Correcting for the Sampling Bias Problem in Spike Train Information Measures

Stefano Panzeri, Riccardo Senatore, Marcelo A. Montemurro and Rasmus S. Petersen

J Neurophysiol 98:1064-1072, 2007. First published 5 July 2007;
doi: 10.1152/jn.00559.2007

Invited Review

J Neurophysiol 98: 1064–1072, 2007.
First published July 5, 2007; doi:10.1152/jn.00559.2007.

Correcting for the Sampling Bias Problem in Spike Train Information Measures

Stefano Panzeri, Riccardo Senatore, Marcelo A. Montemurro, and Rasmus S. Petersen
University of Manchester, Faculty of Life Sciences, Manchester, United Kingdom

Submitted 18 May 2007; accepted in final form 21 July 2007

Panzeri S, Senatore R, Montemurro MA, Petersen RS. Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol* 98: 1064–1072, 2007. First published July 5, 2007; doi:10.1152/jn.00559.2007. Information Theory enables the quantification of how much information a neuronal response carries about external stimuli and is hence a natural analytic framework for studying neural coding. The main difficulty in its practical application to spike train analysis is that estimates of neuronal information from experimental data are prone to a systematic error (called "bias"). This bias is an inevitable consequence of the limited number of stimulus-response samples that it is possible to record in a real experiment. In this paper, we first explain the origin and the implications of the bias problem in spike train analysis. We then review and evaluate some recent general-purpose methods to correct for sampling bias: the Panzeri-Treves, Quadratic Extrapolation, Best Universal Bound, Nemenman-Shafee-Bialek procedures, and a recently proposed shuffling bias reduction procedure. Finally, we make practical recommendations for the accurate computation of information from spike trains. Our main recommendation is to estimate information using the shuffling bias reduction procedure in combination with one of the other four general purpose bias reduction procedures mentioned in the preceding text. This provides information estimates with acceptable variance and which are unbiased even when the number of trials per stimulus is as small as the number of possible discrete neuronal responses.

INTRODUCTION

Almost all sensory messages are encoded as temporal patterns of action potentials (spikes), often distributed over populations of neurons. A fundamental problem in neuroscience is to understand the nature of this neural population code. Ideally, as argued by Rieke et al. (1996), we would like a dictionary that, given some snapshot of spiking activity, tells us what sensory signal has occurred. As a first step, we need to know what kind of neural code we are dealing with (Optican and Richmond 1987). For example, is the precise timing of each spike important, or is it just the number of spikes that matters? As a second step, we need to know what specific stimulus features are being encoded. For example: do spike counts encode the amplitude or the frequency of a sinusoidal stimulus (Arabzadeh et al. 2004)? A widely used approach to neural coding is to treat the brain as a communication channel and to use information theory to quantify and compare the information about stimuli available in different candidate codes. In recent years, this approach has led to significant new insight in our understanding of sensory encoding (Averbeck et al. 2006; Borst and Theunissen 1999; Hertz and Panzeri 2002; Rieke et al. 1996).

Address for reprint requests and other correspondence: S. Panzeri, University of Manchester, Faculty of Life Sciences, The Mill, PO Box 88, Manchester M60 1QD,UK (E-mail: s.panzeri@manchester.ac.uk).

However, estimating accurately the information that spike trains convey about external stimuli is fraught with a major practical difficulty: information theoretic measures suffer from a significant systematic error (or "bias") due to the limited amount of stimulus-response data that can realistically be collected in an experimental session. Over the last few years, several advanced methods have been proposed to correct for the bias problem. In this report, we explain the origin of the bias, and review recent bias correction methods. Finally, we present guidelines to help choose an appropriate procedure for computing information and to help understand the conditions under which it will produce accurate results.

Information carried by neural responses

The aim of information theoretic analysis is to get insight into neural coding. For example, we might want to know whether a particular neuron conveys information by millisecond precision spike timing or simply by the total number of emitted spikes (the "spike count"). We assume that, if precise spike timing is important, a timing code will convey more "information" than a spike-count code. If, on the other hand, precise spike timing is not important, a timing code will convey no more information that a count code. To carry out such an analysis, the first step is to "choose" the neural code. This in practice means to choose a way to quantify the neuronal response that reflects our assumption of what is most salient in it. For example, if we think that only spike counts (not the precise temporal pattern of spikes) are important, we choose a spike-count code: we define a poststimulus response interval and count the number of spikes it contains on each repetition ("trial") of a stimulus. The second step is to compute how much information can be extracted from the chosen response quantification. This allows the assessment of how good is the candidate neural code. Whatever the neural system of interest, the problem of information quantification can be characterized in the following way.

Consider an experiment in which the animal is presented with a stimulus s selected with probability $P(s)$ from a stimulus set $\mathbf{S}$ consisting of S elements, and the consequent response (either of a single neuron or an ensemble of neurons) is recorded and quantified in a certain poststimulus time window. In most cases, the neural response is quantified as a discrete, multi-dimensional array $r = \{r_1, \ldots, r_L\}$ of dimension L. For example, to quantify the spike count response of a population of L cells, r_i would be the number of spikes emitted by cell i on a given trial in the response window. Alternatively, to quantify the spike timing response of a single neuron, the response window is divided into L bins of width Δt, so that r_i is the number of spikes fired in the ith time bin (Strong et al. 1998). Here Δt is the assumed time precision of the code and can be

varied parametrically to characterize the temporal precision of the neural code. We denote by **R** the set of possible values taken by the response array.

Having quantified the response, the next step is to characterize the relationship between stimulus and response and assign a number (the information) that quantifies how well different responses discriminate between different stimuli. The more the response of a neuron varies across stimuli, the greater is its ability to transmit information (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 so is through the concept of entropy (Shannon 1948). The response variability (across all possible stimuli and trials) is quantified by the *response entropy*

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

where $P(r)$ is the probability of observing response r across all trials to any stimulus. However, neurons are typically "noisy" in the sense that their responses to repetitions of an identical stimulus differ from trial to trial. **H(R)** reflects both variation of responses to different stimuli and variation due to trial-to-trial noise. Thus **H(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 entropy at fixed stimulus (that is, conditional on s)

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

H(R|S) is known as the noise entropy. $P(r|s)$ is the probability of observing response r given presentation of stimulus s. The noisier is a neuron, the greater is **H(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;R)** between stimuli and responses (in the following abbreviated to "information")

$$\mathbf{I}(\mathbf{S};\mathbf{R}) \equiv \mathbf{H}(\mathbf{R}) - \mathbf{H}(\mathbf{R}|\mathbf{S}) = \sum_{r,s} P(s)P(r|s) \log_2 \frac{P(r|s)}{P(r)} \quad (3)$$

I(S;R) quantifies how much of the information capacity provided by stimulus-evoked differences in neural activity is robust to noise. An alternative but equivalent interpretation of **I(S;R)** is that it quantifies the reduction of uncertainty about the stimulus that can be gained from observation of a single trial of the neural response (Borst and Theunissen 1999; Rieke et al. 1996). When base-two logarithms are used (as in *Eqs. 1–3*), **I(S;R)** is expressed in units of bits: 1 bit of information means that, on average, observation of the neuronal response on one trial reduces the observer's stimulus uncertainty by a factor of two. **I(S;R)** is zero only when the stimulus-response relationship is completely random.

Calculation of information requires accurate estimation of the stimulus-response probabilities $P(r)$, $P(r|s)$, and $P(s)$ and thereby **H(R)** and **H(R|S)**. The problem is that these probabilities are *not known* but 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.

Bias of the plug-in information estimator: what is it and where does it come from?

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 (Fig. 1). The significance of these finite sampling fluctuations is that they lead to both systematic error (bias) and statistical error (variance) in estimates of entropies and information. These errors, particularly the bias, constitute a significant practical problem. If not corrected, bias can lead to serious misinterpretations of neural coding data. Fortunately, a number of useful techniques have recently been developed for addressing the issue—how to do so is the main topic of this review.

The most direct way to compute information and entropies is to estimate the response probabilities as the experimental histogram of the frequency of each response across the available trials and then plug these empirical probability estimates into *Eqs. 1–3*. We refer to this as the "plug-in" method. In the following, N_s denotes the number of trials recorded in response

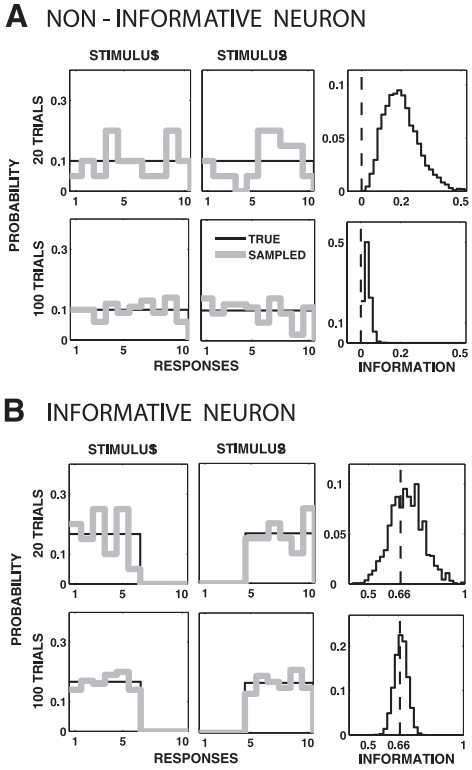

FIG. 1. The effect of limited sampling. *A*: simulation of an uninformative neuron, responding on each trial with a uniform distribution of spike counts ranging from 1 to 10, regardless of which of 2 stimuli was presented. Examples of empirical response probability histograms (gray solid line) sampled from 20 and 100 trials per stimulus (*top* and *bottom rows,* respectively) are shown in the *left* and *middle columns* (responses to stimuli 1 and 2, respectively). The black horizontal line is the true response distribution. *Right*: distribution (over 5,000 simulations) of the plug-in information values obtained with 20 (*top*) and 100 (*bottom*) trials per stimulus respectively. As the number of trials increases, both the information bias and the SD decrease. The dashed vertical line in the *right columns* indicates the true value of the information carried by the simulated neuron. *B*: simulation of an informative neuron, firing (with uniform probability) 1–6 spikes to stimulus 1 and 5–10 spikes to stimulus 2. Results are plotted as in *A*.

to stimulus s and N the total number of trials across all stimuli. Note that $N = \sum_s N_s$. To build intuition on the implications of the plug-in estimation of probabilities, consider a hypothetical neuron that, on each trial, fires 1–10 spikes with equal probability regardless of which of two stimuli was presented. In this case, the true conditional probability $P(r|s)$ is 0.1 (Fig. 1*A*, black horizontal line in the *left* and *middle panels*) for all stimulus-response combinations, which means that $P(r)$ is also 0.1; consequently, *Eq. 3* tells us that the mutual information is precisely zero. Figure 1*A* (top *left* and *middle panels*) shows the probabilities estimated from a simulated experiment where N_s was 20. Due to limited sampling, the estimated probabilities (gray line) differed markedly from 0.1 and from one another. In particular, because stimulus 2 happened to evoke more spikes than stimulus 1, it might naively appear that the neuron was selective for stimulus 2. This was reflected in a nonzero value of the plug-in information estimate (0.2 bits). The result varied from experiment to experiment (*right*): the information distribution was centered at the 0.202 bits—it was *not* centered at the true value of 0 bits. This shows that the plug-in information estimate was biased. In general, *sampling bias* is defined as the difference between the expected value of the information computed from the probability distributions estimated with N trials (here 0.202 bits), and its value computed from the true probability distributions (here 0 bits). The bias is a systematic error that cannot be eliminated just by averaging over repeated experiments under similar conditions or on similar neurons.

To correct for the bias, we need to understand its properties. The first of these is its dependence on the number of trials. The greater the number of trials, the smaller the fluctuations in the estimated probabilities, and consequently the smaller the bias. For example, with $N_s = 100$ (Fig. 1*A*, *bottom*), the bias was 0.033 bits.

However, what makes it difficult to evaluate the bias is that it is not a fixed quantity the magnitude of which only depends on the number of trials: even at fixed number of trials, the bias can be very different for different neurons. This is illustrated in Fig. 1*B*. A second hypothetical neuron fired 1–6 spikes for stimulus 1 and 5–10 spikes for stimulus 2. The stimulus-response probabilities are now stimulus-modulated, and the neuron conveyed 0.666 bits of information (true value). With $N_s = 20$, the distribution of empirical information values was centered at 0.703 bits (Fig. 1*B*, *top right*), giving a bias of 0.036 bits. Thus although the number of trials per stimulus was identical to that of the neuron of Fig. 1*A*, the bias was much less.

The bias of the information comes from the bias of the two entropies (response entropy and noise entropy) that constitute it (Miller 1955). To go deeper into the nature of the sampling bias, it turns out to be useful to study the bias of the entropies directly. We illustrate this by computing both entropies on a set of simulated spike trains the statistical properties of which were based on real spike trains (responses of neurons in rat somatosensory cortex to 13 different sinusoidal stimuli, differing in vibration energy) (Arabzadeh et al. 2004). We simulated two types of responses ("single cell" and "population," respectively). In the single-cell case, we simulated the response of a single neuron over a 0- to 40-ms poststimulus time window and digitized the spike train into $L = 8$ bins of size $\Delta t = 5$ ms (0 or 1 spikes in each bin). In the second case, we simulated a population of $L = 8$ neurons simultaneously responding to a stimulus in a 10- to 15-ms poststimulus window (each neuron

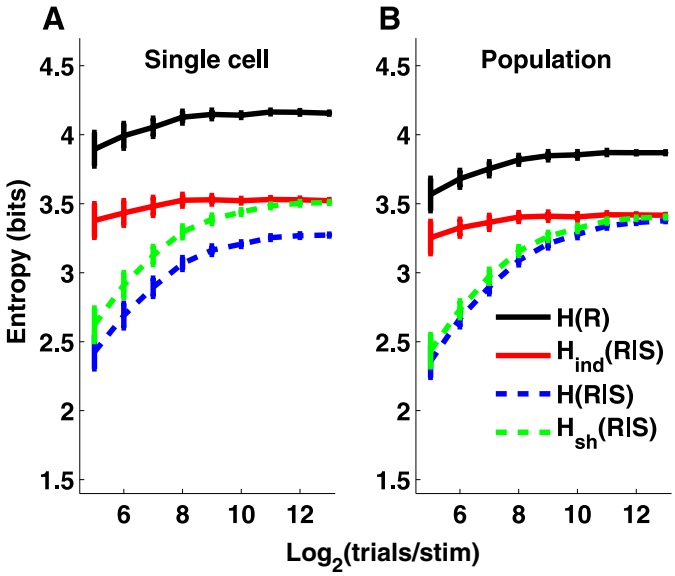

FIG. 2. Sampling properties of plug-in estimation of entropies. The plug-in estimators of entropies and information are plotted as a function of the available number of trials per stimulus. We plot the mean ± SD over 50 simulations) of the plug-in estimators of $\mathbf{H}(\mathbf{R})$, $\mathbf{H}(\mathbf{R}|\mathbf{S})$, $\mathbf{H}_{ind}(\mathbf{R}|\mathbf{S})$, and $\mathbf{H}_{sh}(\mathbf{R}|\mathbf{S})$. *A*: results obtained using realistically simulated single-cortical-cell spike trains (discretized into $L = 8$ time bins each containing 0 or 1 spikes; see text). *B*: results obtained with realistically simulated cortical population responses ($L = 8$ correlated cells each emitting 0 or 1 spike). In both cases, the responses were to 13 different simulated stimuli.

firing 0 or 1 spikes in the window). In both cases, the neural response was a "binary" array consisting of $L = 8$ elements, and the simulated responses matched the lower-order statistics (firing rates and pair-wise auto- or cross-correlations between spikes) of the experimentally recorded neural responses (see APPENDIX for full details).

We performed a series of simulations, systematically varying the number of trials. Figure 2 shows the entropy estimates resulting from the plug-in method, both for the single-cell simulated responses (*left*) and population responses (*right*). In both cases, the estimates of $\mathbf{H}(\mathbf{R})$ and $\mathbf{H}(\mathbf{R}|\mathbf{S})$ *increased* with the number of trials. That is, in contrast to its effect on information estimates, 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 less the number of trials, the less likely we are to fully sample the full range of possible responses. Thus finite sampling makes neuronal responses seem less variable than they really are. Consequently, entropy estimates are lower than their true values, and the effect of finite sampling on entropies is a downward bias. $\mathbf{H}(\mathbf{R})$ is far less biased than $\mathbf{H}(\mathbf{R}|\mathbf{S})$ because the former depends on $P(r)$, which, being computed from data collected across all stimuli, is better sampled than $P(r|s)$.

From *Eq. 3*, the bias of the information is the difference between the bias of $\mathbf{H}(\mathbf{R})$ and that of $\mathbf{H}(\mathbf{R}|\mathbf{S})$. Because the latter is greater (and negative), the net result is that $\mathbf{I}(\mathbf{S};\mathbf{R})$ is typically strongly biased *upward* (Fig. 3, *A* and *B*). Intuitively, this is because finite sampling can introduce spurious stimulus-dependent differences in the response probabilities, which make the stimuli seem more discriminable than they actually are (Fig. 1) and hence the neuron more informative than it really is.

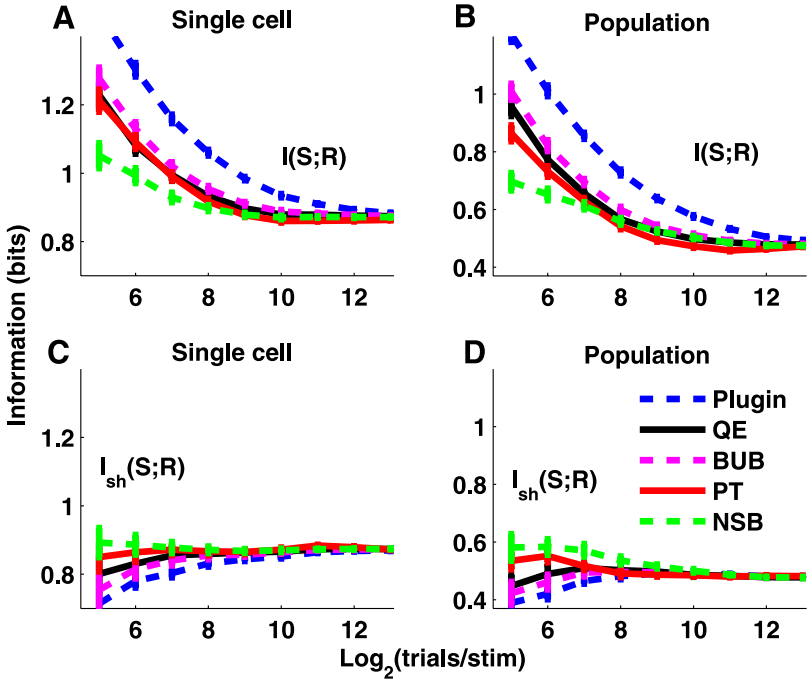

FIG. 3. Comparison of the performance of different bias correction methods. The information estimates $\mathbf{I(S;R)}$ and $\mathbf{I}_{sh}\mathbf{(S;R)}$ are plotted as a function of the available number of trials per stimulus. *A* and *B*: mean ± SD (over 50 simulations) of $\mathbf{I(S;R)}$. *C* and *D*: mean ± SD (over 50 simulations) of $\mathbf{I}_{sh}\mathbf{(S;R)}$. Various methods were used to correct for the bias: plug-in estimation (i.e., no bias correction), PT, QE, BUB, and NSB (see text). *A* and *C* and *B* and *D* report results using realistically simulated single-cell and population cortical spike trains, respectively (see main text).

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." Roughly speaking this is when 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, $N_s\ P(r|s) \gg 1$ for each stimulus-response pair s,r such that $P(r|s) > 0$. 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 and Treves 1996). The leading terms in the biases are, respectively

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

where $\bar{R}_s$ denotes the number of relevant responses for the stimulus conditional response probability distribution $P(r|s)$ (i.e., the number of different responses r with nonzero probability of being observed when stimulus s is presented) and $\bar{R}$ denotes the number of relevant responses for $P(r)$ (i.e., the number of different responses r with nonzero probability of being observed across all stimuli). In practice, $\bar{R}$ is usually going to be equal to the number of elements constituting the response space $\mathbf{R}$ (if a response never happens across all stimuli, it can be removed from the sum over r in *Eqs. 1–3* and thus removed from the response set $\mathbf{R}$). However, $\bar{R}_s$ may be different from $\bar{R}$. For example, take the neuron simulated in Fig. 1*B*, when stimulus 1 evokes a low-firing response with high probability and high-firing response with zero probability, and stimulus 2 elicits the opposite type of response. In this case, we have $\bar{R}_s = 6$ for both stimuli and $\bar{R} = 10$.

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 the bias of $\mathbf{H(R|S)}$ is approximately S times bigger than that of $\mathbf{H(R)}$. This means that, in the presence of many stimuli, the bias of $\mathbf{I(S;R)}$ is similar to that of $\mathbf{H(R|S)}$. However, $\mathbf{I(S;R)}$ is a difference of entropies, and its typical values are much smaller than those of $\mathbf{H(R|S)}$. 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 the information (see Figs. 2 and 3 for an illustration). Second, *Eq. 4* shows that the bias is small when the ratio $N_s/\bar{R}$ is big, i.e., more trials per stimulus than possible responses. This is because, assuming that the number of trials per stimulus is approximately constant and $\bar{R}$ is approximately equal to $\bar{R}$, the bias of $\mathbf{H(R|S)}$ is approximately $-\bar{R}/[2N_s\ln(2)]$. Thus $N_s/\bar{R}$ is *the* crucial parameter for the sampling problem. For example, in the simulations of Fig. 2, with $\bar{R} = 2^8$, the bias of $\mathbf{I(S;R)}$ became negligible for $N_s \geq 2^{13}$ (i.e., $N_s/\bar{R} \geq 32$). Second, *Eq. 4* shows that, even if N_s is constant, the bias increases with the number of responses $\bar{R}$. This has important implications for comparing neural codes. For example, a response consisting of a given number of spikes can arise from many different possible temporal patterns of spikes. Thus $\bar{R}$ is typically much greater for a spike timing code than for a spike count code, and it follows from *Eq. 4* that the information conveyed by a spike count code is more biased than that measured for the same neurons through a spike timing code. If bias is not eliminated, there is therefore a danger of concluding that spike timing is important even when it is not.

A further important feature of *Eq. 4* is that, although the bias is not the same for all probability distributions, in the asymptotic sampling regime, it depends on some remarkably simple details of the response distribution (the number of trials and the number of relevant responses). Thus *Eq. 4* makes it is possible to derive simple rules of thumb for estimating the bias magnitude and compare the relative bias in different situations. For example, *Eq. 4* can predict very effectively the two very different bias properties of the two simulated neurons in Fig. 1.

Both neurons in Fig. 1 have $\bar{R} = 10$. However, $\bar{R} = 10$ for the neuron in Fig. 1*A* and $R_s = 6$ for the neuron in Fig. 1*B*. As a consequence, *Eq. 4* predicts that the bias of the neuron in Fig. 1*A* is $\sim 9/[2N_s \ln(2)]$ bits and the bias of the neuron in Fig. 1*B* is $\sim 1/[2N_s \ln(2)]$ bits (i.e., 9 times smaller). As detailed in the next section, the simplicity of *Eq. 4* can also be exploited to correct effectively for the bias (Panzeri and Treves 1996).

Comparing procedures to correct for the bias

The plug-in estimate of information $\mathbf{I}(\mathbf{S};\mathbf{R})$ tends to require large numbers of trials (N_s larger than $\bar{R}$ by at least a factor of 32 in Fig. 3*A*) to become unbiased and is therefore of limited experimental utility. Over the last 10 years, several bias-correction procedures have been developed to reduce the number of trials necessary. In the following, we review and compare four bias correction procedures that are applicable to any spike train (whatever its statistics) and to any discrete quantification of the neural response. These methods, which are among those most widely used in the literature, were selected in this review because they are in our experience the most effective ones, and because they have the property that they are guaranteed to converge to the true value of information (or entropy) of a given quantification of the neuronal response as the number of trials N goes to infinity. We compare their performance, study the conditions in which they are accurate and consider their relative advantages.

PANZERI-TREVES (PT) BAYESIAN ESTIMATION OF THE NUMBER OF RELEVANT RESPONSES. *Equation 4* provides a simple asymptotic expression that can be used to estimate the bias, provided that one can evaluate the number of relevant responses R_s. However, estimating R_s is not straightforward. The simplest approach is to approximate R_s by the number of responses that are observed at least once—the "naïve" count. This leads to the so-called Miller-Madow bias estimate (Miller 1955). The naïve 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 [taking into account that higher-order terms in $1/N$ of the information bias are all positive (Treves and Panzeri 1995)], the Miller-Madow estimate is usually an underestimate of the bias. To alleviate this problem, Panzeri and Treves (1996) have developed a Bayesian procedure to estimate the number of relevant bins. 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. Figure 3, *A* and *B,* shows that PT bias correction substantially improves the estimates of $\mathbf{I}(\mathbf{S};\mathbf{R})$, which, in this simulation became accurate when $N_s = 2^{10}$ trials per stimulus were available (that is, when $N_s/\bar{R} \geq 4$, compared with $N_s/\bar{R} \geq 32$ for pure plug-in). The advantages of the PT correction are that it performs well and is straightforward to implement (1 Matlab routine that is available at http://stefano.panzeri.googlepages.com/informationbiascorrections). The disadvantage of PT correction is that, by design, it cannot work in the undersampled regime ($N_s/\bar{R} < 1$).

QUADRATIC EXTRAPOLATION (QE). Like the PT method, this procedure (Strong et al. 1998) assumes we are in the asymptotic sampling regime, so that the bias of the entropies and information can be accurately approximated as second order expansions in $1/N$ (Treves and Panzeri 1995). That is $\mathbf{I}_{\text{plug-in}}(\mathbf{S};\mathbf{R}) = \mathbf{I}_{\text{true}}(\mathbf{S};\mathbf{R}) + a/n + b/N^2$, where a and b are free parameters that depend on the stimulus-response probabilities. Unlike the PT procedure, the parameters a and b are not given by analytic formulae but are estimated from the data. This is done by re-computing the information from fractions ($N/2$ and $N/4$) of the data available and then fitting (using e.g., a least-square-error procedure) the plug-in information values obtained with fractions of data to the preceding quadratic function of $1/N$. This provides the best-fit estimates of the parameters a and b and consequently the estimate of $\mathbf{I}_{\text{true}}(\mathbf{S};\mathbf{R})$. Figure 3, *A* and *B,* shows that the results of the QE procedure were very similar to the PT procedure. This is because they both make the same asymptotic sampling assumption.

BEST UNIVERSAL BOUND. Unlike PT and QE, the recently introduced BUB procedure (Paninski 2003) does not rely on the assumption that there are enough data to be in the asymptotic sampling regime. Paninski's (2003) idea is to rewrite the entropy estimation problem as a linear estimation problem in the so-called "histogram order statistics" and then use results from polynomial approximation theory to compute bounds on the entropy estimation errors. These bounds permit the derivation of what Paninski (2003) termed "Best Universal Bounds" (BUBs) on both bias and variance. The BUBs depend strongly on a "degrees of freedom" parameter $k_{\max}$, which gives the method theoretical flexibility. However, finding the optimal $k_{\max}$ for a given dataset may require a considerable effort and may be a practical complication in the empirical use of this method. A potential advantage of BUB is that it may outperform the asymptotic estimators described in the preceding text in conditions when data are scarce but this bound is tight (Paninski 2003). In particular, theoretical considerations show that the BUB procedure is very competitive when both N and $\bar{R}$ are very large (Paninski 2003). However, with our simulated data, the BUB gave a rather similar performance to PT and QE methods even using the optimal $k_{\max}$ (Fig. 3, *A* and *B*). BUB was simple to implement (1 Matlab routine available at http://www.stat.columbia.edu/~liam/research/info_est.html).

NEMENMAN-SHAFEE-BIALEK (NSB) ENTROPY ESTIMATION METHOD. The NSB entropy estimation method (Nemenman et al. 2002, 2004) is rooted in the Bayesian inference approach to entropy estimation and (like BUB) does not rely on the assumption that there are enough data to be in the asymptotic sampling regime. The Bayesian approach makes some prior assumptions on the response probability distributions. The NSB method uses a novel type of prior assumption on the probabilities, which is designed to produce an approximately uniform distribution of the expectation of the entropy value before any stimulus-response data are sampled (so that the entropy estimate is not too much biased by the prior assumptions). As data become available, the entropy estimation is updated by integrating over the hypothetical prior probability distributions weighted by their conditional probability given the data. The NSB procedure is essentially parameter-free (like PT and QE). We found that NSB generally gave the least biased estimate of $\mathbf{I}(\mathbf{S};\mathbf{R})$ in our simulations (Fig. 3, *A* and *B*). Thus NSB has a performance advantage. This result is consistent with simulations presented by Montani et al. (2007). NSB was demanding to implement (it requires a substantial amount of numerical integration and function inversion). A code for NSB can be found at http://nsb-entropy.sourceforge.net/.

Compared with plug-in estimation, all four correction procedures were very useful and greatly reduced the information bias although at the price of a moderate increase in variance (Fig. 3, *A* and *B*). The moderate increase invariance is more than compensated by strong decrease of bias, making the bias correction worthwhile. The increase in variance is due to subtracting the correction term, which has its own variability. Rigorous studies of the trade-off between bias and variance are reported in Paninski (2003) and Nemenman et al. (2004).

In summary, even if some of the bias correction procedures (such as NSB or BUB) did not rely on the asymptotic sampling regime to be valid and thus might potentially work even in deeply undersampled conditions, in practice no bias correction procedure was sufficient to obtain unbiased information estimates in the undersampled regime ($N_s \leq \bar{R}$). They all required $N_s \geq 2$–$4\ \bar{R}$ to work. This is consistent with a number of reported simulation studies (Endres and Foldiak 2005; Nelken et al. 2005; Pola et al. 2003). Recently, however, it has been found that further improvements in bias performance are possible.

Role of correlations in the bias problem and the shuffling bias reduction procedure

The fundamental problem with understanding the neural code is that it may be high dimensional: many neurons, time bins, etc. could contribute to the neuronal population code making the size of the response array L potentially very large. Importantly, the number of possible responses becomes exponentially large as L 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 $\leq$20 spike per stimulus presentation generate $\sim 20^{10}$ ($\sim 10^{13}$) possible responses.

The bias problem for large L is exacerbated by the fact that, in real neuronal recordings, the elements of the response array are often statistically correlated. For example, nearby cortical neurons often have correlated trial-to-trial response variability and a significant fraction of their spikes occurs synchronously (see e.g., Averbeck et al. 2006). Correlations may either increase or decrease the information (Averbeck et al. 2006; Panzeri et al. 1999) and thus cannot be neglected in the information computation. 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 (Montemurro et al. 2007).

Consider the noise entropy that would be obtained if the response in each element of the array was independent of the others at fixed stimulus: that is $P(r|s)$ equals $P_{\text{ind}}(r|s) = \Pi_i P(r_i|s)$. This noise entropy can be estimated in two ways. First, by direct substitution of $P_{\text{ind}}(r|s)$ into *Eq. 2*. The entropy of this "independent" distribution is called $\mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S})$

$$\mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S}) = -\sum_r \sum_s P(s)P_{\text{ind}}(r|s) \log_2 P_{\text{ind}}(r|s)] \quad (5)$$

Because $\mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S})$ depends only on the marginal probabilities of the response array, it typically has very small bias (Fig. 2). Second, correlations between response variables can be removed by "shuffling" the data at fixed stimulus by constructing pseudo response arrays obtained by combining r_i values each taken (randomly and without repetition) from different trials in which the stimulus s was presented 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 known as $P_{\text{sh}}(r|s)$ can be computed. This results in the noise entropy $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$

$$\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S}) = -\sum_r \sum_s P(s)P_{\text{sh}}(r|s) \log_2 P_{\text{sh}}(r|s)] \quad (6)$$

$\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$ has the same value of $\mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S})$ for infinite number of trials N, but it has a much higher bias than $\mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S})$ for finite N. In fact, Fig. 2 shows that the bias of $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$ is approximately of the same of order of magnitude as the bias of $\mathbf{H}(\mathbf{R}|\mathbf{S})$. Intuitively, this is expected because $P_{\text{sh}}(r|s)$ is sampled with the same number of trials as $P(r|s)$ 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 $\mathbf{I}(\mathbf{S};\mathbf{R})$ but through the following formula

$$\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R}) = \mathbf{H}(\mathbf{R}) - \mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S}) + \mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S}) - \mathbf{H}(\mathbf{R}|\mathbf{S}) \quad (7)$$

$\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ has the same value of $\mathbf{I}(\mathbf{S};\mathbf{R})$ for infinite number of trials but has a much smaller bias for finite N due to the bias cancellation created by the entropy terms added to the right hand side of *Eq. 7*. When the biases of $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$ and $\mathbf{H}(\mathbf{R}|\mathbf{S})$ approximately cancel out, the bias of $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ is dominated by the bias of $\mathbf{H}(\mathbf{R}) - \mathbf{H}_{\text{ind}}(\mathbf{R}|\mathbf{S})$, which is much smaller than the bias of $\mathbf{H}(\mathbf{R}|\mathbf{S})$ [which in turn dictates the bias of $\mathbf{I}(\mathbf{S};\mathbf{R})$].

Figure 3, *C* and *D*, confirms that as a result of the bias cancellations in *Eq. 7*, when considering the plug-in estimates, there a huge bias reduction of $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ with respect to $\mathbf{I}(\mathbf{S};\mathbf{R})$. Moreover, Fig. 3, *C* and *D*, shows that the bias of the plug-in estimate of $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ is negative. This is a very typical finding (Montemurro et al. 2007). The reason why the plug-in $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ tends to be often biased downward is that [as shown in Fig. 2, and in more detail in Montemurro et al. (2007)] $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$ is usually slightly more downward biased than $\mathbf{H}(\mathbf{R}|\mathbf{S})$. To understand why, Montemurro et al. (2007) computed the bias of $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$ in the "asymptotic sampling" regime. They found that the asymptotic bias of $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$ has the same expression as that $\mathbf{H}(\mathbf{R}|\mathbf{S})$ in *Eq. 4*, after replacing R_s with $R_{\text{sh}-\text{s}}$, the number of bins relevant to $P_{\text{sh}}(r|s)$. Because $P_{\text{ind}}(r|s) = 0$ implies $P(r|s) = 0$ and because the shuffled responses are generated according to $P_{\text{ind}}(r|s)$, then it must be that $R_{\text{sh}-\text{s}} \geq R_s$. Therefore $\mathbf{H}(\mathbf{R}|\mathbf{S})$ is usually less downward biased than $\mathbf{H}_{\text{sh}}(\mathbf{R}|\mathbf{S})$.

In the previous section, the four bias correction techniques were applied to $\mathbf{I}(\mathbf{S};\mathbf{R})$. However, they can also be applied to $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$. Figure 3 illustrates that, with all four bias correction procedures, there is a considerable bias reduction when using $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ rather than $\mathbf{I}(\mathbf{S};\mathbf{R})$ (compare Fig. 3, *A* to *C*, and *B* to *D*). The result of using shuffling in combination with PT, QE, or BUB is that the estimates of $\mathbf{I}_{\text{sh}}(\mathbf{S};\mathbf{R})$ become unbiased even down to 2^8 trials per stimulus (i.e., for $N_s/R \approx 1$). This is a factor of four better than the best performing bias correction of $\mathbf{I}(\mathbf{S};\mathbf{R})$. Shuffling in combination with NSB worked about as

well as PT or QE in some cases (Fig. 3*C*) but occasionally worked less well as in Fig. 3*D*. This is probably because the shuffled and real probability distribution did not match the prior assumptions of the NSB method equally well in all cases.

A potential short-coming of the shuffling method is that because the information estimate includes two extra terms, its variance might be higher. However, our simulations in Fig. 3, *C* and *D*, illustrate that the variance of $\mathbf{I}_{sh}(\mathbf{S;R})$ is essentially equal to that of $\mathbf{I}(\mathbf{S;R})$. This is because fluctuations in the values of the marginal probabilities of each element of the response array have a major impact on the fluctuations of both $\mathbf{H}(\mathbf{R}|\mathbf{S})$ and $\mathbf{H}_{sh}(\mathbf{R}|\mathbf{S})$ [the 2 most variable terms of $\mathbf{I}_{sh}(\mathbf{R;S})$] and are reflected with the same sign in both $\mathbf{H}(\mathbf{R}|\mathbf{S})$ and $\mathbf{H}_{sh}(\mathbf{R}|\mathbf{S})$ (Montemurro et al. 2007). Thus the statistical fluctuations of these entropies largely cancel out, and the resulting variance of $\mathbf{I}_{sh}(\mathbf{S;R})$ remains under control. Thus using $\mathbf{I}_{sh}(\mathbf{S;R})$ to estimate information is an efficient and simple way to reduce the bias without significantly increasing the variance. The code for the calculation of information $\mathbf{I}(\mathbf{S;R})$ and $\mathbf{I}_{sh}(\mathbf{S;R})$ (Plug-in, PT, and QE) can be found at http://stefano.panzeri.googlepages.com/informationbiascorrections.

Considerations on the neuronal statistics and experimental conditions

The performance of bias corrections methods on the simulated data reported in the preceding text is representative of their performance on the type of the cortical neural data that the simulation was designed to reproduce. However, the performance of the bias correction method does depend on the statistics of the neuronal data and the experimental design, and thus it is difficult to make completely general statements on the range of validity of each method. Despite this, using analytical considerations as well as our own experience on simulated and real data, we can give some rules of thumb on how the performance of the methods may vary.

AMOUNT OF CORRELATION. The correlations in the simulated dataset were set to reproduce typical cortical within- and cross-cell correlations. What happens if we consider datasets with different degrees of correlation? If we consider uncorrelated or weakly correlated data (such as a noninteracting population), the computation of $\mathbf{I}_{sh}(\mathbf{S;R})$ becomes extremely accurate down to very small number of trials because in this case the bias of $\mathbf{H}(\mathbf{R}|\mathbf{S})$ and that of $\mathbf{H}_{sh}(\mathbf{R}|\mathbf{S})$ cancel out perfectly (Montemurro et al. 2007). If we increase the strength of correlation significantly (e.g., introducing an absolute refractory period or a very high degree of synchrony), then $\mathbf{I}_{sh}(\mathbf{S;R})$ becomes more strongly downward biased. The simulations in Montemurro et al. (2007) suggest that the bias of $\mathbf{I}_{sh}(\mathbf{S;R})$ remains small for a wide range of typical correlations.

NUMBER OF STIMULI. As the number of stimuli S becomes larger, for fixed number of trials per stimulus N_s, the overall number of trials N across all stimuli gets larger, thus $P(r)$ becomes better sampled and the bias of $\mathbf{H}(\mathbf{R})$ decreases. However, N_s being fixed, there is no sampling change for $P(r|s)$ and thus the bias of $\mathbf{H}(\mathbf{R}|\mathbf{S})$, $\mathbf{H}_{ind}(\mathbf{R}|\mathbf{S})$, and $\mathbf{H}_{sh}(\mathbf{R}|\mathbf{S})$ remain approximately the same. (Due to summing over more stimuli, their variance does reduce). Because the bias of $\mathbf{I}_{sh}(\mathbf{S;R})$ is dominated by that of $\mathbf{H}(\mathbf{R}|\mathbf{S})$, the sampling behavior of $\mathbf{I}_{sh}(\mathbf{S;R})$ therefore changes little when increasing the number of stimuli. However, because the bias of $\mathbf{H}(\mathbf{R})$ is one of the main components to the bias of $\mathbf{I}_{sh}(\mathbf{S;R})$, the latter becomes less biased (Montemurro et al. 2007).

RELIABILITY OF NEURAL RESPONSES. The preceding simulated data were generated to match the statistics and variability of cortical responses to simple stimuli. As is often the case with cortical responses to simple stimuli (Gershon et al. 1998), the variance of the spike counts was approximately proportional to their mean. However, in some cases, and especially at the sensory periphery (e.g., Arabzadeh et al. 2005), the response variance may be much less than the mean. In general, the more reliable the neural responses, the fewer the relevant responses to a stimulus, and consequently the less the bias. (See the example in Fig. 1). Thus when analyzing reliable neurons, the bias problem is typically less relevant than when analyzing unreliable neurons.

For all these reasons, it is valuable to test information estimation methods on simulated data with statistical properties as similar as possible to the actual experimental data of interest.

Other procedures for information estimation

There are other approaches that either cannot be applied to general datasets, compute the bias of continuous (rather than discrete) neural responses or do not always converge to the true value of information when N becomes infinite. These methods are a useful complement to what presented in the preceding text and are briefly mentioned in the following text. Code for a number of information algorithms can be found at http://neuroanalysis.org/.

ANALYTICAL APPROXIMATION. In certain cases, it may be possible to approximate the stimulus-response probabilities parametrically, for example with a Gaussian distribution (Abbot et al. 1996; Gershon et al. 1998). Or it may be possible to make other assumptions that strongly restrict the complexity of the neural response and thus allow an estimation of information though a simplified analytical expression (e.g., the power series expansion approach) (Panzeri et al. 1999). In such cases, the information depends on relatively few parameters and estimation is therefore simple and data robust. Such procedures can be very valuable but are only applicable in specific situations.

RESPONSE COMPRESSION BY STIMULUS DECODING. When the number of possible responses is high (e.g., a large neural population), a stimulus-decoding procedure can be used to compress the response space into the "predicted stimulus" (see e.g., Rieke et al. 1996; Rolls et al. 1998; Victor and Purpura 1996) as follows. In each trial, a stimulus s is presented and a stimulus s^P is predicted by a decoding algorithm, and the corresponding probability $P(s^P|s)$ (of predicting stimulus s^P when stimulus s is presented) is computed. The "decoded" information $\mathbf{I}(\mathbf{S;S^P})$ is then computed as follows

$$\mathbf{I}(\mathbf{S};\mathbf{S}^{\mathrm{P}}) = \sum_{s,s^{\mathrm{P}}} P(s)P(s^{\mathrm{P}}|s)\log\frac{P(s^{\mathrm{P}}|s)}{P(s^{\mathrm{P}})} \quad (8)$$

Information theoretic inequalities ensure that $\mathbf{I}(\mathbf{S;S^P}) \leq \mathbf{I}(\mathbf{S;R})$ (Shannon 1948). The decoding procedure may be based on a biologically plausible algorithm or on a more abstract but

statistically efficient procedure. If the number of stimuli is much smaller than the number of responses, stimulus-decoding can be an effective and simple way to reduce the space of responses. However, the stimulus decoding algorithm may sometimes perform poorly (for example, the decoding algorithm may predict poorly when operating on a large response array containing many uninformative elements). In such circumstances, unless there is compelling evidence for the decoding algorithm's biological plausibility, not much about the neural code can be learned from the decoded information estimation.

BINLESS STRATEGIES. An interesting and ambitious challenge is to quantify the information carried by a continuous representation of the neural responses, such as considering the spike times with infinite temporal precision rather than binning them (as explained above) into a "binary" word using bins of finite temporal precision. The approach of Victor (Victor 2002; Victor and Purpura 1996) allows an estimation of the information without relying on response discretization. This "binless" approach can potentially address questions than cannot be tackled by discretization methods (e.g.: how much information could be extracted from spikes if we could measures their time with infinite precision?), but it will not work well for small number of trials (Nelken et al. 2006) unless the underlying probability distribution are sufficiently smooth.

MODEL SELECTION TECHNIQUES. As explained in the preceding text, what makes it difficult to sample response probabilities is the presence of correlations between different neurons or spikes. A way to limit the sampling problems posed by correlations is to simplify the correlation structure by fitting it to a simple, low-dimensional model. For example, autocorrelations between spikes emitted by the same neurons could be simplified by assuming they are described by a Markov or suffix-tree model (Kennel et al. 2005; Montemurro et al. 2007). When considering a neuronal population, one may concentrate on pair-wise correlations and ignore higher-order interactions between neurons (Panzeri et al. 1999; Schneidman et al. 2006). This procedure has the potential to lead to huge advantage in sampling properties. The potential disadvantage, analogous to that of the analytic approximation methods, is that it can be difficult to make sure that the selected model is sufficient to describe the whole information content of the spike trains. Also, the sampling advantage is only realized if the structure of correlation is simple, which may not always be the case.

Conclusions

In conclusion, using bias correction procedures is generally essential for obtaining accurate estimates of the information conveyed by spike trains. In the absence of knowledge about some specific feature of the neural response statistics that favor one specific procedure over another, our recommendation is to estimate information using $\mathbf{I}_{sh}(\mathbf{S;R})$ in combination with one of the general purpose bias correction procedures reviewed in the preceding text. The use of $\mathbf{I}_{sh}(\mathbf{S;R})$ makes the results largely independent of the specific bias correction method used; it allows the use of easy-to-implement asymptotic bias corrections such as QE or PT; it provides information estimates with acceptable variance and that are unbiased even when the number of trials per stimulus is as small as the number of possible discrete neuronal responses (i.e., $N_s/\bar{R} \approx 1$). By comparison, when using $\mathbf{I}(\mathbf{S;R})$, even with very good bias correction, a factor of 4 more trials is typically required ($N_s/\bar{R} \geq 4$). To better understand and evaluate any residual errors, we also recommend that simple simulations of neural responses with statistics similar to that of the actual neural data of interest are performed.

APPENDIX: SIMULATION OF CORTICAL SOMATOSENSORY NEURAL RESPONSES

In this APPENDIX, we describe briefly the procedure used to create the simulated data that we used in the main text to test and validate the information analysis methods.

We simulated two types of synthetic spike trains (single cell and population) to 13 stimuli according to the simple processes described in the following text. In both cases, we measured the parameters describing the simulated process from real simultaneous recordings obtained (Arabzadeh et al. 2004) from a population of neurons located in the rat somatosensory cortex in response to 13 different stimuli consisting of sinusoidal whisker vibrations with different vibration energy (200 trials per stimulus were available). We refer to Arabzadeh et al. (2004) for full details on the experimental procedures. The two types of responses were generated as follows.

Single-cell responses

We generated simulated single-neuron responses as binary words made of $L = 8$ binary letters of 0 and 1 s (silence/spike, respectively) that represented the neural response over a 0- to 40-ms poststimulus time window digitized using 5-ms-long time bins. To obtain the simulated sequences, we assumed that the spikes were generated according to a Markov process of order 3 (i.e., the spike in each time bin depended on the activity in the previous 3 time bins). This Markov model is defined in terms of the probability of emitting a spike in each bin and the transition probabilities relating the current response in each time bin to the responses in the previous three time bins. These probabilities were measured, over the corresponding poststimulus window and independently for each stimulus, from the real somatosensory recordings described in the preceding text (see Montemurro et al. 2007 for more details).

Population responses

We simulated a population of $L = 8$ neurons simultaneously responding to the 13 stimuli in the 10- to 15-ms poststimulus window. This time window was chosen because it was the most informative one in the experiments of Arabzadeh et al. (2004). First, using the real spike trains from Arabzadeh et al. (2004), we constructed the "typical" probability of a neuron to fire one spike in the 10- to 15-ms poststimulus window by taking the mean probability across all recorded neurons and across all trials for each stimulus. Second, we assumed that all individual neurons in the simulated population had the same mean firing probability obtained as explained in the preceding text. This assumption was justified because all recorded neurons had similar response profiles to the stimuli considered (Arabzadeh et al. 2004). Third, using again real data, we measured the average level of Pearson correlation coefficient in the considered poststimulus window. This average Pearson cross-correlation value was obtained averaging across all available simultaneously recorded pairs. Then we used the procedure of Mikula and Niebur (2003) to generate correlated spike trains that matched exactly the average mean firing probability of a neuron and the average pair-wise Pearson cross-correlation value. This generated a binary array with $L = 8$ elements (0/1 meaning silence/spike from each of the L neuron).

ACKNOWLEDGMENTS

We are grateful to S. N. Baker for organizing the Engineering and Physical Sciences Research Council (EPSRC)-funded Newcastle workshop on Spike Train Analysis, which inspired the writing of this review. We thank M. Diamond and E. Arabzadeh for sharing data, and P. E. Latham, L. Paninski, and J. D. Victor for useful discussions and insightful comments.

GRANTS

Our research was supported by Pfizer Global Development to S. Panzeri and R. Senatore, EPSRC EP/C010841, EP/E002331, and EP/E057101 to S. Panzeri, and an Medical Research Council Fellowship in Neuroinformatics to M. A. Montemurro.

REFERENCES

Abbott LF, Rolls ET, Tovee MJ. Representational capacity of face coding in monkeys. *Cereb Cortex* 6: 498–505, 1996.

Arabzadeh E, Panzeri S, Diamond ME. Whisker vibration information carried by rat barrel cortex neurons. *J Neurosci* 24: 6011–6020, 2004.

Arabzadeh E, Zorzin E, Diamond ME. Neuronal encoding of texture in the whisker sensory pathway. *PLoS Biol* 3: e17, 2005.

Averbeck BB, Latham PE, Pouget A. Neural correlations, population coding and computation. *Nat Rev Neurosci* 7: 358–366, 2006.

Borst A, Theunissen FE. Information theory and neural coding. *Nat Neurosci* 2: 947–957, 1999.

de Ruyter van Steveninck R, Lewen GD, Strong SP, Koberle R, Bialek W. Reproducibility and variability in neural spike trains. *Science* 275: 1805–1808, 1997.

Endres D, Foldiak P. Bayesian bin distribution inference and mutual information. *IEEE Trans Inform Th* 51: 3766–3779, 2005.

Gershon ED, Wiener MC, Latham PE, Richmond BJ. Coding strategies in monkey V1 and inferior temporal cortices. *J Neurophysiol* 79: 1135–1144, 1998.

Hertz J, Panzeri S. Sensory coding and information transmission. In: *The Handbook of Brain Theory and Neural Networks* (2nd ed.), edited by Arbib MB. Cambridge, MA: MIT Press, 2002, p. 1023–1026.

Kennel MB, Shlens J, Abarbanel HDI, Chichilnisky EJ. Estimating entropy rates with Bayesian confidence intervals. *Neural Comput* 17: 1531–1576, 2005.

Mikula S, Niebur E. The effects of input rate and synchrony on a coincidence detector: analytical solution. *Neural Comput* 15: 539–547, 2003.

Miller GA. Note on the bias of information estimates. In: *Information Theory in Psychology II-B*, edited by Quastler H. Glencoe, IL: Free Press, 1955, p. 95–100.

Montani F, Kohn A, Smith MA, Schultz SR. The role of Correlations in direction and contrast coding in the primary visual cortex. *J Neurosci* 27: 2338–2348, 2007.

Montemurro MA, Senatore R, Panzeri S. Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput* In press.

Nelken I, Chechik G, Mrsic-Flogel TD, King AJ, Schnupp JWH. Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. *J Comput Neurosci* 19: 199–221, 2005.

Nemenman I, Shafee F, Bialek W. Entropy and inference. In: *Revisited Advances in Neural Information Processing Systems,* edited by Dietterich TG, Becker S, and Ghahramani Z. Cambridge, MA: MIT Press, 202, vol. 14, p. 95–100, 2002.

Nemenman I, Bialek W, de Ruyter van Steveninck R. Entropy and information in neural spike trains: progress on the sampling problem, *Phys Rev E* 69: 056111, 2004.

Nirenberg S, Carcieri SM, Jacobs AL, Latham PE. Retinal ganglion cells act largely as independent encoders. *Nature* 411: 698–701, 2001.

Optican LM, Richmond BJ. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. Information theoretic analysis. *J Neurophysiol* 57: 162–178, 1987.

Paninski L. Estimation of entropy and mutual information. *Neural Comput* 5: 1191–1253, 2003.

Panzeri S, Schultz SR, Treves A, Rolls ET. Correlations and the encoding of information in the nervous system *Proc Roy Soc Lond B Biol Sci* 266: 1001–1012, 1999.

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

Pola G, Thiele A, Hoffmann KP, Panzeri S. An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network* 14: 35–60, 2003.

Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W. *Spikes: Exploring the Neural Code.* Cambridge, MA: MIT Press, 1996.

Rolls ET, Treves A, Robertson RG, Georges-Francois P, Panzeri S. Information about spatial view in an ensemble of primate hippocampal cells. *J Neurophysiol* 79: 1797–1813,1998.

Shannon C. A mathematical theory of communication. *Bell Syst Tech J* 27: 379–423, 1948.

Schneidman E, Berry MJ, Segev R, BialekW. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440: 1007–1012, 2006.

Strong SP, Koberle R, de Ruyter van Steveninck R, Bialek W. Entropy and information in neural spike trains. *Phys Rev Lett* 86: 197–200, 1998.

Treves A, Panzeri S. The upward bias in measures of information derived from limited data samples. *Neural Comput* 7: 399–407, 1995.

Victor JD. Binless strategies for estimation of information from neuronal data. *Phys Rev E* 66: 51903–51918, 2002.

Victor JD, Purpura KP. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *J Neurophysiol* 76: 1310–1326, 1996.