* are:

$$\kappa_1 = \sum_i^{n_i>1} \frac{\psi_0(n_i) - \psi_0(1)}{K_1/\kappa_0^2 - \psi_1(\kappa_0) + \psi_1(\kappa_0 + N)} \tag{14}$$

$$\kappa_2 = \frac{\left[\frac{K_1}{\kappa_0^3} + \frac{\psi_2(\kappa_0) - \psi_2(\kappa_0+N)}{2}\right]\kappa_1^2 + \sum_i^{n_i>1} \kappa_0 \left[\psi_1(n_i) - \psi_1(1)\right]}{K_1/\kappa_0^2 - \psi_1(\kappa_0) + \psi_1(\kappa_0 + N)} \tag{15}$$

We have calculated additional higher order terms. However, when $K \gg 1$, as is common in applications, these terms are rarely needed.

We now solve Equation (13). For $\kappa_0 \to 0$ and $N > 0$, the r. h. s. of the equation is approximately $1/\kappa_0$ [24]. For $\kappa_0 \to \infty$, it is close to N/κ_0. Thus if $N = K_1$, and the number of coincidences among data, $\Delta \equiv N - K_1$, is zero, then the l. h. s. majorates the r. h. s., and Equation (13) has no solution. That is, there is no saddle point in the integrand. If there are coincidences, a unique solution exists, and $\Delta \to 0$ means $\kappa_0 \to \infty$. Thus we search for κ_0 of the form $\kappa_0 \sim 1/\Delta + O(\Delta^0)$.

It is useful to define:

$$f_N(j) \equiv \sum_{m=0}^{N-1} \frac{m^j}{N^{j+1}} \tag{16}$$

where each of f_N's scales as N^0. Using properties of polygamma functions [24] and defining $\delta = \Delta/N$, we rewrite Equation (13) as

$$1 - \delta = \sum_{j=0}^{\infty} (-1)^j \frac{f_N(j)}{(\kappa_0/N)^j} \tag{17}$$

Combined with the previous observations, Equation (17) suggests that we look for κ_0 of the form

$$\kappa_0 = N\left(\frac{b_{-1}}{\delta} + b_0 + b_1\delta + \dots\right) \tag{18}$$

where each of b_j's is independent of δ and scales as N^0.

Substituting Equation (18) into Equation (17), we find the series expansion self-consistent, and

$$b_{-1} = f_N(1) = \frac{N-1}{2N} \tag{19}$$

$$b_0 = -\frac{f_N(2)}{f_N(1)} = \frac{-2N+1}{3N} \tag{20}$$

$$b_1 = -\frac{f_N^2(2)}{f_N^3(1)} + \frac{f_N(3)}{f_N^2(1)} = \frac{N^2 - N - 2}{9(N^2 - N)} \tag{21}$$

Again, more terms have been calculated and are used in the software implementation of the estimator.

The obtained expressions present the saddle point value β^* (or κ^*, or ξ^*) as a power series in $1/K$ and δ. To complete the evaluation of Equation (7), we now calculate the curvature at this saddle point:

$$\left.\frac{\partial^2 \mathcal{L}}{\partial \xi^2}\right|_{\xi(\beta^*)} = \left[\frac{\partial^2 \mathcal{L}}{\partial \beta^2} \frac{1}{(d\xi/d\beta)^2}\right]_{\beta^*} = \Delta + N O(\delta^2) \tag{22}$$

Notice that the curvature *does not* scale as a power of N as was suggested in [13]. The uncertainty in ξ^* is determined to the first order only by coincidences. One can understand this by considering $K \gg 1$

with $q_i \ll 1$ for most of the bins. Then counts of $n_i = 1$ are not informative for entropy estimation since they can correspond to massive bins, as well as to some random bins from the sea of the negligible ones. However, coinciding counts likely corresponds to high-probability bins, which should influence the entropy estimation. Note also that, to the first order in $1/K$, the exact positioning of coincidences does not matter: for a fixed Δ, a few coincidences in many bins or many coincidences in a single one produce the same saddle point and the same curvature around it. While this is an artifact of the specific choice of the prior $\mathcal{P}(H(\mathbf{q}))$, the similarity to Ma's coincidence counting [10] is intriguing.

In summary, if the number of coincidences $\Delta \gg 1$, then the saddle point analysis is self-consistent. A specific value β^* is selected a posteriori, and the variance of the entropy is small.

Numerical Implementation

The series expansions calculated above form the basis for a numerical implementation of the NSB algorithm. To calculate the posterior mean and variance of the NSB entropy estimator using Equation (7), evaluation of three integrals is required numerically (normalization, the first, and the second moments of H). The algorithm for this is as follows. When $\Delta = 0$, the integrands are not peaked, and the integrals can be evaluated by simple Gaussian quadratures or other user-selected methods. If instead $\Delta > 1$, the integrands will be peaked, strongly if $\Delta \gg 1$. Identification of the location of the peaks is then essential before numerical integration is done. We proceed as follows:

1. The saddle point (the maximum of $\rho(\xi|\mathbf{n})$) is found numerically by:
   (a) evaluating an approximation for κ_0 using the first few terms of the series, Equation (18);
   (b) using the approximate value as a starting point for the Newton-Raphson iterative algorithm to solve for κ_0 from Equation (13);
   (c) plugging the solution into the series expansion for the saddle κ^*, Equation (12);
   (d) and, finally, using the latter solution as a starting point for the Newton-Raphson search of a more accurate value of κ^* in Equation (11).
2. Each of the integrands in Equation (7) is divided by the value of $\rho(\xi|\mathbf{n})$ at the saddle point, so that the maximum of the integrands is $O(1)$.
3. Curvature around the saddle point (and hence the posterior variance) is evaluated numerically.
4. The integrals are evaluated numerically over the range that spans a few standard deviations on both sides of the saddle point; the range is controlled by the user-specified desired accuracy.

The above algorithm has been implemented in Octave/Matlab and C++. It is available from [16]. The input to the routines is either the histogram of counts (Octave/Matlab and C++), or a series of samples (C++ only). The output of the routines is either the posterior mean and the standard deviation of the entropy, and the position of the saddle point, or a variety of diagnostics information if the integration fails for any reason. The C++ version is implemented specifically to allow estimation of entropies on alphabets with arbitrary large cardinalities. It is limited only by the ability of the data series to fit in the computer memory.

4. Choosing a Value for K?

We are interested in the regime $N \ll K$, when the number of pseudocounts in occupied bins, $K_1\beta$, is negligible compared to their number in empty bins, $(K - K_1)\beta \approx K\beta$. Then Equations (3) and (8) show that selecting β (*i.e.*, integrating over it) means balancing N, the number of actual counts versus $\kappa = K\beta$, the number of pseudocounts, or, equivalently, the scaled number of unoccupied bins. K is often unknown in real-life applications, or the number of possible outcomes is a countable infinity. Estimation of K from data has proven to be a hard problem, only solved completely for uniform distributions [10,11]. One can consider varying K (instead of β) to find its maximum a posteriori value when performing Bayesian integration over κ.

To see that this will not work, we note that smaller K leads to a higher maximum likelihood since the total number of pseudocounts is less. Unfortunately, since there are fewer bins (degrees of freedom) available, smaller K also means smaller volume in the distribution space. Thus Bayesian averaging over K is trivial: the smallest possible number of bins (*i.e.*, no empty bins) dominates. This can be seen from Equation (8): only the first ratio of Γ-functions in the posterior density depends on K, and it is maximized for $K = K_1$. Thus straightforward selection of the value of K is not an option. However, the next section suggests a way around this hurdle.

5. Unknown or Infinite K

Often the true value of K is unknown because its simple estimate is intolerably large. For example, consider measuring entropy of ℓ-gramms in printed English [25] using an alphabet with 29 characters: 26 different letters, one symbol for digits, one space, and one punctuation mark. Then for $\ell = 10$, a naive estimate of K is $29^{10} \sim 10^{14}$. Only very few of all possible 10-gramms are allowed by the grammar, but one does not know how many exactly. Thus one has to work in the space of full cardinality, which is ridiculously undersampled.

As shown in Section 3, NSB is well defined even for finite N and extremely large K, provided $\Delta \gg 1$. Moreover, if $K \to \infty$, then the expressions simplify since only the first term in Equation (12) needs to be kept. Even more interestingly, for an increasing K and $\beta \gg 1/K$, $\mathcal{P}_\beta(H)$ becomes closer to a delta function since the a priori variance of entropy drops to zero as $1/K$, Equation (5). Thus NSB becomes more "certain" as K increases. Correspondingly, a possible solution to the problem of unknown cardinality is to use an upper bound estimate for K. It is better to overestimate K than to underestimate it. Even $K \to \infty$ can be used. Insensitivity of the method to the value of K was explored empirically in [14].

Which assumptions allow NSB to use a few data points to specify entropy of a variable with even an infinite cardinality? A typical distribution in the Dirichlet family has a specific rank ordered (Zipf) plot [13]: the number of bins with the probability less than some q is given by an incomplete B-function, I,

$$\nu(q) = KI(q; \beta, \kappa - \beta) \equiv K \frac{\int_0^q dx x^{\beta-1}(1-x)^{\kappa-\beta-1}}{B(\beta, \kappa - \beta)} \tag{23}$$

where B is the usual complete B-function. NSB estimates the best value for β using bins with coincidences, the head of the rank ordered plot. But knowing β defines the tails, where no data has been observed yet, allowing entropy estimation. Thus NSB relies on the rank-ordered tail of the studied distribution to be not too far away from the form in Equation (23). If the Zipf plot of the studied distribution has a substantially longer tail, then one should not trust the results of the method. An empirical procedure for detecting this case has been suggested in [14,15].

With this warning in mind, we can analytically calculate the entropy estimate and its variance for a very large K. We want the results that hold even if the saddle point analysis, Section 3, fails when $\Delta \sim 1$. Following Equations (12) and (18), $\beta^* \to 0$, but $\kappa^* = K\beta^* \sim N^2/\Delta \gg N \gg 1$. The range of entropies is $0 \le H \le \ln K \to \infty$, so the prior on H produced by $\mathcal{P}(\mathbf{q};\beta)$ is (almost) uniform over a semi-infinite range and thus is ill-defined. Similarly, there is a problem normalizing $\mathcal{P}_\beta(\mathbf{q})$. However, both problems are resolved by an appropriate limiting procedure, and we disregard them in what follows.

To perform the integrals in Equation (7), we point out that, for $K \to \infty$, $\delta = \Delta/N \to 0$, and $\kappa \to \kappa^*$, we have $\kappa^* \sim N^2/\Delta$, and then $\left[\langle H(\mathbf{n})\rangle_\kappa - \xi(\beta)\right]\Big|_{\kappa\approx\kappa^*} = O(\delta) + O(1/K) \equiv O(\delta, 1/K)$. A similar relation holds for $\langle H(\mathbf{n})\rangle_\kappa$. That is, the posterior averages of the entropy and its square are almost indistinguishable from ξ and ξ^2, their respective a priori averages. Since now we are interested in small Δ (otherwise we can use the saddle point analysis), we replace $\langle H^m\rangle_\beta$ by ξ^m in Equation (7). The error of this approximation is $O\left(\delta, \frac{1}{K}\right) = O\left(\frac{1}{N}, \frac{1}{K}\right)$.

We transform the Lagrangian in Equation (10). First, we drop terms that do not depend on κ since they appear in the numerator and denominator of Equation (7) and thus cancel. Second, we expand around $1/K = 0$. This gives

$$\mathcal{L}_K(\mathbf{n},\kappa) = -\sum_i^{n_i>1} \ln\Gamma(n_i) - K_1 \ln\kappa - \ln\Gamma(\kappa) + \ln\Gamma(\kappa+N) + O\left(\frac{1}{K}\right) \tag{24}$$

We note that κ is large in the vicinity of the saddle if δ is small and N is large, *cf.* Equation (18). Thus by definition of ψ-functions, $\ln\Gamma(\kappa+N) - \ln\Gamma(\kappa) \approx N\psi_0(\kappa) + N^2\psi_1(\kappa)/2$. Further, $\psi_0(\kappa) \approx \ln\kappa$, and $\psi_1(\kappa) \approx 1/\kappa$ [24]. Finally, since $\psi_0(1) = -C_\gamma$, where C_γ is the Euler's constant, Equation (4) says that $\xi - C_\gamma \approx \ln\kappa$. Combining all of these expressions, we get

$$\mathcal{L}_K(\mathbf{n},\kappa) \approx -\sum_i^{n_i>1} \ln\Gamma(n_i) + \Delta(\xi - C_\gamma) + \frac{N^2}{2}\exp(C_\gamma - \xi) \tag{25}$$

where $\approx$ means the precision of $O(1/N, 1/K)$.

We write:

$$\widehat{H} \approx C_\gamma - \frac{\partial}{\partial\Delta} \ln \int_0^{\ln K} \mathrm{e}^{-\mathcal{L}} d\xi \tag{26}$$

$$(\widehat{\delta H})^2 \approx \left(\frac{\partial}{\partial\Delta}\right)^2 \ln \int_0^{\ln K} \mathrm{e}^{-\mathcal{L}} d\xi \tag{27}$$

The integrals in these expressions are calculated by substituting $\exp(C_\gamma - \xi) = \tau$ and replacing the limits of integration $1/K\exp(C_\gamma) \le \tau \le \exp(C_\gamma)$ by $0 \le \tau \le \infty$. This introduces errors of $\sim (1/K)^\Delta$ at the lower limit and $\sim \delta^2\exp(-1/\delta^2)$ at the upper limit. Both errors are within the precision of interest $O(1/K, 1/N)$ if there is, at least, one coincidence. Thus

$$\int_0^{\ln K} \mathrm{e}^{-\mathcal{L}} d\xi \approx \Gamma(\Delta)\left(\frac{N^2}{2}\right)^{-\Delta} \tag{28}$$

Finally, substituting Equation (28) into Equation (26) and (27), we get

$$\widehat{H} \approx (C_\gamma - \ln 2) + 2\ln N - \psi_0(\Delta) \tag{29}$$

$$\widehat{(\delta H)^2} \approx \psi_1(\Delta) \tag{30}$$

These equations are valid to zeroth order in $1/K$ and $1/N$. They provide a simple, yet nontrivial, estimate of the entropy that can be used even if the cardinality of the variable is unknown. However, one always must analyze for a possible bias when using the estimator. Note that Equation (30) agrees with Equation (22) since, for large Δ, $\psi_1(\Delta) \approx 1/\Delta$. The similarity between the coincidence counting in Equations (29) and (30) and in Ma's analysis [10] is also clear.

6. Conclusions

We have calculated various asymptotic properties of the NSB estimator for estimation of entropies of discrete random variables. First, the posterior expectations have been evaluated in terms of power series in $1/K$ and $\delta = \Delta/N$, but for the number of coincidences $\Delta \gg 1$. Evaluation is done using the saddle point expansion. Convergence of the series depends on the number of coincidences rather than on the total number of samples. This elucidates the similarity to Ma's argument [10] and verifies the intuition of [13,14] that counting coincidence is what makes the method work in the severely undersampled regime. We have then discussed the limit when $\Delta \sim 1$, and the saddle point analysis is not applicable. Here we have shown that the estimator has a finite asymptote for the case of infinitely many bins, $K \to \infty$, or of an unknown number of bins. We obtained a closed form solutions for the estimate of the entropy and its variance in this regime. As for $\Delta \gg 1$, to the first order, both depend on the number of coincidences rather than on the total number of samples.

The NSB estimator has been implemented in software, using the current asymptotic analysis as one of the steps in numerical evaluation of posterior integrals. Armed with empirical tests for the absence of bias in the estimator suggested in [14,15], the software brings us one step closer to a reliable, model independent estimation of entropy of discrete probability distributions in the severely undersampled Ma regime. The method is proving to be particularly powerful in a variety of biological applications.

Acknowledgments

I thank Jonathan Miller, Chris Wiggins, and William Bialek for stimulating discussions and for encouragement to complete the manuscript. I also thank Christian Mendl for noticing mistakes in the earlier drafts of the manuscript. This work was done in part at Kavli Institute for Theoretical Physics, supported by NSF Grant No. PHY99-07949, and finished at Emory University with a support of NIH/NCI grant No. 7R01CA132629-04 and HFSP grant No. RGY0084/2011 .

References

1. Paninski, L. Estimation of entropy and mutual information. *Neural Comp.* **2003**, *15*, 1191–1253.
2. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Netw. Comput. Neural Syst.* **1996**, *7*, 87–107.

3. Strong, S.; Koberle, R.; de Ruyter van Steveninck, R.; Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200.
4. Victor, J. Binless strategies for estimation of information from neural data. *Phys. Rev. E* **2002**, *66*, 051903.
5. Antos, A.; Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithm.* **2002**, *19*, 163–193.
6. Batu, T.; Dasgupta, S.; Kumar, R.; Rubinfeld, R. The complexity of approximating the entropy. *SIAM J. Comput.* **2005**, *35*, 132–150.
7. Grassberger, P. Entropy estimates from insufficient samples. *arXiv* **2003**, physics/0307138v2.
8. Wyner, A.; Foster, D. On the lower limits of entropy estimation. Available online: http://www-stat.wharton.upenn.edu/~ajw/lowlimitsentropy.pdf (accessed on 16 December 2011).
9. Kennel, M.; Shlens, J.; Abarbanel, H.; Chichilnisky, E. Estimating entropy rates with bayesian confidence intervals. *Neural Comp.* **2005**, *17*, 1531–1576.
10. Ma, S. Calculation of entropy from data of motion. *J. Stat. Phys.* **1981**, *26*, 221–240.
11. Orlitsky, A.; Santhanam, N.; Vishwanathan, K. Population estimation with performance guarantees. In Proceedings of the IEEE International Symposium on Information Theory, Nice, France, 24th–29th June 2007; pp. 2026–2030.
12. Orlitsky, A.; Santhanam, N.; Vishwanathan, K.; Zhang, J. Limit results on pattern entropy. *IEEE Trans. Inf. Ther.* **2006**, *52*, 2954–2964.
13. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. In Advances in Neural Information Processing Systems 14; Dietterich, T.G., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002.
14. Nemenman, I.; Bialek, W.; de Ruyter van Steveninck, R. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E* **2004**, *69*, 056111.
15. Nemenman, I.; Lewen, G.; Bialek, W.; de Ruyter van Steveninck, R. Neural coding of natural stimuli: Information at sub-millisecond resolution. *PLoS Comput. Biol.* **2008**, *4*, e1000025.
16. NSB Entropy Estimation. Available online: http://nsb-entropy.sourceforge.net/ (accessed on 16 December 2011).
17. Wolpert, D.; Wolf, D. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841–6854.
18. Sjölander, K.; Karplus, K.; Brown, M.; Hughey, R.; Krogh, A.; Mian, I.S.; Haussler, D. Dirichlet mixtures: A method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* **1996**, *12*, 327–345.
19. Jeffreys, H. Further significance tests. *Proc. Camb. Phil. Soc.* **1936**, *32*, 416–445.
20. Schwartz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
21. Clarke, B.; Barron, A. Information—Theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Thy.* **1990**, *36*, 453–471.
22. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comp.* **1997**, *9*, 349–368.
23. Nemenman, I. Fluctuation-dissipation theorem and models of learning. *Neural Comp.* **2005**, *17*, doi:10.1162/0899766054322982.

24. Gradshteyn, I.; Ryzhik, I. *Tables of Integrals, Series and Products*, 6th ed.; Academic Press: Burlington, MA, USA, 2000.
25. Schurmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*, 414–427.