NOTE Communicated by Stefano Panzeri

Information in the Nonstationary Case

Vincent Q. Vu
vqv@stat.berkeley.edu
Bin Yu
binyu@stat.berkeley.edu
Department of Statistics, University of California, Berkeley, CA 94720, U.S.A.

Robert E. Kass
kass@stat.cmu.edu
Department of Statistics and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Information estimates such as the direct method of Strong, Koberle, de Ruyter van Steveninck, and Bialek (1998) 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. While this is an effective estimation strategy, it tempts the practitioner to ignore the role of the stimulus and the meaning of mutual information. We show here that as the number of trials increases indefinitely, the direct (or plug-in) estimate of marginal entropy converges (with probability 1) to the entropy of the time-averaged conditional distribution of the response, and the direct estimate of the conditional entropy converges to the time-averaged entropy of the conditional distribution of the response. Under joint stationarity and ergodicity of the response and stimulus, the difference of these quantities converges to the mutual information. When the stimulus is deterministic or nonstationary the direct estimate of information no longer estimates mutual information, which is no longer meaningful, but it remains a measure of variability of the response distribution across time.

1 Introduction

Information estimates are used to characterize the amount of information that a spike train contains about a stimulus (Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998; Borst & Theunissen, 1999). They are motivated by information theory (Shannon, 1948) and widely believed to estimate the mutual information (or mutual information rate) between stimulus and spike train response. They are frequently calculated using data from experiments where the stimulus and response are dynamic and time varying

(Hsu, Woolley, Fremouw, & Theunissen, 2004; Reich, Mechler, & Victor, 2001; Reinagel & Reid, 2000; Nirenberg, Carcieri, Jacobs, & Latham, 2001).

For mutual information to be properly defined (see, e.g., Cover & Thomas, 1991), the stimulus and response must be considered random, and when the estimates are obtained from time averages, they should also be stationary and ergodic. In practice, these assumptions are usually tacit, and information estimates, such as the direct method proposed by Strong et al. (1998), can be made without explicit consideration of the stimulus. This can lead to misinterpretation.

The purpose of this note is to show that the direct method information estimate can be reinterpreted as the average divergence across time of the conditional response distribution from its overall mean; in the absence of stationarity and ergodicity, information estimates do not necessarily estimate mutual information, but potentially useful interpretations can still be made by referring back to the time-varying divergence. Although our results are specialized to the direct method with the plug-in entropy estimator, they should hold more generally regardless of the choice of entropy estimator. (See Victor (2006) for a recent review of existing entropy estimators.)

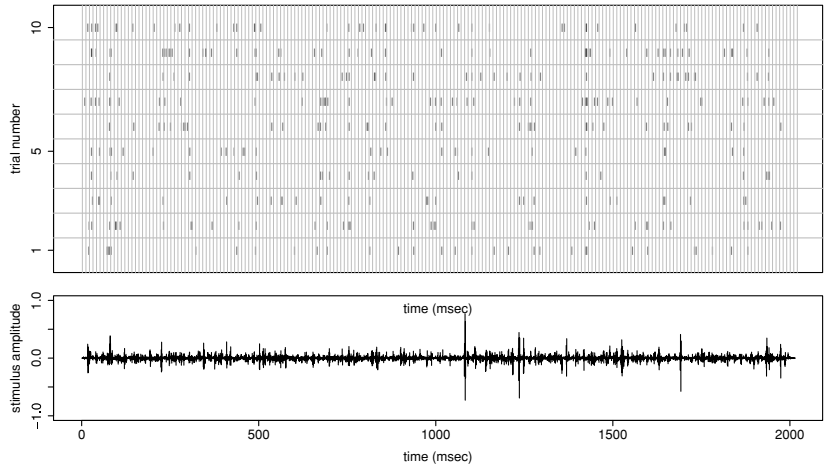
The fundamental issue concerns stationarity: methods that assume stationarity are unlikely to be appropriate when stationarity appears to be violated. In the nonstationary case, our second result should be of use, as would be other methods that explicitly consider the dynamic and nonstationary nature of the stimulus and response (see, e.g., Barbieri et al., 2004).

We begin with a brief review of the direct method and plug-in entropy estimator. This is followed by results showing that the information estimate can be recast as a time average. This characterization leads us to the interpretation that the information estimate is actually a measure of variability of the stimulus-conditioned response distribution. This observation is first made in the finite number of trials case and then formalized by a theorem describing the limiting behavior of the information estimate as the number of trials tends to infinity. Following the theorem is discussion about the interpretation of the limit and examples that illustrate the interpretation with a proposed graphical plot.

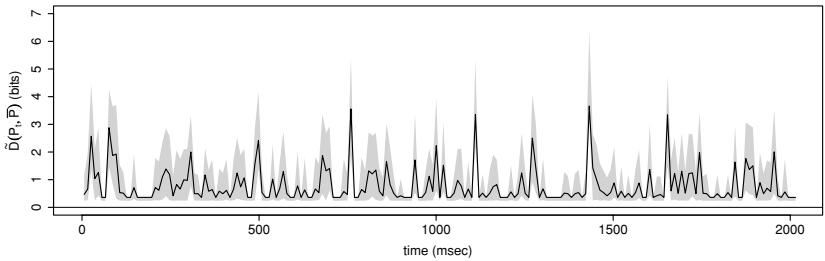
2 Review of the Direct Method

In the direct method, a time-varying stimulus is chosen by the experimenter and then repeatedly presented to a subject over multiple trials. The observed responses are conditioned by the same stimulus. Two types of variation in the response are considered: variation across time (potentially related to the stimulus) and trial-to-trial variation.

Figure 1a shows an example of data from such an experiment. The upper panel is a raster plot of the response of a field L neuron of an adult male zebra finch during synthetic song stimulation. The lower panel is a plot of the



(a) Stimulus and response



(b) Divergence plot

Figure 1: (a) Raster plot of the response of the a field L neuron of an adult male zebra finch (above) during the presentation of a synthetic audio stimulus (below) for 10 repeated trials. The vertical lines indicate boundaries of $L = 10$ millisecond (msec) words formed at a resolution of $dt = 1$ msec. The data consist of 10 trials, each of duration 2000 msec. (b) The coverage adjusted estimate (solid line) of $D(P_t, \bar{P})$ from the response shown above with 10 msec words. Pointwise 95% confidence intervals are indicated by the shaded region and obtained by bootstrapping the trials 1000 times. The information estimate, 0.77 bits (per 10 msec word, or 0.077 bits/msec), corresponds to the average value of the solid curve.

audio signal corresponding to the synthetic song. Details of the experiment can be found in Hsu et al. (2004).

Let us consider the random process $\{S_t, R_t^k\}$ representing the value of the stimulus and response at time $t = 1, \ldots, n$ during trial $k = 1, \ldots, m$. The response is made discrete by dividing time into bins of size dt and then considering words (or patterns) of spike counts formed within intervals

(overlapping or nonoverlapping) of L adjacent time bins. The number of spikes that occur in each time bin become the letters in the words. R_t^k corresponds to these words and may belong to a countably infinite set (because the number of spikes in a bin is theoretically unbounded). In the raster plot of Figure 1a, the time bin size is $dt = 1$ millisecond, and the vertical lines demarcate nonoverlapping words of length $L = 10$ time bins.

Given the responses $\{R_t^k\}$, the direct method considers two entropies:

1. The total entropy H of the response.
2. The local noise entropy H_t of the response at time t.

The total entropy is associated with the stimulus-conditioned distribution of the response across all times and trials. The local noise entropy is associated with the stimulus-conditioned distribution of the response at time t across all trials. These quantities are calculated directly from the neural response, and the difference between the total entropy and the average (over t) noise entropy is what Strong et al. (1998) call "the information that the spike train provides about the stimulus."

H and H_t depend implicitly on the length L of the words. Normalizing by L and considering large L leads to the total and local entropy rates that are defined to be $\lim_{L\to\infty} H(L)/L$ and $\lim_{L\to\infty} H_t(L)/L$, respectively, when they exist. The direct method of Strong et al. (1998) prescribed an extrapolation for estimating these limits; however, they do not necessarily exist when the stimulus and response process are nonstationary. When there is stationarity, estimation of entropy for large L is potentially difficult, and extrapolation from a few small choices of L can be suspect. Since we are primarily interested in the nonstationary case, we do not address these issues and refer readers to Kennel, Shlens, Abarbanel, and Chichilnisky (2005) and Gao, Kontoyiannis, and Bienenstock (2006) for a larger discussion on the stationary case. For notational simplicity, the dependence on L will be suppressed in the remainder of the text.

Strong et al. (1998) proposed estimating H and H_t by plug-in with the corresponding empirical distributions:

$$\hat{P}(r) := \frac{1}{mn} \sum_{t=1}^{n} \sum_{k=1}^{m} 1_{\{R_t^k = r\}} \tag{2.1}$$

and

$$\hat{P}_t(r) := \frac{1}{m} \sum_{k=1}^{m} 1_{\{R_t^k = r\}}. \tag{2.2}$$

Note that $\hat{P}$ is also the average of $\hat{P}_t$ across $t = 1, \ldots, n$. So the direct method plug-in estimates[1] of H and H_t are

$$\hat{H} := -\sum_r \hat{P}(r) \log \hat{P}(r) \tag{2.3}$$

and

$$\hat{H}_t := -\sum_r \hat{P}_t(r) \log \hat{P}_t(r), \tag{2.4}$$

respectively. The direct method plug-in information estimate is

$$\hat{I} := \hat{H} - \frac{1}{n} \sum_{t=1}^{n} \hat{H}_t. \tag{2.5}$$

3 Results

The direct method information estimate is not only the difference of entropies shown in equation 2.5, but also a time average of divergences. The empirical distribution of response across all trials and times (see equation 2.1) is equal to the average of $\hat{P}_t$ over time. That is, $\hat{P}(r) = n^{-1} \sum_{t=1}^{n} \hat{P}_t(r)$ and so

$$\hat{I} = \hat{H} - \frac{1}{n} \sum_{t=1}^{n} \hat{H}_t \tag{3.1}$$

$$= \frac{1}{n} \sum_{t=1}^{n} \sum_r \hat{P}_t(r) \log \hat{P}_t(r) - \sum_r \left[\frac{1}{n} \sum_{t=1}^{n} \hat{P}_t(r) \right] \log \hat{P}(r) \tag{3.2}$$

$$= \frac{1}{n} \sum_{t=1}^{n} \sum_r \hat{P}_t(r) \log \hat{P}_t(r) - \frac{1}{n} \sum_{t=1}^{n} \sum_r \hat{P}_t(r) \log \hat{P}(r) \tag{3.3}$$

$$= \frac{1}{n} \sum_{t=1}^{n} \sum_r \hat{P}_t(r) \log \frac{\hat{P}_t(r)}{\hat{P}(r)}. \tag{3.4}$$

The quantity that is averaged over time in equation 3.4 is the Kullback-Leibler divergence between the empirical time t response distribution $\hat{P}_t$ and the average empirical response distribution $\hat{P}$.

Since the same stimulus is repeatedly presented to the subject and there is no evolution in the response, over multiple trials, the following repeated

[1] Strong et al. (1998) used the name *naive estimates*.

trial assumption is natural: Conditional on the stimulus $\{S_t\}$, the m trials $\{S_t, R_t^1\}, \ldots, \{S_t, R_t^m\}$ are independent and identically distributed (i.i.d.). Under this assumption, $1_{\{R_t^1=r\}}, \ldots, 1_{\{R_t^m=r\}}$ are conditionally i.i.d. for each fixed t and r. Furthermore, the law of large numbers guarantees that as the number of trials m increases, the empirical response distribution $\hat{P}_t(r)$ converges to its conditional expected value for each fixed t and r. Thus, $\hat{P}_t(r)$ and $\hat{P}(r)$ can be viewed as estimates of $P_t(r \mid S_1, \ldots, S_n)$, defined by

$$P_t(r \mid S_1, \ldots, S_n) := P(R_t^k = r \mid S_1, \ldots, S_n) = E\{\hat{P}_t(r) \mid S_1, \ldots, S_n\}, \quad (3.5)$$

and $\bar{P}(r \mid S_1, \ldots, S_n)$, defined by

$$\bar{P}(r \mid S_1, \ldots, S_n) := \frac{1}{n} \sum_{t=1}^{n} P_t(r \mid S_1, \ldots, S_n), \quad (3.6)$$

respectively. $\bar{P}$ is the average response distribution across time $t = 1, \ldots, n$ conditional on the entire stimulus $\{S_1, \ldots, S_n\}$.

The quantity that is averaged over time in equation 3.4 should be viewed as a plug-in estimate of the Kullback-Leibler divergence between P_t and $\bar{P}$. We emphasize this by writing

$$\hat{D}(P_t \| \bar{P}) := \sum_r \hat{P}_t(r) \log \frac{\hat{P}_t(r)}{\hat{P}(r)}. \quad (3.7)$$

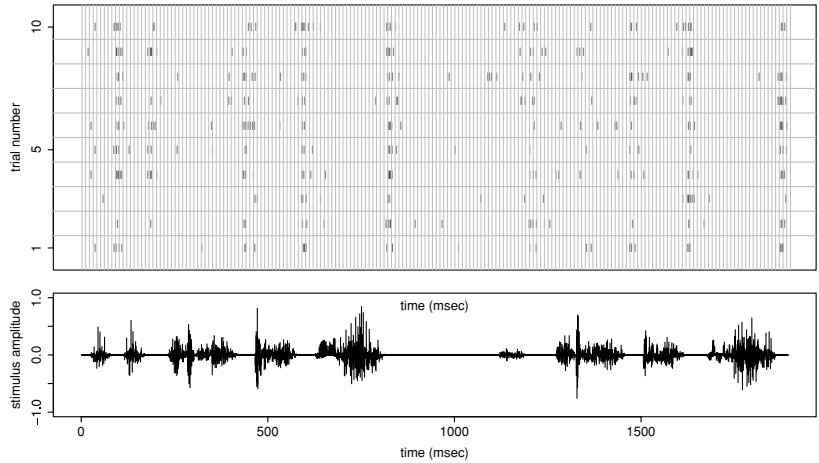
This observation will be formalized by the theorem in the next section. For now, we summarize the above with a proposition:

Proposition 1. *The information estimate is the time-average* $\hat{I} = \frac{1}{n} \sum_{t=1}^{n} \hat{D}(P_t \| \bar{P})$.

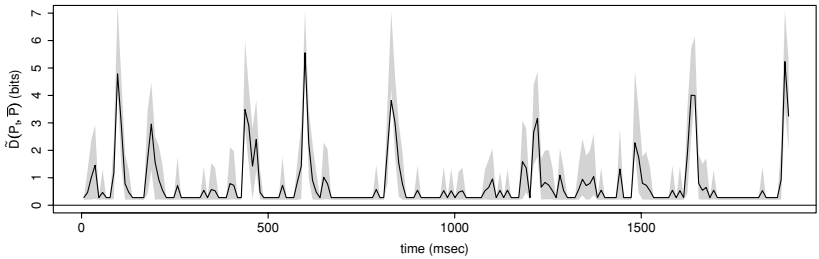
This decomposition of the information estimate is analogous to the decomposition of mutual information that Deweese and Meister (1999) call the "specific surprise," while "specific information" is analogous to the alternative decomposition,

$$\hat{I} = \frac{1}{n} \sum_{t=1}^{n} [\hat{H} - \hat{H}_t]. \quad (3.8)$$

An important difference is that here, the stimulus itself is a function of time, and the decompositions are given in terms of time-dependent quantities. It is possible that these quantities can reveal dynamic aspects of the stimulus and response relationship. This will be explored further in sections 3.2 and 3.3.



(a) Stimulus and response



(b) Divergence plot

Figure 2: (a) Same as in Figure 1, but in this set of trials, the stimulus is a conspecific natural song. (b) The coverage adjusted estimate (solid line) of $D(P_t, \bar{P})$ from the response shown above. Pointwise 95% confidence intervals are indicated by the shaded region and obtained by bootstrapping the trials 1000 times. The information estimate, 0.76 bits (per 10 msec word or 0.076 bits/msec), corresponds to the average value of the solid curve.

3.1 What Is Being Estimated? There are two directions in which the number of observed response data can be increased: length of time n and number of trials m. The information estimate is the average of $\hat{D}(P_t \| \bar{P})$ over time and may not necessarily converge as n increases. This could be due to $\{S_t, R_t^k\}$ being nonstationary or highly dependent in time, or both. Even when convergence may occur, the presence of serial correlation in $\hat{D}(P_t \| \bar{P})$ (see the autocorrelation in Figure 2b, for example) can make assessments of uncertainty in $\hat{I}$ difficult.

Assuming that the stimulus and response process is stationary and not too dependent in time could guarantee convergence, but this could be

unrealistic. On the other hand, the repeated trial assumption is appropriate if the same stimulus is repeatedly presented to the subject over multiple trials. It is also enough to guarantee that the information estimate converges as the number of trials m increases. We prove the following theorem in the appendix:

Theorem 1. *Suppose that P_t has finite entropy for all $t = 1, \ldots, n$. Then under the repeated trial assumption,*

$$\lim_{m\to\infty} \hat{I} = H(\bar{P}) - \frac{1}{n}\sum_{t=1}^{n} H(P_t) = \frac{1}{n}\sum_{t=1}^{n}[H(\bar{P}) - H(P_t)] = \frac{1}{n}\sum_{t=1}^{n} D(P_t \| \bar{P})$$

with probability 1, and in particular the following statements hold uniformly for $t = 1, \ldots, n$ with probability 1:

1. *$\lim_{m\to\infty} \hat{H} = H(\bar{P})$,*
2. *$\lim_{m\to\infty} \hat{H}_t = H(P_t)$, and*
3. *$\lim_{m\to\infty} \hat{D}(P_t \| \bar{P}) = D(P_t \| \bar{P})$ for $t = 1, \ldots, n$,*

where $D(P_t \| \bar{P})$ is the Kullback-Leibler divergence defined by

$$D(P_t \| \bar{P}) := \sum_r P_t(r \mid S_1, \ldots, S_n) \log \frac{P_t(r \mid S_1, \ldots, S_n)}{\bar{P}(r \mid S_1, \ldots, S_n)},$$

and $H(P)$ is the entropy of the distribution P, defined by

$$H(P) := -\sum_r P(r) \log P(r).$$

Note that if stationary and ergodicity do hold, then P_t for $t = 1, \ldots, n$ is also stationary and ergodic.[2] So its average, $\bar{P}(r)$, is guaranteed by the ergodic theorem to converge pointwise to $P(R_1^1 = r)$ as $n \to \infty$. Moreover, if R_1^1 can take on only a finite number of values, then $H(\bar{P})$ also converges to the marginal entropy $H(R_1^1)$ of R_1^1. Likewise, the average of the conditional entropy $H(P_t)$ also converges to the expected conditional entropy: $\lim_{n\to\infty} H(R_n^1 \mid S_1, \ldots, S_n)$. So in this case, the information estimate does indeed estimate mutual information.

However, the primary consequence of the theorem is that in the absence of stationarity and ergodicity, the information estimate $\hat{I}$ does not necessarily estimate mutual information. The three particular statements show that the time-varying quantities $[\hat{H} - \hat{H}_t]$ and $\hat{D}(P_t \| \bar{P})$ converge individually to

[2] P_t and $\bar{P}$ are stimulus conditional distributions, and hence random variables potentially depending on $S_1, \ldots, S_n$.

the appropriate limits and justify our assertion that the information estimate is a time average of plug-in estimates of the corresponding time-varying quantities. Thus, the information estimate can always be viewed as an estimate of the time average of either $D(P_t \| \bar{P})$ or $[H(P) - H(P_t)]$—stationary and ergodic or not.

3.2 The Information Estimate Measures Variability of the Response Distribution. The Kullback-Leibler divergence $D(P_t \| \bar{P})$ has a simple interpretation: it measures the dissimilarity of the time t response distribution P_t from its overall average $\bar{P}$. So as a function of time, $D(P_t \| \bar{P})$ measures how the conditional response distribution varies across time relative to its overall mean. This can be seen in a more familiar form by considering the leading term of the Taylor expansion,

$$D(P_t \| \bar{P}) = \frac{1}{2} \sum_r \frac{[P_t(r \mid S_1, \ldots, S_n) - \bar{P}(r \mid S_1, \ldots, S_n)]^2}{\bar{P}(r \mid S_1, \ldots, S_n)} + \cdots. \quad (3.9)$$

Thus, its average is in this sense a measure of the average variability of the response distribution.

It is, of course, possible that characteristics of the response are due to confounding factors rather than the stimulus. Furthermore, the presence of additional noise in either process would weaken a measured relationship between stimulus and response, compared to its strength if the noise were eliminated. Setting these concerns aside, the variation of the response distribution P_t about its average provides information about the relationship between the stimulus and the response. In the stationary and ergodic case, this information may be averaged across time to obtain mutual information. In more general settings, averaging across time may not provide a complete picture of the relationship between stimulus and response. Instead, we suggest examining the time-varying $D(P_t \| \bar{P})$ directly, via graphical display as discussed next.

3.3 Plotting the Divergence. The plug-in estimate $\hat{D}(P_t \| \bar{P})$ is an obvious choice for estimating $D(P_t \| \bar{P})$, but it turns out that estimating $D(P_t \| \bar{P})$ is akin to estimating entropy. Since the trials are conditionally i.i.d., the coverage adjustment method described in Vu, Yu, and Kass (2007) can be used to improve estimation of $D(P_t \| \bar{P})$ over the plug-in estimate. The appendix contains the details of this.

Figures 1 and 2 show the responses of the same field L neuron of an adult male zebra finch under two stimulus conditions. Details of the experiment and the statistics of the stimuli are described in Hsu et al. (2004). Figures 1a and 2a show the stimulus and response data. In Figure 1 the stimulus is synthetic and stationary by construction, while in Figure 2 the stimulus is a natural song. Figures 2a and 2b show the coverage adjusted estimate of

the divergence $D(P_t \| \bar{P})$ plotted as a function of time. Ninety-five percent confidence intervals were formed by bootstrapping entire trials; that is, an entire trial is either included in or excluded from a bootstrap sample.

The information estimate going along with each divergence plot is the average of the solid curve representing the estimate of $D(P_t \| \bar{P})$. It is equal to 0.77 bits (per 10 millisecond word) in Figure 1b and 0.76 bits (per 10 millisecond word) in Figure 2b. Although the information estimates are nearly identical, the two plots are very different.

In the first case, the stimulus is stationary by construction, and it appears that the time-varying divergence is too. Its fluctuations appear to be roughly of the same scale across time, and its local mean is relatively stable. The average of the solid curve seems to be a fair summary.

In the second case, the stimulus is a natural song. The isolated bursts of the time-varying divergence and relatively flat regions in Figure 2b suggest that the response process (and the divergence) is nonstationary and has strong serial correlations. The local mean of the divergence also varies strongly with time. Summarizing $D(P_t \| \bar{P})$ by its time-average hides the time-dependent features of the plot.

More interestingly, when the divergence plot is compared to the plot of the stimulus in Figure 2, there is a striking coincidence between the location of large isolated values of the estimated divergence and visual features of the stimulus waveform. They tend to coincide with the boundaries of the bursts in the stimulus signal. This suggests that the spike train may carry information about the onset and offset of bursts in the stimulus. We discussed this with the Theunissen Lab, and they confirmed from their spatiotemporal receptive field models that the cell in the example is an offset cell. It tends to fire at the offsets of song syllables—the bursts of energy in the stimulus waveform. They also suggested that a word length within the range of 30 to 50 milliseconds is a better match to the length of correlations in the auditory system. We regenerated the plots for words of length $L = 40$ (not shown here) and found that the isolated structures in the divergence plot became even more pronounced.

4 Discussion

Estimates of mutual information, including the plug-in estimate, may be viewed as measures of the strength of the relationship between the response and the stimulus when the stimulus and response are jointly stationary and ergodic. Many applications, however, use nonstationary or even deterministic stimuli, so that mutual information is no longer well defined. In such nonstationary cases, do estimates of mutual information become meaningless? We think not, but the purpose of this note has been to point out the delicacy of the situation and to suggest a viable interpretation of information estimates, along with the divergence plot, in the nonstationary case.

In using stochastic processes to analyze data, there is an implicit practical acknowledgment that assumptions cannot be met precisely: the mathematical formalism is, after all, an abstraction imposed on the data. The hope is simply that the variability displayed by the data is similar in relevant respects to that displayed by the presumptive stochastic process. "Relevant respects" involve the statistical properties deduced from the stochastic assumptions. The point we are trying to make is that highly nonstationary stimuli make statistical properties based on an assumption of stationarity highly suspect; strictly speaking, they become void.

To be more concrete, let us reconsider the snippet of natural song and response displayed in Figure 2. When we look at the less than 2 seconds of stimulus amplitude given there, the stimulus is not at all time invariant: instead, it has a series of well-defined bursts followed by periods of quiescence. Perhaps, on a very much longer timescale, the stimulus would look stationary. But a good stochastic model on a long timescale would likely require long-range dependence. Indeed, it can be difficult to distinguish nonstationarity from long-range dependence (Kunsch, 1986), and the usual statistical properties of estimators are known to break down when long-range dependence is present (Beran, 1994). Given a short interval of data, valid statistical inference under stationarity assumptions becomes highly problematic. To avoid these problems, we have proposed the use of the divergence plot and a recognition that the "bits per second" summary is no longer mutual information in the usual sense. Instead, we would say that the estimate of information measures magnitude of variation of the response as the stimulus varies and that this is a useful assessment of the extent to which the stimulus affects the response as long as other factors that affect the response are not confounded with the stimulus. In other deterministic or nonstationary settings, the argument for the relevance of an information estimate should be analogous. Under stationarity and ergodicity, and indefinitely many trials, the stimulus sets that affect the response—whatever they are—will be repeatedly sampled, with appropriate probability, to determine the variability in the response distribution, with time invariance in the response being guaranteed by the joint stationarity condition. This becomes part of the intuition behind mutual information. In the deterministic or nonstationary settings, information estimates do not estimate mutual information, but they may remain intuitive assessments of strength of effect.

Appendix

A.1 Coverage Adjusted Estimate of $D(P_t \| \bar{P})$. The main idea behind coverage adjustment is to adjust estimates for potentially unobserved values. This happens in two places: estimation of P_t and estimation of $D(P_t \| \bar{P})$. In the first case, unobserved values affect the amount of weight that $\hat{P}_t$, defined in equation 2.2, places on observed values. In the second case, unobserved values correspond to missing summands when plugging $\hat{P}_t$

into the Kullback-Leibler divergence. Vu et al. (2007) give a more thorough explanation of these ideas. Let

$$N_t(r) := \sum_{k=1}^{m} 1_{\{R_t^k = r\}}. \tag{A.1}$$

The sample coverage, or total P_t-probability of observed values r, is estimated by $\hat{C}_t$ defined by

$$\hat{C}_t := 1 - \frac{\#\{r : N_t(r) = 1\} + .5}{m + 1}. \tag{A.2}$$

The number in the numerator of the fraction refers to the number of singletons—patterns that were observed only once across the m trials at time t. Then the coverage-adjusted estimate of P_t is the following shrunken version of $\hat{P}_t$:

$$\tilde{P}_t(r) = \hat{C}_t \hat{P}_t(r). \tag{A.3}$$

$\bar{P}$ is estimated by simply averaging $\tilde{P}_t$:

$$\tilde{P}(r) = \frac{1}{n} \sum_{t=1}^{n} \tilde{P}_t(r). \tag{A.4}$$

The coverage-adjusted estimate of $D(P_t \| \bar{P})$ is obtained by plugging $\tilde{P}_t$ and $\tilde{P}$ into the Kullback-Leibler divergence, but with an additional weighting on the summands according to the inverse of the estimated probability that the summand is observed:

$$\tilde{D}(P_t \| \bar{P}) := \sum_{r} \frac{\tilde{P}_t(r)\{\log \tilde{P}_t(r) - \log \tilde{P}(r)\}}{1 - (1 - \tilde{P}_t(r))^m}. \tag{A.5}$$

The additional weighting is to correct for potentially missing summands. (This is also explained in detail in Vu et al., 2007.) Confidence intervals for $D(P_t \| \bar{P})$ can be obtained by bootstrap-sampling entire trials and applying $\tilde{D}$ to the bootstrap replicate data.

A.2 Proofs. We use the following extension of the Lebesgue dominated convergence theorem in the proof of theorem 1.

Lemma 1. *Let f_m and g_m for $m = 1, 2, \ldots$ be sequences of measurable, integrable functions defined on a measure space equipped with measure μ, and*

with point-wise limits f and g, respectively. Suppose further that $|f_m| \leq g_m$ and $\lim_{m\to\infty} \int g_m d\mu = \int g \, d\mu < \infty$. Then

$$\lim_{m\to\infty} \int f_m \, d\mu = \int \lim_{m\to\infty} f_m \, d\mu.$$

Proof. By linearity of the integral,

$$\begin{aligned}&\liminf_{n\to\infty} \int (g + g_m) \, d\mu - \limsup_{n\to\infty} \int |f - f_m| \, d\mu \\ &\quad = \liminf_{n\to\infty} \int (g + g_m) - |f - f_m| \, d\mu.\end{aligned}$$

Since $0 \leq (g + g_m) - |f - f_m|$, Fatou's lemma implies

$$\liminf_{n\to\infty} \int (g + g_m) - |f - f_m| d\mu \geq \int \liminf_{n\to\infty} (g + g_m) - |f - f_m| \, d\mu.$$

The limit inferior on the inside of the right-hand integral is equal to $2g$ by assumption. Combining with the previous two displays and the assumption that $\int g_m \, d\mu \to \int g \, d\mu$ gives

$$\limsup_{n\to\infty} \left| \int f d\mu - \int f_m \, d\mu \right| \leq \limsup_{n\to\infty} \int |f - f_m| \, d \leq 0.$$

Proof of Theorem 1. The main statement of the theorem is implied by the three numbered statements together with proposition 1. We start with the second numbered statement. Under the repeated trial assumption, $R_t^1, \ldots, R_t^m$ are conditionally i.i.d. given the stimulus $\{S_t\}$. So corollary 1 of Antos and Kontoyiannis (2001) can be applied to show that

$$\lim_{m\to\infty} \hat{H}_t = \lim_{m\to\infty} -\sum_r \hat{P}_t(r) \log \hat{P}_t(r) \tag{A.6}$$

$$= -\sum_r P_t(r \mid S_1, \ldots, S_n) \log P_t(r \mid S_1, \ldots, S_n) \tag{A.7}$$

$$= H(P_t) \tag{A.8}$$

with probability 1. This proves the second numbered statement.

We use lemma 1 to prove the first numbered statement. For each r, the law of large numbers asserts $\lim_{m\to\infty} \hat{P}_t(r) = P_t(r \mid S_1, \ldots, S_n)$ with probability 1. So for each r,

$$\lim_{m\to\infty} -\hat{P}_t(r) \log \hat{P}(r) = -P_t(r \mid S_1, \ldots, S_n) \log \bar{P}(r \mid S_1, \ldots, S_n) \tag{A.9}$$

and

$$\lim_{m\to\infty} -\hat{P}_t(r)\log\hat{P}_t(r) = -P_t(r \mid S_1,\ldots,S_n)\log P_t(r \mid S_1,\ldots,S_n) \quad \text{(A.10)}$$

with probability 1. Fix a realization where equations A.6 to A.10 hold, and let

$$f_m(r) := -\hat{P}_t(r)\log\hat{P}(r)$$

and

$$g_m(r) := -\hat{P}_t(r)[\log\hat{P}_t(r) - \log n].$$

Then for each r,

$$\lim_{m\to\infty} f_m(r) = -P_t(r \mid S_1,\ldots,S_n)\log\bar{P}(r \mid S_1,\ldots,S_n) =: f(r)$$

and

$$\lim_{m\to\infty} g_m(r) = -P_t(r)[\log P_t(r) - \log n] =: g(r).$$

The sequence f_m is dominated by g_m because

$$0 \le -\hat{P}_t(r)\log\hat{P}(r) = f_m(r) \quad \text{(A.11)}$$

$$= -\hat{P}_t(r)\left[\log\sum_{u=1}^{n}\hat{P}_u(r) - \log n\right] \quad \text{(A.12)}$$

$$\le -\hat{P}_t(r)[\log\hat{P}_t(r) - \log n] \quad \text{(A.13)}$$

$$= g_m(r) \quad \text{(A.14)}$$

for all r, where equation A.13 uses the fact that $\log x$ is an increasing function. From equation A.6, we also have that $\lim_{m\to\infty}\sum_r g_m(r) = \sum_r g(r)$. Clearly, f_m and g_m are summable. Moreover, $H(P_t) < \infty$ by assumption. So

$$\sum_r g(r) = \sum_r -P_t(r)\log P_t(r) + \log n\sum_r P_t(r) = H(P_t) + \log n < \infty, \quad \text{(A.15)}$$

and the conditions of lemma 1 are satisfied. Thus,

$$\lim_{m\to\infty}\sum_r -\hat{P}_t(r)\log\hat{P}(r)$$

$$= \lim_{m\to\infty}\sum_r f_m(r) = \sum_r f(r) = \sum_r -P_t(r)\log\bar{P}(r). \quad \text{(A.16)}$$

Averaging over $t = 1, \ldots, n$ gives

$$\hat{H} = \lim_{m\to\infty} \sum_r -\hat{P}(r) \log \hat{P}(r) = \sum_r -\bar{P}(r) \log \bar{P}(r) = H(\bar{P}) \tag{A.17}$$

for realizations where equations A.6 to A.10 hold. This proves the first numbered statement because the probability of all such realizations is 1.

For the third numbered statement, we begin with the expansions

$$\hat{D}(P_t \| \bar{P}) = \sum_r \hat{P}_t(r) \log \hat{P}_t(r) - \hat{P}_t(r) \log \hat{P}(r) \tag{A.18}$$

and

$$D(P_t \| \bar{P}) = \sum_r P_t(r) \log P_t(r) - P_t(r) \log \bar{P}(r). \tag{A.19}$$

The second numbered statement and equation A.16 imply

$$\begin{aligned} &\lim_{m\to\infty} \sum_r \hat{P}_t(r) \log \hat{P}_t(r) - \hat{P}_t(r) \log \hat{P}(r) \\ &\quad = \sum_r P_t(r) \log P_t(r) - \sum_r P_t(r) \log \bar{P}(r) \end{aligned} \tag{A.20}$$

with probability 1. This proves the third numbered statement.

Acknowledgments

We thank the Theunissen Lab at the University of California, Berkeley, for providing the data set and helpful discussion. We also thank an anonymous reviewer for comments that greatly improved the manuscript. V.Q.V. was supported by an NSF VIGRE Graduate Fellowship and NIDCD grant DC 007293. B.Y. was supported by NSF grants DMS-03036508, DMS-0605165, and 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 R.E.K. 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.

References

Antos, A., & Kontoyiannis, I. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, *19*, 163–193.

Barbieri, R., Frank, L. M., Nguyen, D. P., Quirk, M. C., Solo, V., Wilson, M. A. et al. (2004). Dynamic analyses of information encoding in neural ensembles. *Neural Computation*, *16*(2), 277–307.

Beran, J. (1994). *Statistics for long-memory processes*. London: Chapman & Hall.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature Neuroscience*, *2*(11), 947–957.

Cover, T., & Thomas, J.(1991). *Elements of information theory*. New York: Wiley.

Deweese, M., & Meister, M. (1999, January). How to measure the information gained from one symbol. *Network: Computation in Neural Systems*.

Gao, Y., Kontoyiannis, I., & Bienenstock, E. (2006, July). From the entropy to the statistical structure of spike trains. In *Proceedings of the 2006 IEEE International Symposium on Information Theory* (pp. 645–649). New York: IEEE.

Hsu, A., Woolley, S. M. N., Fremouw, T. E., & Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *J. Neuro.*, *24*(41), 9201–9211.

Kennel, M. B., Shlens, J., Abarbanel, H. D. I., & Chichilnisky, E. J. (2005). Estimating entropy rates with Bayesian confidence intervals. *Neural Computation*, *17*(7), 1531–1576.

Kunsch, H. (1986). Discrimination between monotonic trends and long-range dependence. *Journal of Applied Probability*, *23*(4), 1025–1030.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., & Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, *411*(6838), 698–701.

Reich, D. S., Mechler, F., & Victor, J. D. (2001). Formal and attribute-specific information in primary visual cortex. *Journal of Neurophysiology*, *85*(1), 305–318.

Reinagel, P., & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *Journal of Neuroscience*, *20*(14), 5392–5400.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters*, *80*(1), 197–200.

Victor, J. D. (2006). Approaches to information-theoretic analysis of neural activity. *Biological Theory*, *1*, 302–316.

Vu, V. Q., Yu, B., & Kass, R. E. (2007). Coverage adjusted entropy estimation. *Statistics in Medicine*, *26*(21), 4039–4060.

Received January 30, 2008; accepted July 27, 2008.