SOME STATISTICAL ISSUES IN ESTIMATING INFORMATION IN NEURAL SPIKE TRAINS

Vincent Q. Vu[*], *Bin Yu*[†]

University of California, Berkeley
Department of Statistics

Robert E. Kass[‡]

Carnegie Mellon University
Department of Statistics
Center for the Neural Basis of Cognition

ABSTRACT

Information theory provides an attractive framework for attacking the neural coding problem. This entails estimating information theoretic quantities from neural spike train data. This paper highlights two issues that may arise: non-parametric entropy estimation and non-stationarity. It gives an overview of these issues and some of the progress that has been made.

Index Terms— entropy, estimation, nervous system, information theory

1. INTRODUCTION

Neurons propagate signal by generating sequences of electrical pulses, or more simply, spike trains. Understanding how information is represented and transformed in the nervous system is the central concern of *neural coding*. One appealing approach to neural coding (summarized in [1]) borrows concepts from Shannon's theory of communication and makes precise the vague notion of 'information,' in terms of entropy and mutual information [2].

2. DIRECT METHOD

A widely used method for estimating information directly from neural spike train data (see for instance [3, 4, 5]) is the aptly named 'direct method' [6]. A typical neurophysiology experiment where the direct method is applied consists of repeated presentations of a time-varying stimulus while simultaneously recording neural activity. The stimulus can be physical, for example sound and movies, or it can be abstract like an arm movement task. Figure 1 shows an example of data from such an experiment.

Though neural spike trains are naturally represented by point processes, the direct method makes them discrete by quantizing time at a precision of Δt and considering the pattern of spike counts in time windows of size T, either overlapping or non-overlapping. The number of spikes that occur in each bin become letters in a K-letter word, where $K = T/\Delta t$. See Figure 2 for an illustrated example in the non-overlapping window case.

Consider the stochastic process $\{S_t, X_t^k\}$ representing the value of the stimulus (S_t) and windowed response (X_t^k) at time $t = 1, \ldots, n$ during presentation (or trial) $k = 1, \ldots, m$. Given the responses $\{X_t^k\}$, the direct method considers two different entropies: (1) the *total entropy* H of the response, and (2) the local *noise entropy* H_t of the response at time t. The total entropy,

$$H = -\sum_x \bar{P}(x) \log \bar{P}(x),$$

is associated with the distribution $\bar{P}(x)$ of words across the entire experiment—the entire raster plot, while the local noise entropy,

$$H_t = -\sum_x P_t(x) \log P_t(x),$$

is associated with the distribution $P_t(x)$ of words across trials at a fixed time t—a column of the raster plot. These entropies are estimated directly from the neural response, and the direct method information estimate is the difference between the total entropy and the average (over t) noise entropy.

H and H_t depend implicitly on the size T of the time window and the time resolution Δt. Normalizing by $K = T/\Delta t$ and considering large T leads to the total and local entropy rates (at precision Δt) that are defined to be $\lim_{T\to\infty}(H/T)\Delta t$ and $\lim_{T\to\infty}(H_t/T)\Delta t$, respectively, when they exist. Their rate of convergence depends implicitly on the range of dependence in the response process. Furthermore, the phenomenon under investigation will often involve a fine time resolution Δt. Often an approximation

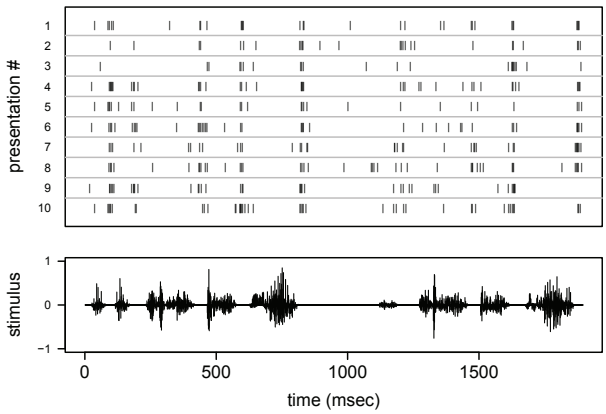

Fig. 1. Each row of the raster plot (above) shows the pattern of spiking of a Field L neuron of an adult male Zebra Finch in response to 1 of 10 repeated presentations of the natural bird song (below). Data from [3].

[*]Supported by a NSF VIGRE Graduate Fellowship and NIDCD grant DC 007293.

[†]Supported by NSF grants DMS-03036508, DMS-0605165, DMS-0426227, ARO grant W911NF-05-1-0104, NSFC grant 60628102, and a fellowship from the John Simon Guggenheim Memorial Foundation.

[‡]This work began while Kass was a Miller Institute Visiting Research Professor at the University of California, Berkeley. Support from the Miller Institute is greatly appreciated. Kass's work was also supported in part by NIMH grant RO1-MH064537-04 and NIBIB grant 5R01EB005847-03.

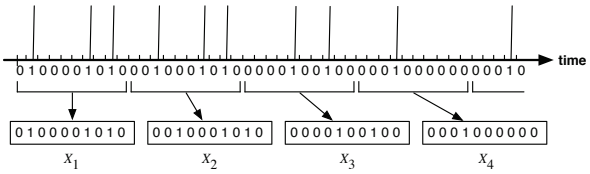

Fig. 2. A spike train is discretized into $\Delta t = 1$ msec bins and represented by a binary sequence. The bits in the binary sequence are grouped into non-overlapping windows of length $T = 10$ msec and divided into K-letter words $X_1, X_2, \ldots$ where $K = T/\Delta$.

is made by choosing sufficiently small Δt, large T, and then extrapolating. Large T and small Δt necessitate large K. As the number of potential words is exponential in K, the estimation of entropy can be challenging for large K.

3. ENTROPY ESTIMATION

Consider the conceptually simpler problem of estimating the entropy of a discrete distribution $P(x)$ over a set of words of unknown cardinality $s = \#\{x : P(x) > 0\}$ (potentially infinite). In the most basic case the observations (X_t) are assumed to be independent and identically distributed (i.i.d.) according to $P(x)$.

An apparent method of estimating the entropy is to apply the entropy formula after estimating $P(x)$, but estimating a discrete probability distribution is, in general, a difficult nonparametric problem. The maximum likelihood estimate (MLE), also referred to as the plug-in estimate, takes the empirical distribution $\hat{P}(x)$ (given by the observed frequencies) and plugs it into the entropy formula so that

$$\hat{H}_{MLE} = -\sum_x \hat{P}(x) \log \hat{P}(x).$$

This approach seems intuitively obvious and is consistent as $n \to \infty$ for fixed, finite s [7]. However, for small samples, it can lead to severely biased estimates [8, 9, 10]. This phenomenon, illustrated in Figure 3, is pronounced in the large s, small n regime endemic in spike train data. It is the dominating component of the mean squared error of the MLE.

3.1. Bias correction

Several proposals have been made to improve on the MLE by reducing its bias, while hopefully accumulating only a small increase in variance. For fixed and finite s, the leading term in the large n asymptotic bias of the MLE is $-(s-1)/(2n)$ [8]. This suggests that the bias of the MLE can be corrected by adding $-(\hat{s}-1)/(2n)$, where $\hat{s}$ is an estimate of the unknown s. However, the problem of estimating s is potentially more difficult. Taking $\hat{s}$ to be the observed cardinality leads to the Miller-Maddow (MM) correction. Bias correction based on the jackknife technique (JK) was proposed by [11]. It is more computationally intensive than the MLE as it requires that the MLE be recomputed n times.

The MLE belongs to a larger class of estimators that are linear in the statistics $f_j := \#\{x : \hat{P}(x) = j/n\}$. (In other words, f_j is the number of words that appeared exactly j times in $X_1, \ldots, X_n$.) [12] recognized this and proposed an estimator (BUB) based on numerical optimization of an upper bound on the bias of estimators in

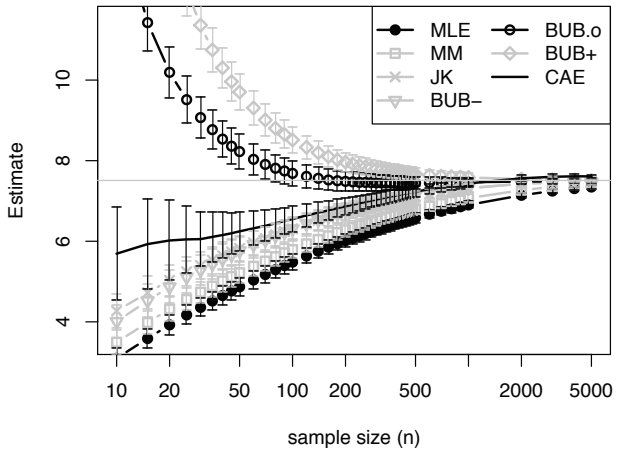

Fig. 3. Comparison of entropy estimators on simulated data sets of various sizes, drawn i.i.d. from a power-law distribution ($p(k)\ k^{-1}$) on 2^{10} words. The lines are averages over 1000 realizations and the bars indicate $\pm$ 1 standard deviation. The true entropy (≈ 7.5 bits) is indicated by the horizontal line. The MLE has the worst bias and consistently underestimates the entropy.

this class. Unfortunately, the upper bound depends the unknown s, but it seems to be less sensitive to underestimating s than the Miller-Maddow correction. Figure 3 shows the performance of BUB for 3 different choices of $\hat{s}$: observed cardinality (BUB-), twice the true cardinality (BUB+), and the true cardinality (BUB.o).

[13], in collaboration with Turing, also considered the statistics f_j, but in the more general context of estimating $P(x)$. They showed that the sample coverage $\sum\{P(x) : \hat{P}(x) > 0\}$ could be estimated by $\hat{C} = 1 - f_1/n$. The sample coverage is typically $\ll 1$ when $s > n$. Thus, estimates that do not account for unseen words, like the MLE, will have a large negative bias because low probability words that are frequently not covered by the sample contribute to the entropy of the distribution in a non-negligible way. [14] observed this and used the Good-Turing formula to correct $\hat{P}$. They suggest scaling $\hat{P}$ by $\hat{C}$ and then plugging into the Horvitz-Thompson estimator for a population total to correct for missing summands in the entropy formula. [15] studied this estimator in the context of spike train data and called it the coverage adjusted entropy estimator (CAE).

Bayesian methods for estimating the entropy have also been proposed. One approach [16] is to apply the entropy formula to a Bayesian estimate of $P(x)$. One example is the Laplace estimate that results from a Dirichlet prior. It accounts for missing words by adding 1 to the observed frequency of each word. A more direct albeit analytically difficult approach is to place a prior on the entropy itself (see, e.g., [17, 18]). [18] observed that the MLE corresponds to an Bayesian estimate with a nearly singular prior on the entropy. They proposed an estimator that attempts to induce an approximately flat prior on entropy. The approach is computationally intensive and depends on the unknown s.

The difficulty of entropy estimation when s is much larger than n is prototyped by the $s = \infty$ case, where the convergence rate of entropy estimates can be very slow [19]. Indeed, [20] showed that for P with finite entropy variance $Var[\log P(X_1)]$ the minimax asymptotic rate of convergence is $O_P(1/\log n)$ and that, surprisingly, the

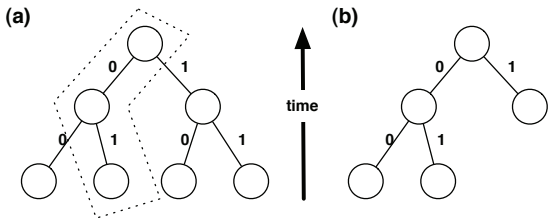

Fig. 4. (a) The suffix tree of binary words of length 2 ($K = 2, d = 2$). The path enclosed by the dashes corresponds to the word 10. (b) The context tree of a VLMC of order $K = 2$ on $d = 2$ letters. This particular model projects all words that end with a 1 down to 1 state.

MLE attains this rate. Asymptotic minimax analysis is not enough, as [15] showed that the CAE also attains this rate, but numerical simulations suggest that the finite sample performance of the CAE and other estimators is much better than that of the MLE (see Figure 3).

3.2. Dependence and context tree methods

When overlapping time-windows are considered, X_t is of the form $X_t = (Z_{t-K+1}, \ldots, Z_t)$. A simple generalization of the i.i.d. assumption is to instead assume that (X_t) is a stationary Markov chain, i.e. $P(X_t = x) = P(x)$ and $P(X_t = x|X_{t-1}, \ldots) = P(X_t = x|X_{t-1})$ for all t. Equivalently, (Z_t) is a stationary Markov chain of order K. This approach imposes structure by recognizing that the word X_t is the concatenation of letters $(Z_{t-K+1}, \ldots, Z_t)$. When the number of possible letters d is finite, there is a natural representation of the maximal state space of (Z_t) as a d-ary suffix tree of depth K, with paths from the top to the bottom of the tree corresponding to K letter words in time-reversed order. Figure 4 (a) shows an example for $K = 2, d = 2$.

Similarly to the i.i.d. case, the entropy of $P(x)$ can be estimated by applying the entropy formula to an estimate of the stationary distribution of the Markov chain. This can be done by first estimating the transition probabilities, and then computing the stationary distribution numerically. Still, this requires the estimation of an exponential number of parameters and is prone to overfitting. This difficulty can be circumvented by assuming that the memory length of the Markov chain is variable (VLMC, see [21]) so that many branches of the full suffix tree can be projected down to a smaller number of states, resulting in a *context tree* (see Figure 4 (b)). This is the approach taken by methods (see, e.g., [22, 23]) based on the context tree algorithm [24, 25]. The algorithm works by growing a maximal tree and then pruning branches according to a model selection criterion. The estimation of transition probabilities can still be problematic, and different proposals vary in their strategy. It may be fruitful to explore the bias correction ideas from i.i.d. entropy estimators with the context tree methods.

4. NON-STATIONARITY

A crucial assumption made in all of the entropy estimation methods described above is stationarity of the joint stimulus and response process $\{S_t, X_t\}$. Many applications, however, use non-stationary or even deterministic stimuli, so that entropy and mutual information are no longer well defined. Consider the the natural song stimulus in Figure 1 for example. The bursts of energy in the signal suggest that its statistical properties are not stationary in time.

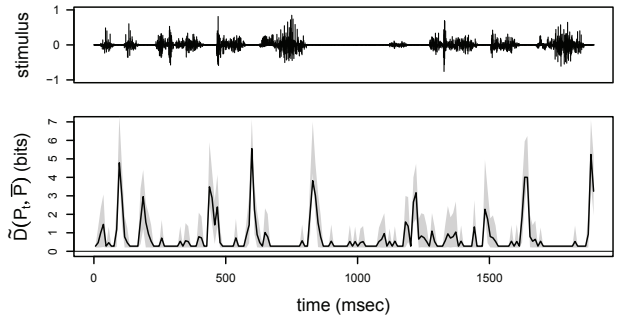

Fig. 5. Coverage adjusted estimate (below, solid line) of $D(P_t, \bar{P})$ from the response shown in Figure 1 with $K = 10$ and $\Delta t = 1$msec. The shaded region indicates pointwise 95% confidence intervals obtained by bootstrapping the trials 1000 times. The information estimate, 0.77 bits (per 10msec word), corresponds to the average value of the solid line. The stimulus is shown above for comparison.

Stationary methods may still be applied, but their results must be interpreted carefully. [26] gave an alternative interpretation of the direct method information estimate. The quantity that it estimates can be written as the time-average of the Kullback-Leibler divergence between the time t response distribution P_t and average response distribution $\bar{P}$ across the entire experiment [26]. To be precise, while holding n fixed, as $m \to \infty$, the direct method information estimate converges to

$$\frac{1}{n}\sum_{t=1}^{n} D(P_t||\bar{P}) = \frac{1}{n}\sum_{t=1}^{n}\left[\sum_x P_t(x)\log\frac{P_t(x)}{\bar{P}(x)}\right] \qquad (1)$$

with probability 1. This holds under mild assumptions, regardless of stationarity. When there is stationarity and ergodicity, this quantity coincides with mutual information as $n \to \infty$.

In the non-stationary case the information estimate no longer estimates mutual information in the usual sense, but this interpretation suggests that it instead measures magnitude of variation of the response as the stimulus varies. This may still be a useful assessment of the extent to which the stimulus affects the response as long as other factors that affect the response are themselves time-invariant. [26] proposed plotting $D(P_t||\bar{P})$, rather than just reporting its average as in (1). See Figure 5 for an example. Other methods that explicitly consider the dynamic and non-stationary nature of the stimulus and response should be used instead; see for instance [27].

5. CONCLUSION

The information estimates discussed in this paper sidestep the difficult problem of estimating the joint distribution of response and stimulus by instead estimating the difference between the marginal and conditional entropies of the response. Although the potentially high-dimensional problem of estimating the stimulus distribution is avoided, high-dimensionality reappears in considering the exponentially many possible patterns of spiking in the response. This makes entropy estimation challenging. Moreover, this approach tempts the practitioner into ignoring the role of the stimulus and the meaning of mutual information. This can lead to misinterpretation. Information theoretic approaches can and should be used, but the methods should explicitly consider the dynamic and non-stationary nature of the stimulus.

6. REFERENCES

[1] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nature Neuroscience*, vol. 2, no. 11, pp. 947–957, 1999.

[2] C. E. Shannon, "The mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[3] A. Hsu, S. M. N. Woolley, T. E. Fremouw, and F. E. Theunissen, "Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons," *J. Neuroscience*, vol. 24, no. 41, pp. 9201–9211, 2004.

[4] P. Reinagel and R. C. Reid, "Temporal coding of visual information in the thalamus," *J. Neuroscience*, vol. 20, no. 14, pp. 5392–5400, 2000.

[5] D. S. Reich, F. Mechler, and J. D. Victor, "Formal and attribute-specific information in primary visual cortex," *J. Neurophysiology*, vol. 85, no. 1, pp. 305–318, 2001.

[6] S. P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, "Entropy and information in neural spike trains," *Phys. Rev. Letters*, vol. 80, no. 1, pp. 197–200, 1998.

[7] G. P. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory of Probability and its Applications*, vol. 4, pp. 333–336, 1959.

[8] G. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology: Problems and Methods II-B*, H. Quastler, Ed., pp. 95–100. Free Press, Glencoe, IL, 1955.

[9] J. D. Victor, "Asymptotic bias in information estimates and the exponential (bell) polynomials," *Neural Computation*, vol. 12, pp. 2797–2804, 2000.

[10] A. Treves and S. Panzeri, "The upward bias in measures of information derived from limited data samples," *Neural Computation*, vol. 7, no. 2, pp. 399–407, 1995.

[11] S. Zahl, "Jackknifing an index of diversity," *Ecology*, vol. 58, pp. 907–913, 1977.

[12] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.

[13] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. 237–264, 1953.

[14] A. Chao and T. Shen, "Nonparametric estimation of shannon's index of diversity when there are unseen species in sample," *Environmental and Ecological Statistics*, vol. 10, pp. 429–443, 2003.

[15] V. Q. Vu, B. Yu, and R. E. Kass, "Coverage adjusted entropy estimation," *Statistics in Medicine*, vol. 26, no. 21, pp. 4039–4060, 2007.

[16] T. Schurmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos*, vol. 6, no. 3, pp. 414–427, 1996.

[17] D. Wolpert and D. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Phys. Rev. E*, vol. 52, no. 6, pp. 6841–6853, 1995.

[18] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Phys. Rev. E*, vol. 69, no. 5, pp. 056111, 2004.

[19] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures and Algorithms*, vol. 19, pp. 163–193, 2001.

[20] A. J. Wyner and D. Foster, "On the lower limits of entropy estimation," *Unpublished manuscript*, 2003.

[21] P. Bühlmann and A. J. Wyner, "Variable length Markov chains," *Annals of Statistics*, vol. 27, no. 2, pp. 480–513, 1999.

[22] M. B Kennel, J. Shlens, H. D. I. Abarbanel, and E. J. Chichilnisky, "Estimating entropy rates with Bayesian confidence intervals," *Neural Computation*, vol. 17, no. 7, pp. 1531–1576, 2005.

[23] Y. Gao, I. Kontoyiannis, and E. Bienenstock, "From the entropy to the statistical structure of spike trains," *IEEE ISIT*, pp. 645–649, 2006.

[24] J. Rissanen, "A universal data compression system," *IEEE IT*, vol. 29, no. 5, pp. 656—664, 1983.

[25] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "Context tree weighting: A sequential universal source coding procedure for FSMX sources," in *IEEE ISIT*, 1993, p. 59.

[26] V. Q. Vu, B. Yu, and R. E. Kass, "Information in the non-stationary case," *Neural Computation*, vol. 21, no. 3, 2009.

[27] R. Barbieri, L. M. Frank, D. P. Nguyen, M. C. Quirk, V. Solo, M. A. Wilson, and E. N. Brown, "Dynamic analyses of information encoding in neural ensembles," *Neural Computation*, vol. 16, no. 2, pp. 277–307, 2004.