LETTER Communicated by Jonathan Victor

Tight Data-Robust Bounds to Mutual Information Combining Shuffling and Model Selection Techniques

M. A. Montemurro
m.montemurro@manchester.ac.uk
R. Senatore
riccardo.senatore@postgrad.manchester.ac.uk
S. Panzeri
s.panzeri@manchester.ac.uk
University of Manchester, Faculty of Life Sciences, Manchester M60 1QD, U.K.

The estimation of the information carried by spike times is crucial for a quantitative understanding of brain function, but it is difficult because of an upward bias due to limited experimental sampling. We present new progress, based on two basic insights, on reducing the bias problem. First, we show that by means of a careful application of data-shuffling techniques, it is possible to cancel almost entirely the bias of the noise entropy, the most biased part of information. This procedure provides a new information estimator that is much less biased than the standard direct one and has similar variance. Second, we use a nonparametric test to determine whether all the information encoded by the spike train can be decoded assuming a low-dimensional response model. If this is the case, the complexity of response space can be fully captured by a small number of easily sampled parameters. Combining these two different procedures, we obtain a new class of precise estimators of information quantities, which can provide data-robust upper and lower bounds to the mutual information. These bounds are tight even when the number of trials per stimulus available is one order of magnitude smaller than the number of possible responses. The effectiveness and the usefulness of the methods are tested through applications to simulated data and recordings from somatosensory cortex. This application shows that even in the presence of strong correlations, our methods constrain precisely the amount of information encoded by real spike trains recorded in vivo.

1 Introduction

A recent fundamental insight from system neuroscience is that information about the external sensory world is often encoded by the precise timing of action potential (spikes). The timing of individual spikes has been shown to encode precisely and reliably the occurrence of certain stimulus features (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1996;

de Ruyter van Steveninck, Lewen, Strong, Koberle, & Bialek, 1997; Buracas, Zador, DeWeese, & Albright, 1998; Panzeri, Petersen, Schultz, Lebedev, & Diamond, 2001; DeWeese, Wehr, & Zador, 2003; Arabzadeh, Zorzin, & Diamond, 2005). However, one question still unanswered is how individual spikes, from either the same or a different neuron, combine together to give rise to perception. One fundamental observation is that spike times are correlated: for example, nearby cortical cells tend to fire in synchrony more than expected by chance. The presence of correlations has suggested that they are a fundamental ingredient of the neural code. The computational advantage of such a representation may be that correlations add an information channel that can be used to either represent more sensory and behavioral features (Abeles, Bergman, Margalit, & Vaadia, 1993; Dan, Alonso, Usrey, & Reid, 1998) or bind together groups of features (Gray, König, Engel, & Singer, 1989; von der Malsburg, 1999). However, whether correlations are a crucial part of the neural code is still highly controversial (Shadlen & Movshon, 1999).

A principled and rigorous way to address how the messages carried by individual spike times are integrated together is to use information theory to quantify and compare different ways and timescales at which spike times may convey information (Rieke et al., 1996; Borst & Theunissen, 1999; Dimitrov & Miller, 2001; Panzeri et al., 2001). The use of information theory allows an estimate of how reliably stimuli are encoded in single trials and which features of the neuronal response, such as independent spikes or the correlations, contribute to stimulus discriminability.

A problem with this approach is that quantifying reliably the information conveyed by spike timing often requires the collection of unpractically large samples of data. This is mainly due to correlations: if correlations did not exist, then the statistics of spike times would be completely characterized by the time-dependent firing rate of each neuron. However, one also needs to measure the correlations among all possible groups of spikes. A complete characterization of these correlations requires a number of parameters that are difficult to sample with realistic amounts of neuronal data. Thus, spike timing information measures suffer from a significant sampling bias problem (Panzeri & Treves, 1996).

Generalizing previous work of Reich and colleagues (Reich, Mechler, Purpura, & Victor, 2000), we have recently proposed an approach to alleviate the sampling bias problem by developing data-robust lower bounds to the spike timing information that neglect long-lag stimulus modulations of correlations (Pola, Petersen, Thiele, Young, & Panzeri, 2005). These bounds can establish if there is information conveyed in spike times above and beyond that conveyed by spike counts. However, these methods cannot be used to test the importance of correlations in coding because they explicitly neglect a potentially important part of the correlation structure.

In this letter, we overcome this limitation by presenting several methodological advances that lead to a radical improvement of the sampling

properties of information measures and provide very tight and data-robust upper and lower bounds to spike timing information. These advances also permit constraining precisely the role of correlations in decoding. Overall, this progress provides the basis for a better determination of the role of correlations in information transmission and at the same time significantly expands the domain of applicability of information-theoretic techniques to the analysis of neural signals.

The letter is organized as follows. We first review basic concepts of information theory applications to spike trains; we then discuss how to quantify the importance of correlations in decoding; we next address the sampling properties of these information quantities and provide bounds that are biased either upward or downward; and we discuss how to use model selection techniques to give virtually unbiased and tight estimations of information. Finally, we apply the new techniques to real neuronal spike trains recorded from rat somatosensory cortex.

2 The Information Carried by Neuronal Population Responses

We consider a time period of duration T, associated with a dynamic or static sensory stimulus s (chosen with probability $P(s)$ from a stimulus set $\mathcal{S}$ with S elements), during which the activity of one neuron is observed.[1] We assume that the spike arrival times are binned with a timing precision Δt and transformed into a sequence of spike counts in each time bin. L denotes the number of time bins (i.e., $T = L\Delta t$). The neuronal response is denoted by a one-dimensional array $\mathbf{r} = \{r(1), r(2), \ldots, r(L)\}$, where $r(t)$ is the number of spikes emitted by the neuron in the tth time bin. The maximum number of spikes that can be observed in a single time bin in any trial is denoted by M. (If Δt is very short, M is 1 and $r(t)$ is binary.) We indicate the response space by $\mathcal{R}$ ($\mathcal{R}$ contains $(M+1)^L$ elements).

Following Shannon (1948), we write the mutual information $I(\mathcal{R}; \mathcal{S})$ (often abbreviated as I in the following) transmitted by the population response about the whole set of stimuli as

$$I(\mathcal{R}; \mathcal{S}) = H(\mathcal{R}) - H(\mathcal{R}|\mathcal{S}), \tag{2.1}$$

where $H(\mathcal{R})$ and $H(\mathcal{R}|\mathcal{S})$ are the response entropy (stimulus-unconditional) and the noise entropy (stimulus-conditional), respectively. They are defined

[1] In this letter, we consider one neuron only in order to keep notations simple. However, the generalization to neuronal populations is relatively straightforward and does not present conceptual difficulties. The main step in generalizing to populations is a slight change needed in the definitions of the Markov models described below; see Pola et al. (2005) for an example of how to carry out this generalization.

(Cover & Thomas, 1991) as

$$H(\mathcal{R}) = -\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}) \log_2 P(\mathbf{r}), \tag{2.2}$$

$$H(\mathcal{R}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}|s) \log_2 P(\mathbf{r}|s). \tag{2.3}$$

The response entropy quantifies how neuronal responses vary with the stimulus and thus sets the capacity of the spike train to convey information. The noise entropy quantifies the irreproducibility of the neuronal responses at fixed stimulus. Thus, mutual information quantifies how much of the information capacity provided by stimulus-evoked differences in neural activity is robust to the presence of trial-by-trial response variability (de Ruyter van Steveninck et al., 1997). In equations 2.2 and 2.3, the summation over $\mathbf{r}$ is over all possible neuronal responses. The summation over s is over all possible stimuli. $P(\mathbf{r}|s)$ is the probability of observing a particular response $\mathbf{r}$ conditional to stimulus s. Experimentally, $P(\mathbf{r}|s)$ is determined by repeating each stimulus on many trials while recording the neuronal responses. The probability $P(s)$ is usually chosen by the experimenter. $P(\mathbf{r}) = \langle P(\mathbf{r}|s)\rangle_s$ is its average across all stimuli (the angular brackets indicate an average over stimuli, $\langle F(s)\rangle_s \equiv \sum_{s\in\mathcal{S}} P(s)F(s)$). We assume that there are enough stimuli in the presented set so that $P(\mathbf{r})$ (which is computed across all trials to all stimuli) is better sampled than $P(\mathbf{r}|s)$. (In practice, this amounts to the requirement that more than a handful of stimuli is presented.)

Estimating the information carried by spike times of real neuronal populations is difficult because each stimulus-response probability has to be measured from a limited number of data. The statistical errors in estimating the response probabilities lead to a downward systematic error (bias) in both noise and response entropy (Miller, 1955). $H(\mathcal{R})$ depends on only $P(\mathbf{r})$, which is sampled across all trials to all stimuli. Under our assumptions, its bias is much smaller than that of $H(\mathcal{R}|\mathcal{S})$, which depends on $P(\mathbf{r}|s)$. This results in an overall upward bias when estimating mutual information (Panzeri & Treves, 1996). This makes it difficult to estimate the information directly from equation 2.1, especially for long time windows or precise spike time discretizations (large L).

3 Simplified Models of Correlation

Having defined the information that neuronal responses transmit about sensory stimuli, we consider how correlations in the responses affect information transmission.

The first step is to define precisely what we mean by correlations. In this letter, when we say that the spike trains are correlated, we mean that, for

some stimulus s, the "true" stimulus-response probability $P(\mathbf{r}|s)$ is different from the probability $P_{ind}(\mathbf{r}|s)$ obtained if spikes were independent at fixed stimulus. By definition, the independent probability model $P_{ind}(\mathbf{r}|s)$ is the product of the stimulus-conditional marginal probabilities $P(r(t)|s)$ of responses in each time bin t:

$$P_{ind}(\mathbf{r}|s) = \prod_{t=1}^{L} P(r(t)|s), \tag{3.1}$$

Thus, when we refer to correlations, we mean correlations at fixed stimulus. These correlations are usually called noise correlations (Gawne & Richmond, 1993; Nirenberg & Latham, 2003; Pola, Thiele, Hoffmann, & Panzeri, 2003). For brevity, in the rest of this letter, when we use the term *correlation*, we mean "noise correlation."

After correlations have been defined, the next step is to characterize how they affect information transmission. Correlations can affect neural information transmission in different ways in terms of both encoding and downstream decoding of neuronal messages. Here, following previous work (Latham & Nirenberg, 2005; Nirenberg & Latham, 2003), we specifically focus on whether correlations must be taken into account to decode the neuronal response. We consider a downstream neural system that bases its decoding decisions on the assumption that the spikes are generated by a simplified response model $P_{simp}(\mathbf{r}|s)$, which neglects certain aspects of the spike train correlation structure (e.g., it considers only correlations between spikes close together in time).[2] We ask how much information is lost because the decoding operation is performed assuming that responses $\mathbf{r}$ are generated with $P_{simp}(\mathbf{r}|s)$ rather than with $P(\mathbf{r}|s)$.

The choice of the mathematical form of $P_{simp}(\mathbf{r}|s)$ will depend on the question that the experimenter wants to address about correlations. If, for example, one is interested in whether correlations of any form are important for decoding, then one considers how much information is lost when the independent model $P_{ind}(\mathbf{r}|s)$ is used for decoding. If instead one is interested in the more specific question of whether correlations within a specified time range are important for decoding, then one considers how much information would be lost when using simplified response models $P_{simp}(\mathbf{r}|s)$ that neglect correlations at timescales outside the specified range. When considering neuronal population recordings, a similar strategy could be used to study the spatial scale at which correlations influence information transmission. In this case, $P_{simp}(\mathbf{r}|s)$ will take into account only correlations among a specific subset of neurons.

[2] For example, the downstream system may decode the stimulus using, via Bayes' rule, a posterior probability based on the simplified model: $P_{simp}(s|\mathbf{r}) = P(s)P_{simp}(\mathbf{r}|s)/P_{simp}(\mathbf{r})$.

3.1 Assumptions on the Simplified Models of Correlation. Although in the remainder of the letter, we will focus on a particular class of simplified response models, in this section it is useful to keep the simplified probability model $P_{simp}(\mathbf{r}|s)$ as general as possible and spell out the minimal requirements to $P_{simp}(\mathbf{r}|s)$ that are necessary to develop our information-theoretic formalism. We will require only that $P_{simp}(\mathbf{r}|s)$ satisfies two assumptions. These assumptions are listed below, and their importance will be made clear in the rest of the letter.

Before we list the assumptions, we will describe our formalism by having in mind simplified models $P_{simp}(\mathbf{r}|s)$ that depend on many fewer parameters than $P(\mathbf{r}|s)$, and are thus much easier to sample than $P(\mathbf{r}|s)$. For example, the simplest possible case of $P_{simp}(\mathbf{r}|s)$ is a parameter-free probability distribution that is uniformly flat across all times and stimuli. Another example of a model that is simple to sample is $P_{simp}(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s)$. In fact, while estimating $P(\mathbf{r}|s)$ requires an evaluation of $(M+1)^L - 1$ parameters for each stimulus s, estimating $P_{ind}(\mathbf{r}|s)$ needs only ML parameters for each stimulus. Although we have in mind very simple models for $P_{simp}(\mathbf{r}|s)$, it is important to note that the formalism developed in this letter would be well defined even if $P_{simp}(\mathbf{r}|s)$ is approximately as complex to sample as $P(\mathbf{r}|s)$. In fact, the family of Markov models that we will study in detail below interpolates parametrically from low to high model complexity.

We require that the simplified model $P_{simp}(\mathbf{r}|s)$ to be used satisfies the following two assumptions:

Assumption 1. We require that the method used for transforming $P(\mathbf{r}|s)$ to $P_{simp}(\mathbf{r}|s)$ operates separately and independently on the responses conditioned to each stimulus. Thus, we require that the transformation from $P(\mathbf{r}|s)$ to $P_{simp}(\mathbf{r}|s)$ is independent of $P(s)$, or of $P(\mathbf{r}|s')$ and $P(s')$ for any $s' \neq s$. This property is useful because the resampling (or "shuffling") techniques that, as detailed in section 4, can reduce the bias of information estimates can be applied only to $P_{simp}(\mathbf{r}|s)$ that, for each stimulus s, are constructed only from responses collected in response to that stimulus.

Assumption 2. We require that for each stimulus s with nonzero probability, the simplified response model P_{simp} satisfies the following condition:

$$\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}|s)\log_2 P_{simp}(\mathbf{r}|s) = \sum_{\mathbf{r}\in\mathcal{R}} P_{simp}(\mathbf{r}|s)\log_2 P_{simp}(\mathbf{r}|s). \tag{3.2}$$

Assumption 2 is important to our analysis for three reasons. The first is that (taking the point of view that $0\log(0)$ is zero and $c\log(0)$ is ill defined for any $c \neq 0$) assumption 2 enforces the condition that if for some $\mathbf{r}$ and s, $P_{simp}(\mathbf{r}|s)$ is zero, then $P(\mathbf{r}|s)$ must also be zero. This fact is crucial in the present context because, as we will see in the next section, it ensures that the information-theoretic quantities to be introduced below are well defined. The second reason is that, as also shown in the next section, assumption 2 ensures that we can rewrite the information-theoretic quantities in a way

that is easier to sample. The third reason is that assumption 2 is satisfied by all maximum entropy models constrained to preserve selected features of the full probability model $P(\mathbf{r}|s)$. This will be demonstrated in section 3.3. Since maximum entropy smoothing is a principled way to fill the unconstrained details of a simplified model, it is useful to ensure that our formalism is applicable to all such models.

3.2 Measures of the Information Lost in Decoding with the Simplified Model. Now we turn to determining how much information is lost when decoding the neural response with a mismatched decoding model $P_{simp}(\mathbf{r}|s)$ instead of with the true model $P(\mathbf{r}|s)$. This problem has been well studied in the information-theoretic literature (Merhav, Kaplan, Lapidoth, & Shamai Shitz, 1994; Latham & Nirenberg, 2005). Although this information loss cannot be expressed through a general and simple analytical expression, Latham and Nirenberg (2005) recently derived a simple closed-form expression that is an upper bound to it, as follows:

$$\Delta I_{simp} \equiv D(P(s|\mathbf{r})||P_{simp}(s|\mathbf{r})) \equiv \sum_{\mathbf{r}} P(\mathbf{r}) \sum_{s} P(s|\mathbf{r}) \log_2 \frac{P(s|\mathbf{r})}{P_{simp}(s|\mathbf{r})}, \quad (3.3)$$

where D is conditional Kullback-Leibler (KL) distance (see Cover & Thomas, 1991, p. 22, eq. 2.65). Assumption 2 ensures that if for some $\mathbf{r}$ and s, $P_{simp}(\mathbf{r}|s)$ is zero, then $P(\mathbf{r}|s)$ must also be zero, and this in turn ensures that ΔI_{simp} is a nonnegative and nondivergent information-theoretic measure.

An important problem in the practical estimation of ΔI_{simp} in the case in which $P_{simp}(s|\mathbf{r})$ is described by many fewer parameters than $P(s|\mathbf{r}))$ is that it is heavily biased, approximately as much as the mutual information I (Pola et al., 2005). In the next sections, we will show how to reduce the bias problem of ΔI_{simp} and thus allow its estimation in practice.

A second quantity of interest is $I_{LB\text{-}simp}$ (see Pola et al., 2005), the difference between the mutual information I and ΔI_{simp}:

$$\begin{aligned} I_{LB\text{-}simp} &= I - \Delta I_{simp} \\ &= -\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}) \log_2 P_{simp}(\mathbf{r}) + \sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}|s) \log_2 P_{simp}(\mathbf{r}|s) \\ &= \chi_{simp}(\mathcal{R}) + \sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}|s) \log_2 P_{simp}(\mathbf{r}|s), \end{aligned} \quad (3.4)$$

where

$$\chi_{simp}(\mathcal{R}) \equiv -\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}) \log_2 P_{simp}(\mathbf{r}). \quad (3.5)$$

Since ΔI_{simp} is nonnegative and is an upper bound to the information lost when decoding the neuronal responses with the mismatched response model P_{simp}, $I_{LB\text{-}simp}$ has a well-defined meaning: it quantifies information that can be decoded by using P_{simp}. As shown by Pola et al. (2005), this quantity is of practical importance because it is much less biased than the mutual information I; therefore, it provides a useful data-robust lower bound to the information decodable with P_{simp}.

A further simplification to both ΔI and I_{LB} can be obtained by making use of assumption 2, which permits rewriting $I_{LB\text{-}simp}$ and ΔI_{simp} as

$$I_{LB\text{-}simp} = \chi_{simp}(\mathcal{R}) - H_{simp}(\mathcal{R}|\mathcal{S}) \tag{3.6}$$

$$\Delta I_{simp} = H_{simp}(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S}) + H(\mathcal{R}) - \chi_{simp}(\mathcal{R}), \tag{3.7}$$

where $H_{simp}(\mathcal{R}|\mathcal{S})$ is the noise entropy of the simple response model:

$$H_{simp}(\mathcal{R}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P_{simp}(\mathbf{r}|s) \log_2 P_{simp}(\mathbf{r}|s). \tag{3.8}$$

The advantage of this rewriting is that now the stimulus-conditional functionals of the simplified model are expressed as the noise entropy of $P_{simp}(\mathbf{r}|s)$. This property is important to improving the sampling properties of the information quantities, because, as we will see in the next section, $H_{simp}(\mathcal{R}|\mathcal{S})$ can be corrected for limited sampling by very effective techniques (Nemenman, Bialek, & de Ruyter van Steveninck, 2004), and its presence in the expression for ΔI_{simp} will allow us to cancel out the bias of the latter with appropriate procedures.

3.3 Maximum Entropy Models. We now consider in more detail the problem of how to construct simplified correlation models that satisfy the assumptions needed by our information-theoretic framework.

In constructing simplified models of correlations, it is natural to ask our model to preserve only some properties of the true probability $P(\mathbf{r}|s)$. A way to formalize this is to require our simplified model to satisfy, apart from the usual requirements of nonnegativity and normalization to one, a certain number m of constraints that are also satisfied by $P(\mathbf{r}|s)$, as follows:

$$\begin{aligned} &P_{simp}(\mathbf{r}|s) > 0 \\ &\sum_{\mathbf{r}} P_{simp}(\mathbf{r}|s) = 1 \\ &\sum_{\mathbf{r}} P_{simp}(\mathbf{r}|s) g_i(\mathbf{r}) = \sum_{\mathbf{r}} P(\mathbf{r}|s) g_i(\mathbf{r}) \; i = 1, \cdots, m, \end{aligned} \tag{3.9}$$

where $g_i(\mathbf{r})$ are arbitrary functions on $\mathcal{R}$.[3] Once the constraints in equation 3.9 have been chosen, it is then desirable to simplify the response model by removing all types of correlation in the data apart from those enforced by the features preserved from the original distribution.

A principled way to choose a $P_{simp}(\mathbf{r}|s)$ that satisfies the constraints in equation 3.9 and adds no further relationship between the data is to choose $P_{simp}(\mathbf{r}|s)$ as the distribution with the maximum entropy allowed by our constraints. This maximum entropy distribution is unique and has the following expression (Cover & Thomas, 1991):

$$P_{simp}(\mathbf{r}|s) = \exp\left\{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{r})\right\}, \tag{3.10}$$

where the parameters $\lambda_0, \lambda_i, \cdots, \lambda_m$ are fixed (independently for each conditional distribution to each stimulus) so as to satisfy the constraints in equation 3.9. The maximum entropy distribution is in some way the most reasonable choice of simplified model of correlations given the constraints: to choose a distribution with lower entropy would correspond to assume some additional structure that we do not know; to choose one with a higher entropy would necessarily violate the constraints that we wish to enforce.

It is important to note that any maximum-entropy simplified model of the form in equation 3.10 that satisfies constraints of equation 3.9 is a suitable simplified model for our analysis; in fact, any such maximum-entropy model satisfies by construction the two assumptions of our formalism. In particular, by using equation 3.10, equationq 3.2 of assumption 2 becomes:

$$\sum_{\mathbf{r}} P_{simp}(\mathbf{r}|s)\left\{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{r})\right\} = \sum_{\mathbf{r}} P(\mathbf{r}|s)\left\{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{r})\right\}, \tag{3.11}$$

which is obviously satisfied if $P_{simp}(\mathbf{r}|s)$ meets the constraints set by equation 3.9. Another demonstration of the relationship between assumption 2 and the maximum-entropy principled is reported in appendix B.

An important class of maximum entropy distributions is made of the distributions that preserve some marginal probabilities of $P(\mathbf{r}|s)$, such as the independent model $P_{ind}(\mathbf{r}|s)$ of equation 3.1, the Markov models considered in the next subsection, and the hierarchical probability models of Amari (2001). Models preserving marginals are obtained from equation 3.10 by constraining P_{simp} and P to have equal sum on a number of subsets

[3]The functions $g_i(\mathbf{r})$ could in principle be different for each stimulus conditional distribution.

$\mathcal{A}_i, \mathcal{B}_i, \cdots$ of the responses space $\mathcal{R}$ (each subset corresponding to the responses that have to be summed to compute the marginal probability to be preserved). This corresponds to choosing for each subset a function $g_i(\mathbf{r})$ in equations 3.9 and 3.10 with value one on the subset under consideration and zero elsewhere.

For example, the independent model of equation 3.1 is obtained from equation 3.10 by partitioning the response space into the disjoint union of the "marginal" subsets $\mathcal{A}_1$, $\mathcal{A}_2$ (the subset of all responses with respectively 1 or zero spikes in time bin 1), $\mathcal{B}_1$, $\mathcal{B}_2$ (the subset of all responses with respectively 1 or zero spikes in time bin 2), and so on for all time bins, and then by enforcing P_{simp} and P to have equal sum on each of the subsets.

More complex maximum entropy models can be obtained by extending the procedure to constrain P_{simp} also on intersections of subsets. For example, Markov models of order 1 can be obtained by constraining the sum of P_{simp} not only on the intersections of the above marginal subsets corresponding to each time bin, but also on the pairwise intersections corresponding to adjacent time bins. Amari's hierarchical models of purely pairwise interactions (Amari, 2001) can be obtained by the stricter requirement of constraining the sum of P_{simp} on all the pairwise intersections of marginal subsets, not only to that corresponding to adjacent time bins.

All of these transformations can be applied recursively. The recursion leads to constructing Markov chains of arbitrary length, as well as hierarchical models containing higher-order interactions (Amari, 2001). It is interesting to note that in this way, one can prove that all suffix trees models satisfy our assumptions and therefore are valid choices of P_{simp}. In fact, suffix tree models can be constructed with this recursive partitioning by constraining P_{simp} on a smaller number of intersections than the Markov model of corresponding length. This observation helps in understanding the relationship between the work presented here and the recent work of London, Schreibman, Hausser, Larkum, & Segev (2002) and Kennel, Shlens, Abarbanel, & Chichilnisky (2005), which use suffix tree models to estimate entropy rates.

In summary, in this section we have established that all maximum-entropy models satisfy our assumptions and can thus be used to estimate information with our formalism. These model include Markov chains, hierarchical distributions, and suffix tree models.

We have also provided a general and explicit way to construct such simplified models from the data through equations 3.9 and 3.10. This construction can be successfully applied whatever the statistics of neuronal firing described by the response probability $P(\mathbf{r}|s)$. In fact, for any given choice of the constraints in equation 3.9, the maximum entropy model fulfilling these constraints will automatically satisfy (by construction) our assumptions, whatever the form of $P(\mathbf{r}|s)$. An important implication of this result is that, in practice, our assumptions will not restrict in any way the applicability of the method to data sets with specific statistical properties.

3.4 Markov Models. Despite the generality of the above constructions of P_{simp}, we will illustrate and develop the main idea behind our formalism by focusing on a specific class of simplified correlation models: the Markov models with finite memory. We choose to illustrate our ideas using this particular class of maximum entropy simplified model because (1) by tuning the order of the Markov process, we can vary parametrically the complexity of the model and thus illustrate clearly how the sampling behavior of the information-theoretic functional depends on the number of parameters describing the simplified model, and (2) these models are easy to construct and apply to data.

A neurophysiological motivation of the use of Markov models stems from the fact that in many neural systems, correlations are significant only between spikes that are separated by a short time lag, in the range 1 to 15 ms (Gray et al., 1989; Brosch, Bauer, & Eckhorn, 1997; Dan et al., 1998; Nirenberg, Carcieri, Jacobs, & Latham, 2001; Golledge et al., 2003). In such cases, to preserve the whole information, it is sufficient to take into account only correlations extending over a short lag. Thus, one can approximate the real probability of current response $r(t)$ given the past firing with a finite-memory Markov model that looks back only q time steps, as follows:

$$P_q(\mathbf{r}|s) = P(r(1)|s) \prod_{t=2}^{L} P(r(t)|r(t-q), \ldots, r(t-1); s), \quad \text{if } q = 1, \ldots, L-1,$$
$$P_0(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s), \quad \text{if } q = 0. \tag{3.12}$$

The probability conditional on the response in the previous time steps at fixed stimulus in the above equation can be computed from the experimental probabilities via

$$P(r(t)|r(t-q), \ldots, r(t-1); s) = \frac{P(r(t-q), \ldots, r(t-1), r(t)|s)}{P(r(t-q), \ldots, r(t-1)|s)}, \tag{3.13}$$

where $P(r(t-q), \ldots, r(t-1), r(t)|s)$ and $P(r(t-q), \ldots, r(t-1)|s)$ are marginal distributions of the full model $P(\mathbf{r}|s)$, computed by integrating away the dependence on all the response variables that do not enter in their argument. This simple procedure to construct the Markov model is equivalent to the maximum-entropy one described above.

Markov models interpolate parametrically between the independent model ($q = 0$) and the full probability model (obtained for $q = L - 1$, because $P_{L-1}(\mathbf{r}|s) = P(\mathbf{r}|s)$). $P_q(\mathbf{r}|s)$ preserves all correlations extending up to q time bins in the past, and it neglects all correlations of range longer than q. Thus, it is a perfect description of neuronal firing if correlations extend to a lag shorter than or equal to q time bins.

The information-theoretic probability functionals corresponding to the choice $P_{simp}(\mathbf{r}|s) = P_q(\mathbf{r}|s)$ will be indicated by a subscript q in place of the subscript *simp*. For completeness, their expression is reported below:

$$I_{LB-q} = \chi_q(\mathcal{R}) - H_q(\mathcal{R}|\mathcal{S}) \tag{3.14}$$

$$\Delta I_q = H_q(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S}) + H(\mathcal{R}) - \chi_q(\mathcal{R}), \tag{3.15}$$

where $H_q(\mathcal{R}|\mathcal{S})$ is the noise entropy of the simple response model,

$$H_q(\mathcal{R}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}\in\mathcal{R}} P_q(\mathbf{r}|s) \log_2 P_q(\mathbf{r}|s), \tag{3.16}$$

and $\chi_q(\mathcal{R})$ is

$$\chi_q(\mathcal{R}) = -\sum_{\mathbf{r}\in\mathcal{R}} P(\mathbf{r}) \log_2 P_q(\mathbf{r}). \tag{3.17}$$

4 Bias Cancellations Obtained by Shuffling the Responses

As discussed above, mutual information has been broken down into two terms, $I_{LB\text{-}simp}$ and ΔI_{simp}, with radically different sampling properties, the former easy to sample and the latter very difficult to sample. Here we examine in detail the sampling behavior of ΔI_{simp} and show how to reduce its bias dramatically without increasing its variance. Since ΔI_{simp} is the most biased part of I, this will also improve the sampling properties of the mutual information. Since I, $I_{LB\text{-}simp}$ and ΔI_{simp} consist of four quantities—$H(\mathcal{R}|\mathcal{S})$, $H_{simp}(\mathcal{R}|\mathcal{S})$, $H(\mathcal{R})$, and $\chi_{simp}(\mathcal{R})$—the relative sampling properties of I, $I_{LB\text{-}simp}$, and ΔI_{simp} can be established by considering the sampling properties of the above four quantities.

4.1 The Independent Decoder: Ignoring All Correlations. For clarity, we start by considering in detail the sampling properties of the quantities corresponding to the $q = 0$, independent probability model $P_{simp}(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s) = P_0(\mathbf{r}|s)$. The corresponding information-theoretic functionals are I_{LB-0} and ΔI_0, taken from equations 3.16 and 3.17 with $q = 0$.

4.1.1 Uncorrected Estimators. There are two relevant aspects of the sampling properties of a probability functional: its bias and its variance. We will consider the bias first and the variance later. The bias of a functional $F(P)$ of the probabilities P is defined as the difference between $\langle F(P_N)\rangle_N$, the ensemble-averaged value of the functional computed from the probability distributions P_N empirically obtained from N trials, and the true value of the functional $F(P)$ computed with the true probability distribution P. Thus, the bias is a systematic error that cannot be eliminated just by averaging.

We illustrate the sampling behavior of the entropy functionals by computing them on a set of realistically simulated neuronal data, generated as follows. We simulated the spiking response of one neuron in somatosensory cortex to 49 different stimuli consisting of sinusoidal whisker vibrations with different amplitude and frequency (Arabzadeh, Petersen, & Diamond, 2003; Arabzadeh, Panzeri, & Diamond, 2004). We simulated neuronal responses over a 0 to 50 ms poststimulus time window. We then digitized, with time precision Δt equal to 5 ms, these responses into $L = 10$ binary words. The simulations were performed using a Markov process with order $q = 3$, with all the marginal probabilities of order 3 or less and the transition probabilities taken from real responses of a cortical somatosensory neuron.[4] The spike train simulated in this way retains a faithful description of all the real neuronal marginal distributions and their correlations up to $q = 3$ time bins.

In Figure 1A we report the sampling behavior of the four quantities $\chi_0(\mathcal{R})$, $H(\mathcal{R})$, $H_0(\mathcal{R}|\mathcal{S})$, and $H(\mathcal{R}|\mathcal{S})$ as a function of the number of trials per stimulus. For each simulation with a fixed number of trials, we computed the plug-in value of the functional by plugging into their equations the empirical estimates of the probability (without application of any bias correction procedure) and then averaging all simulations with the same number of trials.

We first consider the noise entropy $H(\mathcal{R}|\mathcal{S})$. Figure 1A shows that this is by far the most downward biased of the four functionals considered. This is because it requires the simultaneous measure of all $P(\mathbf{r}|s)$ to all stimuli. To understand better its sampling behavior, it is useful to find analytical approximations to the bias. These can be easily derived in the asymptotic sampling regime. The latter is defined as the case in which the number of trials N_s to each stimulus s is so large that each response bin with nonzero probability is observed many times, $N_s P(\mathbf{r}|s) \gg 1$ (Panzeri & Treves, 1996). In this asymptotic sampling regime, the bias of $H(\mathcal{R}|\mathcal{S})$ (and, similarly, of all other functionals) can be expanded in inverse powers of $1/N$ (N being the total number of trials across all stimuli), as follows:

$$Bias[H(\mathcal{R}|\mathcal{S})] \approx \frac{C_1}{N} + \frac{C_2}{N^2} + \cdots, \tag{4.1}$$

[4] For each stimulus condition s, N_s binary spike trains were generated with a q-order Markov model as follows (see equation 3.13). The response in the first bin was assigned to be a 1 or a 0 (spike or no spike) according to $P(r(1)|s)$, the latter being computed from the real data. Responses in successive time bins were generated one after the other, one by one, using the corresponding transition probabilities. For instance, the response at bin k was generated according to the real-data probability $P(r(k)|r(k-j), \ldots, r(k-1); s)$, where $j \leq \min(q, k-1)$.

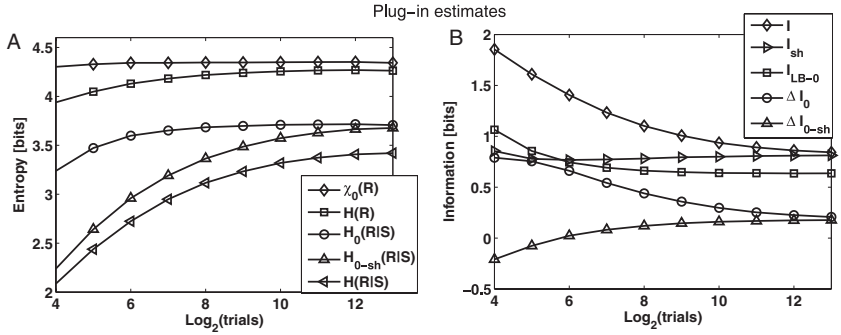

Figure 1: Comparison of the sampling properties of plug-in estimations of different probability functionals. The plug-in estimators of the probability functionals are plotted as a function of the number of trials per stimulus available. Results were averaged over a number of repetitions of the simulation (decreasing from 200 to 10 as the number of trials per stimulus available increased). We simulated a neuron responding to 49 different stimuli. We considered a time precision of 5 ms and a poststimulus time window of 50 ms; thus, L was equal to 10. The spike train was simulated with a Markov process with the same mean firing rates and up-to-third-order marginal probabilities as a real spike train recorded in Arabzadeh et al. (2003) in response to 49 different sinusoidal whisker vibrations. (A) Average values of plug-in estimators of $\chi_0(R)$, $H(R)$, $H_0(R)$, $H(R|S)$, and $H_{0-sh}(R|S)$ obtained without any bias correction. (B) Average values of plug-in estimators of $I(R, S)$, I_{LB-0}, ΔI_0, and ΔI_{0-sh} obtained without any bias correction.

The leading term in $1/N$ has a simple analytical expression,

$$Bias[H(\mathcal{R}|\mathcal{S})] \approx -\frac{1}{2N \ln 2} \sum_s (\tilde{R}(s) - 1), \tag{4.2}$$

where $\tilde{R}(s)$ denotes the number of "relevant" responses of the stimulus conditional response probability distribution $P(\mathbf{r}|s)$, that is, the number of different responses $\mathbf{r}$ with nonzero probability of being observed when stimulus s is presented (Panzeri & Treves, 1996). $\tilde{R}(s)$ is of order $(M+1)^L$ for each stimulus. Thus, it follows that for the bias in equation (4.2) to be small, N should be much bigger than $S \times (M+1)^L$.

Let us now consider $H_0(\mathcal{R}|\mathcal{S})$. It is apparent from Figure 1A that it is almost unbiased. The reason is that it can be expressed as the sum of single-bin entropies (see Pola et al., 2005). As a consequence, its bias is $\approx SML/2N\log(2)$ and is thus much smaller than that of $H(\mathcal{R}|\mathcal{S})$.

Figure 1A shows that the noise entropy $H(\mathcal{R})$ is considerably simpler to sample than $H(\mathcal{R}|\mathcal{S})$. The reason is that since $H(\mathcal{R})$ depends only on $P(\mathbf{r})$,

its bias is approximately S times smaller than the bias of $H(\mathcal{R}|\mathcal{S})$. This is an advantage when many different stimuli are presented.

Figure 1A shows that the bias of $\chi_0(\mathcal{R})$ is much smaller than the bias of $H(\mathcal{R})$. The reason is as follows. Bias arises from the logarithmic form of entropy functionals. The log in $\chi_0(\mathcal{R})$ depends on $P_0(\mathbf{r})$. Since $P_0(\mathbf{r})$ is better sampled than $P(\mathbf{r})$, $\chi_0(\mathcal{R})$ has less bias than $H(\mathcal{R})$, whose log depends on $P(\mathbf{r})$.

Now we study how the properties of the four functionals combine together to give rise to the sampling properties of $I(\mathcal{R};\mathcal{S})$, I_{LB-0} and ΔI_0.

The bias of the mutual information $I(\mathcal{R};\mathcal{S})$ is the difference between the biases of $H(\mathcal{R})$ and $H(\mathcal{R}|\mathcal{S})$. As the most biased term is $H(\mathcal{R}|\mathcal{S})$, the mutual information is upward biased (Panzeri & Treves, 1996).

The bias of I_{LB-0} (see equation 3.14 with $q = 0$) is the difference between the biases of $\chi_0(\mathcal{R})$ and $H_0(\mathcal{R};\mathcal{S})$. As both these quantities are virtually unbiased, so is I_{LB-0}.

Let us now consider the bias of ΔI_0 (see equation 3.15 with $q = 0$). When many stimuli are presented, the contribution of the $\chi_0(\mathcal{R})$ and $H(\mathcal{R})$ is only very mildly biased and determined by the bias of $H(\mathcal{R})$. Thus, most of the bias comes from the stimulus-conditional term $H_0(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S})$. Since $H(\mathcal{R}|\mathcal{S})$ is very strongly biased downward and $H_0(\mathcal{R}|\mathcal{S})$ is essentially unbiased, the two biases do not cancel out, and as a result, ΔI_0 is biased upward and behaves like $-H(\mathcal{R}|\mathcal{S})$.

An important practical problem is how to reduce the bias of $H_0(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S})$ and thus improve the sampling properties of ΔI_0. A solution to this problem is to compute $H_0(\mathcal{R}|\mathcal{S})$ not only directly from the single bin marginal probability as in equation 3.16 but by randomly shuffling the responses and then computing their entropy. In the $q = 0$ case considered here, we can generate a new set of shuffled responses to stimulus s by randomly permuting, for each time bin, the order of trials collected in response to the stimulus s considered, and then joining together the shuffled responses in different time bins into a response vector $\mathbf{r}_{0-sh}$. This shuffling operation leaves each single-time-bin marginal probability unchanged (because responses in each bin are just randomly permuted), while destroying any within-trial correlation between different time bins. We define $H_{0-sh}(\mathcal{R}|\mathcal{S})$ as the noise entropy of the shuffled distribution:

$$H_{0-sh}(\mathcal{R}|\mathcal{S}) = -\sum_{s\in\mathcal{S}} P(s) \sum_{\mathbf{r}_{sh}\in\mathcal{R}} P_{0-sh}(\mathbf{r}|s) \log_2 P_{0-sh}(\mathbf{r}|s), \tag{4.3}$$

where $P_{0-sh}(\mathbf{r}|s)$ is the distribution of response values obtained from $\mathbf{r}_{0-sh}$. The asymptotic large N value of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ is the same of that of $H_0(\mathcal{R}|\mathcal{S})$, but its scaling with the number of trials is much different. This is shown in Figure 1A. Unlike $H_0(\mathcal{R}|\mathcal{S})$, $H_{0-sh}(\mathcal{R}|\mathcal{S})$ scales with N approximately as $H(\mathcal{R}|\mathcal{S})$, but with a slightly more negative slope (i.e., slightly more

downward bias) as the number of trials decreases. The fact that the biases of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ are of similar size can be intuitively understood from the fact that $P_{0-sh}(\mathbf{r}|s)$ is sampled with the same number of trials as $P(\mathbf{r}|s)$ from responses with the same dimensionality. To understand why $H_{0-sh}(\mathcal{R}|\mathcal{S})$ is more biased downward than $H(\mathcal{R}|\mathcal{S})$, we computed the bias of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ in the "asymptotic sampling" regime. We found that the asymptotic bias of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ has the same expression as that of $H(\mathcal{R}|\mathcal{S})$ in equation 4.2, after replacing $\tilde{R}(s)$ with $\tilde{R}_{0-sh}(s)$, the number of bins relevant to $P_{0-sh}(\mathbf{r}|s)$. Since $P_0(\mathbf{r}|s) = 0$ implies $P(\mathbf{r}|s) = 0$ and since the shuffled responses are generated according to $P_0(\mathbf{r}|s)$, then it must be that $\tilde{R}_{0-sh}(s) \geq \tilde{R}(s)$. Thus, at the leading order in the asymptotic regime, $H_{0-sh}(\mathcal{R}|\mathcal{S})$ is never less downward biased than $H(\mathcal{R}|\mathcal{S})$.[5] Therefore, if we are in the asymptotic sampling regime (or at least the number of trials is enough to make the asymptotic equations a decent estimate of the size of the bias) and if $\tilde{R}_{0-sh}(s)$ and $\tilde{R}(s)$) are roughly similar, $H_{0-sh}(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S})$ will either (1) have a leading-order $1/N$ bias term that is negative and arises from a partial cancellation of roughly similar leading bias terms of two entropies, or (2) (in case $\tilde{R}_{0-sh}(s) = \tilde{R}(s)$) it will have a very small bias, only at the order $1/N^2$ or higher.

The sampling behavior of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ suggests a simple strategy to reduce the bias of ΔI_0: use $H_{0-sh}(\mathcal{R}|\mathcal{S})$ instead of $H_0(\mathcal{R}|\mathcal{S})$. We call this estimator ΔI_{0-sh}:

$$\Delta I_{0-sh} = H_{0-sh}(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S}) + H(\mathcal{R}) - \chi_0(\mathcal{R}). \tag{4.4}$$

Simulations in Figure 1B show that ΔI_{0-sh} is much less biased than ΔI_0, but it converges to the same asymptotic large-N value. The biases of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ almost cancel each other, leaving an overall small negative bias.[6]

The good sampling properties of ΔI_{0-sh} suggest a new, alternative way to estimate mutual information, as follows:

$$I_{sh} = I_{LB-0} + \Delta I_{0-sh}. \tag{4.5}$$

[5] There are different types of resampling methods that have been used to generate surrogate data sets for the validation of information-theoretic calculations (Johnson, Gruner, Baggerly, & Seshagiri, 2001). It is important to note that $\tilde{R}_{0-sh}(s) \geq \tilde{R}(s)$ holds for the resampling procedure "without replacement" used here to generate the surrogate data, but it does not necessarily hold for other resampling techniques "with replacement."

[6] The idea of using shuffling to cancel out the biases in the stimulus-dependent part of ΔI was first proposed by Nirenberg et al. (2001). However, the fact that $H_{0-sh}(\mathcal{R}|\mathcal{S})$ is more downward biased than $H(\mathcal{R}|\mathcal{S})$, and thus ΔI_0 computed in this way is mildly downward biased, has not been reported before.

Since I_{LB-0} is virtually unbiased and ΔI_{0-sh} is mildly biased downward, I_{sh} is also mildly biased downward. This is confirmed by the simulation results in Figure 1B: using I_{sh} compares very favorably with computing mutual information directly as I in equation 2.1.

In summary, we now have two classes of estimators of ΔI_0 and therefore of I: the direct estimator ΔI_0 and I in equations 3.3 and 2.1, which are strongly biased upward, and the shuffled estimators ΔI_{0-sh} and I_{sh}, which are mildly biased downward. The difference in the sign of the bias of the two different estimators is very important in practice. In fact, computing ΔI_0 and I with both upward- and downward-biased algorithms provides a mean to bound from both above and below. This can greatly improve the confidence in estimates obtained from experimental data. In the rest of this letter, we devote our attention to how to make both upper and lower bounds tighter than they are in Figure 1.

4.1.2 Effect of Different Bias Correction Methods. The above shows the behavior of the probability functionals when estimated by plugging in the probabilities estimated directly from the data. However, these estimates can be improved by using available bias correction techniques. In this section, we apply different bias correction techniques to the probability functionals in order to understand how best to evaluate each functional from a limited number of data.

We first evaluated the performance of a quadratic extrapolation procedure (Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998), performed computing the quantities from fractions of the data available and then fitting the resulting data-scaling behavior to a quadratic function of $1/N$ as in equation 4.1. This procedure should work well in the asymptotic regime where the number of trials N is large.[7] Simulations suggest that in practice, a good bias correction from quadratic extrapolations or other asymptotic requires the number of trials per stimulus to be at least as big as the number of possible responses R (Panzeri & Treves, 1996).

Results are shown in Figures 2A and 2B. Figure 2A shows that $\chi_0(\mathcal{R})$ after the quadratic extrapolation procedures become completely unbiased, even with as few as 32 trials per stimulus. After applying the quadratic extrapolation, $H_0(\mathcal{R}|\mathcal{S})$ also becomes almost unbiased even at 32 trials per stimulus. This can be understood by remembering that $H_0(\mathcal{R}|\mathcal{S})$ can be expressed as a sum of single bin entropies, and the sampling of each single bin entropy can be considered to be in asymptotic regime even for a small number of trials. Thus, the quadratic extrapolation performs very well on $H_0(\mathcal{R}|\mathcal{S})$. As

[7] We also considered the performance of other bias correction methods developed to work in the asymptotic regime, such as computing analytically the coefficients of the $1/N$ expansion of the bias as a function of probabilities (Panzeri & Treves, 1996; Pola et al., 2005). As we found no overall increase in performance with respect to the quadratic extrapolation, we decided to work with the latter, which is easier to implement.

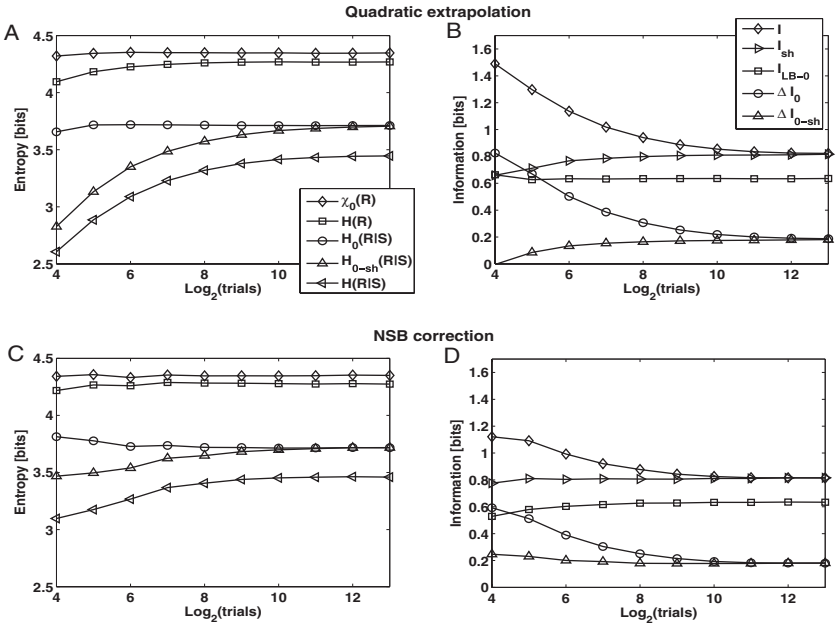

Figure 2: Comparison of the sampling properties of estimators computed with the quadratic extrapolation and the NSB bias correction procedure. Results are plotted as a function of the number of trials per stimulus and were averaged over a number of repetitions of the simulation. The simulated data were obtained exactly as in Figure 1, again considering a poststimulus window of 50 ms discretized into $L = 10$ bins of size $\Delta t = 5$ ms. (A, B) Quadratic extrapolation. (A) Averaged estimated values of $\chi_0(R)$, $H(R)$, $H_0(R)$, $H(R|S)$, and $H_{0-sh}(R|S)$. (B) Averaged estimated values of $I(R, S)$, I_{LB-0}, ΔI_0, and ΔI_{0-sh}. (C, D) NSB estimation. (C) Averaged estimated values of $\chi(R)$, $H(R)$, $H_0(R)$, $H(R|S)$, and $H_{0-sh}(R|S)$. (D) Averaged estimated values of $I(R, S)$, I_{LB-0}, ΔI_0, and ΔI_{0-sh}.

a result, the extrapolated I_{LB-0} estimator (see Figure 2B) becomes almost unbiased, even with as few as 32 trials per stimulus.

Figure 2A shows that the response entropy $H(\mathcal{R})$ is still mildly biased after correction, requiring at least 128 trials per stimulus for good estimation. In contrast, the noise entropies $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ remain highly biased even after the extrapolation procedure, requiring at least $2^{10} = 1024$ trials per stimulus for unbiased estimation. The reason is that $P(\mathbf{r}|s)$ and $P_{0-sh}(\mathbf{r}|s)$ are high dimensional, and thus the number of trials needed to get them into the asymptotic sampling regime is much higher. As a consequence, ΔI_0 is still substantially biased upward, and applying a quadratic extrapolation procedure is still not enough: we still need at least 2^{10} trials per stimulus for unbiased estimation (see Figure 2B). However, the good

news from Figure 2B is that the behavior of ΔI_{0-sh} is better than the one of ΔI_0. Because of the partial bias cancellation in $H_{0-sh}(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S})$, the quadratic extrapolation is able to remove most of the downward bias of ΔI_{0-sh} from at least 128 trials per stimulus. Similarly, estimating mutual information as I_{sh} (rather than directly as I in the mutual information definition in equation 2.1) leads to a much less biased estimation (see Figure 2B).

In summary, we simulated data with $L = 10$ time bins and 2^{10} possible responses, and we use the quadratic extrapolation bias correction. Estimating ΔI_0 and I directly requires approximately 2^{10} trials per stimulus, of the order of the size of the response space. However, estimating the same quantities through the ΔI_{0-sh} and I_{sh} shuffling procedure required only 2^7 trials per stimulus, eight times smaller than the size of the responses space. To check how these results scaled with L, we simulated data with the same procedure as in Figures 1 and 2, but using different sizes of poststimulus windows (ranging from $L = 8$ to $L = 14$). The results were always comparable to the $L = 10$ case: estimating ΔI_0 and I directly requires approximately a number of trials per stimulus of the order of the size of the response space. Estimating ΔI_{0-sh} and I_{sh} shuffling procedure required only a number of trials per stimulus eight times smaller than the size of the response space (results not shown).

We now consider the effect of using a more sophisticated Bayesian entropy bias correction proposed by Nemenman et al. (2004). This procedure (called NSB) was designed to operate beyond the asymptotic regime[8] and is briefly reviewed in appendix A. As shown in (Nemenman et al. (2004), it works well even when data are scarce and the response space is high-dimensional, so that each response is observed in no more than handful of trials. Figures 2C and 2D report the values of the functional corrected with the NSB procedure and show that the NSB method in general performs well. The response entropy $H(\mathcal{R})$ (see Figure 2C) is even better behaved than with the quadratic correction and is now essentially unbiased, requiring only 32 trials per stimulus for good estimation. The noise entropies $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ are also better evaluated with the NSB procedure than with the quadratic extrapolation procedure. However, both $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ still remain substantially biased. As a consequence, Figure 2D shows that ΔI_0 and I are still substantially biased upward, requiring at least 2^9 trials per stimulus for unbiased computation. However, the unbiased computation of ΔI_{0-sh} and of I_{sh} requires only 2^6 trials per stimulus, because the residual biases of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $-H(\mathcal{R}|\mathcal{S})$ cancel out almost exactly. Estimating mutual information as I_{sh} (see equation 4.5) works much

[8] We also considered the performance of the procedure of Paninski (2003), which also aims at working beyond the asymptotic sampling regime. We found that on the simulated data and information-theoretic quantities, the Paninski procedure helped to reduce the bias but did not perform as well as the NSB and the quadratic extrapolation. Thus, we omitted the presentation of its results.

better than estimating I directly. However, it should be noted that as a result of applying the NSB procedure, the residual bias for I_{sh} is not always negative. Thus, using the Nemenman et al. (2004) procedure to estimate I_{sh} gives an almost unbiased estimator for small numbers of trials, but not necessarily a lower bound to the true information (as it instead happens when correcting I_{sh} with the quadratic extrapolation).

The only entropy term that performed worse when corrected with the NSB procedure is the independent entropy $H_0(\mathcal{R};\mathcal{S})$. As discussed in appendix A, the reason is that the Nemenman et al. (2004) procedure is tailored to work for high-dimensional response spaces, and thus its assumptions do not work well for $H_0(\mathcal{R};\mathcal{S})$, which is essentially a sum of single-bin entropies. This problem is particularly serious for low firing rates (see appendix A).

Finally, we note that the procedure of Nemenman et al. (2004) has been developed so far only for entropy quantities. Therefore, it cannot be applied to $\chi_0(\mathcal{R})$. Thus, $\chi_0(\mathcal{R})$ will always be corrected with the extrapolation procedure, which, as demonstrated before, works extremely well for this quantity.

It is now useful to come back to discuss why assumption 2 on the probability model, the one that enabled us to express $I_{LB\text{-}simp}$ and ΔI_{simp} as in equations 3.6 and 3.7, is very important to improve the sampling properties. There are two reasons behind it. The first is that the term outside the logarithm in the stimulus-conditional part of $I_{LB\text{-}simp}$ now is $P_{simp}(\mathbf{r}|s)$ rather than $P(\mathbf{r}|s)$, and this is expected to reduce statistical fluctuations. The second, and more important, reason is that now the stimulus-dependent part of $\Delta I_{LB\text{-}simp}$ can be expressed as a difference of entropies. This has a big impact on the ways the sampling bias of this quantity can be corrected. In fact, the sampling bias correction techniques of Nemenman et al. (2004) and the shuffling bias relationships used above both depend crucially on being able to express the stimulus-conditional part of $I_{LB\text{-}simp}$ and of ΔI_{simp} entirely in terms of entropies.

4.1.3 Variance of Shuffling-Based Estimators. We now consider whether the reduction in bias of the shuffled estimators ΔI_{0-sh} and I_{sh} comes at the expense of an increase in their variance. Equation 4.5 indicates that the variance of I_{sh} is largely determined by that of ΔI_{0-sh}, its worst-sampled component. The most biased terms in the computation of ΔI_{0-sh} are the stimulus-conditional entropies $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$, with their bias almost canceling out (see equation 4.4). The amount of variance of ΔI_{0-sh} will depend on the degree of correlation (across independent realizations with the same number of trials) between the values of $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$. If these two quantities were positively correlated, then their statistical fluctuations would tend to cancel out, and the resulting variance of ΔI_{0-sh} would be under control. If instead $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$ were either uncorrelated or negatively correlated across different realizations, then their

statistical fluctuations would either not cancel out or even increase, thus making the variance of ΔI_{0-sh} larger. The scatter plot in Figure 3A shows that $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ are strongly positively correlated when taken from the same realization of simulations. (The data were simulated exactly as in Figure 2 and were obtained using 128 trials per stimulus in each simulation). The fact that $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$ are correlated stems from the fact that fluctuations in the values of single bin marginal probabilities have a major impact on the fluctuations of entropy values (Schultz & Panzeri, 2001). These fluctuations are reflected with the same sign in both $H_{0-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$. Thus, the correlation between the entropies ensures that the variances of I_{sh} and $I(\mathcal{R};\mathcal{S})$ remain comparable. This is clearly demonstrated in Figure 3B (quadratic extrapolation correction procedure used).

To better understand the importance of the correlation of $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$ across different realizations, we paired values of $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$ taken from randomly chosen realizations of the simulation. The scatter plot of these randomly paired entropies is reported in Figure 3C. Computing information I_{sh} from the randomly paired entropies $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$ leads to a considerable increase in invariance of the information I_{sh} (results reported in Figure 3D).

Thus, we conclude that the reduction of bias in estimating information with I_{sh} rather than directly with I does not come at the expense of increased variance, because the computation of I_{sh} involves a cancellation of fluctuations between similarly biased terms. I_{sh} is a very efficient estimator of information in terms of both bias and variance.

4.1.4 Performance of Shuffling-Based Estimators in the Presence of Strong Correlation. An important question is how the bias properties of $H(\mathcal{R}|\mathcal{S})$, $H_{0-sh}(\mathcal{R}|\mathcal{S})$, and I_{sh} are affected by the strength of correlations between spikes. As it was discussed above, when the responses are shuffled, the new value of relevant responses, $\tilde{R}_{0-sh}(s)$, will be equal to or larger than the original $\tilde{R}(s)$. Thus, because of equation 4.2, if correlations are not strong enough to induce radical differences of shape and support between $P(\mathbf{r}|s)$ and $P_{ind}(\mathbf{r}|s)$, ΔI_{0-sh} is expected to have a downward bias that is much smaller in absolute magnitude than the upward bias of ΔI_0. The previously shown simulations (based on a correlation structure taken from a real cortical neuron) confirmed this expectation and suggest that ΔI_{0-sh} and I_{sh} are only very mildly biased in realistic neurophysiological conditions.

However, this argument also suggests that the magnitude of the downward bias may be affected by the overall strength of correlation. High correlation tends to favor certain patterns in the response and thus decrease the number of possible responses. The stronger the correlations, the larger the number of new response words that may appear after shuffling and the larger the resulting downward bias. If correlations are strong enough to make the number of shuffled relevant responses $\tilde{R}_{0-sh}(s)$ much larger

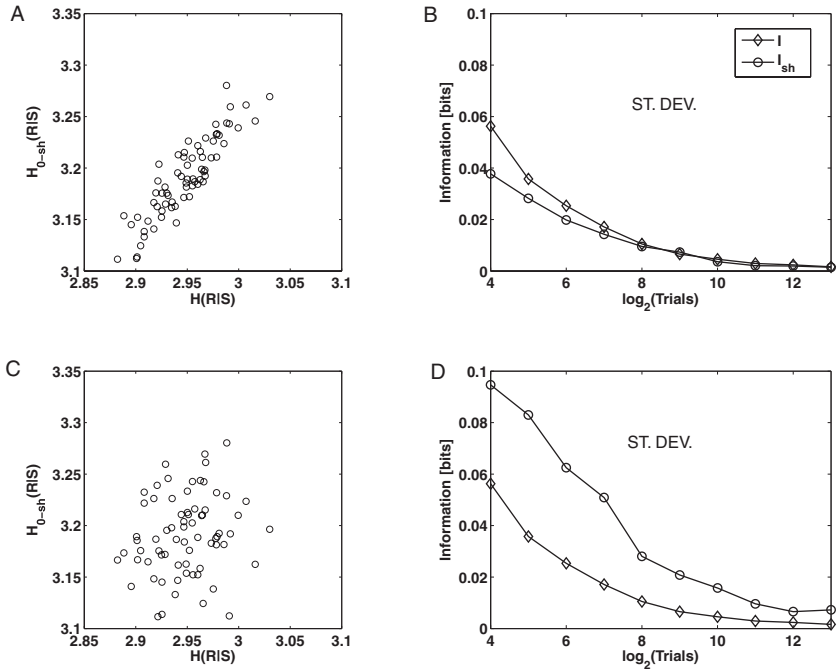

Figure 3: Variance of shuffling estimators. The plug-in estimates of the probability functionals were computed (without any bias corrections) from exactly the same simulated neural data as in Figure 1, again with $L = 10$. (A) Scatter plot of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ versus $H(\mathcal{R}|\mathcal{S})$ computed from 100 random realizations of the simulated spike trains, each realization of the simulation consisting of 128 simulated trials per stimulus condition. Each point in the scatter plot represents a data pair taken from the same realization of the numerical simulation. (B) The variance (across realizations of the simulation) of the estimators of I and I_{sh} plotted as a function of the number of trials per stimulus condition. In this case, I_{sh} was estimated using values of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ from the same realization of the simulation. (C) Scatter plot of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ versus $H(\mathcal{R}|\mathcal{S})$ computed from 100 random realizations of the simulated spike trains, each realization being made of 128 trials per stimulus condition. Now each point in the scatter plot represents a data pair taken from different randomly paired realizations of the numerical simulation. (D). The variance (across realizations of the simulation) of the estimators of I and I_{sh} plotted as a function of the number of trials per stimulus condition. Now I_{sh} was estimated using values of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ from different randomly paired realizations of the numerical simulation.

than the original $\tilde{R}(s)$, then ΔI_{0-sh} may become strongly downward biased.

It is thus important to assess numerically the dependence of the sampling properties of ΔI_{0-sh} on the correlation strength. To address this

issue, we next simulated spike trains as a first-order ($q = 1$) Markov process with a firing rate ρ to each stimulus and a given coefficient c quantifying Pearson correlation of spikes across adjacent time bins (both parameters were taken as constant over time). Reference values for the parameters ρ_0 and c_0 were, similar to previous simulations, measured (as an average over a poststimulus time window of 20–40 ms poststimulus onset) from the experimental cortical responses recorded in response to 49 different whisker vibration stimuli by Arabzadeh et al. (2004). In the simulations, we then took the same rate ρ_0 as the real data and produced a simulated correlation strength c that was f times stronger than the real one: $c = fc_0$. Figure 4 reports how the bias of $H(\mathcal{R}|\mathcal{S})$ and of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ depends on the correlation strength (no bias correction procedure was applied). Figure 4A shows the results obtained for $f = 0$ (i.e., absence of correlations). In this case, the number of possible responses is, on average, unchanged by shuffling. Thus, $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$ have exactly the same bias, and I_{sh} is unbiased. For $f = 1$ (i.e., real neuronal correlation value, shown in Figure 4B), the bias of $H_{0-sh}(\mathcal{R}|\mathcal{S})$ is still very close to that of $H(\mathcal{R}|\mathcal{S})$. For correlation four times stronger than that of the real neuron (see Figure 4C), there is some difference between the biases of $H(\mathcal{R}|\mathcal{S})$ and of $H_{0-sh}(\mathcal{R}|\mathcal{S})$. However, with 64 trials per stimulus or more, the bias of the two entropies cancels out and I_{sh} is unbiased. $H_{0-sh}(\mathcal{R}|\mathcal{S})$ becomes considerably more biased downward than $H(\mathcal{R}|\mathcal{S})$ only for correlation eight times stronger than that of the real neuron (see Figure 4D). Thus, we conclude that although their sampling properties will worsen as the amount of correlation in the data grows, the shuffled estimators ΔI_{0-sh} and I_{sh} can be very useful for the analysis of a wide range of neurophysiological experiments.

4.2 The Markov Model Decoder. The above section dealt with the use of the simplest decoding model, $P_0(\mathbf{r}|s)$. How do the sampling properties of I_{LB-q} and ΔI_q change when using more detailed decoding models such as $P_q(\mathbf{r}|s)$ with $q > 0$?

To address this issue, we consider the properties of $\chi_q(\mathcal{R})$ and $H_q(\mathcal{R}|\mathcal{S})$ with $q > 0$. As q increases, $\chi_q(\mathcal{R})$ is expected to become more biased because it depends on $P_q(\mathbf{r}|s)$ logarithmically. The bias of $H_q(\mathcal{R}|\mathcal{S})$ is expected to increase even more significantly with q. In fact, it can be decomposed into a sum of stimulus-conditional entropies of the marginal probability distributions of up to $q + 1$ consecutive time bins (Pola et al., 2003). For this reason, although the bias of $H_q(\mathcal{R}|\mathcal{S})$ is smaller than that of $H(\mathcal{R}|\mathcal{S})$, it grows for larger q values. The bias of I_{LB-q} is given by the difference between the biases of $\chi_q(\mathcal{R})$ and $H_q(\mathcal{R}|\mathcal{S})$. Thus, as q increases, I_{LB-q} will be more biased.

These expectations are confirmed by the simulations shown in Figure 5. These simulations were performed as in Figure 1, by creating a somatosensory-like, simulated spike train over $L = 10$ time bins responding to 49 different stimuli with realistic firing rates and correlations described

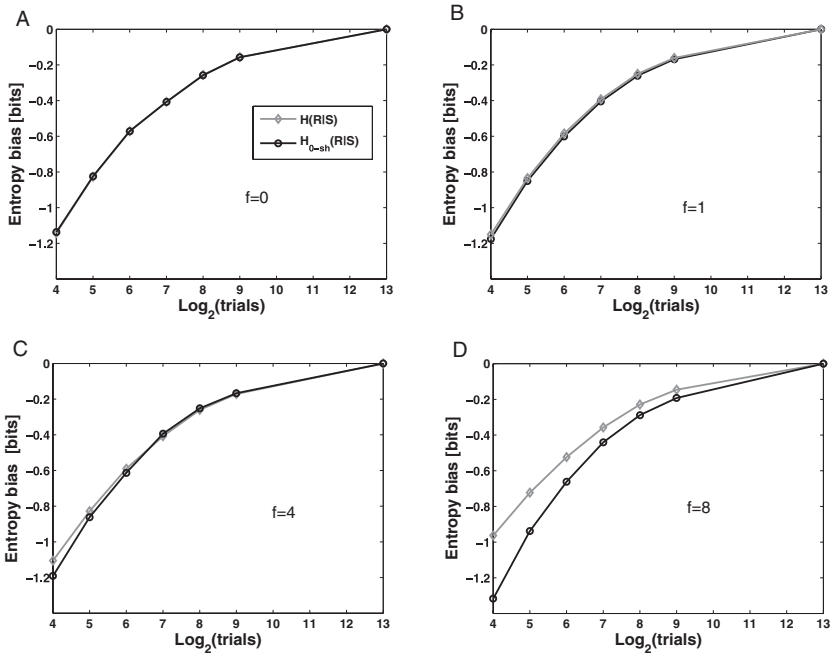

Figure 4: Effect of correlations on the performance of the shuffling estimators. We simulated spike trains as a first-order ($q = 1$) Markov process with constant rate ρ and a given Pearson correlation coefficient c. The value of ρ was taken from experimental data (Arabzadeh et al., 2004) as an average over a poststimulus time window of 40 ms after stimulus onset. The correlation coefficient c was then adjusted to produce different correlation strengths commensurate with correlation value found in the real data, c_0. The relationship between the real and the simulated correlation strength was given by a factor f as $c = fc_0$, where f is a multiplicative factor. (A–D) Plots of the bias of $H(\mathcal{R}|\mathcal{S})$ and $H_{0-sh}(\mathcal{R}|\mathcal{S})$ as a function of the number of trials per stimulus, for different values of the correlation strength f. No bias correction was used to compute these entropies.

by a $q = 3$ Markov model. A quadratic extrapolation procedure was used to correct for the bias. I_{LB-q}, ΔI_q, and ΔI_{q-sh} were computed for $q = 1$, 3, and 6 and plotted in Figures 5A, 5C, and 5E, respectively. Figure 5 confirms the intuition that as the value of q increases, I_{LB-q} becomes more biased. The asymptotic value of I_{LB-1} can be computed well even with as few as 32 trials (see Figure 5A). However, since the actual length of the Markov model used to generate the data was equal to $q = 3$, the asymptotic value of I_{LB-1} is only 0.69 bits, which is less than the true value of the full mutual information carried by the spike train (0.82 bits). This correction reflects the fact that some information would be lost by a decoder neglecting correlations

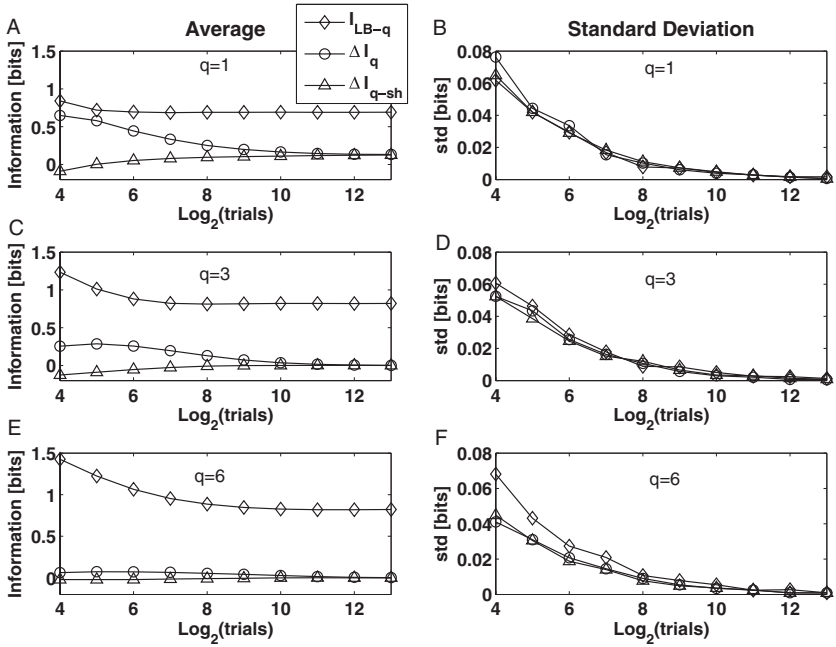

Figure 5: Bias and standard deviation of estimators making use of simplified Markov models of different orders q. We computed I_{LB-q}, ΔI_q, and ΔI_{q-sh} for q 1, 3, and 6 as a function of the number of trials per stimulus. Results are plotted as a function of the number of trials per stimulus. (A, C, E) Report of the average of these quantities over a number of repetitions of the simulation. (B, D, F) Report of the standard deviation across simulations. The simulated data were obtained exactly as in Figure 1, again considering a poststimulus window of 50 ms discretized into $L = 10$ bins of size $\Delta t = 5$ ms. It is worth noticing that the underlying Markov process to generate the simulated data was of order 3.

of range longer than one bin. On the contrary, I_{LB-3} and I_{LB-6} can reach the correct asymptotic value of information (see Figures 5C and E). However, they are considerably more biased than I_{LB-1}.

Figures 5A, 5C, and 5E show that the bias of ΔI_q has a completely different dependence on q: the higher q, the smaller the bias of ΔI_q. This is because as q increases, the difference $H(\mathcal{R}|\mathcal{S}) - H_q(\mathcal{R}|\mathcal{S})$ becomes smaller, and thus the overall bias of ΔI_q decreases. Similar to the $q = 0$ case considered in the previous section, the bias in ΔI_q can be further reduced by using a shuffling procedure. In fact, it is easy to construct a "q-shuffling" procedure that generate simulated responses $\mathbf{r}_{q-sh}$ preserving all marginals and correlations up to $q + 1$ consecutive time bins, but destroying all the

higher length correlations.[9] As in equation 4.3, we can construct from the probabilities of q-shuffled responses $\mathbf{r}_{q-sh}$ a "shuffled noise entropy of order q" $H_{q-sh}(\mathcal{R}|\mathcal{S})$, which converges to the same asymptotic value of $H_q(\mathcal{R}|\mathcal{S})$ for large numbers of trials but presents a higher bias for small trial numbers. The introduction of $H_{q-sh}(\mathcal{R}|\mathcal{S})$ allows us to define, in analogy to equation 4.4, a shuffled estimator ΔI_{q-sh}, as follows:

$$\Delta I_{q-sh} = H_{q-sh}(\mathcal{R}|\mathcal{S}) - H(\mathcal{R}|\mathcal{S}) + H(\mathcal{R}) - \chi_q(\mathcal{R}). \tag{4.6}$$

ΔI_{q-sh} must be less biased than ΔI_q, because there are substantial bias cancellations between $H_{q-sh}(\mathcal{R}|\mathcal{S})$ and $H(\mathcal{R}|\mathcal{S})$. This expectation is investigated numerically in Figure 5. Figures 5A, 5C, and 5E show the sampling behavior of ΔI_{q-sh} for $q = 1, 3$, and 6, respectively. The bias of ΔI_{q-sh} has a negative sign, for the same reasons given above for the case of ΔI_{0-sh}. However, for any q value, the asymptotic value of ΔI_{q-sh} for large trial numbers is the same as ΔI_q, but they differ significantly for low trial numbers. ΔI_{q-sh} is biased downward and is much less biased than ΔI_q.

Figures 5B, 5D, and 5F consider the behavior of the variances of the information-theoretic quantities as the value of q increases. The most notable finding is that like their bias, the variances of both ΔI_q and ΔI_{q-sh} also decrease with q. Moreover, for the same reasons presented above for $q = 0$, the variance of ΔI_{q-sh} is never higher than that of ΔI_q. Thus, the decrease of bias of ΔI_{q-sh} does not come at the expense of an increase of variance. This stresses the competitiveness of the shuffling procedures for the estimation of ΔI_q.

5 Tighter Upper and Lower Information Bounds Using Model Selection

We introduced information-theoretic quantities ΔI_q that quantify, for any value of memory length q, the information-theoretic cost of using a simplified decoding model described by a Markov model of order q rather than by the full-response probability distribution. Intuition suggests that the shorter the memory needed to decode the neuronal spike trains, the

[9] The "q-shuffled" responses $\mathbf{r}_{q-sh}$ can be constructed as follows. The response in the first time bin is chosen as a random permutation of the first time bins across all trials. Then the response in each bin $r(k)$ for $k = 2, \ldots, L$ is taken randomly without replacement from the subset of trials that satisfy $r'(k-j) = r(k-j)$ for $j = 1, \ldots, m$, where $m = \min(k-1, q)$. In other words, q-shuffled bins are concatenated with others taken at random without replacement from the subset of trials that have the same state of up to q past bins. Naturally, as q approaches the total length L of the time window, it could happen that if data are scarce, the q-shuffling procedure becomes trivial by regenerating exactly the original data set for each random shuffling. Thus, care should be taken to verify that this is not happening with the data under analysis

fewer data are needed to compute its information content. However, we have not explored yet the specific advantages offered by knowing that to decode all information, we need to consider only response chains extending over a number w of time bins, which is shorter than the L time bins making up the time window used to analyze spike trains.

Suppose that we know that the shorter history depth needed to decode all information is w time bins. This amounts to requiring that $\Delta I_q = 0$ for any $q \geq w$ and will happen if $P(s|\mathbf{r})/P_q(s|\mathbf{r})$ is not stimulus modulated for each response and each $q \geq w$. This condition will be met if the stimulus-conditional probability of the neuronal responses $P(\mathbf{r}|s)$ is generated according to a Markov process with length not higher than w for every stimulus. What are the further advantages offered to us in computing information quantities when we know the actual value of w?

If the shortest Markov order needed to decode all the information in the spike train is equal to w (i.e., $\Delta I_q = 0$ for any $q \geq w$), then any I_{LB-q} (with $q \geq w$) will be equal to the total mutual information I. Therefore, this knowledge offers a huge advantage when computing I. In fact, in this case, it will be most convenient to compute the true value of I by using I_{LB-w}, which will be equal to I in the asymptotic sampling regime, but it will be much less biased than I if w is much shorter than the window length L.[10]

The knowledge that the shortest Markov order needed to decode all the information encoded in the spike train is equal to w offers an advantage in the estimation of the quantities ΔI_q, which may still be larger than zero if $q < w$. Using the fact that $\Delta I_w = 0$, it is easy to show that all ΔI_q for $q < w$ are in this case equal to a simpler and much less biased quantity, called δI_q, and defined as follows:[11]

$$\delta I_q \equiv \Delta I_q - \Delta I_w = H_q(R|S) - H_w(R|S) + \chi_w(R) - \chi_q(R) \quad \text{if } q < w. \quad (4.7)$$

Since $\Delta I_w = 0$, it is clear than $\delta I_q = \Delta I_q$. However, when estimating the quantities from finite samples, δI_q is much less biased than ΔI_q. If $w < L - 1$, then term $H_q(R|S) - H_w(R|S)$ (the dominant term for the bias in equation 4.7) will have a resulting upward bias that will be smaller than the bias in the corresponding term in the definition of ΔI_q, namely, $H_q(R|S) - H(R|S)$. The smaller w is with respect to L, the better the sampling advantage is in using δI_q.

Along the same lines explained above, we can used the q-shuffling procedure to obtain a very weakly downward-biased estimator of δI_q. By

[10] In this case, one could in principle use any other I_{LB-q} (with $q > w$) to compute information I. However it is more efficient to use I_{LB-w} because the bias of I_{LB-q} grows with q. See Figure 5.

[11] The quantity δI_q should have a further index w, which we omit for compactness of notation.

computing the noise-entropy difference $H_q(R|S) - H_w(R|S)$ by using the "q-shuffled" distribution, we can define the following quantity:

$$\delta I_{q-sh} = H_{q-sh}(R|S) - H_{w-sh}(R|S) + \chi_w(R) - \chi_q(R). \tag{4.8}$$

In this case, the difference of entropies $H_{q-sh}(R|S) - H_{w-sh}(R|S)$ will have only a very small downward bias. $H_{q-sh}(R|S) - H_{w-sh}(R|S)$ will be less biased than its counterpart $H_{q-sh}(R|S) - H(R|S)$ in ΔI_q. Thus, δI_{q-sh} will be a tighter downward-biased estimator than ΔI_{q-sh}.

In summary, knowing the true shortest length of the Markov model w that can decode all information encoded by the spike train is useful because it leads to computing ΔI_q through two tight estimators, δI_q and δI_{q-sh}, which bound precisely the information from above and below, respectively, and that are tighter and less biased than the corresponding estimators ΔI_q and ΔI_{q-sh} obtained in the absence of knowledge about w.

In the next two sections we illustrate how the knowledge about w can reduce the sampling bias in a dramatic way. We consider two cases: when we are given some independent knowledge of the value of w and when we do not have this knowledge and have to guess w from the data.

5.1 Using a Prior Knowledge of the Shortest Markov Order Needed to Decode All Information. There may be situations in which there is precise knowledge about the memory length w needed to decode all the information encoded in the spike train. For example, when analyzing the information properties of some simulated data, we often have theoretical insights into the value of w. Alternatively, for some well-studied neural systems, there may be data available indicating that there is no correlation between spike times exceeding a certain time separation of w time bins. In this case, we are entitled to use straightforwardly the quantities δI_q and δI_{q-sh} as bounds to the true value of ΔI_q.

In Figure 6A we analyze again the simulated cortical responses to 49 stimuli (generated as in Figures 1 and 2 over $L = 10$ time bins) and show the behavior of various information estimators. (Here, a quadratic extrapolation bias correction was used.) As usual, I and ΔI_0 are strongly biased upward and require at least 2^{10} trials per stimulus to give estimates close to the asymptotic values. In contrast, ΔI_{0-sh}, which is also computed without any assumption on the order w of the Markov model underlying the real spike train, gives an accurate estimate at around 256 trials per stimulus. What is the effect of making use of the knowledge that the true order w of the underlying Markov model is equal to 3? This amounts to using δI_q and δI_{q-sh} in equations 4.7 and 4.8 with $w = 3$. Figures 6A and 6B show that δI_0, δI_{0-sh} provide much tighter and data-robust upper- and lower-bound estimations than ΔI_q, ΔI_{q-sh}. As expected by the above considerations, δI_0 is a much tighter upward-biased estimator than the direct use of ΔI_0. This

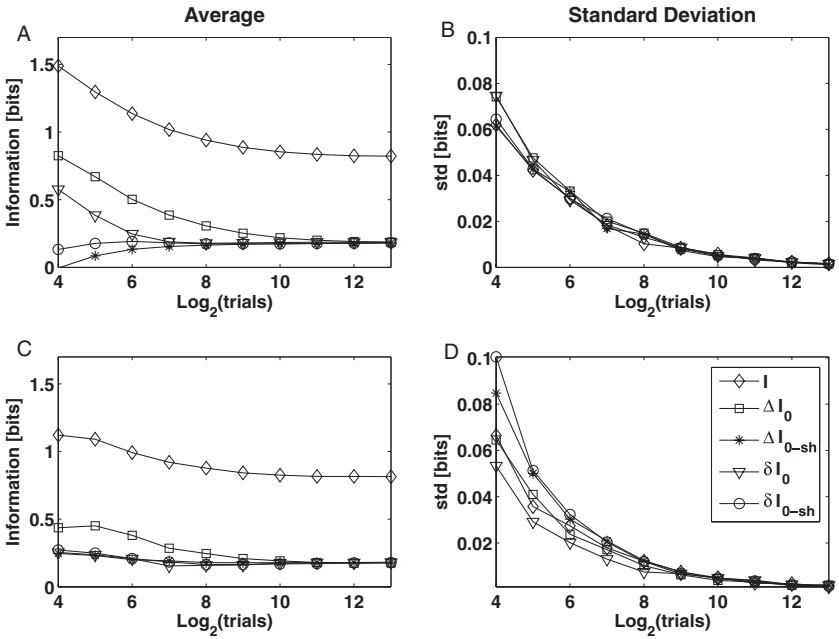

Figure 6: Performance of estimators δI_0 and δI_{0-sh} when the Markov order of the underlying model is known in advance. In addition to δI_0 and δI_{0-sh}, for comparison we also show the average values of $I(R, S)$, ΔI_0, ΔI_{sh} as a function of the number of trials per stimulus. (A, C) Report of the average of these quantities over a number of repetitions of the simulation. (B, D) Report of the standard deviation across simulations. The simulated data were obtained exactly as in Figure 1, again considering a poststimulus window of 50 ms discretized into $L = 10$ bins of size $\Delta t = 5$ ms. The underlying Markov process used to generate the simulated data was of order 3. (A) Data corrected with the quadratic extrapolation method. (B) Standard deviations of the data shown in *B*. (C) Estimation based on the method of Nemenman et al. (2004). (D) Standard deviations of the data shown in *C*.

is because the knowledge of $w = 3$ ensures that we compute the functionals in equation 4.7 on a response space with 2^4 different responses, a number much smaller than the 2^{10} possible responses characterizing the probability distribution over 10 bins. The downward-biased shuffled estimator δI_{0-sh} is even tighter than ΔI_{0-sh}, whcih provides reliable estimates of information with as few as 64 trials per stimulus.

The standard deviation of the estimators δI_0 and δI_{0-sh} is considered in Figure 6B. It is apparent that the reduction in bias of δI_0 and δI_{0-sh}, with respect to ΔI_0 and ΔI_{0-sh}, is not obtained at the expense of an increase in variance.

In Figure 6C we show the performance of the same estimators obtained with the NSB method bias-correction method rather than with the quadratic extrapolation procedure. The comparison with Figure 6A shows that the direct estimation of I and ΔI_0 is less biased than that obtained with the quadratic extrapolation. However, this increase in performance is greatly enhanced when using the δI_0 and δI_{0-sh}. The NSB method is in general effective, and ΔI_{0-sh} and δI_{0-sh} gave a good estimate of the asymptotic value of ΔI_0, even down to 128 trials per stimulus. A potential problem with using the NSB method in this context is that although it performs well, in general it does not preserve the sign of the residual bias. Therefore, for very low trial values, it could happen that the downward bias quantities ΔI_{0-sh} and δI_{0-sh} occasionally become higher than the asymptotic values of ΔI_0.

Figure 6D compares the standard deviations of the estimates obtained with the NSB method. A comparison of the estimates and the performance of the quadratic extrapolation suggests that the reduction of bias obtained using the NSB method comes at the expense of a relatively small but appreciable increase in variance and that this increase is higher for the shuffled estimators.

5.2 Selecting the Order of the Decoding Model with a Nonparametric Test. In general, when analyzing real spike trains recorded during a neurophysiological experiment, we do not have precise prior knowledge of the temporal extent of correlations between spikes. How can we take advantage of the sampling properties offered by δI_0 and δI_{0-sh} to the case in which we do not know the value of w a priori? In this section, we address this problem and show that nonparametric statistical techniques (Efron & Tibshirani, 1993) can be used to compute empirically an effective shortest Markov order w needed to decode all information encoded by the spike train.

The crucial property in the derivation of the expression for δI_q was that $\Delta I_q = 0$ for all $q \geq w$. Therefore, we can suggest a simple way to determine effectively a statistical procedure to estimate the value of the memory length w to be inserted in the definition of δI_q. This value can be defined as the w representing the the highest value of Markov order above that all ΔI_q are zero for any $q > w$. In practice, this is not easy to determine from data as individual values of ΔI_q that should be zero in the asymptotic large-N limit may be estimated as greater than zero because of statistical fluctuations or because of a residual bias not entirely corrected for (see, e.g., Figure 5). However, the likelihood of an individual outcome of ΔI_q to be significantly greater than zero could be established easily with a statistical test evaluating the null hypothesis that ΔI_q is zero. Once a particular statistical test has been chosen, the value of w can be determined as follows. We start from the largest possible value of q for which ΔI_q can be > 0 (i.e., $q = L - 2$), and we use the statistical test to determine whether ΔI_q is significantly more than

0. If this is the case, we cannot discard any correlation of any length, and we set $w = L - 1$. If instead the null hypothesis that ΔI_{L-2} is zero cannot be discarded, we repeat the process by decreasing q at each step and test again whether ΔI_q is significantly different from zero. The parameter q is then decreased until we find that ΔI_q is significantly more than 0. At this point, the procedure is stopped, and we take w as the last value of q for which the null hypothesis could not be discarded.

A simple way to test the null hypothesis that $\Delta I_q = 0$ for some q is to use the following nonparametric "bootstrap" test (Efron & Tibshirani, 1993). We first generated a set of bootstrapped data sets obtained by the q-shuffled procedure described above by destroying all correlations longer than $q - 1$ bins. We then use these q-shuffled responses to compute ΔI_q. For each random realization of these shuffled responses, the corresponding ΔI_q must be zero. However, because of statistical fluctuations and errors in fully eliminating the bias, the distribution of ΔI_q will peak above zero. We can use this empirically generated distribution to set boundaries on accepting the null hypothesis that the value of ΔI_q computed with the original data set is zero.

In our simulations, we found that the distribution of values of ΔI_q computed on the q-shuffled distribution was approximated by a gaussian (data not shown). Thus, our statistical criterion for rejection was simply that the actual value of ΔI_q computed with the original data set was higher than two standard deviations above the mean of the bootstrapped distribution (the cutoff at two standard deviation was chosen because it gave the best estimates of information quantities when applied to the simulated data shown below).

In Figures 7A and 7B, we show the resulting histograms after applying the bootstrap test to simulated data. The simulations were again the same as in Figures 1 and 2, simulating a somatosensory neuron with underlying memory length $w = 3$ and a response word of length $L = 10$. In Figure 7A, we plotted the distribution of estimated w values obtained with the bootstrap test when 128 trials per stimulus were available. This distribution is broad and peaked at the correct order of the simulations. As the number of trials per stimulus increases to 256 (see Figure 7B), the test becomes more effective as more of the estimated values of w are correct. The trend continues as the number of trials increases, and the test almost always reports the correct w for number of trials larger than 512 (results not shown).

In Figures 8A and 8B, we plotted δI_0 and δI_{sh-0}, computed from equations 4.7 and 4.8, with w determined for each simulation through the test described above (values were corrected with the quadratic extrapolation method). Figure 8A shows that even in the case in which w is not known a priori but must be determined through a statistical test, the lower bound given by δI_{0-sh} has already reached the asymptotic value of ΔI for as small as 2^7 trials per stimulus. The upper bound, δI_0, converges more slowly toward the asymptote; it represents a significant improvement with respect

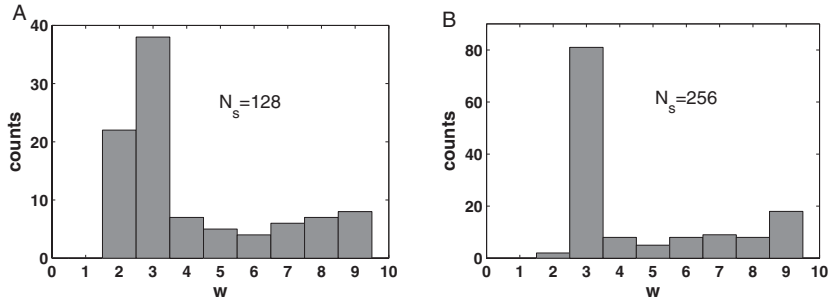

Figure 7: Bootstrap test to estimate model complexity. We simulated spike trains as in Figure 1 using a Markov process of order 3. The poststimulus time window used in the analysis was of 50 ms and was discretized into $L = 10$ bins with resolution $\Delta t = 5$ ms. For each realization (out of a total of 100) of the simulated spike trains, we estimated w using the bootstrap test described in the text. (A, B) Histograms of the distribution of w values obtained in this way for 128 and 256 trials per stimulus, respectively.

to the direct estimation of ΔI_q. δI_{0-sh} and δI_0 bound the true value of ΔI with an error of less than 10% for 256 trials per stimulus, and $L = 10$. Thus, δI_{0-sh} and δI_0 behave much better than $I(R, S)$, ΔI, and ΔI_{0-sh}, quantities that do not depend on the determination of the order of the process.

In Figure 8C we report the information values estimated using the NSB method. It is apparent from Figure 8C that both δI_{0-sh} and δI_0 perform much better than for the quadratic extrapolation case shown in Figure 8A. The estimates bound the exact value of ΔI_0 less than 5% for as low as 128 trials per stimulus.[12]

The above results were obtained using simulated data generated by a Markov process of order $w = 3$. To check that this procedure also worked for data generated by higher-order Markov processes, we repeated the analysis on synthetic data simulating a somatosensory neuron with different values of Markov order ranging from $w = 3$ to $w = 8$, and the usual response word of length $L = 10$. We found results consistent with that of Figures 7 and 8 (data not shown). In brief, the bootstrap test continued to give estimates of w peaked around the correct highest value of w with $\Delta I_w > 0$ (which in this simulation was constructed to be equal to the w value used to generate the data). The distributions of values obtained with the bootstrap test were

[12]The NSB method performed less well than the quadratic extrapolation only when computing low-dimensional entropies such as $H_0(R|S)$ (see appendix A). Thus, a useful practical consideration is that when pairing the bootstrap test with the NSB correction method, it is safer to use a quadratic extrapolation to compute entropies such as $H_0(R|S)$ that appear in equations 4.7 and 4.8 when the statistical test suggests a value of $w = 0$ as the most likely.

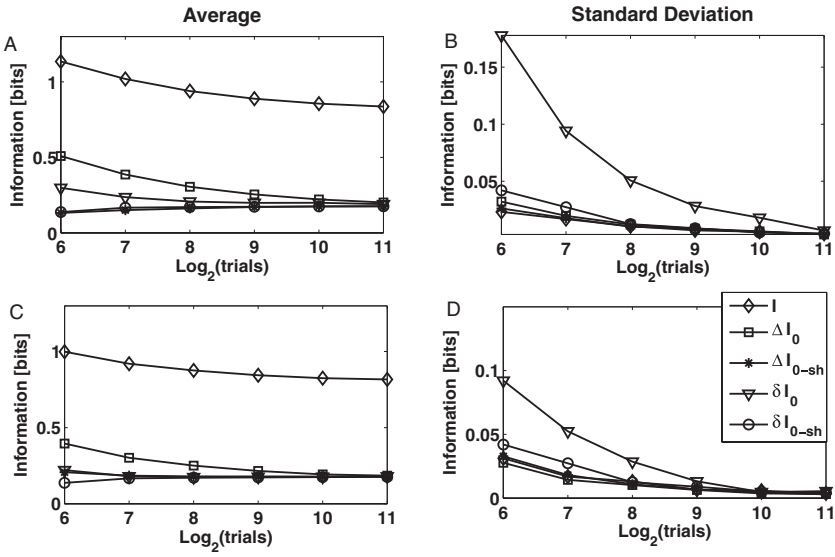

Figure 8: Performance of estimators δI_0 and δI_{0-sh} when the Markov order of the underlying model is selected with the bootstrap test. In addition to δI_0 and δI_{0-sh}, for comparison we also show the values of $I(R, S)$, ΔI_0, ΔI_{0-sh} as a function of the number of trials per stimulus. The data correspond to simulated spike trains as in Figure 1 using a Markov process of order $w = 3$. However, the knowledge of the true Markov order w used to generate the data was not used; w was instead estimated on each simulation using the bootstrap test described in the text. (A, C) Report of the average of these quantities over a number of repetitions of the simulation. (B, D) Report of the standard deviation across simulations. (A) Data corrected with the quadratic extrapolation method. (B) Standard deviations of the data shown in *A*. (C) Estimation based on the method of Nemenman et al. (2004). (D) Standard deviations of the data shown in *C*.

slightly narrower for simulated processes with higher w values, probably because ΔI_q has a smaller variance for higher q values (see Figure 5). As the number of trials increased, the information quantities δI_0 and δI_{0-sh} always converged (from above and below, respectively) to their correct asymptotic value and were less biased than ΔI_0 and ΔI_{0-sh}. Obviously, the lower the Markov order used to generate the process, the bigger was the bias advantage of δI_0 and δI_{0-sh} over ΔI_0 and ΔI_{0-sh}.

Using a simplified model of the Markov family and selecting the most economic decoding model with a statistical test leads to drastic reductions of the bias of the information quantities. However, it is likely that future extensions of this work to include other families of simplified models may lead to even better performance. For example, it is likely that in the

presence of serial correlation of adjacent but long interspike intervals, selecting among a class of suffix tree model would perform much better than selecting only among full Markov models, because the depth of history of a suffix tree could be made to depend on when and if the previous spike occurred (Kennel et al., 2005).

6 Analysis of Real Data

In this section we illustrate the application of the methods developed above to two different data sets of real neuronal recordings from the whisker representation in somatosensory ("barrel") cortex of rats anesthetized with urethane.

The first data set consisted of 21 neural clusters, each recorded from a different electrode in a silicon array made up of a total of 100 electrodes (see Petersen & Diamond, 2000, for further details). Spike times from each electrode were determined by a voltage threshold set to a value 2.5 times the root mean square voltage. Since it was not possible to sort well-isolated units from each channel, spikes from the same recording channel were all considered together as a single neural cluster. It has been estimated that each cluster captured the spikes of approximately two to five neurons (see Petersen & Diamond, 2000). Neural activity was recorded in response to individual stimulation of one of nine different whiskers (whisker D2 and its eight nearest neighbors); individual whiskers were stimulated near their base by a piezoelectric wafer, controlled by a voltage generator. The stimulus was an up-down step function of 80 μm amplitude and 100 msec duration, delivered once per second. The trials per stimulus available were 200 to 500 for this first data set.

The time course of the estimates of the information transmitted about stimulus location by spike times from a neuronal cluster in response to instantaneous whisker deflection is reported in Figure 9A. For the information analysis, we considered the response $\mathbf{r}$ to be a spike timing code (binarized with temporal resolution $\Delta t = 5$ ms) defined in a poststimulus time window that began at 5 ms poststimulus and whose end was gradually increased in 5 ms steps up to 55 ms poststimulus. The 21 clusters in the data set were analyzed separately and then averaged. The estimation of all the entropy quantities was done using the NSB method (similar conclusions were reached using the quadratic extrapolation; results not shown). The full spike timing mutual information was computed through the estimator I_{sh} and was found to increase smoothly along the entire time range analyzed. Consistently with the fact that under these conditions neurons typically stop firing after 40 to 50 ms, I_{sh} also saturated after 40 to 50 ms. We also computed the temporal evolution of ΔI_0 and ΔI_{0-sh}. ΔI_{0-sh} was small over all time ranges and accounted for no more than 8% of the total information I_{sh}. ΔI_0 was also small, but unlike ΔI_{0-sh}, it increased markedly for longer time windows. This indicates that ΔI_0, unlike ΔI_{0-sh}, suffers

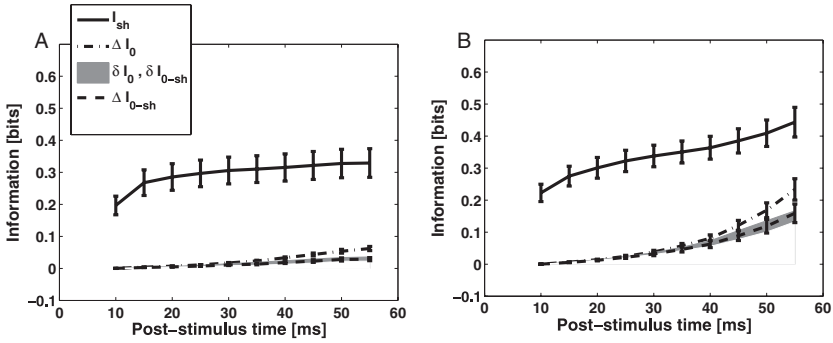

Figure 9: Analysis of rat somatosensory cortex data. We show the total mutual information I_{sh}, ΔI_0, and ΔI_{0-sh}. The gray area shows the region delimited by the upper and lower bounds δI_0 and δI_{0-sh}. The NSB method was used to correct for finite sampling. (A) Information about stimulus location conveyed by individual neural clusters recorded in rat somatosensory cortex in response to instantaneous whisker deflections (Petersen & Diamond, 2000). Results reported as average (± SEM) over 21 different clusters. (B) Information about amplitude and frequency of sinusoidal whisker vibration conveyed by individual neural clusters recorded in rat somatosensory cortex (Arabzadeh et al., 2004). Results reported as average (± SEM) over 24 different clusters.

sampling problems when the number of bins L is more then 8 to 10. The fact that ΔI_{0-sh} is a good estimator of the asymptotic value of ΔI_0 is confirmed by the computation of the tight upper and lower bounds δI and δI_{sh}. The latter quantities were computed by selecting the w value required for their computation with the bootstrap test. The mean value of w across channels was 3.7 ± 2.8 (mean ± SEM). In Figure 9A the gray region represents the area enclosed by the upper and lower bounds δI_0 and δI_{0-sh}. ΔI_{0-sh} remained within the two bounds along the whole timescale considered. In contrast, ΔI_0 overshot the data-robust upper bound δI_0 for longer time windows, indicating that it was suffering from sampling problems in that time range. To summarize, this analysis shows that the use of I_{sh} and ΔI_{0-sh} gives precise estimates of the asymptotic values of I and ΔI_0 for up to 10 time bins. This an excellent performance considered that the number of trials per stimulus was 200 to 500. It is much beyond what could be achieved, for example, with a standard direct method to estimate ΔI_0.

The second data set consisted of 24 recordings of neural clusters, again from rat barrel cortex, obtained with the same type of electrodes and anesthesia as above. In this case, the stimulation protocol was different. The set of stimuli consisted of 49 different types of sinusoidal whisker vibrations, each defined by a unique combination of amplitude and frequency of vibration and delivered for 500 ms (see Arabzadeh et al., 2004, for full

details). The trials per stimulus available were 200. As before, we considered the response $\mathbf{r}$ to be a spike timing code (resolution $\Delta t = 5$ ms) defined in a poststimulus time window starting at 5 ms poststimulus and gradually increasing in 5 ms steps up to 55 ms poststimulus.

The time courses of the estimates of the information transmitted by spike times from a single neural cluster (averaged over the set of available clusters) about the parameters of whisker vibration are reported in Figure 9B. In contrast to the previous case, the spike timing mutual information (computed as I_{sh}) continued to grow over time, consistent with the fact that neurons show stimulus-evoked activity for several hundred ms in response to these sinusoidal vibrations (Arabzadeh et al., 2003). As in the previous example, both I_{sh} and ΔI_{0-sh} increased smoothly over time (with no sudden upward jump reflecting a potential bias problem). The accuracy of ΔI_{0-sh} in estimating the asymptotic value of ΔI_0 was substantiated by the fact that it was tightly bound by δI_0 and δI_{0-sh} (gray area) up to 55 time bins. (δI_0 and δI_{0-sh} were computed by selecting the w value with the bootstrap test. The estimated mean value of w across channels was 6.2 ± 1.6.) In contrast, ΔI_0 overshot the gray area for long windows, indicating it suffers from an upward bias problem in this time range. A potentially interesting neurophysiological observation is that unlike in the case of instantaneous whisker deflection, ΔI_{0-sh} is now considerable (29% of the total information I_{sh} at 50 ms poststimulus), and it grows supralinearly over time. This suggests that correlation may be useful in decoding complex whisker vibrations from cortical neuronal activity and is consistent with the approximated analytical prediction of Panzeri and Schultz (2001), which suggests that when neurons keep firing over a sustained period of time, ΔI_0 is expected to grow at least quadratically as a function of time.

7 Discussion

Using information theory to probe the neural code over fine time resolutions, large populations or large time windows remain technically difficult because the analysis of long sequences of spikes requires the collection of large amounts of data to sample the probability of occurrence of each possible spike sequence. Under this condition, information measures suffer from serious bias problems, which impose severe limitations to the range of timescales and population sizes available for analysis (Panzeri & Treves, 1996; Strong et al., 1998).

In this letter, we have developed a new procedure to estimate the information carried by spike trains that drastically alleviates its sampling problems. The starting point of the procedure is the observation that if we break down the mutual information I into a Kullback-Leibler divergence ΔI_{simp} (which bounds the information lost by decoding when ignoring stimulus-dependent correlations) and the rest (called $I_{LB\text{-}simp}$), then almost all the bias usually comes from ΔI_{simp}. The second key observation is that the bias of

ΔI_{simp} (and, as a consequence, that of the total mutual information) can be drastically reduced (and be made negative) at no increase of variance by an appropriate shuffling procedure. The third key observation is that the bias of ΔI_{simp} (and thus that of the mutual information) can be further reduced by the use of a nonparametric test to find the minimal complexity within a class of models of correlation that still permits computing ΔI_{simp} correctly.

From a practical point of view, the overall reduction of the bias is useful because it extends the domain of applicability of information theory. To appreciate how this domain is extended, we note that our numerical examples, based on a response space made of 2^{10} different responses, showed that our techniques eliminated the bias even with number of samples as low as 32 trials per stimulus, a number of the order of the square root of the number of different responses. This is a significant advance over the previous requirement that the number of trials is comparable to the number of different responses. This effectively allows one to double the timing precision, the time window length, or the population size that can be analyzed. This fact is timely because large-scale recording techniques are rapidly becoming available, and because there are theoretical arguments (reviewed in Averbeck, Latham, & Pouget, 2006) that suggest that the impact of correlations to the neural code may be particularly important when considering larger populations. The bias-reduction techniques open up the possibility of analyzing populations twice as big than those previously considered in information-theoretic studies of neural codes. Although this is a significant step forward, it is important to recognize that due to the "dimensionality curse," even this advance will not ultimately be enough for the direct analysis of very large populations or very long temporal sequences at very fine timescales. Alternative approaches that may work better in these more extreme situations include algorithms not relying on binning (Victor & Purpura, 1997; Victor, 2002).

The reduction of the bias problem may also help in extending the information analysis to the domain of graded brain signals, such as fMRI or local field potentials (Logothetis, 2003). These graded signals are more difficult to analyze because, unlike spikes, they cannot simply be converted into a binary word sampled with a certain temporal precision. The techniques presented here may lend themselves to the analysis of LFPs/fMRI, by first discretizing the graded signals into a number of different levels, characterizing the correlation structure among them, and finally fitting it to a low-dimensional stochastic model.

A useful property of the shuffled information estimator presented here is that besides reducing the overall magnitude of the bias when compared to a direct procedure, it makes the bias negative rather than positive. This downward bias property is useful in practical studies of neural codes because a finding of significant extra information in spike timing obtained with this new method will ensure that this additional spike timing information is genuine and not an artefict due to sampling problems.

An important technical step in the reduction of the mutual information bias was the selection, within a predefined class, of the minimal complexity of the simplified model of correlation that still captures the correct value of ΔI_{simp}. This was achieved in practice by introducing a parametric family of Markov models to approximate the correlation structure of the real data and by using a nonparametric bootstrap statistical test to select the order q of the Markov model that best described the neural response. While we showed that the simple nonparametric test and the simple class of models described here are effective at reducing the data constraints in information calculations, it is important to note that neither of these steps must necessarily be performed exactly in this form. Particularly interesting families of maximum-entropy simplified correlation models are those considered by Amari (2001) and Schneidman, Berry, Segev, and Bialek (2006). The model selection can be also performed in different ways, for example, through log-likelihood ratio model selection or other types of inference (see, e.g., Cover & Thomas, 1991; Kennel et al., 2005). An important step of future research is to understand which class of models describes neural data more economically and which statistical model selection technique is more powerful under different circumstances.

In recent years, there has been a debate on how best to measure the role of correlations in neural coding (see Nirenberg & Latham, 2003; Pola et al., 2003; Schneidman, Bialek, & Berry, 2003 for different points of view). The measure used in this letter is ΔI, which was proposed by Nirenberg and Latham (2003). It has an interpretation as an upper bound to the information lost by a decoder that neglects correlation. In this letter, ΔI was used to break-down the information into mildly biased and strongly biased components and to obtain a more data-robust estimator of the total mutual information through this breakdown. The considerable sampling advantages in the computation of the mutual information obtained in this way would also be available to situations when the computation of ΔI is not of interest, either because the only purpose is to quantify mutual information or because other measures of the importance of correlation are used. In fact, the other measures proposed (Pola et al., 2003; Schneidman et al., 2003) quantify the importance of correlation as differences of certain mutual information quantities; therefore, the bias of each such information quantity could be substantially reduced with the techniques presented here. The bias reduction procedure presented here is thus useful to better compute other measures of the importance of correlations on coding.

In summary, the combination of the simplified models and the shuffling methods allowed us to extend the range of applicability of information theory to neuronal responses. Usually computing information accurately with a straight application of the best bias correction methods available requires a sample size comparable to the number of possible different neural responses. By applying the same bias correction techniques to the shuffled estimators after a model selection procedure, it is now possible to estimate

accurately information quantities with amounts of data one order of magnitude smaller.

Appendix A: The NSB Method

In this appendix we sketch some theoretical considerations and present some additional numerical simulations on the performance of the NSB bias correction method (Nemenman et al., 2004) when entropies are computed from response spaces of different sizes and reflecting different firing rates. While these considerations follow straight from Nemenman et al., they are helpful to understand why this method is not suited for correcting I_{LB-q} at low q values, especially at low firing rates. For simplicity, we focus on correcting the response entropy $H(\mathcal{R})$. However, similar considerations hold for the stimulus-conditional entropies.

The NSB method (Nemenman et al., 2004) is rooted in the Bayesian inference approach to estimate the entropy $H(\mathcal{R})$ (a function of a generally unknown underlying probability $P(\mathbf{r})$). The Bayesian approach assumes that there exists some a priori probability density function $\mathcal{P}_{pr}(P)$ in the continuum space of all the possible probability distributions on the response space $\mathcal{R}$ with R elements. The Bayesian estimation of the entropy after the observation of $\{n(\mathbf{r})\}$ (the experimental number of times $\{n(\mathbf{r})\}$ in which each response is observed) can be computed as an average of the corresponding entropy over all the possible hypothetical probability distributions weighted by their conditional probability given the data.

Unless we have some other criteria to select a preferred value in the entropy range $[0, \log_2 R]$, we would like to have a flat a priori distribution of entropy. However, the choice of the prior has a strong impact on the value of H^{Bayes} unless very many data are observed. Nemenman et al. (2004) have addressed this problem by using a mixture of Dirichlet priors, the latter being defined in terms of a parameter β ranging between 0 and ∞. Nemenman et al. have shown that after fixing β (i.e., after choosing the prior within the family), the Bayesian estimate of the entropy is sharply defined and monotonically dependent on the parameter β (naturally, until the number of samples becomes large, in which case the likelihood dominates the estimate). It can be shown that at fixed β, the variance of entropy estimation before any observation (i.e., when $n(\mathbf{r}) = 0$) scales as $1/R$ as R grows, and it is thus small compared to the range of possible entropy values $[0, \log_2 R]$. Therefore, the goal of constructing a prior on the space of probability distributions that generates a nearly uniform distribution of entropies can be approximately achieved by an average over the Bayesian estimation over all the one-parameter Dirichlet family of priors labeled by β.

While this procedure is designed to work well for large R, it may work less well for small values of R. In fact, if R is small, then the variance of a priori entropy estimation is not small anymore with respect to the range of possible entropies. Thus, the result of integration over β will not be flat

but will typically be smaller near the edges 0 and $\log_2 R$ than in the central region of possible entropies. Thus, the method is likely to give problems in the estimation of low entropy values for low R values. In particular, the NSB method is likely to give problems when estimating the entropy of processes generated with very low firing rates. In this case, since the probability of observing a spike in each bin is low, the entropy value would be much nearer zero than to $\log_2(R)$; since in this case, the NSB prior distribution of entropy values is instead higher in the central part of the interval $[0, \log_2 R]$, the resulting NSB estimation of a low-firing rate, the entropy for low R may strongly overestimate the entropy unless many experimental observations are available. In the latter case, asymptotic bias correction procedures might work better anyway.

We tested these considerations by applying the NSB method to a homogeneous Poisson process. The spike times generated by the Poisson processes were binned using a bin size of $\Delta t = 5$ ms. In Figure 10 we report the performance of the NSB by comparing it to the quadratic extrapolation procedure and the uncorrected measure of the entropy. We tested two different values of R and two different firing rates. Figures 10A and 10B compare the estimations of the entropy using $R = 2$ for a firing rate of 2 Hz in Figure 10A and 40 Hz in Figure 10B. In both cases the estimation with the NSB method performs worse than the extrapolation procedure of Strong et al. (1998). Figures 10C and 10D show the estimators applied to Poisson processes with the same firing rates as with Figures 10A and 10B, respectively, and using $R = 2^6$. It is apparent now that the NSB performs substantially better than the quadratic extrapolation procedure. We consistently found that for higher R values, the NSB method was always more competitive than the quadratic extrapolation (data not shown, but see Figure 2 for the $L = 10$ case).

When the NSB estimator is clearly the best estimator (low N and large R; see Figures 10C and 10D), it is dominated by the prior. Thus, a potential concern is that this superb performance might be specific to the Poisson process used in the simulation, perhaps because this process is a good match to a distribution within the family of Dirichlet priors. It is conceivable that for some class of strongly correlated response distributions, there may not be a good match in the Dirichlet family, and thus the superiority of the NSB method in the large-R regime may suffer. To test for this potential problem, we repeated the analysis in Figure 10 using the same correlated processes used to generate the data in Figures 1 and 4 (and described in the main text). We found results consistent with those plotted in Figures 10C and 10D. Although these results cannot rule out completely the above concern, they suggest that the NSB method will perform well in the large R regime on a wide range of processes with realistic neuronal statistics.

Since the Markov noise entropies $H_q(\mathcal{R}|\mathcal{S})$ of order q can be written as sums of entropies defined over q adjacent time bins, it follows that the NSB method is not suited for correcting I_{LB-q} at low q values, especially at low

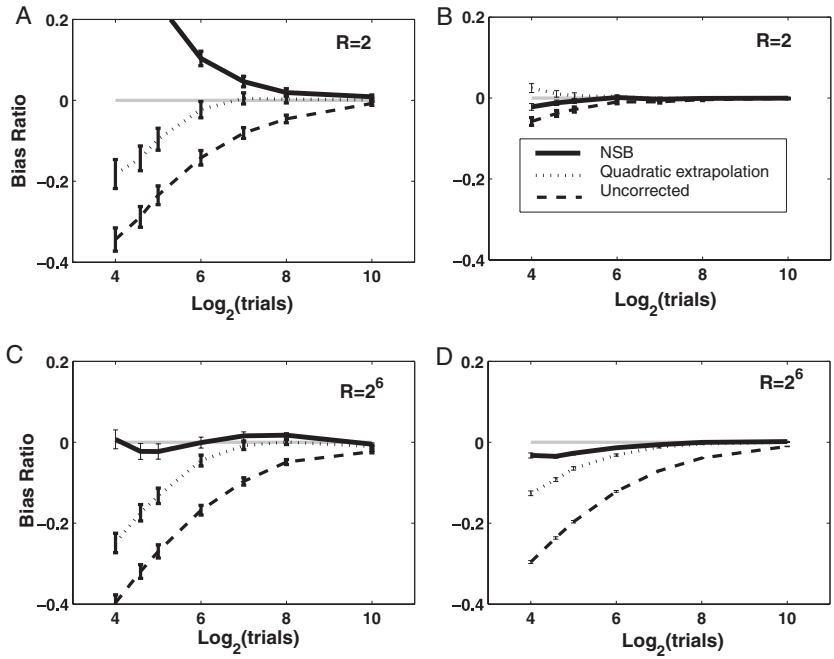

Figure 10: Bias in the NSB method. Average and standard error of the bias ratio (defined as the bias divided by the true asymptotic value of the entropy) for the NSB, quadratic extrapolation, and uncorrected estimators of the entropy. The estimators were applied to homogeneous Poisson process. The generated spike times were binned using a bin size of $\Delta t = 5$ ms. The average and standard error were computed over 1000 realizations of the simulations. (A, B) Panels correspond to $R = 2$, and the rates of the homogeneous Poisson process were 2 Hz in A and 40 Hz in B. In both A and B the quadratic extrapolation outperforms the NSB method. (C, D) Panels correspond to $R = 2^6$, and the rates of the Poisson processes were respectively the same as in A and B. For larger R, the NSB method performs better than the quadratic extrapolation.

firing rates. However, it appears to be an excellent method for correcting the quantities involving entropies defined over long response times, such as I_{LB-q} at high q values.

Appendix B: A Link Between Our Assumptions and the Maximum-Entropy Principle

A direct contact between assumption 2 and the maximum-entropy principle can be made as follows. Consider a one-dimensional family of simplified

models $P_{simp}(\mathbf{r}|s)$ that are obtained from the true $P(\mathbf{r}|s)$ as a one-dimensional trajectory in probability space parameterized by λ:

$$P_{simp}(\mathbf{r}|s) = P(\mathbf{r}|s) + \lambda u(\mathbf{r}|s), \tag{B.1}$$

with the "modulator" u satisfying the normalization $\sum_{\mathbf{r}} u(\mathbf{r}|s) = 0$. By computing the zeroes of the derivative of the entropy of $P_{simp}(\mathbf{r}|s)$ with respect to λ, it is easy to show that the entropy of such $P_{simp}(\mathbf{r}|s)$ is maximized (as a function of λ) when λ is chosen so that

$$\sum_{\mathbf{r}} u(\mathbf{r}|s) \log_2(P(\mathbf{r}|s) + \lambda u(\mathbf{r}|s)) = 0. \tag{B.2}$$

Using equation B.1, the maximum entropy condition in equation B.2 can be rewritten as

$$\sum_{\mathbf{r}} (P(\mathbf{r}|s) - P_{simp}(\mathbf{r}|s)) \log_2(P_{simp}(\mathbf{r}|s)) = 0 \tag{B.3}$$

which is exactly the condition 3.2 requested by our assumption 2. Thus, for any simplified model belonging to the family defined in equation B.1, the only one that satisfies our assumption 2 is the model with the highest entropy within the family.

This parametric entropy maximization can be related to the construction of the classes of maximum-entropy models developed in section 3.3, for example by considering the extremization of entropy with respect to a large number of modulator functions. Thus, assumption 2 is related to a maximum entropy principle.

Acknowledgments

We are grateful to R. Petersen, M. E. Diamond, and E. Arabzadeh for many useful discussions and for kindly making available to us the example data used in Figure 9. Gianni Pola contributed to the early stages of this work. We are indebted to the anonymous referees for useful insights, particularly on the relation between our work and the maximum entropy principle. This research was supported by the International Human Frontier Science Program (M.A.M.), Pfizer Global Development (R.S.), Wellcome Trust 066372, and the Royal Society.

References

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). Spatio-temporal firing patterns in the frontal cortex of behaving monkeys. *J. Neurophysiol.*, *70*, 1629–1638.

Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Trans. Inform. Theory*, *47*, 1701–1711.

Arabzadeh, E., Panzeri, S., & Diamond, M. E. (2004). Whisker vibration information carried by rat barrel cortex neurons. *J. Neurosci.*, *24*(26), 6011–6020.

Arabzadeh, E., Petersen, R. S., & Diamond, M. E. (2003). Encoding of whisker vibration by rat barrel cortex neurons: Implications for texture discrimination. *J. Neurosci.*, *23*(27), 9146–9154.

Arabzadeh, E., Zorzin, E., & Diamond, M. E. (2005). Neuronal encoding of texture in the whisker sensory pathway. *PLOS Biology*, *3*(1), 0155–0165.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, *7*, 358–366.

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

Brosch, M., Bauer, R., & Eckhorn, R. (1997). Stimulus dependent modulations of correlated high frequency oscillations in cat visual cortex. *Cerebral Cortex*, *7*, 70–76.

Buracas, G. T., Zador, A. M., DeWeese, M. R., & Albright, T. D. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron*, *20*, 959–969.

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

Dan, Y., Alonso, J.-M., Usrey, W. M., & Reid, R. C. (1998). Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nature Neuroscience*, *1*, 501–507.

de Ruyter van Steveninck, R., Lewen, G., Strong, S., Koberle, R., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, *21*, 1805–1808.

DeWeese, M. R., Wehr, M., & Zador, A. M. (2003). Binary spiking in auditory cortex. *J. Neurosci.*, *23*, 7940–7949.

Dimitrov, A. G., & Miller, J. P. (2001). Neural coding and decoding: Communication channels and quantization. *Network: Comput. Neural Syst.*, *12*, 441–472.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Gawne, T. J., & Richmond, B. J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.*, *13*, 2758–2771.

Golledge, H. D. R., Panzeri, S., Zheng, F., Pola, G., Scannell, J. W., Giannikopoulos, D. V., Mason, R. J., Tovee, M. J., & Young, M. P. (2003). Correlations, feature binding and population coding in primary visual cortex. *Neuroreport*, *14*, 1045–1050.

Gray, C. M., König, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, *338*, 334–337.

Johnson, D. H., Gruner, C. M., Baggerly, K., & Seshagiri, C. (2001). Information-theoretic analysis of neural coding. *J. Comput. Neurosci.*, *10*, 47–69.

Kennel, M. B., Shlens, J., Abarbanel, H. D. I., & Chichilnisky, E. J. (2005). Estimating entropy rates with Bayesian confidence intervals. *Neural Comp.*, *17*(7), 1531–1476.

Latham, P. E., & Nirenberg, S. (2005). Synergy, redundancy, and independence in population codes, revisited. *J. Neurosci.*, *25*(21), 5195–5206.

Logothetis, N. K. (2003). The underpinnings of the bold functional magnetic resonance imaging signal. *J. Neurosci.*, *23*, 3963–3971.

London, M., Schreibman, A., Hausser, M., Larkum, M. E., & Segev, I. (2002). The information efficacy of a synapse. *Nature Neurosci.*, *5*, 332–340.

Merhav, N., Kaplan, G., Lapidoth, A., & Shamai Shitz, S. (1994). On information rates for mismatched decoders. *IEEE Trans. Inform. Theory*, *40*, 1953–1967.

Miller, G. A. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology: Problems and methods* (pp. 95–100). New York: Free Press.

Nemenman, I., Bialek, W., & de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, *69*(5), 056111.

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

Nirenberg, S., & Latham, P. E. (2003). Decoding neuronal spike trains: How important are correlations. *Proc. Natl. Acad. Sci., USA*, *100*, 7348–7353.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*, 1191–1253.

Panzeri, S., Petersen, R. S., Schultz, S. R., Lebedev, M., & Diamond, M. E. (2001). The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron*, *29*, 769–777.

Panzeri, S., & Schultz, S. (2001). A unified approach to the study of temporal, correlational and rate coding. *Neural Computation*, *13*, 1311–1349.

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network*, *7*, 87–107.

Petersen, R. S., & Diamond, M. (2000). Spatio-temporal distribution of whisker-evoked activity in rat somatosensory cortex and the coding of stimulus location. *J. Neurosci.*, *20*, 6135–6143.

Pola, G., Petersen, R., Thiele, A., Young, M. P., & Panzeri, S. (2005). Data-robust tight lower bounds to the information carried by spike times of a neural population. *Neural Comp.*, *17*, 1962–2005.

Pola, G., Thiele, A., Hoffmann, K.-P., & Panzeri, S. (2003). An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network*, *14*, 35–60.

Reich, D. S., Mechler, F., Purpura, K. P., & Victor, J. D. (2000). Interspike intervals, receptive fields, and information encoding in primary visual cortex. *J. Neurosci.*, *20*, 1964–1974.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., & Bialek, W. (1996). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Schneidman, E., Berry, M. J., Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, *440*, 1007–1012.

Schneidman, E., Bialek, W., & Berry, M. J., II. (2003). Synergy, redundancy, and independence in population codes. *J. Neurosci.*, *23*(37), 11539–11553.

Schultz, S., & Panzeri, S. (2001). Temporal correlations and neural spike train entropy. *Phys. Rev. Lett.*, *86*, 5823–5826.

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony unbound: A critical evaluation of the temporal binding hypothesis. *Neuron*, *24*, 67–77.

Shannon, C. E. (1948). A mathematical theory of communication. *AT&T Bell Labs. Tech. J.*, *27*, 379–423.

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

Victor, J. D. (2002). Binless strategies for estimation of information from neuronal data. *Physical Review E*, *66*, 51903–51918.

Victor, J. D., & Purpura, K. P. (1997). Metric-space analysis of spike trains: Theory, algorithms, and application. *Network*, *8*, 127–164.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, *24*, 95–104.

Received May 30, 2006; accepted October 20, 2006.