Information theoretic methods for studying population codes

Robin A. A. Ince[a], Riccardo Senatore[a], Ehsan Arabzadeh[b], Fernando Montani[c,d,e], Mathew E. Diamond[f,g], Stefano Panzeri[c,*]

[a]*Faculty of Life Sciences, University of Manchester, Manchester, UK*
[b]*School of Psychology, University of New South Wales, Sydney, Australia*
[c]*Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology, Via Morego 30, 16163 Genoa, Italy*
[d]*Instituto de Física La Plata (IFLP), Universidad Nacional de La Plata, La Plata, Argentina*
[e]*Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), La Plata, Argentina*
[f]*SISSA Unit, Italian Institute of Technology, Trieste, Italy*
[g]*Cognitive Neuroscience Sector, International School for Advanced Studies, Trieste, Italy*

Abstract

Population coding is the quantitative study of which algorithms or representations are used by the brain to combine together and evaluate the messages carried by different neurons. Here, we review an information-theoretic approach to population coding. We first discuss how to compute the information carried by simultaneously recorded neural populations, and in particular how to reduce the limited sampling bias which affects calculation of information from a limited amount of experimental data. We then discuss how to quantify the contribution of individual members of the population, or the interaction between them, to the overall information encoded by the considered group of neurons. We focus in particular on evaluating what is the contribution of interactions up to any given order to the total information. We illustrate this formalism with applications to simulated data with realistic neuronal statistics and to real simultaneous recordings of multiple spike trains.

Keywords: mutual information, sampling bias, population coding,

[*] Corresponding author
Email address: stefano.panzeri@iit.it (Stefano Panzeri)

1. Introduction

The central nervous system supports highly reliable and fast perception of sensory events. In most conditions, animals can perceive a sensory stimulus based on a single presentation. Yet responses of individual neurons in the central nervous system of mammals are often highly variable: repeated presentations (*trials*) of the same stimulus elicit a different single-neuron response each time. As a result, single neuron messages are ambiguous and difficult to interpret. From the point of view of off-line analysis, it is easy to reduce the effect of this variability by averaging responses over repeated trials, as often done by Neurophysiologists. However, the trial averaging strategy cannot be used by the brain, because the brain usually processes information and takes decisions based on single events. It is widely believed that the strategy used by the brain to make sense of single trials of the noisy responses of individual neurons is to evaluate the simultaneous activity of large neural populations. In other words, it is believed that the brain uses a neural population code (rather than a single neuron code) to transmit information about sensory stimuli. However, exactly how the brain puts together the information from several neurons remains largely unknown.

Since it is commonly found that the neurons within local networks are correlated, *i.e.* that the response of a neuron does not depend only upon the stimulus but also upon the activity of other neurons (Li, 1959; Mastronarde, 1983; Abeles et al., 1993; Luczak et al., 2009), several authors have hypothesized that such interactions between neurons play a crucial part in forming unambiguous population responses. For example, interactions among neurons may be used to coordinate their relative firing time to tag particular features to be bound together (von der Malsburg, 1999), may stabilize the temporal relationships between cells against the detrimental effect of trial-to-trial variability (Chase & Young, 2007; Gollisch & Meister, 2008), or may be exploited to implement strategies for error correction (Schneidman et al., 2006).

In this Review, we will consider one particular mathematical analysis approach to population coding, based on information theory (Quian Quiroga & Panzeri, 2009). One advantage of this approach is that information theory quantifies stimulus discriminability based on single trials (rather than on an average across trials), and this makes it biologically relevant to characterizing population codes, because (as discussed above) brains recognize sensory stimuli and take decisions on single trials. After introducing the main concepts of information theory in the context of sensory neuroscience, we will discuss ways to reduce the limited sampling bias which plagues estimation of information measures from experimentally recorded neural populations, extending the feasibility of such analysis to larger populations. We will then discuss how to quantify the contribution of the interactions between groups of neurons to the overall information carried by the neuronal population. We will focus in particular on evaluating what is the contribution of interactions up to any given order to the total information transmitted by the population, and how this contribution scales with population size. We will validate and demonstrate this formalism by applying it to simulated data with realistic neuronal statistics, with the aim of exploring the robustness of the methods to data sampling. We will also illustrate the methodology by computing the information about whisker stimuli carried by real simultaneously recorded populations from the rat somatosensory cortex in order to demonstrate what type of neurophysiological conclusion can be reached with it.

2. The information carried by neuronal population responses

Consider an experiment in which an animal is presented with a stimulus s selected with probability $P(s)$ from a stimulus set S, and the consequent response of a population of C neurons is recorded and quantified in a certain poststimulus time window. We assume that the neural population response is quantified as a discrete, multi-dimensional array $\mathbf{r} = r_1, \ldots, r_C$ of dimension C, where r_c is the response of neuron c on a given trial in the response window. To simplify the presentation, we assume that r_c is the number of spikes emitted by neuron c during the trial in the response window although the method could in principle be easily extended to consider more detailed quantifications of single neuron responses, for example those which include the temporal response patterns of single neurons. The maximum number of spikes that can be observed from an individual neuron in any trial is denoted by M. (If the considered time window is very short, M is 1 and r_c is binary). We indicate the response space by $\mathbf{R}$ ($\mathbf{R}$ contains $(M+1)^C$ elements). In all examples considered here, M equals 1 (binary responses), but the formalism is generic and is well defined for any M value.

Having discussed how to quantify the response, the second step is to compute how much information can be extracted from the chosen response quantification. The more the response of a neuron varies across a set of stimuli, the greater its ability to transmit information about those stimuli (de Ruyter van Steveninck et al., 1997). The first step in measuring information is thus to measure the response variability. The most general way to do this is through the concept of Shannon entropy, referred to hereafter as entropy, which is a measure of the uncertainty associated with a random variable. Intuitively one can posit some desirable properties of any uncertainty measure. The first is that small changes in the underlying probabilities should result in small changes in the uncertainty. The second is that the measure should not depend on the labelling or ordering of the variables and outcomes. The third is that the measure should take its maximum value when all outcomes are equally likely and for systems with uniform probabilities, the measure should increase with the number of outcomes. The fourth is that the measure should be additive; that is it should be independent of how the system is grouped or divided into parts. It can be shown (Cover & Thomas, 2006) that any measure of uncertainty about the neural responses satisfying these properties has the form

$$H(\mathbf{R}) \quad = \quad -\sum_{\mathbf{r}\in\mathbf{R}} P(\mathbf{r}) \log_2 P(\mathbf{r}) \tag{1}$$

where $P(\mathbf{r})$ is the probability of observing response $\mathbf{r}$ across all trials to all stimuli. In the neuroscience literature, the quantity in Eq. (1) is usually called the response entropy, and it quantifies how neuronal responses vary with the stimulus and thus sets the capacity of the spike train to convey information. In Eq. (1) (and in the following equations) the summation over $\mathbf{r}$ is over all possible neuronal responses.

However, neurons are typically noisy; their responses to repetitions of an identical stimulus differ from trial to trial. Therefore $H(\mathbf{R})$ reflects both variation of responses to different stimuli and variation due to trial-to-trial noise. Thus $H(\mathbf{R})$ is not a pure measure of the stimulus information actually transmitted by the neuron. We can quantify the variability specifically due to noise, by measuring the so called *noise entropy*, which is the entropy conditional on stimulus presentation:

$$H(\mathbf{R}|S) \quad = \quad -\sum_{s\in S} P(s) \sum_{\mathbf{r}\in\mathbf{R}} P(\mathbf{r}|s) \log_2 P(\mathbf{r}|s) \tag{2}$$

where in the above the summation over s is over all possible stimuli, and $P(\mathbf{r}|s)$ is the probability of observing a particular response $\mathbf{r}$ given that stimulus s is presented. Experimentally, $P(\mathbf{r}|s)$ is determined by repeating each stimulus on many trials, while recording the neuronal responses. The noise entropy quantifies the irreproducibility of the neuronal responses at fixed stimulus. The noisier is a neuron, the greater is $H(\mathbf{R}|S)$.

The information that the neuronal response transmits about the stimulus is the difference between the response entropy and the noise entropy. This is known as the mutual information $I(S;\mathbf{R})$ between stimuli and responses (in the following sometimes abbreviated to information).

$$\begin{aligned} I(S;\mathbf{R}) &= H(\mathbf{R}) - H(\mathbf{R}|S) \\ &= \sum_{s\in S} P(s) \sum_{\mathbf{r}\in\mathbf{R}} P(\mathbf{r}|s) \log_2 \frac{P(\mathbf{r}|s)}{P(\mathbf{r})} \end{aligned} \tag{3}$$

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). Alternatively, it quantifies the reduction of uncertainty about the stimulus that can be gained from observation of a single trial of the neural response. When the logarithms used are base 2, as in Eq. (3), the entropy and information have units of *bits.* 1 bit of mutual information corresponds to an average reduction in uncertainty about the stimulus by a factor of 2 after observation of a single response.

The mutual information has a number of important qualities that make it well suited to characterizing how well a neural response is modulated by the stimulus (recently reviewed for example by Borst & Theunissen (1999); Fuhrmann Alpert et al. (2007); Panzeri et al. (2008); Quian Quiroga & Panzeri (2009)). First, as outlined above, information theoretic techniques quantify information gains in single trials (rather than on average across trials) and this makes them biologically relevant, because brains recognize sensory stimuli and take decisions on single trials. Second, with respect to other single trial analysis techniques (such as decoding or reconstruction of the most likely stimulus that elicited the neural response) information theory has the advantage that it naturally takes into account all possible ways in which neurons can convey information, for example, by predicting the most likely stimulus, by reporting the uncertainty of the prediction, or by ruling out very unlikely stimuli (Quian Quiroga & Panzeri, 2009). Third, $I(S;\mathbf{R})$ is the most general measure of correlation between the stimuli and the neural responses, because it automatically takes into account contributions of all interactions among neurons at all orders. This property is central to

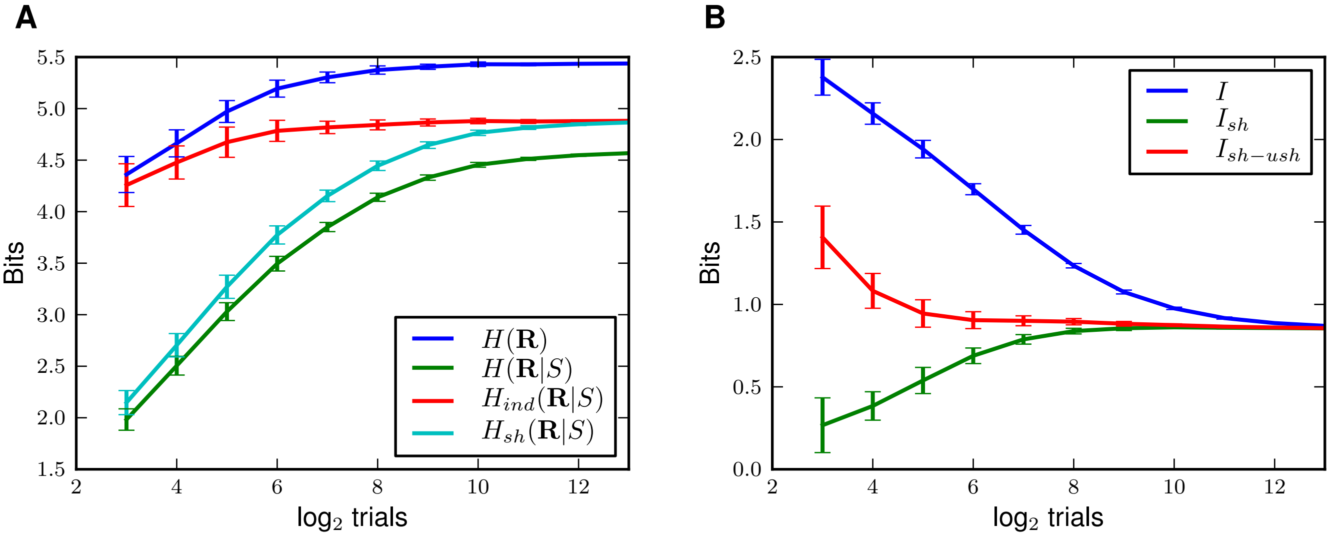

Figure 1: *The limited sampling bias.* Plugin estimates of entropy quantities (panel **A**) and mutual information estimators (panel **B**) are shown as function of number of trials per stimulus used for the estimation. Data is generated from a model of 8 cells from rodent somatosensory cortex, responding to whisker stimulation at 13 different velocities (ranging between 0.15 mm/s and 47.7 mm/s). Stimulus-conditional individual and pair-wise marginal probabilities are equal to those observed experimentally (see Section 5) but no higher order correlations are present. Each point represents the average over 50 simulations of the system; error bars show ±1 SD.

evaluating (as we will do below) the information losses due to the simplification of the structure of interactions among neurons. Fourth, computing information does not require specifying a stimulus–response model; it only requires computing the response probabilities conditional on each stimulus. Therefore, the calculation of information does not require spelling out which stimulus features (e.g., contrast, orientation, etc.) are encoded. This makes this formalism not only adaptable to different experimental designs but also suited to the analysis of neural responses to complex, rapidly varying stimuli (de Ruyter van Steveninck et al., 1997).

3. Computing information from limited datasets

3.1. Origin and properties of the limited sampling bias

Calculation of information requires accurate estimation of the stimulus-response probabilities $P(\mathbf{r})$, $P(\mathbf{r}|s)$ and $P(s)$ and thereby $H(\mathbf{R})$ and $H(\mathbf{R}|S)$. These probabilities, however, are not known a priori and have to be measured experimentally from the available neurophysiological data. This is the key practical issue for the accurate application of Information Theory to the study of neural codes. If we had an infinite amount of data, we could measure the true stimulus-response probabilities precisely. However, any real experiment only yields a finite number of trials from which these probabilities must be estimated. The estimated probabilities are subject to statistical error and necessarily fluctuate around their true values. These finite sampling fluctuations lead to a systematic error (bias) in estimates of entropies and information. If not corrected, bias can lead to serious misinterpretations of neural coding data. In this subsection, following and extending the work of Panzeri et al. (2007), we illustrate and investigate the nature of this problem, and we present a number of useful techniques that have recently been developed for addressing this issue.

The most direct way to compute information and entropies is to estimate the response probabilities as the histogram of the experimental frequency of each response across the available trials. Plugging these empirical probability estimates into Eqs. (1,2,3) results in a direct estimate that we refer to as the 'plug-in' method. In the following, N_s denotes the number of trials recorded in response to stimulus s and N is the total number of trials across all stimuli.

To understand better the effects of plugging experimentally determined probabilities into the information functional, we performed a series of simulations of a realistic population of 8 cells (see Appendix A for details), systematically varying the number of trials. Fig. 1A shows the entropy estimates resulting from the plug-in method. In both cases, the estimates of $H(\mathbf{R})$ and $H(\mathbf{R}|S)$ increased with the number of trials. That is, finite sampling makes plug-in entropy estimates biased downward. This is the case for any stimulus-response probability distribution (Paninski, 2003). Intuitively, the reason is that entropy is a measure of variability. The fewer the number of trials available, the less likely we are to fully sample the range of possible responses. Consequently, entropy estimates are lower than their true values, and the effect of finite sampling on entropies is a downward bias. $H(\mathbf{R})$ is far less biased than $H(\mathbf{R}|S)$ because the former depends on $P(\mathbf{r})$, which, being computed from data collected across all stimuli, is better sampled than $P(\mathbf{r}|s)$. From Eq. (3), the bias of the information is the difference between the bias of $H(\mathbf{R})$ and that of $H(\mathbf{R}|S)$. Because the latter is greater (and negative), the net result is that $I(S;\mathbf{R})$ is strongly biased upward (Fig. 1B). Intuitively, this is because finite sampling can introduce spurious stimulus dependent differences in the response probabilities, which make the stimuli seem more discriminable and hence the neurons more informative than they actually are.

To understand the sampling behavior of information and entropy better, it is useful to find analytical approximations to the bias. This can be done in the so-called asymptotic sampling regime where, roughly speaking, the number of trials is large. More rigorously, the asymptotic sampling regime is defined as N being large enough that every possible response occurs many times: that is, $N_s P(\mathbf{r}|s) \gg 1$ for each stimulus-response pair s, r such that $P(\mathbf{r}|s) > 0$. In this regime, the bias of the entropies and information can be expanded in inverse powers of $1/N$ and analytical approximations obtained (Miller, 1955; Panzeri & Treves, 1996). The leading terms in the biases are, respectively

$$\begin{aligned}
BIAS\left[H(\mathbf{R})\right] &= \frac{-1}{2N\ln 2}\left[\bar{R}-1\right] \\
BIAS\left[H(\mathbf{R}|S)\right] &= \frac{-1}{2N\ln 2}\sum_s\left[\bar{R}_s-1\right] \\
BIAS\left[I(S;\mathbf{R})\right] &= \frac{1}{2N\ln 2}\left\{\sum_s\left[\bar{R}_s-1\right]-\left[\bar{R}-1\right]\right\} \qquad (4)
\end{aligned}$$

where $\bar{R}_s$ denotes the number of relevant responses for the stimulus conditional response probability distribution $P(\mathbf{r}|s)$ (i.e. the number of different responses $\mathbf{r}$ with nonzero probability of being observed when stimulus s is presented) and $\bar{R}$ denotes the number of relevant responses for $P(\mathbf{r})$ (i.e. the number of different responses $\mathbf{r}$ with nonzero probability of being observed across all stimuli).

Although valid only in the asymptotic regime, Eq. (4) sheds valuable light on the key factors that control the bias. First, Eq. (4) shows that if N_s is constant, the bias increases with the number of responses R. This means that the bias increases exponentially with the population size, and this is what makes the application of information theory to the analysis of neural populations so hard. Second, Eq. (4) shows that the bias of $H(\mathbf{R}|S)$ is approximately S times bigger than that of $H(\mathbf{R})$. This means that, in the presence of many stimuli, the bias of $I(S;\mathbf{R})$ is similar to that of $H(\mathbf{R}|S)$. However, $I(S;\mathbf{R})$ is a difference of entropies, and its typical values are much smaller than those of $H(\mathbf{R}|S)$. This implies that spike train analysis methods must be validated on the performance of information and not only on entropies, because, in many cases, the bias may be proportionally negligible for entropies but not for the information. Third, Eq. (4) shows that the bias is small when the ratio N_s/R is big, i.e. there are more trials per stimulus than possible responses. Thus N_s/R is the crucial parameter for the sampling problem. For example, in the simulations of Fig. 1B, with R equal to $2^8 = 256$, the bias of $I(S;\mathbf{R})$ became negligible for $N_s = 2^{13}$ (i.e. $N_s/R \approx 32$).

3.2. Bias correction techniques

The plug-in estimate of information $I(S;\mathbf{R})$ tends to require large numbers of trials ($N_s/R \approx 32$ as in Fig. 1B) to become unbiased and is therefore of limited experimental utility. However, over the last decade, several bias correction procedures have been developed to reduce the number of trials required to obtain accurate unbiased estimates (see Panzeri et al. (2007) and Victor (2006) for reviews).

3.2.1. Quadratic Extrapolation

One such correction is the so called *quadratic extrapolation* (QE). In the asymptotic sampling regime, the bias of entropies and information can be approximated as second order expansions in $1/N$, where N is the number of trials (Strong et al., 1998; Treves & Panzeri, 1995). For example, for the information:

$$I_{\text{plugin}}(S;\mathbf{R}) = I_{\text{true}}(S;\mathbf{R}) + \frac{a}{N} + \frac{b}{N^2} \tag{5}$$

This property can be exploited by calculating the estimates with subsets of the original data, with $N/2$ and $N/4$ trials and fitting the resulting values to the polynomial expression above. This allows an estimate of the parameters a and b and hence $I_{\text{true}}(S;\mathbf{R})$. To use all available data, estimates of two subsets of size $N/2$ and four subsets of size $N/4$ are averaged to obtain the values for the extrapolation. Together with the full length data calculation, this requires seven different evaluations of the quantity being estimated. An advantage of QE is that it is simple to implement and that, although designed for the asymptotic regime, it works well also for intermediately sampled regimes ($N_s/R \approx 2-4$ or more). The disadvantage of QE correction is that, by design, it cannot work in the undersampled regime ($N_s/R \ll 1$).

3.2.2. Panzeri-Treves (PT)

Eq. (4) can be used to estimate the bias, provided that one can evaluate the number of relevant responses $\bar{R}_s$. However, estimating $\bar{R}_s$ is not straightforward. The simplest approach is to approximate $\bar{R}_s$ by the number of responses that are observed at least once - this is the naive count. This leads to the so-called Miller-Madow bias estimate (Miller, 1955). The naive count is a lower bound on the actual number of relevant responses because some relevant responses are likely to have been missed due to lack of data. Thus, the Miller-Madow estimate is usually an underestimate of the bias. To alleviate this problem, Panzeri & Treves (1996) have developed a Bayesian procedure to estimate the number of relevant responses. This estimate can be inserted into Eq. (4) to compute the bias and then subtract it from the plug-in information value: we refer to this procedure as PT bias correction. PT, being designed also for the asymptotic sampling regime, has a performance similar to that of QE. It works well also for intermediately sampled regimes ($N_s/R \approx 2-4$ or more), but by design, it cannot work in the undersampled regime ($N_s/R \ll 1$).

3.2.3. Nemenman-Shafee-Bialek (NSB)

The NSB method (Nemenman et al., 2002, 2004) utilises a Bayesian inference approach and does not rely on the assumption of the asymptotic sampling regime. It is based on the principle that when estimating a quantity, the least bias will be achieved when assuming an uniform a priori distribution over the quantity. This method is challenging to implement, involving a large amount of function inversion and numerical integration. However, it often gives a significant improvement in the accuracy of the bias correction and it can potentially work well also in conditions of severe undersampling (Nemenman et al., 2002, 2004; Montemurro et al., 2007).

3.2.4. Comparison of bias subtraction methods

Figure 2A reports the results of the performance of bias correction procedures on the estimates of the information in simulated data reproducing the firing rates and second order interactions of 8 neurons in rat somatosensory cortex (see Section 5 and Appendix A). Figure 2A shows that bias correction procedures improve the estimate of $I(S;\mathbf{R})$ with respect to the plug-in estimator, and the NSB correction is especially effective. When using bias corrections, the estimation of $I(S;\mathbf{R})$ in this simulation became accurate

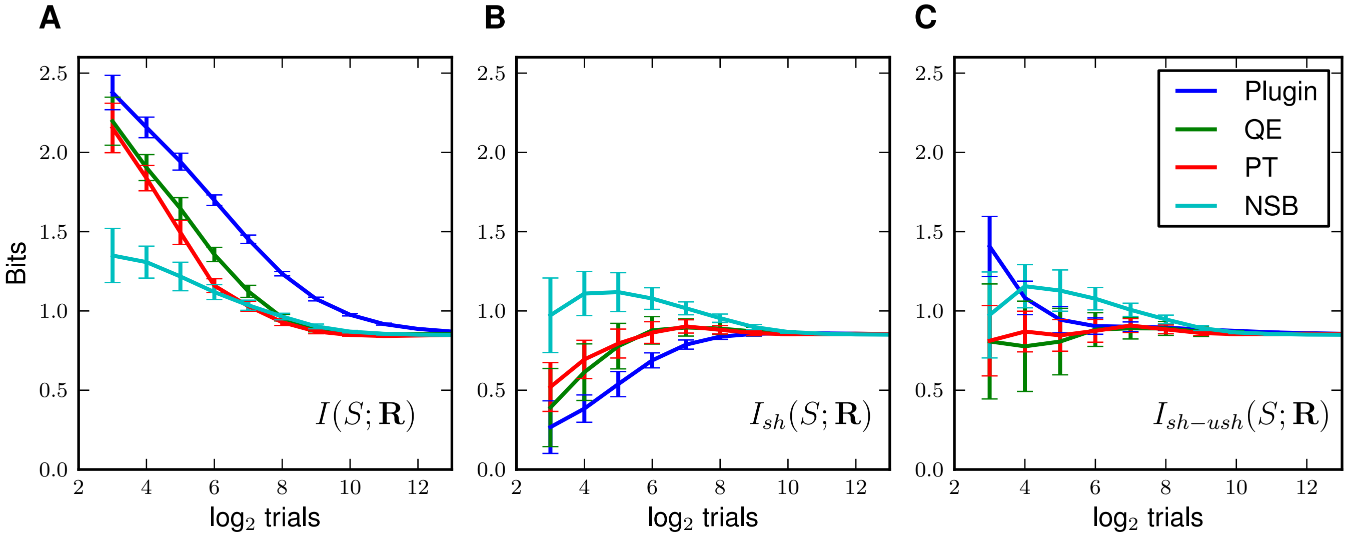

Figure 2: *Bias corrected mutual information estimates.* The effect of different entropy bias correction methods on the mutual information estimators I (panel **A**), I_{sh} (panel **B**) and $I_{\text{sh-ush}}$ (panel **C**) are shown. Data is generated from the same 8 cell model system as in Figure 1. Each point represents the average over 50 simulations of the system; error bars show ±1 SD.

when 2^9 trials per stimulus were available (that is, when $N_s/R \approx 2$, compared with $N_s/R \approx 32$ for the pure plug-in method).

3.3. Reducing the bias by shuffling correlated responses

The fundamental problem with understanding neural population codes is the dimensionality problem: the number of possible responses becomes exponentially large as the number of neurons, C, grows. Thus (via Eq. 4) the bias gets quickly out of control when considering a large response array. For example, 10 'spike-count' neurons emitting up to 4 spikes per stimulus presentation generate 4^{10} ($\approx 10^6$) possible responses. The bias problem for large C is exacerbated by the fact that, in real neuronal recordings, the elements of the response array are often statistically correlated at fixed stimulus. In other words, for some stimulus s the 'true' stimulus-response probability $P(\mathbf{r}|s)$ is significantly different from the probability $P_{ind}(\mathbf{r}|s)$ obtained if neurons 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 $P(r_c|s)$ of responses of each neuron:

$$P_{ind}(\mathbf{r}|s) = \prod_{c=1}^{C} P(r_c|s), \tag{6}$$

These interactions at fixed stimulus are usually referred to in the literature as noise correlations (Gawne & Richmond, 1993; Pola et al., 2003; Latham & Nirenberg, 2005). In sampling terms, the implication is that the sampling of the full probability of a response array cannot be reduced to computing the probabilities of each individual array element (*marginal probabilities*) as would be legitimate if responses were uncorrelated. Thus one has to deal with the full exponentially large response array. However, fortunately there is a way to keep the sampling difficulties introduced by correlations under control, as follows. This discussion closely follows that of Panzeri et al. (2007).

Consider the independent noise entropy that would be obtained if the response in each element of the array was independent of the others at fixed stimulus:

$$H_{ind}(\mathbf{R}|S) \quad = \quad -\sum_{s\in S} P(s) \sum_{\mathbf{r}\in\mathbf{R}} P_{ind}(\mathbf{r}|s) \log_2 P_{ind}(\mathbf{r}|s) \tag{7}$$

Because $H_{ind}(\mathbf{R}|\mathbf{S})$ depends only on the marginal probabilities of each individual neuron, it has very small bias (Figure 1A). Alternatively, correlations between response variables can be removed by ‘shuffling’ the data at fixed stimulus. This is done by constructing pseudo response arrays combining r_c values each taken (randomly and without replacement) from different trials in which the stimulus s was presented. The algorithm for this is as follows. Take all responses in the first element of the response array to trials for a given stimulus and randomize their order across the N_s trials. Repeat for the other elements, randomizing independently across trials each time. This results in a pseudo response array from which shuffled stimulus-response probabilities denoted $P_{sh}(\mathbf{r}|s)$ can be sampled and the shuffled noise entropy $H_{sh}(\mathbf{R}|S)$ computed:

$$H_{sh}(\mathbf{R}|S) \;\; = \;\; -\sum_{s \in S} P(s) \sum_{\mathbf{r} \in \mathbf{R}} P_{sh}(\mathbf{r}|s) \log_2 P_{sh}(\mathbf{r}|s) \qquad (8)$$

$H_{sh}(\mathbf{R}|S)$ has the same value as $H_{ind}(\mathbf{R}|S)$ for large numbers of trials N_s, since the independent shuffling of responses removes any correlations. However, it has a much higher bias than $H_{ind}(\mathbf{R}|S)$ for small N_s. In fact, Fig. 1A shows that the bias of $H_{sh}(\mathbf{R}|S)$ is of the same order of magnitude as the bias of $H(\mathbf{R}|S)$. Intuitively, this is expected because $P_{sh}(\mathbf{r}|s)$ is sampled with the same number of trials as $P(\mathbf{r}|s)$ from responses with the same dimensionality (Montemurro et al., 2007; Nirenberg et al., 2001). This observation has led to the suggestion (Montemurro et al., 2007) to compute information not directly though $I(S;\mathbf{R})$ but through the following formula:

$$I_{sh}(S;\mathbf{R}) = H(\mathbf{R}) - H_{ind}(\mathbf{R}|S) + H_{sh}(\mathbf{R}|S) - H(\mathbf{R}|S) \qquad (9)$$

$I_{sh}(S;\mathbf{R})$ has the same value of $I(S;\mathbf{R})$ for infinite number of trials but has a much smaller bias for finite N due to the approximate bias cancellation created by the entropy terms added in the right hand side of Eq. 9.

Figure 1B confirms that as a result of the bias cancelations in Eq. (9), there a huge bias reduction in the plug-in estimates of $I_{sh}(S;\mathbf{R})$ with respect to $I(S;\mathbf{R})$.

Moreover, Figure 1B shows that the bias of the plug-in estimate of $I_{sh}(S;\mathbf{R})$ is negative. Indeed simulations show that the bias of $I_{sh}(S;\mathbf{R})$ tends to be negative more often that it tends to be positive (Montemurro et al., 2007). It is worth briefly considering the reason of this. The first reason is that the bias cancelation between $H_{sh}(\mathbf{R}|S)$ and $H(\mathbf{R}|S)$ is only exact when correlations are totally absent, and is only an incomplete cancelation in general. In particular, in the presence of correlated firing (as shown in Fig. 1A, and in more detail by Montemurro et al. (2007)) $H_{sh}(\mathbf{R}|S)$ is usually slightly more downward biased than $H(\mathbf{R}|S)$. To understand why, Montemurro et al. (2007) derived the bias of $H_{sh}(\mathbf{R}|S)$ in the asymptotic sampling regime, and found that

$$BIAS\left[H_{sh}(\mathbf{R}|S)\right] = \frac{-1}{2N\ln 2}\sum_{s}\left[\bar{R}_{sh-s} - 1\right] \tag{10}$$

where $\bar{R}_{sh-s}$ denotes the number of relevant responses for the stimulus conditional response probability distribution $P_{sh}(\mathbf{r}|s)$. The number of shuffled relevant stimulus-conditional $\bar{R}_{sh-s}$ is greater than or equal to $\bar{R}_s$, because $P_{ind}(\mathbf{r}|s) = 0$ implies $P(\mathbf{r}|s) = 0$ and the shuffled responses can be considered as samples from $P_{ind}(\mathbf{r}|s)$. Thus the negative bias of $H_{sh}(\mathbf{R}|S)$ is greater than or equal to that of $H(\mathbf{R}|S)$. The second reason why the bias of $I_{sh}(S;\mathbf{R})$ is often negative is that there is a downward bias in $H(\mathbf{R})$. This has a significant negative effect on the bias of $I_{sh}(S;\mathbf{R})$ when there are only very few stimuli, because in this case the negative bias of $H(\mathbf{R})$ (which is computed across all trial to all stimuli) may be large and may outweigh that of $H_{ind}(\mathbf{R}|S)$ (which appears in Eq. (9) with the sign opposite to $H(\mathbf{R})$). When the number of stimuli is large, then $H(\mathbf{R})$ is well sampled and its bias becomes negligible compared to that of $H_{ind}(\mathbf{R}|S)$. However, it should be also noted that the bias of $I_{sh}(S;\mathbf{R})$ can also be mildly positive in some occasions. This can happen for example when the cancelation between the biases of $H_{sh}(\mathbf{R}|S)$ and $H(\mathbf{R}|S)$ is perfect (because there are no correlations) and there are very many stimuli, so that $H(\mathbf{R})$ has effectively no bias and the whole bias of $I_{sh}(S;\mathbf{R})$ comes from that of $-H_{ind}(\mathbf{R}|S)$.

In the previous section, the four bias correction techniques were applied to $I(S;\mathbf{R})$. However, they can also be applied to $I_{sh}(S;\mathbf{R})$. Figure 2B illustrates that, with all three bias correction procedures, there is a bias reduction when using $I_{sh}(S;\mathbf{R})$ rather than $I(S;\mathbf{R})$. The result of using $I_{sh}(S;\mathbf{R})$ in combination with bias correction techniques is that the estimates of information become less dependent on the bias correction method used, and become unbiased even down to 2^6 trials per stimulus (i.e., for $N_s/R \approx 1/4$) for the PT and QE methods. This is a factor of eight better than the best performing bias correction of $I(S;\mathbf{R})$.

Finally we present a novel extension of the $I_{sh}(S;\mathbf{R})$ estimator, that further reduces the bias in cases when the number of stimuli is not very high. Under such conditions, the bias of $H(\mathbf{R})$, while less than that of $H(\mathbf{R}|S)$, may still contribute significantly to the downward bias of $I_{sh}(S;\mathbf{R})$. However, it is possible to cancel the bias of $H(\mathbf{R})$ in a similar way as the bias of $H(\mathbf{R}|S)$ is canceled in $I_{sh}(S;\mathbf{R})$. To do this the responses across all stimuli are shuffled and sampled, obtaining an unconditional shuffled probability distribution $P_{ush}(\mathbf{r})$, with corresponding entropy:

$$H_{ush}(\mathbf{R}) \;\; = \;\; -\sum_{\mathbf{r}\in\mathbf{R}} P_{ush}(\mathbf{r}) \log_2 P_{ush}(\mathbf{r}) \tag{11}$$

The entropy $H_{ush}(\mathbf{R})$ has approximately the same bias as $H(\mathbf{R})$, since it is computed with the same number of samples over a response space of the same size. In the limit of large numbers of trials it converges asymptotically to the entropy of the independent response distribution unconditional to the stimuli $P_{uind}(\mathbf{r}) = \prod_{i=1...C} P(r_i)$:

$$H_{uind}(\mathbf{R}) \;\; = \;\; \sum_{i=1...C} H(R_i) \tag{12}$$

As discussed for $H_{ind}(\mathbf{R}|S)$, $H_{uind}(\mathbf{R})$ has a very small bias since it depends only on the marginal probabilities of each individual neuron. So proceeding exactly as in the derivation of $I_{sh}(S;\mathbf{R})$ it is possible to add and subtract these asymptotically equivalent terms to cancel the bias of $H(\mathbf{R})$. This results in the following estimator:

$$\begin{aligned} I_{sh-ush}(S;\mathbf{R}) \;\; &= \;\; H(\mathbf{R}) - H_{ush}(\mathbf{R}) + \sum_{i=1...C} H(r_i) \\ &\qquad -H_{ind}(\mathbf{R}|S) + H_{sh}(\mathbf{R}|S) - H(\mathbf{R}|S) \end{aligned} \tag{13}$$

Fig 2C considers the performance of this estimator in the case when the number of stimuli is intermediate (13 stimuli). The uncancelled bias in $I_{sh-ush}(S;\mathbf{R})$ comes from the terms $H_{uind}(\mathbf{R}) - H_{ind}(\mathbf{R}|S)$. Bias correction methods such as quadratic extrapolation and the PT analytical approximation perform better for $I_{sh-ush}(S;\mathbf{R})$, resulting in improved performance relative to $I_{sh}(S;\mathbf{R})$ for low numbers of trials. This allows an unbiased estimate even down to 2^5 trials per stimulus (i.e. for $N_s/R \approx 1/8$).

As discussed above, in general the bias of $I_{sh}(S;\mathbf{R})$ tends to be negative in most cases, although it can in principle be either positive or negative. The additional terms added to $I_{sh-ush}(S;\mathbf{R})$ ensure that its bias is always more positive than that of $I_{sh}(S;\mathbf{R})$, because these additional terms were explicitly designed to cancel out the negative bias of $H(\mathbf{R})$. The use of $I_{sh-ush}(S;\mathbf{R})$ is therefore only beneficial in cases where $I_{sh}(S;\mathbf{R})$ is biased downwards because of the negative bias of $H(\mathbf{R})$, which, as mentioned above, takes a prominent role when the number of stimuli is small.

The difference in performance between $I_{sh-ush}(S;\mathbf{R})$ and $I_{sh}(S;\mathbf{R})$ depends mainly on the number of stimuli. We verified that (results not shown) if the number of stimuli were lower, the benefit of $I_{sh-ush}(S;\mathbf{R})$ would be bigger, since the level of bias of $H(\mathbf{R})$ would be closer to that of $H(\mathbf{R}|S)$. Conversely if the number of stimuli would be much larger (e.g. hundreds) then $I_{sh-ush}(S;\mathbf{R})$ would offer no benefit since $H(\mathbf{R})$ would already be very well sampled.

4. Quantifying the effect of interactions on information transmission

4.1. Defining and quantifying the interactions among neurons

Having defined the information that neuronal responses transmit about sensory stimuli, we consider how interactions among neurons affect information transmission.

The first step is to define precisely what we mean by interactions. Here we say that neuronal populations interact if, for some stimulus s, the 'true' stimulus-response probability $P(\mathbf{r}|s)$ is different from the probability $P_{ind}(\mathbf{r}|s)$. In a large population, these interactions may be in general very complex and may be characterized by many parameters. For example, in a network of C binary neurons, $2^C - 1$ parameters are needed to characterize all the possible interactions. A central question in studying neural population codes is to understand which aspects of neural population activity are important and which are not. Therefore, it is important not only to document the presence of a statistically significant correlation structure among responses of neurons within a population, but also to determine which specific aspects of the interaction structure are most important for information transmission. An approach to this question is to consider a simplified response model which neglects certain aspects of the spike train correlation structure (e.g. it considers only correlations among a specific subset of neurons), and test the effects on information transmission of making such simplifications.

A question which has received much attention recently is whether we can describe all interactions between the neurons in terms of interactions between up to two neurons only, or whether there are interactions among groups of more than two neurons which cannot be explained in terms of pairwise interactions. Understanding this is important for building minimal models of neural population responses which still capture the main functional properties. A rigorous way to investigate the effects of different orders of interaction is provided by the technique of *maximum entropy*, which was originally introduced in statistical physics (Jaynes, 1957), and is now beginning to be used in neuroscience (Martignon et al., 2000; Nakahara & Amari, 2002; Schneidman et al., 2006; Shlens et al., 2006; Tang et al., 2008; Nirenberg & Victor, 2007; Montemurro et al., 2007; Montani et al., 2009). In general, the idea of the maximum entropy (ME) principle is to first fix some constraints that are of interest and then seek the simplest, or most random, distribution subject to those constraints. Using entropy as a measure of randomness, asking for the most random distribution corresponds to asking for the distribution with maximal entropy subject to the constraints. This removes all types of correlation or structure in the data that does not result from the constrained features.

The ME formalism can be naturally used to to address the problem of whether we can describe all interactions between neurons in terms of interactions between up to k neurons only, or whether there are higher interactions among more than k neurons which cannot be explained in terms of interactions of order up to k. Measuring all interactions of up to k variables means measuring all the marginal response probabilities involving up to k variables. Therefore any probability matching the observed interactions of up to k elements must preserve the same marginal response probabilities of up to order k as the original distribution.

The probability distribution $P_k^{ME}(\mathbf{r}|s)$ with maximum entropy among those satisfying the constraints of equality of marginal probabilities up to order k, is the one that imposes the absence of any additional interactions of higher order. The case $k = 1$ corresponds to all neurons firing independently at fixed stimulus (i.e. $P_1^{ME}(\mathbf{r}|s) = P_{ind}(\mathbf{r}|s)$).

Following (Amari, 2001; Cover & Thomas, 2006), it can be shown that there is a unique solution to the constrained maximum entropy problem for any order k. This solution takes in general an exponential form. In the specific case of binary response variables (when the maximum number M of spikes per neuron equals 1, and $r_i \in \{0, 1\}$) the k^{th} order ME solution can be written in the following form:

$$\begin{aligned}P_k^{ME}(\mathbf{r}|s) &= \exp\left\{\theta_0(s) + \sum_i r_i\theta_i(s) + \sum_{i_1<i_2} r_{i_1}r_{i_2}\theta_{i_1i_2}(s) + \cdots\right.\\ &\left. + \sum_{i_1<\cdots<i_k} r_{i_k}\cdots r_{i_k}\theta_{i_1\cdots i_k}(s)\right\} \qquad (14)\end{aligned}$$

The set of indices $i_1, \ldots, i_a$ label the subsets of a variables among the total C considered.

In the more general case of variables taking values from any finite alphabet, the k^{th} order exponential ME solution takes the following more complicated form (Amari, 2001):

$$\begin{aligned}P_k^{ME}(\mathbf{r}|s) &= \exp\left\{\theta_0(s) + \sum_i \sum_{a\in\mathbf{A}'} \delta^a(r_i)\theta_i^a(s) + \right.\\ &\sum_{i_1<i_2}\sum_{a_1,a_2\in\mathbf{A}'} \delta^{a_1}(r_{i_1})\delta^{a_2}(r_{i_2})\theta_{i_1i_2}^{a_1a_2}(s) + \\ &\left.\cdots \quad + \sum_{i_1<..<i_k}\sum_{a_1..a_k\in\mathbf{A}'} \delta^{a_1}(r_{i_1})..\delta^{a_k}(r_{i_k})\theta_{i_1..i_k}^{a_1..a_k}(s)\right\} \qquad (15)\end{aligned}$$

where $\mathbf{A}' = \mathbf{A} \setminus \{0\}$ is the set of members of the finite alphabet considered, excluding the 0 value, and (following (Amari, 2001)) we define an indicator function as follows:

$$\delta^a(r_i) = \begin{cases} 1, & r_i = a \\ 0, & r_i \neq a \end{cases} \qquad (16)$$

To quantify whether interactions of up to k neurons in a population are sufficient to describe the probabilities of neural responses to stimuli, we can quantitatively compare the true distribution $P(\mathbf{r}|s)$ of neural responses to the stimulus s to the distribution $P_k^{ME}(\mathbf{r}|s)$. By performing this comparison over a range of values of k, we can empirically determine the minimal k necessary to fit the empirically measured response probability well.

In order to compute the maximum entropy distribution $P_k^{ME}(\mathbf{r}|s)$ of Eq. (14) from real data, we need to find the θ coefficients with up to k indices. These θ coefficients can be determined from the experimentally measured marginal probabilities of up to k elements through a set of algebraic equations which were derived in the work of Amari (Amari & Nagaoka, 2000; Amari, 2001). To solve these equations numerically, we used our recently developed and publicly available[1] `pyentropy` numerical package (Ince et al., 2009b). We refer to (Ince et al., 2009b) for full details of the algorithm and code implementing the numerical solutions.

It is important to note that the maximum entropy models described and used here are not the only way to investigate the presence and effect of high order interactions. We refer the reader for example to Gütig et al. (2003); Staude et al. (2009); Onken et al. (2009) for examples of other techniques based on cumulants or copulas.

4.2. Defining the effects of interactions on information

After correlations have been defined, the next step is to characterize how they affect information transmission. Here, for simplicity we focus only on two specific information theoretic measures, which are designed to address two different specific questions: what is the impact of interactions up to a given order on the total information about stimuli encoded by the population, and what order of interactions a downstream system needs to take into account in order to extract all the information about stimuli available from neural population activity.

4.2.1. Measures of how interactions affect encoding

To understand what is the impact of interactions up to a given order on the total information encoded by the population, it is useful to compare the information $I(S; \mathbf{R})$ available in the population including interactions at all orders with the information $I_k(S; \mathbf{R})$ that would be available if only interactions up to a given (low) order were present. The information that would be available if only interactions up to a given order k were present can be evaluated by calculating the mutual information that would result from a system exhibiting the probability distributions obtained from the maximum entropy solution, as follows:

$$I_k(S; \mathbf{R}) = H_k(\mathbf{R}) - H_k(\mathbf{R}|S) \tag{17}$$

where $H_k(\mathbf{R})$ and $H_k(\mathbf{R}|S)$ are the response and noise entropies respectively of the k-th order maximum entropy model. These entropies are obtained by

[1] http://code.google.com/p/pyentropy/

replacing $P(\mathbf{r}|s)$ and $P(\mathbf{r})$ with $P_k^{ME}(\mathbf{r}|s)$ and $P_k^{ME}(\mathbf{r})$ in Eqs. (1,2), where $P_k^{ME}(\mathbf{r}) = \sum_s P_k^{ME}(\mathbf{r}|s)P(s)$:

$$\begin{aligned} H_k(\mathbf{R}|S) &= -\sum_{s\in S} P(s) \sum_{\mathbf{r}\in\mathbf{R}} P_k^{ME}(\mathbf{r}|s) \log_2 P_k^{ME}(\mathbf{r}|s) \\ H_k(\mathbf{R}) &= -\sum_{\mathbf{r}\in\mathbf{R}} P_k^{ME}(\mathbf{r}) \log_2 P_k^{ME}(\mathbf{r}) \end{aligned} \tag{18}$$

Then

$$I_k(S;\mathbf{R}) = \sum_{\mathbf{r},s} P(s)P_k^{ME}(\mathbf{r}|s) \log_2 \frac{P_k^{ME}(\mathbf{r}|s)}{P_k^{ME}(\mathbf{r})} \tag{19}$$

4.3. Measures of the information lost in decoding with the simplified model

A second, equally important question is whether a downstream system needs to take into account interactions at all orders for extracting all the information available from neural population activity. Following Wu et al. (2000, 2001); Nirenberg et al. (2001), the problem can be formalized by considering a downstream neural system that extracts information about the stimulus by relying on the assumption that the spikes are generated by a simplified response model that contains only correlations up to a given order. For example, the downstream system may decode the stimulus using, via Bayes' rule, a posterior probability based on the simplified maximum entropy model which neglects interactions of order higher than k:

$$P_k^{ME}(s|\mathbf{r}) = \frac{P_k^{ME}(\mathbf{r}|s)P(s)}{P_k^{ME}(\mathbf{r})} \tag{20}$$

An important question is how much information is lost because the information-extracting operation is performed assuming that responses $\mathbf{r}$ are generated with the simplified maximum entropy distribution $P_k^{ME}(\mathbf{r}|s)$ rather than with the true probability distribution $P(\mathbf{r}|s)$. An an upper bound to this information loss is expressed by the following simple closed-form expression(Nirenberg et al., 2001; Latham & Nirenberg, 2005):

$$\begin{aligned} \Delta I_k &\equiv D(P(s|\mathbf{r})||P_k^{ME}(s|\mathbf{r})) \\ &\equiv \sum_{\mathbf{r}} P(\mathbf{r}) \sum_s P(s|\mathbf{r}) \log_2 \frac{P(s|\mathbf{r})}{P_k^{ME}(s|\mathbf{r})} \end{aligned} \tag{21}$$

where D is conditional Kullback-Leibler (KL) distance (Cover & Thomas, 2006, Eq. 2.65). The ME model construction ensures that if, for some $\mathbf{r}$ and

s, $P_k^{ME}(\mathbf{r}|s)$ is zero, then $P(\mathbf{r}|s)$ must also be zero, and this in turn ensures that ΔI_k is a non-divergent information-theoretic measure.

A second quantity of interest is I_{LB-k} (Pola et al., 2005), defined as the difference between the mutual information I and ΔI_k:

$$\begin{aligned} I_{LB-k} &= I - \Delta I_k \\ &= \sum_{\mathbf{r},s} P(\mathbf{r},s) \log_2 \frac{P_k^{ME}(\mathbf{r}|s)}{P_k^{ME}(\mathbf{r})} \end{aligned} \tag{22}$$

Since ΔI_k is non-negative and is an upper bound to the information lost when decoding the neuronal responses with the mismatched response model P_k^{ME}, I_{LB-k} has a well defined meaning: it provides a lower bound to the information that can be decoded by using the simplified probability distribution P_k^{ME}.

The maximal amount of information $\hat{I}_k$ that can be decoded by using the mismatched decoder based on $P_k^{ME}(\mathbf{r}|s)$ (rather than on the true distribution $P(\mathbf{r}|s)$) can be quantified by computing the maximum over the parameter β of the following quantity (Merhav et al., 1994; Oizumi et al., 2009):

$$\begin{aligned} I_k^{(\beta)} &= -\sum_{\mathbf{r}} P(\mathbf{r}) \log_2 \left[\sum_s \left[P_k^{ME}(\mathbf{r}|s) \right]^\beta P(s) \right] \\ &\quad + \sum_{\mathbf{r},s} P(\mathbf{r}|s) P(s) \log_2 \left[P_k^{ME}(\mathbf{r}|s) \right]^\beta \end{aligned} \tag{23}$$

It can be shown (Oizumi et al., 2009, 2010) that $\hat{I}_k \leq I(S;\mathbf{R})$ and that I_{LB-k} equals $I_k^{(\beta)}$ for $\beta = 1$. Since $\hat{I}_k$ is the maximum over β of $I_k^{(\beta)}$, it follows that $\hat{I}_k \geq I_{LB-k}$, confirming that I_{LB-k} is a lower bound. When the value of $\hat{I}_k$ is strictly lower than $I(S;\mathbf{R})$, it follows that including in the decoding model correlations up to order k is not enough to decode the entire information in the neural responses.

The computation of the precise amount of information decodable by taking into account correlations up to order k, $\hat{I}_k$ has not been used in neuroscience until very recently (Oizumi et al., 2009, 2010). Its sampling properties and the best procedures to estimate it from a limited amount of data are still largely unexplored. Given this quantity can in principle provide important answers to characterizing simple but efficient ways to read out a population code, we suggest that investigating in detail the statistical issues regarding the valuation of $\hat{I}_k$ is an important topic for further research.

4.4. Other information theoretic measures of the importance of neural interactions in coding

The quantities $I_k(S;\mathbf{R})$, ΔI_k, I_{LB-k}, $\hat{I}_k$ are by no means the only information theoretic measures of the importance of correlations that have been developed. Although for space reasons we cannot describe them all here, in this subsection we briefly review some of the main ones and (when appropriate) we discuss their relationship with the information theoretic quantities $I_k(S;\mathbf{R})$, ΔI_k, I_{LB-k}, $\hat{I}_k$ that we presented above.

The idea of evaluating the impact of interactions on information by taking the difference between the information $I(S;\mathbf{R})$ available in the population including all interactions and the information $I_1(S;\mathbf{R})$ that would be available if the single-neuron marginal were the same but no interactions were present was introduced some 10 years ago in (Panzeri et al., 1999; Hatsopoulos et al., 1998; Nirenberg & Latham, 1998). The difference between $I(S;\mathbf{R})$ and $I_1(S;\mathbf{R})$ was previously termed I_{cor} by Pola et al. (2003) and ΔI_{noise} in Schneidman et al. (2003), the latter name being due to the fact that interactions and correlations measures at fixed stimulus are usually named noise correlation[2].

Panzeri and colleagues (Panzeri et al., 1999; Pola et al., 2003) introduced a so called "information breakdown" formalism which separates the total impact of neural interactions on encoding I_{cor} (which, as mentioned above, equals $I(S;\mathbf{R})-I_1(S;\mathbf{R})$) into two contributions $I_{cor-ind}$ and $I_{cor-dep}$, reflecting stimulus independent and stimulus dependent interactions respectively. The quantity $I_{cor-dep}$ equals ΔI_1. Therefore the quantify ΔI_k in Eq (21) can also be interpreted as the contribution of stimulus modulations of interactions of order higher than k. More recent generalizations of the information breakdown have focused on how to carefully separate the contributions of interactions between spikes emitted by different neurons from the contributions of interactions among spikes from the same neuron (Scaglione et al., 2008, 2010).

The idea of using ME methods to separate out the information about stimuli carried by different orders of neural interactions was pioneered by Nakahara & Amari (2002), who investigated how to separate information attributable to stimulus modulation of higher order interactions from the

[2] The quantity $I_1(S;\mathbf{R})$ was also called $I_{ind}(S;\mathbf{R})$ or $I_{lin} + I_{sig-sim}$ in (Pola et al., 2003)

information attributable to stimulus modulation of lower order marginals. The idea (reviewed in previous subsections) of computing the effect of interactions of order higher than k on encoding of stimulus-related information by considering the difference between $I(S;\mathbf{R})$ and $I_k(S;\mathbf{R})$ was introduced in (Montani et al., 2009; Ince et al., 2009a).

5. Neurophysiological data

After having described information theoretic techniques to study the effect of interactions of up to any given order on information transmission by neural populations, we illustrate their use by applying them to a population of neurons recorded from the whisker representation in the somatosensory cortex of urethane anaesthetized rats. We first describe the dataset and we then evaluate the effect of the interaction order on the information about the stimuli carried by the neural responses.

The data set, previously published in (Arabzadeh et al., 2003, 2004), consists of 24 simultaneously recorded neural clusters, each sampled with a different electrode with a minimal inter-electrode distance of 400 μm. 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 considered together as a single neural cluster. It has been estimated that, under these recoding conditions, each cluster captured the spikes of approximately two to five neurons located near the tip of the electrode (Petersen & Diamond, 2000). Neural activity was recorded in response to stimulation (with a piezoelectric wafer controlled by a voltage generator) consisting of sinusoidal whisker vibrations, each defined by a different value of vibration velocity and delivered for 500 ms (see (Arabzadeh et al., 2004) for full details). Thirteen different values of vibration velocity were tested, ranging between $Af = 0.15$ mm/s and $Af = 47.7$ mm/s. Each value of vibration velocity was treated as a different stimulus s; there were 13 stimulus classes in total. The number of recorded repetitions for each stimulus (called 'trials' in neurophysiology), from which the probability of response at fixed stimulus is determined, varied between a minimum of 200 and a maximum of 1400 across stimuli. In each trial, the population response $\mathbf{r}$ is computed as follows. It was previously shown (Arabzadeh et al., 2004) that the majority of the information is transmitted very early post stimulus onset. We therefore quantified the response using a post-stimulus time window of $10-15$ ms,

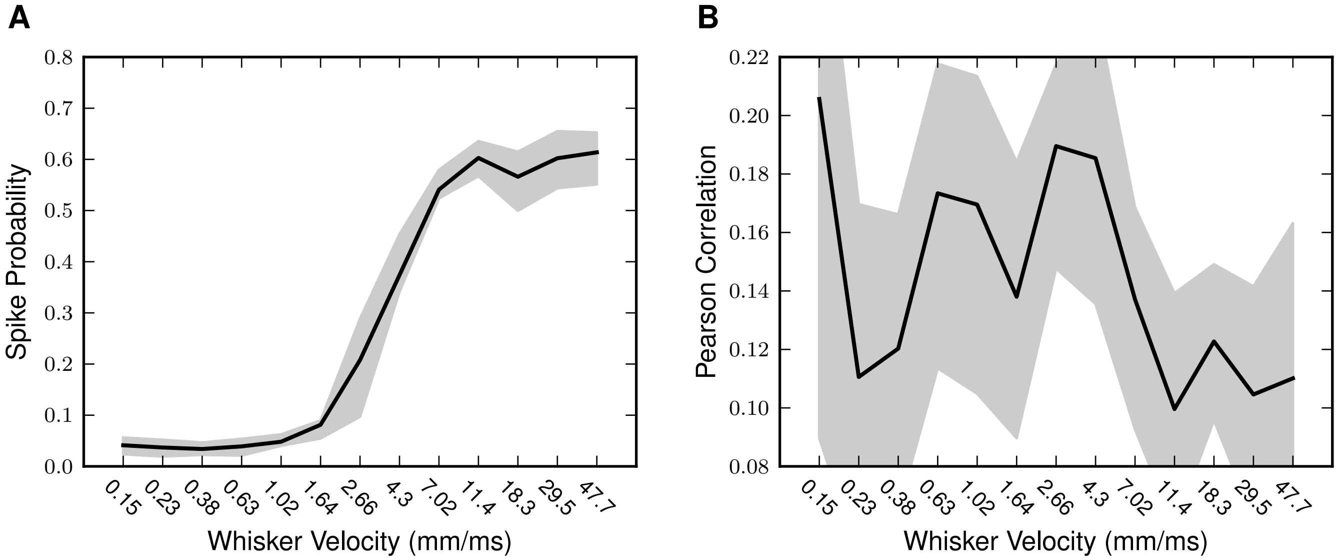

Figure 3: *Mean firing rate and correlation coefficients for the data.* The mean probability of observing a spike in the 10−15ms post stimulus window considered is shown (panel **A**) as a function of the whisker velocity stimulus. Black line shows mean spike probability across the 8 cells considered; grey region shows 25^{th}-75^{th} percentile. The Pearson correlation coefficients of neural responses for each stimulus are also shown (panel **B**). Black line shows mean correlation coefficient across the 28 pairs among the 8 cells considered; grey region shows 25^{th}-75^{th} percentile.

which was the window which maximized the information about the stimulus conveyed by the responses of individual channels. To facilitate the sampling of the population response probabilities, for each recording channel we "binarized" the responses, i.e. we set the response of each channel to 1 if at least one spike occurred in in $10 - 15$ ms post-stimulus window, and 0 otherwise. The reason for the binarization of responses was that, although we had enough data to reliably compute information in binarized responses of up to 8 cells, we did not have enough data to compute information in multi-level population responses. However, we checked that, at the single neuron level, the binarization of neural responses had a small effect on information: the single channel information about the stimuli was 0.216 ± 0.003 bits for the binarized spike count responses and 0.221 ± 0.003 for the unprocessed spike count response (average ± SD across the population).

For performing the information analysis, we decided to consider response distributions of $C = 8$ simultaneously recorded channels (out of the 24 available). The reason was that this population size was big enough to begin observing how the effect of interactions at various orders depends on the population size, while it was small enough to be sampled with the available data and provide sufficiently accurate information analysis. The robustness of the results with respect to sampling issues was verified by dividing the data into two halves, by recomputing each of the considered information quantities from each halved dataset, and by obtaining the result that none of the information quantities obtained with the entire dataset differed more than 2% from the ones obtained from the halved dataset over the entire population size range 2–8 (results not shown).

To get a better feeling for the data, in Fig. 3a we plot the mean and 25–75th percentile spread across the population of the probability of channel firing in the 10–15 ms post-stimulus window as function of the different values of whisker velocity used during the experiment. It is apparent that the firing rate of these cells increases monotonically with increasing stimulus velocity. The value of the mean spike probabilities are of interest in the context of studying the effect of correlations because previous studies have argued that the population is compelled to be well described by just considering pairwise correlations if the product of the mean spiking probability of a single neuron and the size of the population is much smaller than one (Panzeri et al., 1999; Roudi et al., 2009a). Fig 3a shows that this is not the case for our population. In fact, for the upper half of the presented stimulus velocities, firing probabilities are in the range of 0.4 - 0.5, and we will quantify the information carried by a population of up to $C = 8$ channels. This suggests that there is the possibility that we could find some contribution of higher order correlations to the transmitted information in the range of population sizes that we will analyze. In Fig 3b, we report the mean and 25-75th percentile spread across the population of the Pearson correlation coefficient computed from the binarized spike counts in the 10-15 ms post-stimulus window for each value of whisker velocity. These correlation coefficient distribution shows that these neurons are indeed correlated, and that correlations depend on the stimulus. For example, stimuli with high velocity tend to elicit lower response correlation than stimuli with low velocity. The fact that this population is correlated in a stimulus dependent way suggests that such correlations might play some role in this particular population code for stimulus velocity.

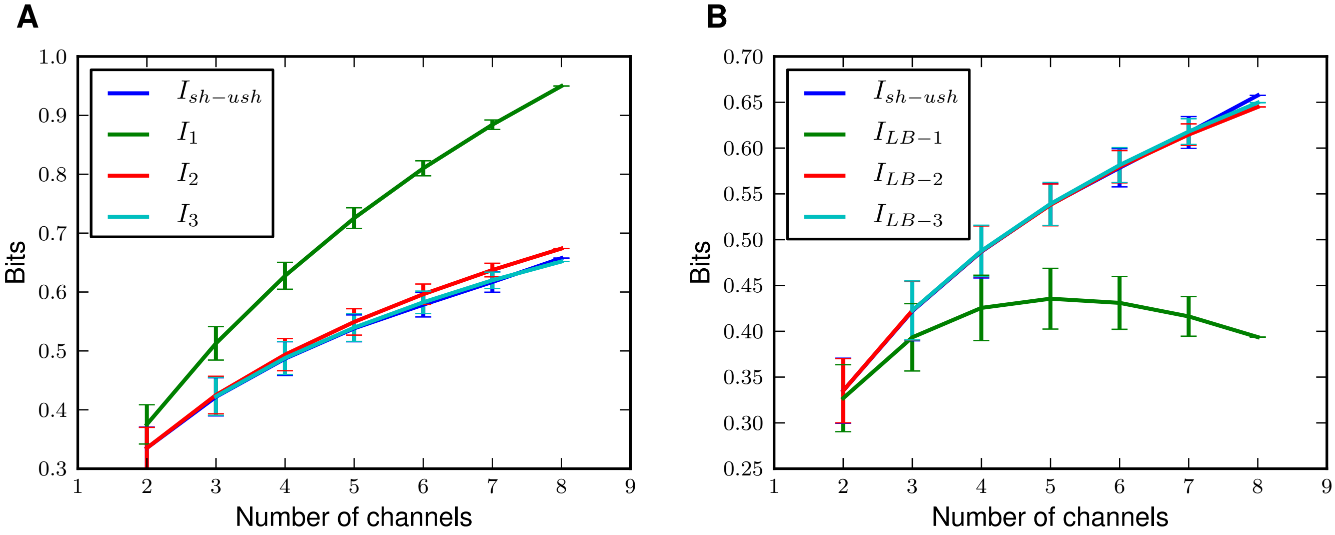

Figure 4: *Quantifying the effect of correlations on information.* The information $I_k(s; \mathbf{R})$ carried by systems containing correlations only up to order k, and shortened as I_k ($k = 1, 2, 3$) (panel **A**), lower bounds on information available to a decoder neglecting correlations of order higher than k, I_{LB-k} ($k = 1, 2, 3$) (panel **B**) and the I_{sh-ush} mutual information estimator are shown as a function of the number of cells considered. 8 channels from the experimental data set (see Section 5) were chosen; each point shows the mean value over all $\binom{8}{c}$ combinations of c channels. Error bars show ±1 SD. All values are corrected using quadratic extrapolation.

We begun the investigation of how this population encodes information by first considering how well $I_k(S;\mathbf{R})$, the information transmitted if interactions only up to order k were present, approximates the true total information carried by the population. To understand how these information theoretic quantities scaled with the population size, we computed the information carried by each m-plet of channels ($m = 2, \ldots, 8$) out of the 8 available channels. We averaged the information values over all $\binom{8}{m}$ m-plet's, and we plotted the mean information values as a function of the population size. Results are reported in Fig 4A. The information $I_1(S;\mathbf{R})$ carried if individual neurons had the observed marginal probabilities but they fired independently at fixed stimulus is much higher than the true information $I(S;\mathbf{R})$ carried by the population. Note that in the following discussion we refer to the full mutual information as $I(S;\mathbf{R})$, although in practice we computed its value from experimental data using the shuffled estimator $I_{sh-ush}(S;\mathbf{R})$ described above. The difference between $I(S;\mathbf{R})$ and $I_1(S;\mathbf{R})$ is small when considering 2-3 cells, but it grows steeply as a function of the population size. When considering 8 cells, $I_1(S;\mathbf{R})$ is almost 50% bigger than $I(S;\mathbf{R})$. This means that, in this system, the presence of interactions has a severely limiting effect on information transmission. This limiting effect grows with the population size. However, the information $I_2(S;\mathbf{R})$ computed taking into account interactions up to order two gave a good approximation to $I(S;\mathbf{R})$. For small population sizes (up to 4 cells), $I_2(S;\mathbf{R})$ was almost exactly equal to $I(S;\mathbf{R})$. For larger population size, the difference between $I_2(S;\mathbf{R})$ and $I(S;\mathbf{R})$ remained small, and reached 3.2% when considering 8 channels. Finally, the the information $I_3(S;\mathbf{R})$ computed taking into account interactions up to order three gave an essentially perfect (within 0.5%) approximation to $I(S;\mathbf{R})$ within the entire population size range considered.

In sum, within the population size range explored here, the mutual information of the system was very well approximated by models containing interactions up to order 2, and was perfectly approximated by models containing interactions of up to order 3. This is a significant simplification since it greatly reduces the parameters required to describe the system. For example, in the case of 8 binary cells as considered here, specification of the full distribution requires $2^8 - 1 = 255$ parameters, while the second and third order models require only 36 or 92 parameters respectively. While it is still challenging to sample second and third order marginals, it is a much more tractable problem than the case where all orders of interaction must be accurately determined. Moreover, we note that, to our knowledge this is the first report of $I_2(S;\mathbf{R})$ being a close approximation to $I(S;\mathbf{R})$ outside of the regime when the probability of observing a spike across all cells in a bin is not small compared to 1 (see (Roudi et al., 2009a) for the reasons why this result is of interest).

We then computed, as a function of the population size, the quantity I_{LB-k}, which is a lower bound to the information that can be extracted by a downstream system assuming that the probabilities of neural response contain only interactions up to order k. Results are plotted in Fig 4B. The quantity I_{LB-3} was equal to $I(S;\mathbf{R})$ over all the population size considered, meaning that a downstream decoder needs only to pay attention to correlation up to order 3 to extract all information available in this population. In fact, the quantity I_{LB-2} was also very close to $I(S;\mathbf{R})$ over all the population size considered, the difference between the two quantities however showing a slight tendency to increase as function of the population size. I_{LB-2} was within 1% of $I(S;\mathbf{R})$ when considering 8 channels, meaning that a downstream decoder operating on populations of up to 8 cells would decode essentially all of the information even when paying attention to correlations only up to order 2.

Interestingly, Fig. 4B also shows that the quantity I_{LB-1} was much smaller than $I(S;\mathbf{R})$, with the difference between the two quantities increasing very steeply with the population size. This is to our knowledge the first time that I_{LB-1} was reported to fail so dramatically to match $I(S;\mathbf{R})$, as previous studies reported a close match between I_{LB-1} and $I(S;\mathbf{R})$ for both neuronal pairs (Nirenberg et al., 2001; Golledge et al., 2003; Petersen et al., 2001; Montani et al., 2007) and for populations of few tens of neurons (Pillow et al., 2008). The fact that I_{LB-1} is much smaller than $I(S;\mathbf{R})$ is interesting because it raises the possibility that downstream receivers of barrel cortex activity must use knowledge of their correlation in order to extract the information that this neurons carry. However, the fact that I_{LB-1} is only a lower bound to the information that can be obtain by neglecting all types of interactions, makes it difficult to assess whether the large difference between I_{LB-1} and $I(S;\mathbf{R})$ comes from the fact that the knowledge of interactions was really necessary to decode all the information, or simply because the lower bound was not tight. We will investigate this point in more detail in later Sections.

A notable fact arising from the study of both I_{LB-k} and $I_k(S;\mathbf{R})$ is that in both cases the effect of neural interactions of information increases steeply with the population size. This means that studies on the role of correlations on information carried out on pairs of neurons do not necessarily generalize to large populations, and it stresses the importance of further developing the information calculation methods in order to evaluate the information content of larger and larger populations.

6. Using the knowledge of the correlation structure to reduce the bias

Another way to reduce the bias and the problems in computing information from limited datasets is to reduce the complexity of the population response space by fitting the probability distributions to simple, low-dimensional models. A prominent example of this strategy is to assume that the structure of the interactions between cells is described by a low dimensional model, for example the maximum-entropy model including correlations up to a given low order k. If this assumption is correct and is sufficient to describe the whole information content of population responses, then the resulting calculation of information from the low order model is potentially much more data robust.

The sampling advantages of information theoretic quantities based on low-dimensional probability models can be appreciated by computing different information theoretic quantities on a simulated population of 8 cells. Figure 5 compares the scaling with the number of trials of the information $I_1(S;\mathbf{R})$ and $I_2(S;\mathbf{R})$ (based on no interactions and on pairwise interactions respectively) to that of the full information $I(S;\mathbf{R})$. It is apparent that $I_1(S;\mathbf{R})$ can be computed without bias from as little as $2^3 = 8$ trials per stimulus, and is thus is much more data robust than $I_2(S;\mathbf{R})$, which requires at least $2^5 = 32$ trials per stimulus for unbiased computation. $I_2(S;\mathbf{R})$ is in turn is more data robust than $I(S;\mathbf{R})$ which is unbiased only with at least 2^{10} trials per stimulus. The reason for this behavior is that for small k fewer parameters are necessary to characterize the responses, and the fewer parameters that have to be sampled, the smaller the bias (see Eq. (4)). The same reasoning also applies to I_{LB-1} and I_{LB-2} which, as shown in Fig. 5, scale with the number of trials in a way similar to $I_1(S;\mathbf{R})$ and $I_2(S;\mathbf{R})$ respectively.

The drawback of using low order models to estimate information is that they do not converge to the correct information value if the wrong assumption is made about the minimal order k which fully captures the interaction in the

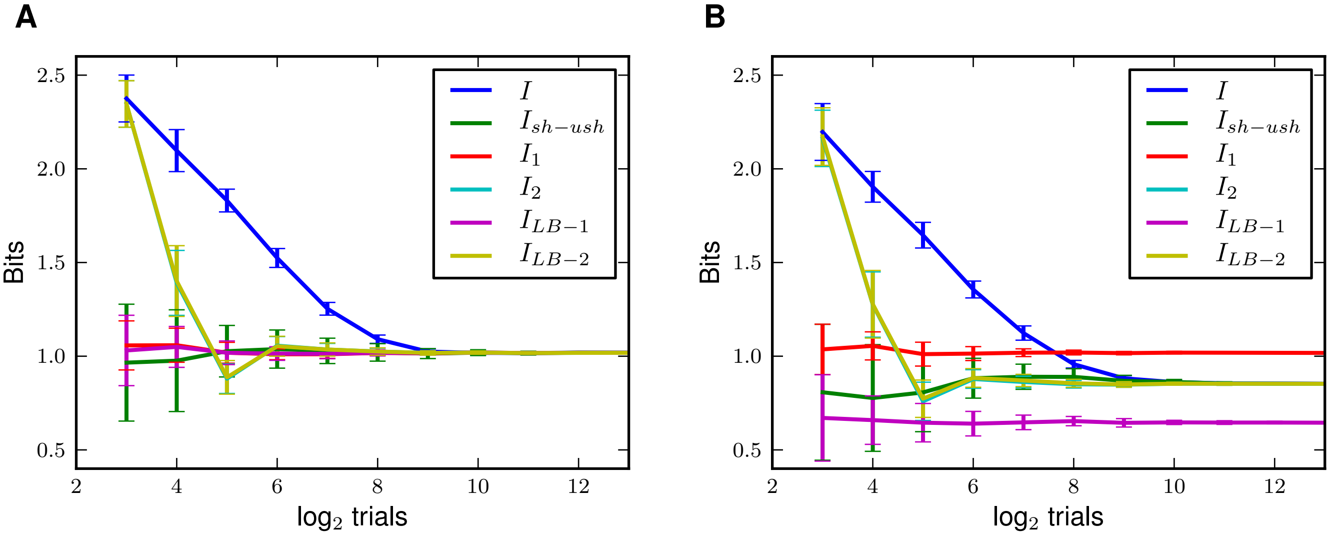

Figure 5: *Bias properties of mutual information bounds.* Information bounds I_k, I_{LB-k}, $(k = 1, 2)$ and mutual information estimates I, I_{sh-ush} are shown as a function of number of trials per stimulus. Two model systems are considered, one with individual marginals matching the experimental data, but no second or higher order correlations (panel **A**) and one with individual and pairwise marginals matching the experimental data as in Figures 1,2,4 (panel **B**). Each point represents the average over 50 simulations of the system; error bars show ±1 SD. All values are corrected using quadratic extrapolation.

data. This can be appreciated by comparing the simulations in Fig 5A and Fig 5B. In Fig 5A, data are simulated in such a way that stimulus-conditional response probabilities are independent (i.e. the responses are described by the $k = 1$ model). In this case, I_{LB-1} and $I_1(S;\mathbf{R})$ have to converge to the correct value of information $I(S;\mathbf{R})$ (because by construction the response probabilities are perfectly described by a $k = 1$ model), and so the sampling advantages of $I_1(S;\mathbf{R})$ and I_{LB-1} can be used at no price of distortion of information calculation. However, in Fig 5B, data are simulated in such a way that stimulus-conditional response probabilities are described by a pairwise model (i.e. the responses are described by the $k = 2$ model). In this case, I_{LB-2} and $I_2(S;\mathbf{R})$ converge to the correct value of information $I(S;\mathbf{R})$, but I_{LB-1} and $I_1(S;\mathbf{R}))$ do not. Using I_{LB-1} and $I_1(S;\mathbf{R}))$ would lead to a highly misleading evaluation of information, because the order $k = 1$ is in this case inadequate to describe the data. This example illustrates that using simplified probability models to compute information is only advisable if a rigorous criterion is used to select the simplest model that fully describes the data. Various parametric and non-parametric statistical procedures for such model selections have been proposed (Martignon et al., 2000; Nakahara & Amari, 2002; Kennel et al., 2005; Montemurro et al., 2007; Ince et al., 2009a), and we refer the reader to such articles for a thorough discussion of various proposals.

Fig. 5 also compares the sampling behavior of the information estimates $I_k(S;\mathbf{R})$ and I_{LB-k} based on lower order models ($k = 1, 2$) to the improved shuffled estimator $I_{sh-ush}(S;\mathbf{R})$ discussed in the previous section. The result is that $I_{sh-ush}(S;\mathbf{R})$ has a sampling behavior which is comparable to that of $I_1(S;\mathbf{R})$ and I_{LB-1}, with the added advantage of converging (for large datasets) to the true information value independently of the order of the interactions in the simulated model. This results stresses that the shuffled estimator remains competitive even when compared to estimators based on model selection.

7. Computing data robust information lower bounds using decoding techniques

Despite the progress with the bias correction procedures described above, when the neuronal population is large it becomes impossible to compute the information in neural responses directly because the number of possible responses $\mathbf{r}$ grows exponentially with the population size (this is sometimes called the curse of dimensionality). At some point even the bias correction procedures discussed above will be ineffective for the quantities of data that can be experimentally collected. Therefore, calculation of information from large populations remains a problem unless highly efficient ways to compress the response space with little information loss can be found.

A promising approach to the information analysis of larger populations is the use of information theory coupled to decoding approaches (Quian Quiroga & Panzeri, 2009). These procedures use a stimulus-decoding procedure to predict the most likely stimulus elicited from a single trial population response, and this makes it possible to compress the population response space into the space of 'predicted stimuli' (Quian Quiroga & Panzeri, 2009). If the number of stimuli is much smaller than the number of responses, stimulus-decoding is an effective and simple way to reduce the space of responses.

In more detail, this approach works as follows. Decoding can be defined as the prediction of which stimulus elicits a particular neuronal response in a single trial. More formally, decoding is a function $f(\mathbf{r})$ operating on the population response in any given trial and giving a prediction s^p of the stimulus that elicited the observed neural population response in that trial:

$$s^p = f(\mathbf{r}) \tag{24}$$

A prominent example of decoding is Bayesian decoding which predicts the most likely stimulus given the response as follows (Cover & Thomas, 2006):

$$s^p = \underset{s}{\operatorname{argmax}}\, P(s|\mathbf{r}) \tag{25}$$

From this decoding procedure performed on each trial, the performance of the decoder across all trials can be summarized by computing the so called confusion matrix $Q(s^p|s)$, which is defined as the fraction of times that a stimulus s^p was predicted in a given trial in which stimulus s was presented. To validate decoding results, some trials can be used to optimize the decoder (the training set) and the rest to test its performance, a procedure called cross-validation (Quian Quiroga & Panzeri, 2009). It is important that trials belonging to the training set are not used to evaluate the decoding performance because this may lead to artificially high values due to overfitting. A common procedure is the 'leave-one-out' validation, in which each trial is predicted based on the distribution of all the other trials. This has the advantage that both optimization and assessment testing are based on the largest possible number of trials (Quian Quiroga & Panzeri, 2009).

Since the distribution of test responses to a given stimulus is given by $P(\mathbf{r}|s)$, $Q(s^p|s)$ can be written down as:

$$Q(s^p|s) = \sum_{\mathbf{r}} \delta(s^p, f(\mathbf{r}))P(\mathbf{r}|s) \tag{26}$$

where δ is the Kronecker delta function. The 'decoded' information $I(S;S^P)$ is then quantified as follows:

$$I(S;S^P) = \sum_{s}\sum_{s^p} P(s)Q(s^p|s)\,\log_2 \frac{Q(s^p|s)}{Q(s^p)} \tag{27}$$

where $Q(s^p) = \sum_s Q(s^p|s)P(s)$.

It is useful to note that information theoretic inequalities ensure that $I(S;S^P) \leq I(S;\mathbf{R})$ (Cover & Thomas, 2006). The reason why $I(S;S^P)$ can be less than $I(S;\mathbf{R})$ even when the decoding algorithm is well constructed and the probability model used for decoding is correct is that the decoding operation captures only one aspect of the information carried by the population response, namely the identity of the most likely stimulus. However, neural populations can carry information by other means than by reporting the most likely stimulus. For example, they can carry information by reporting which stimuli are very unlikely and should be ruled out, or they can carry additional information by reporting the identity of the second most likely stimulus, and so on (Quian Quiroga & Panzeri, 2009). The quantity $I(S;\mathbf{R})$ automatically captures all these ways to carry information, whereas $I(S;S^P)$ does not. However, other information can be added from the decoding procedure by progressively including in the calculation information about less likely or unlikely stimuli. For example, one can extend the information carried by the most likely stimulus prediction $I(S;S^P)$ by adding knowledge of the second most likely stimulus to the decoded information:

$$I(S;S^{P1}S^{P2}) = \sum_{s}\sum_{s^{p1},s^{p2}} P(s)Q(s^{p1}s^{p2}|s)\,\log_2 \frac{Q(s^{p1}s^{p2}|s)}{Q(s^{p1}s^{p2})} \tag{28}$$

where $Q(s^{p1}s^{p2}|s)$ is probability of predicting stimulus s^{p1} and s^{p2} as most and second most likely stimulus respectively when stimulus s is presented. Again, information theoretic inequalities ensure that $I(S;S^P) \leq I(S;S^{P1}S^{P2}) \leq I(S;\mathbf{R})$, and so on for adding more aspects of stimulus likelihood (for example, computing the information $I(S;S^{P1}S^{P2}S^{P3})$ carried by the identity of the three most likely stimuli). Progressively adding more and more knowledge about the order of stimulus likelihood and the relative likelihood of all stimuli should eventually let the procedure to converge to $I(S; \mathbf{R})$.

It is interesting to consider how to evaluate the role of correlations at given order using the decoding procedure outlined above. To address this, one can decode the stimulus using a decoding algorithm based on the simplified posterior probability model containing correlations up to order k, $P_k^{ME}(s|\mathbf{r})$, rather than the true distribution $P(s|\mathbf{r})$, as follows:

$$s_k^p = \underset{s}{\operatorname{argmax}}\, P_k^{ME}(s|\mathbf{r}) \tag{29}$$

Using this decoded stimulus one can compute $I(S; S_k^P)$, the information obtained by Bayesian decoding of the most likely stimulus using the kth order model through Eq. (29). This is done by first computing the confusion matrix $Q(s_k^p|s)$ obtained from Eq. (26) when using the Bayesian k-th order decoder (Eq. (29)) as the decoding function $f(\mathbf{r})$, and by then inserting this decoding matrix into Eq. (27).

We note that to compute the above equation for $Q(s_k^p|s)$, the "decoder" needs to know $P_k^{ME}(\mathbf{r}|s)$, but not $P(\mathbf{r}|s)$. In fact, $P(\mathbf{r}|s)$ is not needed in the single trial decoding operation, and the average over $\mathbf{r}$ in Eq. (26) is done by simply counting how many times a stimulus was presented and then decoded as s_k^p.

As outlined above, one can also extend this calculation to include the information carried by the second most likely stimulus with the k-th order model $I(S; S_k^{P1} S_k^{P2})$, and so on. The difference with respect to decoding based on the full probability is that in the case of k-th order model-based decoding, adding more and more knowledge does not necessarily lead to converge to $I(S; \mathbf{R})$. The value of information to which this iterative procedure converges to can be taken as an estimation, or a definition, of the maximal amount of information that can be extracted through the mismatched k-th order model. It remains to be understood in which conditions this iterative estimation of the maximal amount of information that can be extracted through the mismatched k-th order model is equivalent to the one denoted as $\hat{I}_k$ and discussed above (Merhav et al., 1994; Oizumi et al., 2009), and in which conditions these two procedures differ in estimated value and meaning.

In the following we explore the iterative procedure by applying it to populations of up to 8 somatosensory recording channels responding to whisker stimulations of different velocities from the data set considered in Figs 3,4

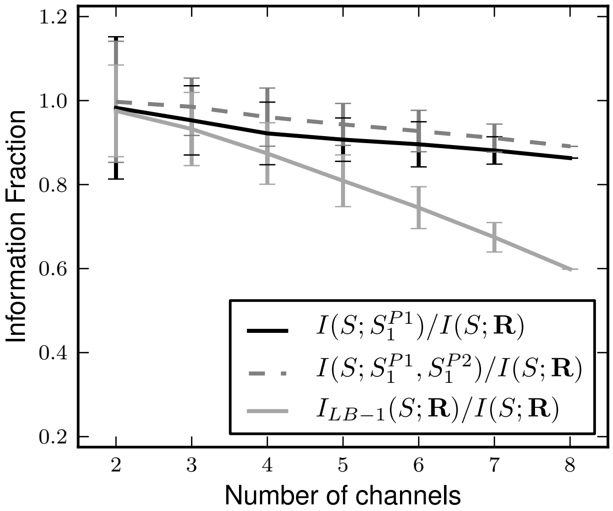

Figure 6: *Decoded information using the independent model.* 8 clusters from the experimental data set (see Section 5) were chosen and for all $\binom{8}{c}$ combinations of c clusters, the decoded information under the independent model considering the most likely stimulus, $I(S;S_1^{P1})$, and considering the two most likely stimuli $I(S;S_1^{P1}S_1^{P2})$ was computed. These values are shown as fractions of the full information $I(S;\mathbf{R})$ computed using the shuffled estimator and quadratic extrapolation. Error bars show ±1 SD.

(see Section 5). We remind the reader that we previously found (Fig 4) that the lower bound to the information that could be decoded by the independent model I_{LB-1} was much lower than the full information $I(S;\mathbf{R})$ carried by the population, and that by using only this lower bound we could not resolve whether the result meant that taking into account interactions was indeed necessary to decode the information, or it meant simply that the lower bound was not tight. Here we explore the use of the decoding methodology to address this issue. We computed the information $I(S;S_1^{P1})$, obtained by maximum likelihood decoding using the independent ($k=1$) model, and then iteratively added knowledge about other less likely stimuli (i.e. computing $I(S;S_1^{P1}S_1^{P2})$ etc.) and we investigated whether this procedure converged to the full information $I(S;\mathbf{R})$.

Fig 6 shows that in this population $I(S;S_1^{P1})$ was higher than I_{LB-1}, demonstrating that the lower bound was indeed not tight, especially when considering more than 3-4 channels. However, $I(S;S_1^{P1})$ remained lower than $I(S;\mathbf{R})$ and the difference between $I(S;S_1^{P1})$ and $I(S;\mathbf{R})$ increased with the population size. For 8 channels, $I(S;S_1^{P1})$ was 86% of $I(S;\mathbf{R})$. This suggests that decoding the most likely stimulus with the independent model is not enough to decode all information, especially when considering larger populations. To understand whether more information can be extracted by the independent model, we computed also the information obtained from the knowledge of the most and second most likely stimulus of the independent model $I(S;S_1^{P1}S_1^{P2})$. This quantity was significantly higher than $I(S;S_1^{P1})$. We verified the statistical significance of the increase by computing $I(S;S_1^{P1}S_{rand})$, the information carried by the most likely stimulus of the $k=1$ model and the addition of a second "dummy" decoded stimulus, S_{rand}, which was chosen at random. This was lower than $I(S;S_1^{P1}S_1^{P2})$ across all population sizes; one-way anova, $p<0.05$. However, $I(S;S_1^{P1}S_1^{P2})$ did not reach $I(S;\mathbf{R})$, which means that knowing both the most likely and the second most likely stimulus with the independent model is still not enough to decode all information. We then added the knowledge of the identity of the third stimulus of the independent model (i.e. we computed $I(S;S_1^{P1}S_1^{P2}S_1^{P3})$), and we found that the addition of the knowledge about the third most likely stimulus did not add any more information (as shown by the statistical test of adding a third random stimulus to the first two, one way anova, $p>0.5$). Similarly, adding more estimates of other types of knowledge (for example, knowledge of the least likely stimulus of the independent model, and so on) did not significantly increase the information with respect to $I(S;S_1^{P1}S_1^{P2})$. Therefore we took $I(S;S_1^{P1}S_1^{P2})$ as an empirical estimate of the maximal amount of information that can be extracted by interpreting the data using an uncorrelated ($k=1$) model. The quantity $I(S;S_1^{P1}S_1^{P2})$ was significantly less than $I(S;\mathbf{R})$ (one-way anova, $p<0.05$). Fig 6 shows that the gap between $I(S;S_1^{P1}S_1^{P2})$ and $I(S;\mathbf{R})$ also increased with the population size. For 8 channels, $I(S;S_1^{P1}S_1^{P2})$ was 89% of $I(S;\mathbf{R})$. We concluded that, although the independent model allowed the extraction of a good fraction of the total information, decoding while ignoring interactions was not enough to extract the whole information about the population, particularly for larger population sizes.

The iterative decoding procedure to compute the maximal amount of information extractable from a given level of correlation is less elegant than a direct calculation of $\hat{I}_k$ (Merhav et al., 1994; Oizumi et al., 2009). However, it has the advantage that it gives an explicit construction of how to extract information from a probability model, and it also is useful to understand which aspects of the posterior probability distributions of neural population responses carry information. Moreover, such iterative decoding evaluation may be convenient in cases in which the population size is too large and so sampling problems prevent a direct calculation of $I(S;\mathbf{R})$ and $\hat{I}_k$.

8. Discussion

Information theoretic tools provide metrics which are useful to understand how populations of neurons encode information in single trials, and they can help to understand how neuronal interactions shape the way in which neural populations represent and transmit messages about the sensory environment. However, information-theoretic calculations are difficult with neuronal populations because of the curse of dimensionality and the resulting sampling bias problem. Because of these problems, until very recently most information theoretic studies of neural codes concentrated only on single neurons or on pairs of neurons. In recent years, however several techniques have been developed to ameliorate the problems caused by the limited sampling bias. One of the most promising of these is the I_{sh} estimator (Montemurro et al., 2007), which we presented here together with a novel extention (I_{sh-ush}) which improves the performance of the estimator for systems where the number of stimuli is small. These techniques now permit the computation of the information carried by populations of up to some 8 neurons. This enables scientists to begin exploring information processing in local networks. One of the findings that the analysis of these networks starts to reveal, and which was highlighted by the somatosensory cortical examples presented here, is that the effect of interactions among neurons increases steeply with the population size. This means that, as previously discussed in theoretical studies (Averbeck et al., 2006; Roudi et al., 2009a), it could be potentially dangerous to assume that conclusions about neural codes obtained with small populations generalize in a straightforward way to larger populations. This, in turn, implies that future work needs to set the bar for analyzable population size even higher than now. For the above reasons, a major and important challenge for computational neuroscientists is to find ways to further extend the feasibility of performing information theoretic computations with larger populations. In this review, we have highlighted two directions which are particularly promising.

First, we have considered how to investigate the interaction structure of cortical population responses through using simplified maximum entropy models preserving interactions up to a given order, but featuring no higher order correlations. The use of maximum entropy models containing only low-order interactions has been the subject of intense studies in neuroscience over the last few years (Martignon et al., 2000; Nakahara & Amari, 2002; Schneidman et al., 2006; Shlens et al., 2006; Tang et al., 2008; Nirenberg & Victor, 2007; Montemurro et al., 2007). These models have been mostly used for either inferring network connectivity (Roudi et al., 2009b; Tatsuno et al., 2009) or for quantifying the effect of interactions on the so called fraction of network information (which measures the reduction of network variability specifically attributable to correlations up to a given order, see Schneidman et al. (2003, 2006)). Here, we extended the maximum entropy approach to quantify the effect of interactions up to a given order to the mutual information that population responses carry about sensory stimuli. The latter is in principle distinct from their effect on the fraction of connected information, because the variability of the population response is not equivalent to the information they carry about the stimuli. In the example analysis of somatosensory data that we reported in this paper, we found that within the population size range ($2-8$ channels) explored here, the mutual information between stimuli and population responses was very well approximated by models containing interactions up to order 2, and was perfectly approximated by models containing interactions of up to order 3. While it is still challenging to sample second and third order marginals, it is a much more tractable problem than the case where all orders of interaction must be accurately determined. Therefore, as well as revealing characteristics of local network processing, accurate analysis based on low order maximum entropy models can be used to effectively reduce the number of parameters describing how neural population responses carry information about the stimuli, thereby greatly reducing the information bias problem.

Second, we have explored the use of decoding techniques as a robust and efficient way to reduce the dimensionality of the response space while losing little information. We have described a set of hierarchial approximations to the information $I(S; \mathbf{R})$ carried by neural responses which are based on decoding various features of the response (e.g. the identity of the most or least likely stimuli). In conditions when the number of stimuli is not excessive and the population size is large, these techniques provide a data robust way to iteratively approximate the mutual information carried by the population activity, and to learn which aspects of the posterior probability distributions of neural population responses are most important for carrying information. As discussed in this Review, these techniques can be used in conjunction with low-interaction-order maximum entropy models to evaluate the performance of such simplified models in extracting information from population activity. This approach, together with those developed in (Merhav et al., 1994; Latham & Nirenberg, 2005; Oizumi et al., 2009), could help in discovering the minimal set of response features needed to decode all information from population responses.

As the development of information theoretic analysis tools becomes increasingly specialised, it is important to make sure they are available to experimental groups to apply to a wide range of data. An excellent way of achieving this goal is to make the code freely and publicly available, a practice known as open source. One of the factors which has (in our view) limited the expansion of the use of information theory in the analysis of neuroscience data has been the lack of such open source analysis packages containing state of the art techniques. Fortunately, in recent years several groups (including ours) have released open source information theoretic packages for the analysis of neuroscience data (Ince et al., 2009b; Magri et al., 2009; Goldberg et al., 2009). All the techniques and calculations implemented in this paper were implemented through calls of the routines of the open source entropy software of (Ince et al., 2009b), and are therefore relatively easy to reproduce. We believe that the continued expansion of such open source effort in information theoretic analysis and in other areas of neuroscience data analysis is important because it holds the promise for a significant advance in the standardization, transparency, quality, reproducibility and variety of techniques used to analyze neurophysiological data.

In summary, it is our hope that data analysis tools such as those described here will help to provide insights into the mechanism of neuronal computation. In addition to revealing features of the biological system through direct analysis of experimental data, such tools can also be used to provide additional metrics for comparing the results of large-scale models with real neural systems.

Appendix A. Simulation of cortical somatosensory neural responses

In this appendix, we describe briefly the procedures used to generate the simulated data used in the figures. We obtained the parameters for the simulated systems from the experimental data described in Section 5. As discussed, this data set consists of simultaneous recordings of neural clusters from somatosensory cortex of urethane anaesthetised rats in response to sinusoidal whisker stimulation at 13 different average velocities. A binary response was obtained for each cell consisting of a 1 if there was at least one spike in a window of $10 - 15$ ms post-stimulus onset, and a 0 if no spike was emitted in this window. This window was chosen since it was the window for which the presence or absence of one or more spikes was most informative about the stimulus presented.

For the responses of these clusters, the independent firing probabilities (first order marginals) and pairwise marginal firing probabilities were obtained for each stimulus condition. The empirically obtained maximum entropy solutions (Section 4.1) preserving these first and second order marginals were used as the model stimulus conditional distributions for generating data in Figs 1,2 and 5B. The maximum entropy solutions preserving only first order marginals were used for Fig 5A.

With the model stimulus conditional distributions defined as above, fixed numbers of trials per stimulus were generated via inverse transform sampling, and the resulting data set was analysed with the information theoretic `pyentropy` package[3] (Ince et al., 2009b)

Acknowledgements

We thank the referees for thoughtful and detailed critiques which significantly improved the revised version of the manuscript. We are indebted to R. Petersen and M. Montemurro for valuable collaborations on the topics discussed here. This research was supported by the BMI project of the Italian Institute of Technology and by the EPSRC CARMEN project (grant EP/E002331/1).

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

Amari, S., & Nagaoka, H. (2000). *Methods of Information Geometry*. Oxford University Press.

[3] http://code.google.com/p/pyentropy/

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

Arabzadeh, E., Panzeri, S., & Diamond, M. E. (2004). Whisker Vibration Information Carried by Rat Barrel Cortex Neurons. *J. Neurosci.*, *24*, 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*, 9146–9154.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neurosci.*, *7*, 358–367.

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

Chase, S. M., & Young, E. D. (2007). First-spike latency information in single neurons increases when referenced to population onset. *Proc. Natl. Acad. Sci. USA*, *104*, 5175–5180.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory, 2nd Ed.*. John Wiley & sons.

Fuhrmann Alpert, G., Sun, F. T., Handwerker, D., D'Esposito, M., & Knight, R. T. (2007). Spatio-temporal information analysis of event-related BOLD responses. *Neuroimage*, *34*, 1545–1561.

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

Goldberg, D. H., Victor, J. D., Gardner, E. P., & Gardner, D. (2009). Spike Train Analysis Toolkit: Enabling Wider Application of Information-Theoretic Techniques to Neurophysiology. *Neuroinformatics*, *7*, 165–178.

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

Gollisch, T., & Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *319*, 1108–1111.

Gütig, R., Aertsen, A., & Rotter, S. (2003). Analysis of higher-order neuronal interactions based on conditional inference. *Biol. Cybern.*, *88*, 352–359.

Hatsopoulos, N. G., Ojakangas, C. L., Paninski, L., & Donoghue, J. P. (1998). Information about movement direction obtained from synchronous activity of motor cortical neurons. *Proc. Natl. Acad. Sci. USA*, *95*, 15706–15711.

Ince, R. A. A., Montani, F., Arabzadeh, E., Diamond, M. E., & Panzeri, S. (2009a). On the presence of high-order interactions among somatosensory neurons and their effect on information transmission. *J. Phys.: Conf. Ser.*, *197*, 012013.

Ince, R. A. A., Petersen, R. S., Swan, D. C., & Panzeri, S. (2009b). Python for information theoretic analysis of neural data. *Front. Neuroinf.*, *3*, 4.

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Phys. Rev.*, *106*, 620–630.

Kennel, M. B., Shlens, J., Abarbanel, H. D. I., & Chichilnisky, E. J. (2005). Estimating Entropy Rates with Bayesian Confidence Intervals. *Neural Comput.*, *17*, 1531–1576.

Latham, P. E., & Nirenberg, S. (2005). Synergy, Redundancy, and Independence in Population Codes, Revisited. *J. Neurosci.*, *25*, 5195–5206.

Li, C.-L. (1959). Synchronization of unit activity in the cerebral cortex. *Science*, *129*, 783–784.

Luczak, A., Bartho, P., & Harris, K. D. (2009). Spontaneous Events Outline the Realm of Possible Sensory Responses in Neocortical Populations. *Neuron*, *62*, 413–425.

Magri, C., Whittingstall, K., Singh, V., Logothetis, N. K., & Panzeri, S. (2009). A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC Neurosci.*, *10*, 81.

Martignon, L., Deco, G., Laskey, K., Diamond, M., Freiwald, W., & Vaadia, E. (2000). Neural Coding: Higher-Order Temporal Patterns in the Neurostatistics of Cell Assemblies. *Neural Comput.*, *12*, 2621–2653.

Mastronarde, D. N. (1983). Correlated firing of cat retinal ganglion cells. I. Spontaneously active inputs to X-and Y-cells. *J. Neurophysiol.*, *49*, 303–324.

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

Miller, G. (1955). Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, (pp. 95–100).

Montani, F., Ince, R. A. A., Senatore, R., Arabzadeh, E., Diamond, M. E., & Panzeri, S. (2009). The impact of high-order interactions on the rate of synchronous discharge and information transmission in somatosensory cortex. *Philos. Transact. A Math Phys. Eng. Sci.*, *367*, 3297–3310.

Montani, F., Kohn, A., Smith, M. A., & Schultz, S. R. (2007). The role of correlations in direction and contrast coding in the primary visual cortex. *J. Neurosci.*, *27*, 2338–2348.

Montemurro, M. A., Senatore, R., & Panzeri, S. (2007). Tight Data-Robust Bounds to Mutual Information Combining Shuffling and Model Selection Techniques. *Neural Comput.*, *19*, 2913–2957.

Nakahara, H., & Amari, S. I. (2002). Information-geometric measure for neural spikes. *Neural Comput.*, *14*, 2269–2316.

Nemenman, I., Bialek, W., & de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *69*, 56111.

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and Inference, Revisited. *Adv. Neural Inf. Process. Syst.*, *14*, 95–100.

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

Nirenberg, S., & Latham, P. E. (1998). Population coding in the retina. *Curr. Opin. Neurobiol.*, *8*, 488–493.

Nirenberg, S., & Victor, J. (2007). Analyzing the activity of large populations of neurons: how tractable is the problem? *Curr. Opin. Neurobiol.*, *17*, 397–400.

Oizumi, M., Ishii, T., Ishibashi, K., Hosoya, T., & Okada, M. (2009). A general framework for investigating how far the decoding process in the brain can be simplified. *Adv. Neural Inf. Process. Syst.*, *21*, 1225–1232.

Oizumi, M., Ishii, T., Ishibashi, K., Hosoya, T., & Okada, M. (2010). Mismatched Decoding in the Brain. *J. Neurosci.*, *30*, 4815–4826.

Onken, A., Grünewälder, S., Munk, M. H. J., & Obermayer, K. (2009). Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS Comput. Biol.*, *5*, e1000577.

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

Panzeri, S., Magri, C., & Logothetis, N. K. (2008). On the use of information theory for the analysis of the relationship between neural and imaging signals. *Magn. Reson. Imaging*, *26*, 1015–1025.

Panzeri, S., Schultz, S. R., Treves, A., & Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proc. R. Soc. Lond. B. Biol. Sci.*, *266*, 1001–1012.

Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the Sampling Bias Problem in Spike Train Information Measures. *J. Neurophysiol.*, *98*, 1064–1072.

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

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

Petersen, R. S., Panzeri, S., & Diamond, M. E. (2001). Population coding of stimulus location in rat somatosensory cortex. *Neuron*, *32*, 503–514.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*, 995–999.

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

Quian Quiroga, R., & Panzeri, S. (2009). Extracting information from neural populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.*, *10*, 173–185.

Roudi, Y., Nirenberg, S., & Latham, P. E. (2009a). Pairwise maximum entropy models for studying large biological systems: when they can work and they can't. *PLoS Comput. Biol.*, *15*, e1000380.

Roudi, Y., Tyrcha, J., & Hertz, J. (2009b). Ising model for neural data: model quality and approximate methods for extracting functional connectivity. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *79*, 051915.

Scaglione, A., Foffani, G., Scannella, G., Cerutti, S., & Moxon, K. A. (2008). Mutual information expansion for studying the role of correlations in population codes: How important are autocorrelations? *Neural Comput.*, *20*, 2662–2695.

Scaglione, A., Moxon, K. A., & Foffani, G. (2010). General poisson exact breakdown of the mutual information to study the role of correlations in populations of neurons. *Neural Comput*, *22*, 1445–1467.

Schneidman, E., Berry II, 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., Still, S., Berry, M. J., & Bialek, W. (2003). Network Information and Connected Correlations. *Phys. Rev. Lett.*, *91*, 238701.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., & Chichilnisky, E. J. (2006). The Structure of Multi-Neuron Firing Patterns in Primate Retina. *J. Neurosci.*, *26*, 8254–8266.

Staude, B., Rotter, S., & Grün, S. (2009). Cubic: cumulant based inference of higher-order correlations in massively parallel spike trains. *J. Computat. Neurosci.*, (pp. 0929–5313).

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

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1998). Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.*, *80*, 197–200.

Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., Prieto, A., Petrusca, D., Grivich, M. I., Sher, A., Hottowy, P., Dabrowski, W., Litke, A. M., & Beggs, J. M. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.*, *28*, 505–518.

Tatsuno, M., Fellous, J., & Amari, S. I. (2009). Information-geometric measures as robust estimators of connection strengths and external inputs. *Neural Comput.*, *21*, 2309–2335.

Treves, A., & Panzeri, S. (1995). The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Comput.*, *7*, 399–407.

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

von der Malsburg, C. (1999). The What and Why of Binding: The Modeler's Perspective. *Neuron*, *24*, 95–104.

Wu, S., Nakahara, H., & Amari, S. I. (2001). Population coding with correlation and an unfaithful model. *Neural Comput.*, *13*, 775–797.

Wu, S., Nakahara, H., Murata, N., & Amari, S. (2000). Population decoding based on an unfaithful model. *Adv. Neural Inf. Process. Syst.*, *12*, 167–173.