Estimation of Entropy and Mutual Information

Liam Paninski [a,1]

[a] *Center for Neural Science, New York University*

Abstract

We present some new results on the nonparametric estimation of entropy and mutual information. The setup is completely general and related to Grenander's method of sieves. First, we prove almost sure consistency and central limit theorems, under conditions on the rate of sieve growth, for three of the most commonly used information estimators. We derive improved descriptions of the bias and variance of these estimators along the way. Second, we prove a converse to these consistency theorems, demonstrating that a misapplication of the most common estimation techniques leads to an arbitrarily poor estimate of the true information, even given unlimited data. This "inconsistency" theorem leads to an analytical approximation of the bias, applicable over a large region of parameter space and valid in surprisingly small sample regimes. The two most practical implications of these results are negative: 1) past information estimates in the literature are likely contaminated by bias, even if "bias-corrected" estimators were used, and 2) confidence intervals calculated by standard techniques drastically underestimate the error of the most common estimation methods. Finally, we introduce an estimator with some very nice properties: the estimator comes equipped with rigorous bounds on the maximum error over all possible underlying probability distributions, and this maximum error turns out to be surprisingly small. We demonstrate the application of this new estimator on both real and simulated data.

Key words: Entropy, mutual information, estimator, bias, variance

1 Introduction

Information-theoretic analyses of neural data have become increasingly popular over the last decade. Two quantities — entropy, H, and mutual informa-

[1] LP is supported by an HHMI predoctoral fellowship. Contact: liam@cns.nyu.edu; for more details, see http://www.cns.nyu.edu/∼liam

tion, I — play a central role in information theory, and applied statisticians have long been interested in how to estimate these quantities from data. It is therefore somewhat surprising that many of the issues involved in this estimation problem are still far from completely understood.

The difficulty stems from the fact that I and H are functionals of probability measures. Thus, to estimate the information $I(X;Y)$ between two random variables X and Y, defined on some arbitrary measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, it seems that we need to estimate the underlying joint probability measure $p(X,Y)$ of X and Y. In many interesting cases, the "parameter space" — the space of probability measures under consideration — can be very large. For example, $\mathcal{X}$ could be a space of time-varying visual stimuli, and $\mathcal{Y}$ the space of spike trains evoked by a given stimulus. Given N samples, $\{x_i, y_i\}, 1 \leq i \leq N$, of stimulus together with the evoked response, how well can we estimate the information this cell provides the brain about the visual scene? Clearly, it is difficult to answer this question as posed; the relationship between stimulus and response could be too complex to be revealed by the available data, even if N is large by neurophysiological standards (in fact, there are general theorems to this effect [1]). Some kind of regularization is needed.

One reasonable approach [6] is indicated by the data processing inequality [3], which states that for any random variables X and Y and any functions S and T on the range of X and Y, respectively,

$$I(X;Y) \geq I(S(X);T(Y)).$$

Thus, instead of (quixotically) attempting to estimate I directly, we estimate a sequence of lower bounds on I: we introduce a sequence of functions, S_N and T_N, such that S_N and T_N preserve as much information as possible given that $I(S_N;T_N)$ can be estimated with some fixed accuracy, and as the size of the available data set increases, the "compression ratio" of S_N and T_N — the amount of information these functions throw away — shrinks and our lower bound grows monotonically toward the true information. This strategy could be considered an instance of the "method of sieves" from the theory of estimation in abstract spaces [4].

The simplest S_N and T_N discretize $\mathcal{X}$ and $\mathcal{Y}$ into a finite number of points, $m_{S,N}$ and $m_{T,N}$, where $m_{S,N}$ and $m_{T,N}$ grow with N. The generality of the data processing inequality means that these discretizations can take arbitrary form; T_N could, say, encode the total number of spikes emitted by the neuron for small N, then the occurrence of more detailed patterns of firing [6] for larger N, until, in the limit, all of the information in the spike train is retained. Thus, for each value of N, our problem reduces to estimating $I(S,T)$, where the joint distribution of the random variables S and T is discrete on $m_{S,N}m_{T,N}$ points, and our parameter spaceis the tractable $m_{S,N}m_{T,N}$-simplex, the set of convex combinations of $m_{S,N}m_{T,N}$ disjoint point masses.

Much is already known about estimating information in the case when m is fixed. The available results can be stated for the most commonly used estimator of I, the maximum likelihood (ML) estimate (also called the "plug-in" [1] or "naive" [6] estimator, denoted here by $\hat{I}_{MLE}$). For example, we know that $\hat{I}_{MLE}$ is asymptotically normal as $N \to \infty$. The asymptotic $\frac{1}{N}$ bias and variance follow from standard normal and chi-square theory [2]: the asymptotic variance rate varies smoothly across the space of underlying probability measures $p(x, y)$, while the bias rate depends only on the number of nonzero elements of p (and is therefore constant on the interior of the m-simplex). The asymptotic behavior of this estimation problem (again, when m is fixed and $N \to \infty$) is thus largely understood. However, it is unclear how applicable these results are in the "sieve" setting, where m (and therefore the discretized probabilities $p(S_N, T_N)$) can vary with N; we address this problem below.

2 Results

Many of the following results are stated in terms of the entropy H; corresponding results for I follow by Shannon's formula for discrete information. See the author's website for proofs of all of the statements given below. We will consider three common estimators. The first, $\hat{H}_{MLE}$, was defined above. Miller and Madow [5] proposed directly subtracting off the MLE's $\frac{1}{N}$ bias term, $\frac{m'-1}{2N}$; here the number of bins with nonzero p-probability, m', is estimated by ML. We will denote the resulting estimator by $\hat{H}_{MM}$. We will also consider the popular jackknife estimator

$$\hat{H}_{JK} \equiv N\hat{H}_{MLE} - \frac{N-1}{N}\sum_{i=1}^{N}\hat{H}_{(i)},$$

where $H_{(i)}$ is the MLE based on all but the i-th sample (something very similar to the jackknife was used, e.g., in [6]). We will let $\hat{H}$ denote a placeholder for which any of $\hat{H}_{MLE}$, $\hat{H}_{MM}$, or $\hat{H}_{JK}$ can be substituted.

Our first result is on the consistency of $\hat{H}$ in the "sieve" setting. We have that if $m_{S,N}$ and $m_{T,N}$ grow with N, but not too quickly, the sieve regularization works: the sieve estimator is almost surely consistent if $m \sim o(N)$ (a central limit theorem is available under slightly stronger conditions). The power of this result lies in its complete generality: we place no constraints whatsoever on either the underlying probability measure, $p(x, y)$, or the sample spaces $\mathcal{X}$ and $\mathcal{Y}$.

Our next results are mathematically simpler and serve to debunk a few myths that seem to have found their way into the neuroscience literature. First, no unbiased estimator exists. Second, while $\hat{H}_{MLE}$ is known to be always

nonpositively biased, $\hat{I}_{MLE}$ is *not* always nonnegatively biased. Third, the bias of H_{JK} and H_{MM} is *not* always $o(1/N)$; in fact, the bias of these two common estimators (and not only the "naive" H_{MLE}) is $\sim (N/m)^{-1}$ if $N/m \sim 1$, while the bias is $\sim \log(N/m)$ if $N/m << 1$. Moreover, the bias of $\hat{H}$ is *not* constant given m (as a function of the underlying distribution p), even approximately, unless $N/m >> 1$: in other words, it is not possible to simply "subtract off the bias" until the bias is already negligible. In short, bias problems plague even the "bias-corrected" estimators H_{JK} and H_{MM} unless $N >> m$.

For the variance, we have that

$$V(\hat{H}) \sim O(\frac{\log(N)^2}{N}) \; \forall N, m \; [1],$$

and

$$V(\hat{H}) \approx V(log(p)) \equiv -(H(p))^2 + \sum_i p_i (\log(p_i))^2 = O(\frac{\log(m)^2}{N}) \; [2].$$

This leads to a "bias-variance balance" function, valid when $N >> m$:

$$V/B^2 \approx \frac{N(\log m)^2}{m^2};$$

if V/B^2 is large, variance dominates the mean-square error (in the "worst-case" sense), and bias dominates if V/B^2 is small. It is not hard to see that if m is at all large, bias dominates until N is relatively huge. The major practical conclusion of all this is that error bars based on sample variance (or bootstrapping techniques) give very bad confidence intervals if m and N are large but comparable; that is, error bars computed by the usual techniques are too small, and miss the true value of H or I with very high probability, because of the large bias of the most common estimators.

Our last result is perhaps the most surprising, mathematically, and helps to explain why the bias of $\hat{H}$ can remain large while the variance always shrinks to zero. The basic idea is that entropy is a symmetric function of $p_i, 1 \leq i \leq m$, in that H is invariant under permutations of the points $\{1, \ldots, m\}$. Most reasonable estimators of H, including $\hat{H}_{MLE}$, $\hat{H}_{MM}$, and $\hat{H}_{JK}$, share this permutation symmetry (in fact, one can show that there is some statistical justification for restricting our attention to this class of symmetric estimators). Thus, the distribution of $\hat{H}_{MLE}$, say, is the same as that of $\hat{H}_{MLE}(p')$, where p' is the rank-sorted empirical measure. This leads us to study the limiting distribution of these sorted empirical measures. We have that the limit sorted empirical histogram exists but is not equal to the true density (Figure 1); as a consequence, $\hat{H}$ is biased, even in the limit of infinite data. It turns out that this asymptotic bias $B(p, c)$ is given by relatively simple formulae for a few

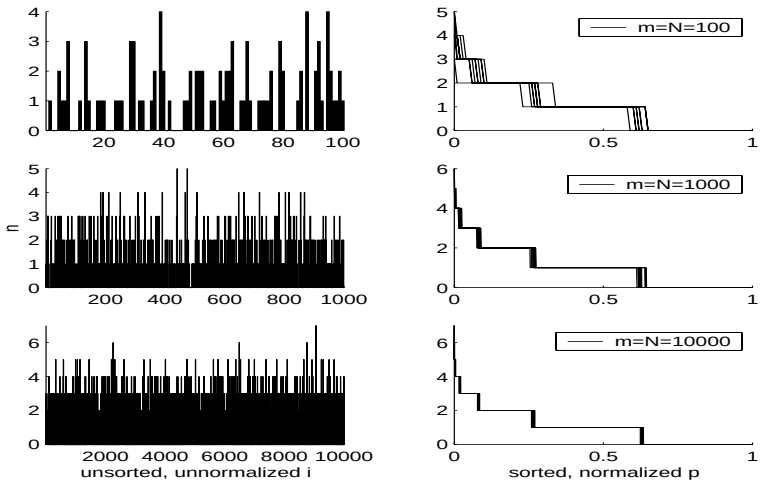

Fig. 1. "Incorrect" convergence of sorted empirical measures. Each left panel shows an example unsorted m-bin histogram of N samples from the uniform density, with $N/m = 1$. Ten sorted sample histograms are overlaid in each right panel, demonstrating the convergence to a nonuniform limit. The analytically derived $p'_{c,\infty}$ is drawn in the final panel, but is obscured by the sample histograms.

special cases of p: for example, when p is the uniform distribution,

$$B(p, c, \hat{H}_{MLE}) = \log(c) - e^{-c} \sum_{j=1}^{\infty} \frac{c^{j-1}}{(j-1)!} \log(j).$$

(Note that $B(p, c, \hat{H}_{MLE})$ behaves as $\log(N) - \log(m)$ as $c \to 0$, as expected given that $\hat{H}_{MLE}$ is supported on $[0, \log(N)]$.)

The result illustrated in Figure 1 leads to a new estimator (denoted $\hat{H}_{BAG}$ here) with several nice properties, e.g., $\hat{H}_{BAG}$ comes equipped with rigorous bounds on its "worst-case" error, and this maximum error turns out to be surprisingly small; some comparisons are made in Figures 2 and 3. Details on the derivation and computation of this estimator are available electronically.

References

[1] Antos, A. & Kontoyiannis, I. To appear, Random Structures and Algorithms; available at http://www.dam.brown.edu/people/yiannis/ (2002).

[2] Basharin, G. Theory of Probability and its Applications 4, 333-336 (1959).

[3] Cover, T. & Thomas, J. Elements of Information Theory. Wiley, NY (1991).

[4] Grenander, U. Abstract Inference. Wiley, NY (1981).

[5] Miller, G. & Madow, W. AFCRC-TR: 54-75 (1954).

[6] Strong, S. *et al.* Physical Review Letters 80: 197-202 (1998).

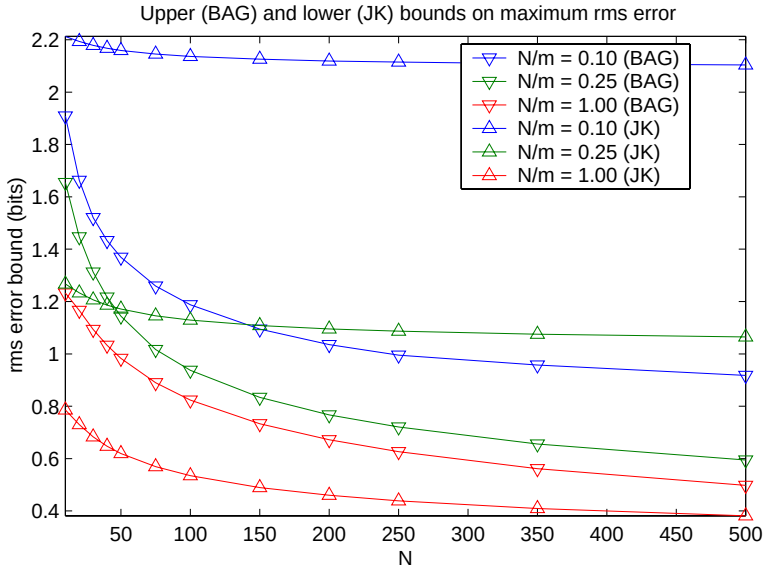

Fig. 2. A comparison of lower bounds on worst-case error for $\hat{H}_{JK}$ (upward-facing triangles) to *upper* bounds on the same for the new estimator $\hat{H}_{BAG}$ (down triangles), for several different values of N/m. The plots indicate that, roughly, $\max(E(\hat{H}_{BAG})) <\sim N^{-\alpha}, \alpha \approx .3$, while we know from the results above that $\max(E(\hat{H})) >\sim B_{\hat{H}}(N/m) + N^{-1/2}\log(\min(N, m))$.

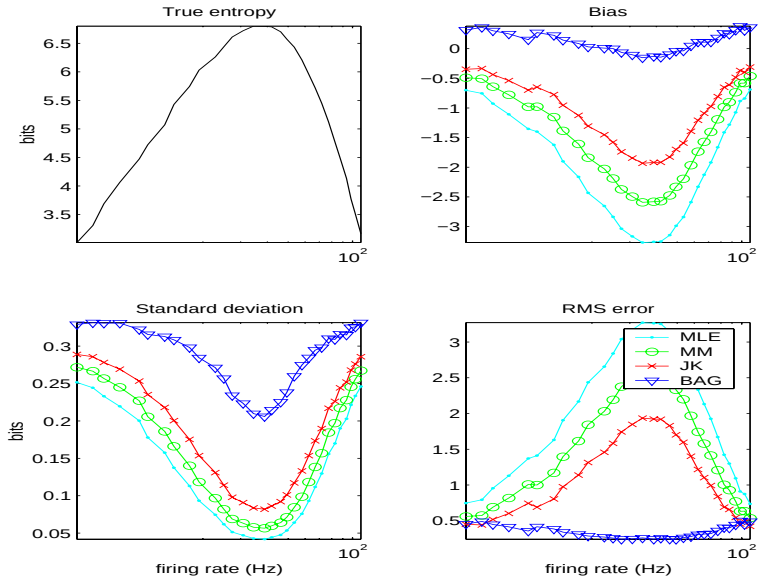

Fig. 3. Error curves for entropy estimators given $N = 100$ i.i.d. instances of binned noise-driven integrate-and-fire spike train data ($T = 200$ ms, $dt = 20$ ms, binary discretization). The firing rate of the IF cell was varied parametrically by changing the DC of input current. Note the large errors of $\hat{H}$ visible over much of the displayed parameter range.