Chapter 10

A practical guide to information analysis of spike-trains

Gianni Pola [1], Simon R. Schultz[2], Rasmus S. Petersen[3] and Stefano Panzeri[1]
[1]Department of Psychology, University of Newcastle upon Tyne, The Henry Wellcome Building for Neuroecology, Medical School, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK [2]Howard Hughes Medical Institute and Center for Neural Science, New York University, 4 Washington Place, New York, New York 10003
[3]Cognitive Neuroscience Sector, International School for Advanced Studies, Via Beirut 2/4, 34014 Trieste, Italy
Correspondence to: stefano.panzeri@ncl.ac.uk

Abstract: Information Theory enables different candidate coding strategies to be quantified and compared, and hence is a natural framework for studying neural coding. The main difficulty is that estimates of information from experimental data are prone to systematic *sampling error*. In this chapter, we present a step-by-step guide to how this error can be addressed, and reliable information estimates obtained.

Key words: Information Theory, Spike Timing, Neural Coding.

1. INTRODUCTION

Almost all sensory messages and motor commands are encoded as temporal patterns of action potentials (spikes), often distributed over thousands of neurons. A fundamental problem in neuroscience is to understand the nature of this neural code. Ideally, as argued by Rieke et al (1996), we would like a dictionary that, given some ensemble of spike trains, tells us what sensory signal has occurred or what motor command is about to be executed. As a first step, we need to know what kind of code we are dealing with. For example, is the precise timing of spikes important, or is it just the number of spikes that matters? A powerful approach is to treat the brain as a communication channel, and to seek to quantify and compare the "information" available in different candidate codes. Information Theory provides a mathematical framework for quantifying information using

rigorous measures such as mutual information (Shannon, 1948; Cover and Thomas, 1991). This approach has recently become a very popular way of studying neural coding (Borst and Theunissen, 1999).

Mutual information quantifies how well an ideal observer reading out the neuronal responses on a given trial is able to determine the stimulus identity. It is a functional of the distribution of joint probabilities between each possible stimulus and each possible response. In a typical neurophysiological study, the experimenters repeatedly present certain stimuli to the animal, whilst recording the spike emission times. To measure the mutual information that such responses convey about the stimulus set, it is necessary to use the recorded data to *estimate* these joint probabilities. The number of trials that can feasibly be collected is limited, and the probability estimates are therefore necessarily imprecise. The key difficulty in applying Information Theory is that these estimation errors lead to an *upward bias* in the mutual information, and since this bias gets worse for more complex codes, there is a very real danger of spurious results.

In this chapter we will focus on two approaches developed and used by the authors to address this problem. The *bias correction* approach is to explicitly estimate the bias, so that it can be subtracted. The *series expansion approach* (Panzeri and Schultz, 2001) is to approximate the mutual information by an expression that depends only on individual-cell mean firing rates and pair-wise correlations between spikes. Since these quantities are much easier to estimate accurately from limited data than the full response probabilities, the series expansion has much better sampling properties than "brute force" estimation of the mutual information. These two approaches are complementary, and indeed we often use them together.

With the aid of this chapter and of the discussed references, the reader should be able to compute basic information quantities (Section 2), subtract bias estimates to improve their accuracy with limited data samples (Sections 3 and 4), and implement the series expansion formalism (Section 5).

2. SHANNON INFORMATION: DEFINITIONS

Consider a time period of duration T, associated with a dynamic or static sensory stimulus, during which the activity of C cells is observed. The neuronal population response to the stimulus in this post-stimulus time window is described by a vector r - each element $r_1, \ldots, r_C$ of the vector indicating the response of an individual cell. The response of each cell can be described in a number of ways depending on the experimental questions to be addressed. For example, the experimenter might be interested in a *spike count code*. In this case r_c would simply be the number of spikes emitted by

cell c in the post-stimulus time window $[0,T]$ on a given trial (see Fig 1). Or else, the experimenter might wish to investigate a *spike timing code*. In this case the response r_c would be a sequence of spike arrival times $\{t^c_j\}$, t^c_j being the time of the j-th spike emitted by the c-th neuron in a given trial. In this case the responses of the cell ensemble are sampled into L time-bins with a finite time precision dt (i.e. $T=Ldt$), small enough to contain at most one spike. Each cell response can be represented by a binary word[1] of length L, the entries of which represent the occupation number of each time-bin (see Fig 1). The response of the entire neuronal population can be represented by a binary word of length LC.

For a given choice of code, following Shannon (1948), we can write down the mutual information transmitted by the population response about the whole set of stimuli as

$$I = \sum_s P(s) \sum_r P(r \mid s) \log_2 \frac{P(r \mid s)}{P(r)} \qquad (1).$$

Summation is over all possible population responses r and stimuli s. $P(r|s)$ is the probability of response r given presentation of stimulus s, and $P(r)$ is its average across all stimulus presentations. The probability $P(r|s)$ is determined experimentally by repeating each stimulus in exactly the same way on many trials, while recording the neuronal responses. The probability

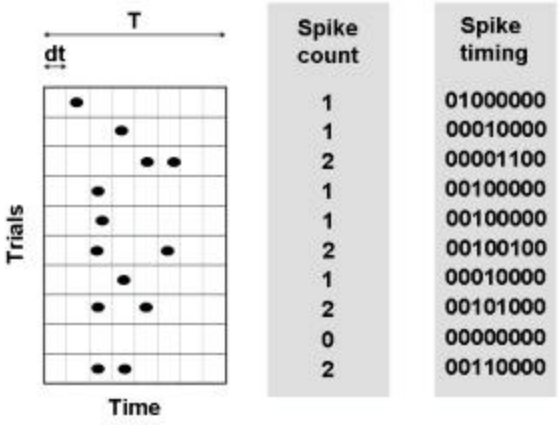

Figure 1. Spike count and spike timing codes. Each trial consists of a number of bins, which can either be occupied by an action potential (indicated by a dot), or not. The way in which this pattern is turned into a number is determined by the *code* – two examples are shown at right.

[1] A binary word is a sequence of two different symbols. In this context we are assuming that there is at most one spike in each time-bin, hence we are considering a $\{0,1\}$ sequence.

$P(s)$ is usually chosen by the experimenter.

A set of neurons can only encode stimuli effectively if each stimulus evokes a distinct response. Mutual information quantifies how informative these distinct response distributions are: it tells us how well an ideal observer of neuronal responses can discriminate between all the different stimuli, based on single trial responses. Mutual information is actually a "Kullback-Leibler distance" between the response distribution for each particular stimulus and the "overall" distribution of responses. This measures how well any given stimulus can be "picked out from the crowd" – see Cover and Thomas (1991).

3. WHAT IS THE BIAS PROBLEM?

Shannon's mutual information (Eq. (1)) is defined in terms of the probabilities of neuronal responses to the stimuli. In order to estimate the mutual information conveyed by a neuron, for example, the key step is to *estimate* these probabilities from the available neurophysiological data. If we had an infinite amount of data, we could measure the true probabilities precisely but, in practice, our estimates will necessarily differ from the true values. As explained below, this leads directly to the key issue to the practical application of Information Theory.

Consider, for example, a "neuron" that responds identically to each of two stimuli. On each trial, the neuron fires from *0* to *9* spikes with equal probability, and this is the same for both stimuli. In this case, $P(r|s) = 0.1$ for all responses and both stimuli, and Eq. (1) tells us that the mutual information is precisely zero. However, consider the effect of limited sampling. Fig 2a shows the probabilities estimated from a random sample of *10* trials per stimulus. The estimated probabilities differ markedly from *0.1* and, as a result, the estimated mutual information is *0.419 bits.* The distribution of information estimates obtained by repeating this calculation for *5000* different samples of *10* trials is shown in the right panel. Although the exact result varies from sample to sample, there is a large, consistent, bias. The lower panels illustrate the effect of increasing the number of trials. With $N=100$, the fluctuations in the estimated probabilities are smaller. There is still a bias, but it is much lower (*0.046 bits*). With $N=1000$, the estimated probabilities are very near to their true values, and the bias has almost disappeared (*0.004 bits*).

In this example, the magnitude of the bias is inversely proportional to the number of trials per stimulus. In fact, this is a general characteristic of information bias (Miller 1955; Treves and Panzeri 1995; Panzeri and Treves 1996; Victor 2000). A second general characteristic is that the bias

magnitude depends on the number of different response categories that are considered. With a spike count code (e.g., the simulated neuron in Fig 2), the number of response bins required scales approximately linearly with spike count. With spike timing codes and population codes, the number of different response categories scales exponentially with the number of neurons and with the number of time bins, and hence the bias problem is much more severe.

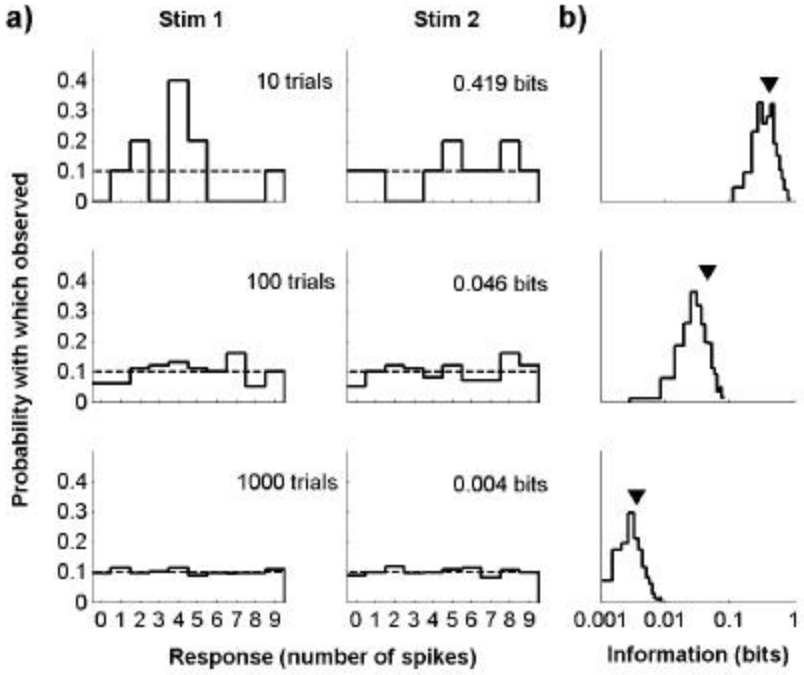

Figure 2. The effect of limited sampling. An uninformative neuron was simulated, which responded on each trial with a uniform distribution of spike counts ranging from *0* to *9*, regardless of which of two stimuli was presented. (a) Examples of response probability histograms for *10*, *100* and *1000* trials. Left: responses to Stimulus 1; Right: responses to Stimulus 2; the dashed line is the true (underlying) response distribution. Also shown are the mutual information values corresponding to each pair of distributions, obtained from Eq. (1). (b) The distribution of information values obtained with a sample of *10*, *100* and *1000* trials per stimulus (see main text). As the number of trials increases, the information bias shifts towards zero and the standard deviation decreases. The arrows in panel (b) indicate the values of the information of the corresponding example in panel (a).

3.1 Calculating the bias correction and estimating the number of relevant bins

We define the *bias* of the mutual information to be the difference between an average of the estimated information I_N across repeated samples of N trials (across all stimuli) and the true information I computed with the true underlying probability distributions

$$Bias(I) = < I_N[P_N(\cdot)] >_N - I[P(\cdot)] \quad (2),$$

where $< \ldots >_N$ stands for an average performed over all possible outcomes of the N trials (Treves and Panzeri, 1995) and I_N stands for the information computed starting with the estimated probabilities. Remarkably, provided that the number of trials is not too small (we make this more precise below), this expected bias is well-approximated by a simple expression. This expression depends only on the total number of trials and on the number of so-called *relevant* response categories. A given stimulus will evoke different responses with different probabilities: some will have high probability, some low probability, and others will have zero probability. A response is *relevant* if its probability, conditional to the stimulus, is non-zero. R_s is the number of responses that are relevant to stimulus s; R is the number of responses that are relevant considering all the stimuli together. The formula for the bias (Panzeri and Treves, 1996) is:

$$Bias(I) = -\frac{1}{2N \ln 2}(R - 1 - \sum_s (R_s - 1)) \quad (3).$$

To evaluate this expression, the key step is to estimate the number of relevant responses. This is non-trivial, since, with real data, we only have an *estimate* of the response probabilities, and hence only an estimate of the number of relevant responses. We will discuss three approaches to this problem.

The simplest approach is to approximate R_s by the *total* number of response categories R^{tot}, which we call the "full" count. In this case, the bias is:

$$Bias(I) = \frac{1}{2N \ln 2}(S - 1)(R^{tot} - 1) \quad (4),$$

where S is the number of the stimuli. Since $R^{tot} = R_s$, the full count bias is an upper bound on the actual bias. If all responses occur, the bound is right, but if the responses are sparse, the bias will be significantly over-estimated.

Another simple approach is to approximate R_s by the number of responses that were observed at least once – the "naïve" count. This will be a *lower* bound on the actual number of relevant bins, and hence the naïve count bias will be a lower bound on the actual bias.

These ideas are illustrated in Fig 3. We generated up to *8192* trials for the responses of a pair of neurons to each of three stimuli. If no bias correction is applied (solid line), the information decreases as the number of trials increases, as expected. The figure shows the effect of correcting the information using both the full count and naïve count estimates of R_s. When the number of trials is small, there is a big difference, which gradually shrinks as the N increases. The true mutual information must lie between the full and naïve curves.

Although these simple strategies are useful for bracketing the true information, we would like a method for making a better estimate of the number of relevant responses. In many cases, good results can be obtained by using a Bayesian procedure for estimating R_s (Panzeri and Treves 1996).

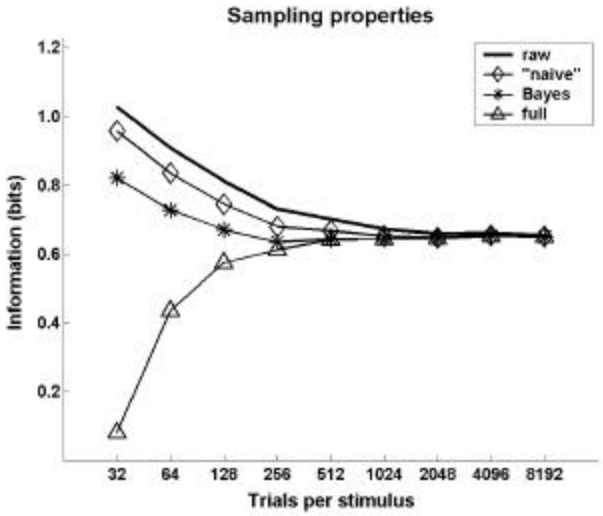

Figure 3. Correcting the bias with different subtraction procedures. In this example we considered the information carried by the joint spike timing of simulated neuronal pairs. We simulated three stimuli and we considered three time bins per cell. Hence the total number of response bins R^{tot} is equal to *64*. The information is plotted as a function of the number of simulated trials per stimulus. Results are averaged over *100* simulations for each fixed number of trials per stimulus. The "raw" information obtained without any bias correction is compared with the values obtained by using three different ways to count the number of relevant bins in the bias correction formula Eq. (3): the "naïve" count, the "full" count, and the "Bayes" count. It is clear that the "Bayes" count produces more accurate estimates than the other procedures, particularly at low number of trials.

In brief, the Bayes counting procedure of Panzeri and Treves (1996) is based on the empirical idea to use some simple prior information about the response probability distribution (such as "the prior probability of bins in which many responses were observed may be higher than that of bins in which no response was observed") to guess the number of relevant bins from the data. According to this procedure, the most likely number of relevant bins given the data and the prior is the one that reproduces more closely the number of response bins with non-zero probability observed experimentally from N trials. (Details on how to implement this procedure are reported in Appendix A). Results of applying this method are also shown in Fig 3. It can be seen that the Bayes-corrected information falls between the bounds given by the full and naïve estimates. The Bayes method has been extensively tested on various datasets and we apply it routinely in our studies of neural coding (Panzeri et al 2001; Petersen et al 2002). Although the Bayes count procedure is usually very effective, there are two potential problems with it. The first problem is that it can go wrong when the assumptions made for the priors (see Appendix A) are violated. The second problem is that the Bayes count procedure, in the form presented here, does not allow to estimate the residual error in the information evaluation. This last problem was to a good extent overcome by the more advanced and recent work of Nemenman et al (2002). A rule of thumb for the amount of data required to compute accurately the information using the "Bayes" procedure is as follows (Panzeri and Treves 1996; Pola et al 2002). When there are more trials per stimulus (at least *2* to *4* times) than population response classes R^{tot}, then the mutual information can be reasonably well corrected by the Bayes "bias" correction term and the estimated information is computed with reasonable accuracy.

We finally note that another method to empirically determine the value of the bias correction term of Eq. (3) from the data was proposed in Strong et al (1998). The procedure of Strong and colleagues consists in dividing the dataset into sub-samples and to determine empirically a best fit polynomial curve to the scaling of information with the data size. The linear term of the fit could be used to estimate the magnitude of the bias term in Eq. (3). We have tested this interesting empirical algorithm extensively, and found that it generally works less well than the Bayes procedure for estimating the bias, although it works generally much better than the "naïve" procedure.

4. SERIES EXPANSION APPROACH

Since the bias increases with the number of response categories R^{tot}, it is a particular problem when estimating the information carried by spike timing

codes or population codes. Even using the correction procedures discussed above, it is usually possible to make useful information estimates only for very short word lengths. However, it is possible to achieve considerably better resolution if the number of spikes per stimulus is small. Since this is the case for many neurophysiological preparations (see, for example, Rieke et al 1996), we developed the *series expansion approach* to address such situations. The idea is that, when the number of spikes in the response time window is small, the mutual information can be well-approximated by a power series expansion in the time window that depends only on low order firing statistics. Hence, the series expansion method tends to be much less susceptible to sampling bias than "brute force" estimation from Eq. (1) and, for a given number of trials, permits information to be estimated at greater temporal resolution. Provided the assumptions of the method are met (see below), we have found a second-order approximation to provide accurate results in a number of cases. The derivation of the equations has been presented elsewhere (Panzeri et al 1999, Panzeri and Schultz 2001), as have its applications to neurophysiological data (Panzeri et al 2001, Petersen et al 2001, Petroni 2002). Our aims here are to explain how to correct the series expansion for bias and how to verify if its assumptions are satisfied.

4.1 The series expansion of the Mutual Information

The series expansion method mainly assumes that (i) the time window is short enough (or rates are low enough) so that there are very few spikes per time window (ii) Spikes are not locked to one another with infinite time precision – in other words the conditional probability of a spike occurring in a time bin given the presence of other patterns of spikes scales for small dt proportionally to dt plus higher order terms. If these assumptions are satisfied (see Panzeri and Schultz 2001 for a more precise formulation), the mutual information can be approximated by a Taylor series expansion. To second order:

$$I \approx It + Itta + Ittb + Ittc \quad (5).$$

An important feature of the method is that each of these terms quantifies different types of coding. *It* and *Itta* together express information that the neurons convey by virtue of the timing of *individual* spikes. Any extra contribution due to correlations between spikes is given by *Ittb* and *Ittc*. For precise details, see Panzeri and Schultz (2001); for a more intuitive description, see Petersen et al (2001). The four terms of the second-order series expansion are as follows:

$$It = \sum_{a,i} \left\langle \bar{n}_{ais} \log_2 \frac{\bar{n}_{ais}}{\left\langle \bar{n}_{ais'} \right\rangle_{s'}} \right\rangle_s \quad (6),$$

$$Itta = \frac{1}{2} \sum_{a,b,i,j} \left[CS_{aibj} \left(1 - \log_2 \frac{CS_{aibj}}{MS_{ai} MS_{bj}} \right) - ECS_{aibj} \right] \quad (7),$$

$$Ittb = -\frac{1}{2} \sum_{a,b,i,j} \left\langle CN_{aibjs} - ECN_{aibjs} \right\rangle \log_2 \frac{CS_{aibj}}{ECS_{aibj}} \quad (8),$$

$$Ittc = \frac{1}{2} \sum_{a,b,i,j} \left\langle CN_{aibjs} \log_2 \left[\frac{CN_{aibjs}}{ECN_{aibjs}} \div \frac{\left\langle CN_{aibjs'} \right\rangle_{s'}}{\left\langle ECN_{aibjs'} \right\rangle_{s'}} \right] \right\rangle_{s'} \quad (9),$$

where n_{ais} is the response in time bin i of cell a to stimulus s on a particular trial. The bar means an average over trials, thus $\bar{n}_{ais}$ is simply the corresponding Post-Stimulus-Time-Histogram (PSTH) normalized in units of mean number of spike per time bin. The angle brackets $< \ldots >_s$ denote an average over stimuli, weighted by the stimulus probabilities $P(s)$. $MS_{ai} = \langle n_{ais} \rangle_s$ is the average of the PSTH over stimuli for time bin i of cell a. $CS_{aibj} = \langle n_{ais} n_{bjs} \rangle_s$ is the *signal correlation* between time bin i of cell a and bin j of cell b; $ECS_{aibj} = MS_{ai} MS_{bj}$ is the expected value of CS_{aibj} for PSTHs that are uncorrelated across stimuli. CN_{aibjs} (*noise correlation)* is the *joint* PSTH of bin i of cell a and bin j of cell b given stimulus s. It is equal to $n_{ais} n_{bjs}$, unless $a=b$ and $i=j$, in which case it is zero. $ECN_{aibjs} = n_{ais} n_{bjs}$ is the expected value of CN_{aibjs} for statistically independent spikes.

4.2 Bias correction for the series expansion

However, it is useful to estimate the bias for each of the four terms of the expansion separately. The first important point is that to a very good approximation, the bias comes only from *It* and *Ittc*. The rationale for correcting bias in these terms remains that of Eq. (3). For the first term, *It*, the number of responses is simply the total number of bins in all the (single cell) PSTHs. Thus $R_{tot}=CL$. R_s is the number of non-zero PSTH bins. For the fourth term, *Ittc*, the number of responses is the number of pairwise combinations of time bins. This has two components – an auto-correlation component from within single cells, and a cross-correlation component from across pairs of cells. Auto-correlation contributes *CL(L-1)/2* responses, cross-correlation $L^2C(C-1)/2$ responses. For a single cell, *Ittc* is characterised by *L(L-1)/2* responses. For a pair of cells, it is characterised by $L^2 + L(L-1)$ responses.

The key point is that, whereas the number of response categories in the brute force method scales *exponentially* with the number of time bins, it scales only *quadratically* with the second order series expansion. Thus, for a given amount of data, the series expansion permits an analysis at higher temporal resolution. For example, by considering the responses of one cell with *10* time bins, some two hundred trials per stimulus should be sufficient to have a good estimation of the information with the Bayes count procedure (out of two thousands for a "brute force" evaluation directly from the mutual information formuale, see Section 2 and 3).

How does the series expansion method compare to the brute force method on real data? We analysed the mutual information conveyed in the responses of single cells recorded from rat barrel cortex in response to stimulation of each of *9* whiskers on the rat's snout. In this case, *50* trials per stimulus were available. We considered the time window [*0-80 ms*] post-stimulus (*0*= time of whisker deflection) and subdivided it into *10 ms* bins. Results of applying the two methods are shown in Fig 4. It is evident that the "brute force" estimation of the full temporal information diverges rapidly after the first four to five time bins. This is due to failure of corrections for finite sampling, as expected by the rule of thumb for sampling corrections discussed above. In fact while at *40 ms* the number of the response bins is

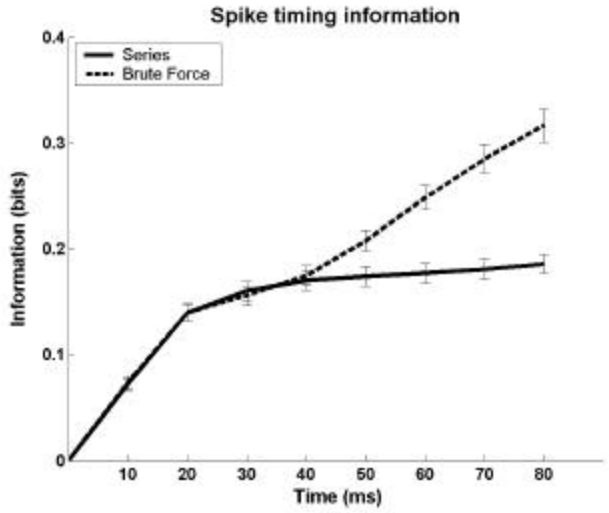

Figure 4. Information analysis of neuronal responses of *106* single units recorded in the barrel-column D2 of the rat somatosensory cortex (see Panzeri et al. (2001) for full details). Both series expansion and "brute force" method were applied. Results averaged across all cells and are plotted as a function of post-stimulus time. The information is about which of 9 whiskers (centred around whisker D2) was stimulated.

$2^4=16$, at *80 ms* the response bins are $2^8=256$. It is clear that *50* trials per stimulus could be enough to estimate the spike timing information until *40 ms* but they do not give credible results for larger time windows. The spike time series expansion is a close match to the "brute force" estimator up to *40* to *50 ms*. This result illustrates the benefit obtained by the superior scaling properties of the series expansion, compared to "brute force" information estimates. A detailed comparison of the sampling properties of brute force and series expansion approaches is reported in Schultz and Panzeri (2001).

4.3 Checking the method

The series expansion approach is an approximation: it is based on some assumptions. So it is necessary to evaluate if the results are valid. Here we discuss how to perform some important checks to validate the results. First, it is necessary to verify that the approximation is sensible: the mathematical assumptions of the method should be satisfied. Second, in order to get an idea of how precise the approximation, it is important to compare the "brute force" estimation of data-robust information quantities (such as spike count information or total response entropies) to the corresponding estimates obtained with the series expansion.

Suppose that we have applied the series expansion method to some data and obtained a certain result. How can we validate whether the results are well-sampled? We suggest that the following points should be checked:

- Is the stimulus average firing rate in all time windows of interest less than one?
- Is $\mathrm{C}\mathrm{S}_{abij}$ / $\mathrm{E}\mathrm{C}\mathrm{S}_{abij}$ not divergent at any time resolution *dt?*
- Do brute force and series expansion estimates match for spike count? Usually, enough data is available with which to estimate information in the spike count using the brute force method. Since the series expansion for spike *count* makes the same assumptions as that for spike timing, a useful check is to compare the two spike count estimates. An example for rat barrel cortex data is shown in Fig 5a.
- Do brute force and series expansion estimates match for the response entropy? The response entropy in spike timing for each cell is well sampled especially with a large number of stimuli: in fact its bias properties depend on the total number of trials available and not on the trials per stimulus (see Cover and Thomas 1991). So this check can address directly the hypothesis of the series expansion in the spike timing. An example for rat barrel cortex data is shown in Fig 5b.

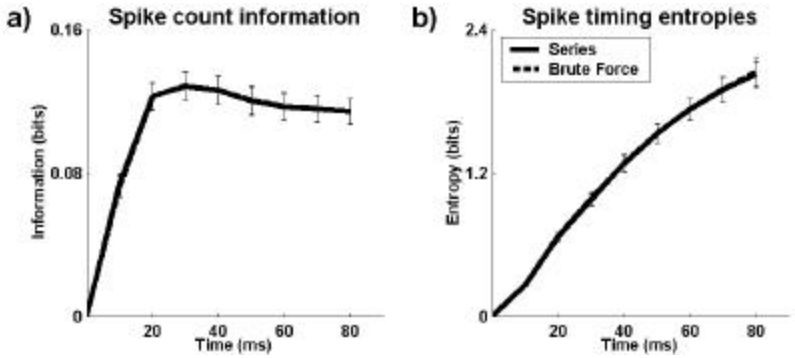

Figure 5. Checking the method on single cells from rat somatosensory cortex (a) A comparison of spike count information with series expansion and "brute force" method. (b) A comparison of spike timing entropy information with series expansion and "brute force" method.

5. SUMMARY

The main practical issue in making accurate information estimations from neurophysiological data is that these estimates are prone to systematic bias. In this chapter we have reviewed some practical aspects on how to solve or at least alleviate this problem, by discussing how to apply the bias subtraction procedure and by presenting the series expansion approach. These methods have permitted a detailed quantitative analysis of the role of both spike timing and spike correlations in cortical sensory coding (Panzeri et al 2001; Petersen et al 2001). Although these methods have been applied so far mainly to real experimental data, where data sampling problem is particularly severe, it is conceivable that they will be fruitfully applied in the future to the analysis of simulated spike trains generated by computer models of neuronal microcircuits (see Chapter 9: Computer models and analysis tools for neural microcircuits, by Natschläger, Markram and Maas), contributing thereby to a better understanding of the relationship between neuronal structures and neuronal coding.

Acknowledgements

We thank Mathew Diamond and Misha Lebedev for valuable scientific collaboration and for making available to us the neuronal data that we used as an illustration. This research was supported by EU, BBSRC, Telethon Foundation, J.S. McDonnell Foundation, Wellcome Trust grant 066372/Z/01/Z, MURST, CNR and Regione Friuli Venezia Giulia. SRS was supported by HHMI and SP by an MRC Research Fellowship. Some matlab codes that implement the procedures described here are available at http://www.staff.ncl.ac.uk/stefano.panzeri/software.html

APPENDIX A

In this Appendix we review the Bayesian procedure to count the number of relevant response bins that was developed by Panzeri and Treves (1996). With respect to what presented in Panzeri and Treves (1996), we try to describe more clearly the practical implementation of this empirical procedure.

In brief, the Bayesian counting procedure is based on the idea of using Bayes' theorem to "guess" the number of relevant bins from the data. The number of relevant bins, i.e. the number of response bins with non-zero probability, is one of the parameters of the prior distribution of probabilities. This parameter is varied in the allowed range, and then the most likely value for number of relevant bins is chosen to be the one that predicts an expected number of occupied bins (calculated from the Bayes estimate of the probabilities) that matches the experimentally observed value.

We focus for simplicity on how to compute the number of bins R_s relevant to a certain stimulus s, and we discuss at the end how this generalizes to the calculation of the number of bins R relevant to the unconditional response probability. In the following we denote by R_s^{naive} the "naïve" count of the relevant bins, i.e. the number of different response bins observed at least once from N trials (see main text), and by R_s^{tot} the total number of response bins. Obviously, the true number of relevant bins R_s is bound between R_s^{naive} and R_s^{tot}. The procedure for the Bayesian estimation of R_s is as follows.

- We select a particular value of R_s, starting from $R_s = R_s^{naive}$, and then increasing by one R_s at each recursive step until the procedure converges (see below);
- For the selected value of R_s, we construct a Bayesian estimate of the true response probabilities $P^B(r|s)$ given the experimental frequencies. This is done as follows. The prior distribution of response probabilities is constant across all the R_s - R_s^{naive} that were non-occupied in the experiment. The sum of the total prior probability allocated across all non-occupied bins is denoted by G. The prior probability is also assumed to be constant across experimentally occupied bin. Of course the total prior probability allocated across the occupied bins is $1-G$. With these flat choice for priors, one can obtain the following expression for the Bayesian estimates of the probabilities as a function of R_s and G (derivation given in Wolpert and Wolf 1995):

$$P^B(r \mid s) = \frac{n+1}{N_s + R_s^{naive}}(1-\Gamma) \quad \text{if } r \text{ is "occupied"}$$

$$P^B(r \mid s) = \frac{\Gamma}{R_s - R_s^{naive}} \qquad \text{if } r \text{ is not "occupied"} \quad \text{(A1)},$$

where n is the number of events observed in the particular bin r.

- For each , we compute the expected value of the number of occupied bins:

$$\left\langle R_s^{naive} \right\rangle = \sum_r \left\{ 1 - \left(1 - P^B(r \mid s)\right)^{N_s} \right\} \quad \text{(A2)};$$

- We then compare this expected value with the one compared when R_s was one less. We stop the procedure when increasing R_s does not bring the expected number of bins any closer to the observed value of the number of occupied bins, or else when R_s reaches its maximum allowed value R^{tot}.
- After stopping the procedure, we choose as our estimate of the number of relevant bins the value of R_s that gives the expected number of bins which is closest to the experimental value of the number of occupied bins to stimulus s.

One arbitrary point of the procedure is the choice of the factor G that describes the flat priors in both the occupied and non-occupied response bins. We tried a range of different choices, and found that they work with roughly similar performance as long as the prior probability per bin is higher for occupied than for non-occupied bins. The particular choice of G that we used in Panzeri and Treves (1996), as well as in several other papers (e.g. Pola et al 2002; Petersen et al 2002; Schultz and Panzeri 2001), is as follows. We chose G so that the Bayes estimate of the response probability $P^B(r|s)$ of a non-occupied bin satisfies the following condition: the probability of that bin being empty in N_s trials is N_s / R_s times larger than the probability of being occupied in the same number of N_s trials. This requirement reflects the idea that if all observed responses are concentrated in a few bins only (i.e. high N_s/R_s), the probability in the non-occupied bins should be less than the probability assigned by a prior function constant across all R_s bins. The resulting equation for G is:

$$\Gamma = \left(R_s - R_s^{naive}\right)\left[1 - \left(\frac{N_s}{N_s + R_s^{naive}}\right)^{1/N_s}\right] \quad \text{(A3)}.$$

The procedure to evaluate the stimulus-unconditional number of bins R is as above, the only difference being that the quantities have to be computed from all N trials, and not from only the N_s trials specific to stimulus s.

References:

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

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

Miller, G. A. (1955) *Note on the bias on information estimates*. *Information Theory in Psychology; Problems and Methods II-B*, 95-100.

Nemenman, I., Shafee, F. and Bialek, W. (2002) *Entropy and Inference, Revisited* in Advances in Neural Information Processing Systems 14. Dietterich, T. G., Becker, S. and Z. Ghahramani (Eds.), MIT Press, Cambridge, MA.

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

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

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

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

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

Petroni, F. (2002) *A computational analysis of the functional role of timing of action potentials in the cerebral cortex.* Unpublished doctoral dissertation, University of Newcastle upon Tyne.

Pola, G., Thiele, A., K. –P. Hoffmann and Panzeri, S. (2002) An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network: Special issue on statistical analysis of spike trains and information theory*, submitted.

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

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

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

Strong, S.P., Koberle R., de Ruyter van Stevenink, R.R. and Bialek, W. (1998) Entropy and information in neuronal spike trains. *Phys. Rev. Lett.* 80, 197-200.

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

Victor, J. D. (2000) Asymptotic bias in information estimates and the exponential (bell) polynomials. *Neural Comput.* 12, 2797-2804.

Wolpert, D. H. and Wolf , D. R. (1995) Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E*, 52, 6841-54.