How Well Can We Estimate the Information Carried in Neuronal Responses from Limited Samples?

David Golomb
Zlotowski Center for Neuroscience and Department of Physiology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

John Hertz
Nordita, Copenhagen, Denmark

Stefano Panzeri
Department of Experimental Psychology, Oxford University, Oxford OX1 3UD, U.K.

Alessandro Treves
SISSA, Biophysics and Cognitive Neuroscience, Trieste, Italy

Barry Richmond
Laboratory of Neuropsychology, NIMH, NIH, Bethesda, MD, USA

It is difficult to extract the information carried by neuronal responses about a set of stimuli because limited data samples result in biased estimates. Recently two improved procedures have been developed to calculate information from experimental results: a binning-and-correcting procedure and a neural network procedure. We have used data produced from a model of the spatiotemporal receptive fields of parvocellular and magnocellular lateral geniculate neurons to study the performance of these methods as a function of the number of trials used. Both procedures yield accurate results for one-dimensional neuronal codes. They can also be used to produce a reasonable estimate of the extra information in a three-dimensional code, in this instance, within 0.05–0.1 bit of the asymptotically calculated value—about 10% of the total transmitted information. We believe that this performance is much more accurate than previous procedures.

1 Introduction

Quantifying the relation between neuronal responses and the events that have elicited them is important for understanding the brain. One way to do this in sensory systems is to treat a neuron as a communication channel (Cover & Thomas 1991) and to measure the information conveyed by the

neuronal response about a set of stimuli presented to the animal. In such experiments (e.g., Gawne & Richmond 1993; McClurkin et al. 1991; Tovée et al. 1993) a set of S sensory (e.g., visual) stimuli is presented to the animal, each stimulus being presented for N_S trials. After the neuronal response is quantified in one of several ways (e.g., the number of spikes in a certain time interval or a descriptor of the temporal course of the spike train), the transmitted information (mutual information between stimuli and responses) is estimated. This approach is useful for investigating issues such as the resolution of spike timing (Heller et al. 1995), the effectiveness of encoding for stimulus sets (Optican & Richmond 1987; McClurkin et al. 1991; Rolls et al. 1996a, 1996b), and the relations between responses of different neurons (Gawne & Richmond 1993).

The equation for transmitted information can be written in several ways, including:

$$I(S;R) = \int d\mathbf{r} \sum_s P(s,\mathbf{r}) \log_2 \left[\frac{P(s,\mathbf{r})}{P(s)P(\mathbf{r})} \right]$$

$$= \left\langle \sum_s P(s\,|\,\mathbf{r}) \log_2 \left[\frac{P(s\,|\,\mathbf{r})}{P(s)} \right] \right\rangle_{\mathbf{r}} . \tag{1.1}$$

Here $P(s,\mathbf{r})$ is the joint probability of stimulus s and response $\mathbf{r}$, $P(s\,|\,\mathbf{r})$ is the conditional probability of stimulus s given response $\mathbf{r}$, and $P(s)$ and $P(\mathbf{r})$ are the unconditional stimulus and response probabilities, respectively. The stimulus index s is discrete, as there is a finite number of stimuli. The response measure $\mathbf{r}$ can be either discrete or continuous, depending on the way the response is quantified.

The notation $\langle \cdots \rangle_{\mathbf{r}}$ in the second form indicates an average over the (unconditional) response distribution. Computing $I(S;R)$ requires estimates of these joint or conditional probabilities and carrying out the appropriate integration over response space or the average over the response distribution.

A large number of trials is needed for accurate information calculations. When the number of data samples is small, there are systematic errors in estimating the transmitted information by direct application of the formal definition, that is, by binning responses and estimating the relevant probabilities as the empirical bin frequencies (Miller 1955; Carlton 1969; Optican et al. 1991; Treves & Panzeri 1995; Abbott et al. 1996; Panzeri & Treves 1996). With a small number of data samples, the information is overestimated; the information is biased upward. Intuitively, this can be understood by noting that the number of bins can exceed the number of data samples, especially as the dimensionality of the response increases. The data samples must be evenly distributed across the bins to yield zero information, a circumstance that cannot occur when the number of bins exceeds the number of data samples. On the other hand, if the number of bins is too small, responses that should be reliably distinguishable will not be, leading to information values

that underestimate the true values. These effects are magnified when the dimensionality of response increases, and thus even more samples are needed to make accurate information estimates when the response dimensionality is larger than one.

In practice, the size of data sets is limited by experimental constraints. Thus it is important to try to correct for these systematic errors and to design experiments so that a sufficient number of data samples are obtained to answer reliably the questions posed. Only then can we make effective use of mutual information measurements. Here, we test procedures designed to overcome the systematic errors that arise with limited data sets by comparing mutual information calculations from large data sets with those calculated from smaller ones. Currently the only way to get such large data sets is through simulations. Artificial spike trains are created using a model of the response of lateral geniculate nucleus (LGN) neurons (Golomb et al. 1994). The model is based on experimentally measured spatiotemporal receptive fields of LGN neurons (Reid & Shapley 1992). Two procedures are tested here. The first is a binning-and-correcting procedure; the information is calculated by direct application of the first definition in equation 1.1, and a correction value is estimated and subtracted (this method attacks the bias problem) (Panzeri & Treves 1996; Treves & Panzeri 1995). The second method uses a neural network to estimate the conditional probabilities in the second definition in equation 1.1 directly (Heller et al. 1995; Hertz et al. 1992; Kjaer et al. 1994). We find that both methods yield accurate results for one-dimensional codes, even for a relatively small number of samples. Moreover, estimates of the extra information carried in three-dimensional codes are also reasonable, within 0.05–0.1 bits (about 10%) of the correct values.

2 Methods

2.1 Producing Simulated Data. The spike trains are created using a model of the response of parvocellular and magnocellular LGN cells, as described in Golomb et al. (1994). In brief, the spatiotemporal receptive fields $R(\vec{r}, t)$ of the two cells types in response to an impulse in space and and time were measured (Reid & Shapley 1992; data presented in Figure 1 of Golomb et al. 1994). The set of stimuli $\{\sigma_s\}, s = 1 \cdots S(=32)$ includes 4×4 flashed Walsh figures in space and their contrast reverse,

$$\sigma_s(\vec{r}, t) = u_s(\vec{r})\Theta(t), \tag{2.1}$$

where $u_s(\vec{r})$ are the spatial Walsh figures and $\Theta(t)$ is the Heavyside function ($\Theta(t) = 1$ if $t > 0$ and is 0 otherwise). The ensemble-average response to the sth figures $Z_s(t)$ is calculated by centering it on the receptive field center, convolving it with the spatiotemporal receptive field, adding the constant

baseline Z_0 corresponding to the spontaneous firing, and rectifying at zero response:

$$Z_s(t) = \Theta\left[Z_0 + \int d\vec{r} \int_{-\infty}^{t} dt' \, R\left(\vec{r}, t - t'\right) \, \sigma\left(\vec{r}, t'\right)\right]. \tag{2.2}$$

The response $Z_s(t)$ for Walsh figures is shown in Figure 5 of Golomb et al. (1994).

Realizations of spike trains are created at random with inhomogeneous Poisson statistics, using the average response as the instantaneous rate. The probability density of obtaining a spike train $\Lambda_s(t)$, with k spikes at times $t_1 \ldots t_k$ during a measurement time T, is

$$\begin{aligned} P\left(\Lambda_s\left(t\right) | \sigma_s\right) &= P\left(t_1 \ldots t_k | \sigma_s(\vec{r}, t)\right) \\ &= \frac{1}{k!}\left[\prod_{i=1}^{k} Z_s\left(t_i\right)\right] \exp\left(-\int_0^T Z\left(t'\right) \, dt'\right). \end{aligned} \tag{2.3}$$

A set of 1024 simulated responses for each of the 32 stimuli is used for testing the information calculation procedures. The asymptotic estimate of transmitted information is calculated using 10^6 trials per stimulus.

2.2 Response Representation. The neuronal response to a stimulus as represented by the spike train is quantified by several variables. One is the number of spikes (NOS) in the response time interval, taken here to be 250 ms. The others are the projection of the spike train into the n principal components (PCs) (Richmond & Optican 1987; Golomb et al. 1994). We concentrate here on the first principal component (PC1) and the first three principal components (PC123). The four-dimensional code composed of the first three principal components and the number of spikes (PC123s) is also considered.

2.3 Information Estimation. We describe here briefly some of the methods that can be used for estimating information from neuronal responses. We simulate an experiment in which a set of S stimuli is presented at random. Each stimulus is shown N_s times; here N_s is the same for all the stimuli. The total number of visual stimuli presented is $N = SN_s$.

2.3.1 Summation over the Poisson Distribution. For each stimulus here, the number of spikes NOS is Poisson distributed with an average $\overline{\mathrm{NOS}} = \int_0^T Z_s(t)dt$. The asymptotic value of the transmitted information carried by the NOS can be calculated directly by summing over the distribution.

Equation 1.1 becomes

$$I(S;\mathrm{NOS}) = -\sum_{\mathrm{NOS}} P(\mathrm{NOS}) \log_2 P(\mathrm{NOS}) + \frac{1}{N_s} \sum_{s} \sum_{\mathrm{NOS}} P(\mathrm{NOS}|s) \log_2 P(\mathrm{NOS}|s) . \tag{2.4}$$

This sum is discrete and is calculated using the Poisson probability distribution $P(\mathrm{NOS}|s)$. The sum over NOS from 1 to ∞ is replaced by a sum from 1 to $\mathrm{NOS}_{\max} = 36$; taking a higher $\mathrm{NOS}_{\max}$ has only a negligible effect on the result. Using this method, the mutual information can be calculated exactly, but only when the firing rate distribution is known. In real experiments, the firing rate distribution is unknown, and therefore the methods described below should be used.

2.3.2 Straightforward Binning (Golomb et al. 1994). The principal components used here were calculated from the covariance matrix $C(t, t')$ formed over all responses in the set under study

$$C(t,t') = \frac{1}{SN_s} \sum_{s=1}^{S} \sum_{\mu=1}^{N_s} \left[\Lambda_{s,\mu}(t) - \bar{\Lambda}(t)\right]\left[\Lambda_{s,\mu}(t') - \bar{\Lambda}(t')\right], \tag{2.5}$$

where $\Lambda_{s,\mu}(t)$ is the μth realization of the response to the sth stimulus and $\bar{\Lambda}(t)$ is the average response over all the stimuli and realizations

$$\bar{\Lambda}(t) = \frac{1}{SN_s} \sum_{s=1}^{S} \sum_{\mu=1}^{N_s} \Lambda_{s,\mu}(t). \tag{2.6}$$

The eigenvalues of the matrix C are labeled according to a decreasing order; the corresponding eigenvectors are $\Phi_1(t), \Phi_2(t) \ldots$. The expansion coefficients of the neuronal response $\Lambda_{s,\mu}(t)$ are given by

$$a_{s,\mu,m} = \frac{1}{T} \int_0^T dt\, \Lambda_{s,\mu} \Phi_m(t). \tag{2.7}$$

Each response is then quantified using the coefficients to the first n principal components, and these are used as the response representation. The number n of coefficients used for quantifying the response is referred to here as the code dimension. The maximal and minimal values for each component are found, and the interval between the minimum and the maximum of the mth component is divided into $R(m)$ bins. The mutual information is calculated from the discrete distribution obtained. The N-dimensional response space is therefore divided into $R = \prod_{m=1}^{n} R(m)$ N-dimensional bins.

For PC123, we choose $R(1) = 36$, $R(2) = 20$, and $R(3) = 20$. The mutual information carried by the first principal component only, PC1, is calculated in a similar way with $R(1) = 36$. Note that binning is a simple form of regularization, and some kind of regularization is always needed when response measurements span a continuous range (the number of spikes is discrete). Regularization results in a downward bias of the calculated information value (see section 1). For PC1 and our simulated data set, using $R(1) = 36$ results in underestimating the mutual information by ~ 0.01 bit in comparison to $R(1) = 300$.

2.3.3 Binning with Finite Sampling Correction. The binning method is improved by doing the following:

1. For each dimension, equipopulated bins are used. For a one-dimensional code (NOS, PC1), the bin size varies across the response dimension, with nonequal spacing, so that each bin gets on average the same number of counts. For a three-dimensional code (PC123), the equipopulated binning is done for each dimension separately (Panzeri & Treves 1996).
2. The systematic error due to limited sampling can be expanded analytically in powers of $1/N$. Treves and Panzeri (1995) have shown that the first-order correction term C_1 carries almost all the error, provided that N is large enough. Therefore, we correct the information estimation by subtracting C_1 from the result calculated by raw equipopulated binning. The term C_1 is expressed as (Panzeri & Treves, 1996)

$$C_1 = \frac{1}{2N \ln 2} \left\{ \sum_s \widetilde{R}_s - R - (S-1) \right\}, \tag{2.8}$$

where $\widetilde{R}_s$ is the number of "relevant" response bins for the trials with stimulus s.

The number of relevant bins $\widetilde{R}_s$ differs from the total number of bins R allocated because some bins may never be occupied by responses to a particular stimulus. Thus, if the term C_1 is calculated using the total number of bins R for each stimulus, the systematic error is overestimated whenever there are stimuli that fail to elicit responses spanning the full response set. The number of relevant bins also differs from the number of bins actually occupied for each stimulus (with few trials), R_s, because more trials might have occupied additional bins. Again, C_1 is underestimated if R_s is used when only a few trials are available (the underestimation becoming negligible for $R/N_s \ll 1$ because R_s tends to coincide with $\widetilde{R}_s$ for all stimuli).

We estimate $\widetilde{R}_s$ (a number between R_s and R) from the data by assuming that the expectation value of the number of occupied bins should be precisely R_s given the number of trials available and the estimate $\widetilde{R}_s$. The

Table 1: Number of Bins R Used for Codes and Numbers of Trials N_s

N_s	16	32	64	128
R for NOS	16	$1 + \mathrm{NOS}_{max}$	$1 + \mathrm{NOS}_{max}$	$1 + \mathrm{NOS}_{max}$
R for PC1	16	36	63	128
$R = R(1) \times R(2) \times R(3)$ for PC123	$4 \times 2 \times 2$	$6 \times 3 \times 2$	$7 \times 3 \times 3$	$8 \times 4 \times 4$

algorithm for achieving this result is explained in Panzeri and Treves (1996). This method for correcting the information values in the face of a limited number of data samples yields reasonable results even up to $R/N_s \simeq 1$, the region in which we use the method for the study presented here.

Equation 2.8 depends on the probability distributions much more weakly than does the mutual information itself because the dependence is only through the parameters $\widetilde{R}_s$. Therefore, although the parameters $\widetilde{R}_s$ have to be estimated from the data, this procedure leads to better accuracy.

The choice of the number of bins $R(m)$ in each dimension for an experiment with S stimuli and N_s trials per stimuli remains somewhat arbitrary. Here we choose $R \sim N_s$, to be at the limit of the region where the correction procedure is expected to work, and thus still be able to control finite sampling, while minimizing the downward bias produced by binning into too few bins. For NOS, however, each response is just an integer ranging from 0 to the maximal number of spikes (NOS_{max}—25 for the parvocellular cell and 34 for the magnocellular cell), so even if we allocate more bins than this maximum, the extra ones will stay empty. For a multidimensional code (e.g., PC123), we allocate a number of bins $R(m)$ in the mth direction in relation to the amount of mutual information carried by this principal component alone, as shown in Table 1. When differences between different codes are calculated, we use the same number of bins in the relevant dimension. When PC1 and NOS are compared, we use the same number, R, as for PC1; in this case, many bins for NOS stay empty. For comparing PC123 and NOS, we use the same number of bins for NOS as for the first principal component, which is the richest in information among the three (e.g., 8 for $N_s = 128$). In this way we compare quantities calculated in a homogeneous way.

2.3.4 Neural Network. A two-layer network is trained by backpropagation to classify the neuron's responses according to the stimuli that elicited them (Hertz et al. 1992; Kjaer et al. 1994). The network uses sigmoidal activation for the nodes in the hidden layer and exponential activations for the nodes in the output layer, with the sum of the outputs normalized to one after each step. The input to the network is the quantified output of the biological neuron: the spike count, the first n principal components, or both. There is one output unit for each stimulus, and, after training,

the value of output unit number s is an estimate $\hat{P}(s|\mathbf{r})$ of the conditional probability $P(s \mid \mathbf{r})$. The mutual information is then calculated from these estimates using the second form of equation 1.1. The average over the response distribution is estimated by sampling over randomly chosen data points.

All other parameters of the algorithm—notably the number of training iterations and the input representation, are controlled by cross-validation. The data are divided into training and test sets, and for each representation, the training is stopped when the test error, defined as

$$E = -\sum_{\mu} \log_2 \hat{P}(s^{\mu}|\mathbf{r}^{\mu}) \tag{2.9}$$

(the negative log-likelihood or crossed-entropy), reaches a minimum. In equation 2.9, the index μ labels the trials in the test set, and s^{μ} is the stimulus that actually evoked the response $\mathbf{r}^{\mu}$ observed in that trial. After carrying out this procedure for four train-test data splits, the optimal response representation is chosen as the one with the lowest average test error. In the present calculations, as in previous work using the neuronal responses from neurons in the monkey visual system (Heller et al. 1995), the optimal representation consists of the spike count plus the first three principal components. Including the spike count leads to smaller crossed-entropy values than those obtained from the three PCs alone.

3 Results

We calculated the information carried about a set of 32 Walsh patterns by the simulated neuronal response quantified by the number of spikes I (S; NOS), the first principal component I (S; PC1), and the first three principal components I (S; PC123) (see Figure 1). For the network technique, we calculated also the information carried by the first three principal components and the spike count together I (S; PC123s) (see Figure 2). The arrows at the right side of the panels in Figure 1 represent the asymptotic values calculated from simple binning using 10^6 trials per stimulus. These figures show the estimated transmitted information for $N_s = 16, 32, 64, 128$, and for two sample cells: magnocellular and parvocellular. The parvocellular cell has sustained activity over an interval of 250 ms, whereas the magnocellular cell is active mostly over the first 100 ms. The magnocellular cell has more phasic responses (Golomb et al. 1994). Thus, we expect the multidimensional codes to capture a larger proportion of the information in its responses. A simple rate code is more likely to be an acceptable zeroth-order description of the parvocellular cell. The results we obtain (compare panels A and C, D and F, respectively, in Figure 1) bear this expectation out.

All of the calculations show that the first principal component is more informative about the stimulus than the spike count. As expected, this effect

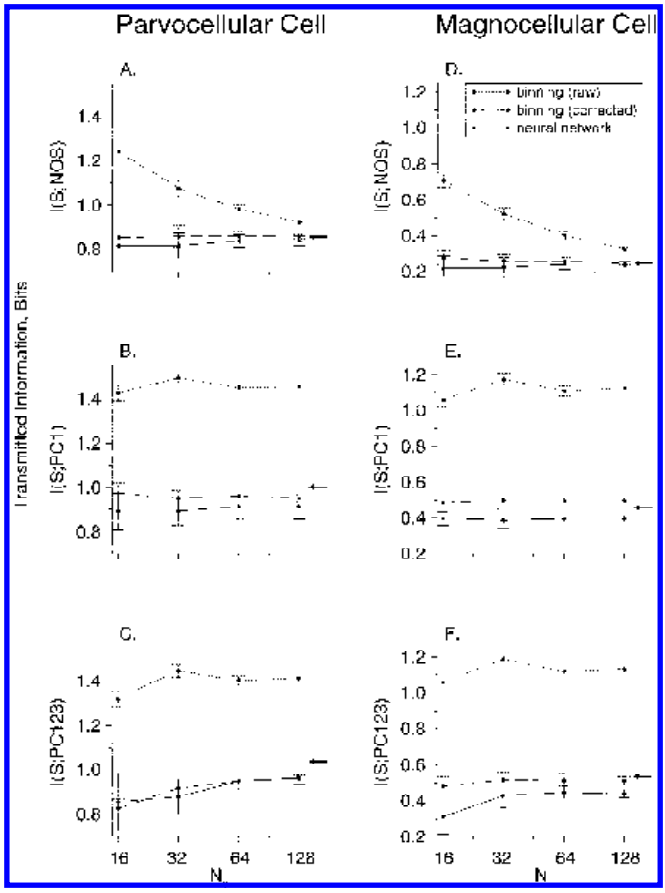

Figure 1: The information about a stimuli set of 32 Walsh figures conveyed by the neuronal response of model parvocellular (A–C) and magnocellular (D–F) cells, as estimated by various methods. The response is quantified by the number of spikes (A,D), the first principal components (B,E), and the first three principal components (C,F). The mutual information is estimated by straightforward equipopulated binning (dotted lines), equipopulated binning with finite sampling correction (dashed lines), and a neural network (solid line). The numbers of bins for each code and N_s are shown in Table 1. The neural network has six hidden units and a learning rate of 0.003. The arrows at the right indicate an asymptotic value (very good approximation for the "true" value) obtained with equispaced binning with $36 \times 20 \times 20$ bins and 10^6 trials per stimulus.

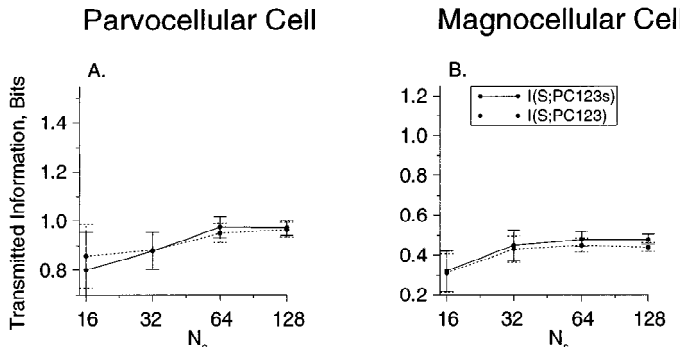

Figure 2: The information about a stimuli set of 32 Walsh figures conveyed by the neuronal response of model parvocellular (A) and magnocellular (B) cells. The response is quantified by the first three principal components with (solid line) and without (dotted line) the number of spikes. The mutual information is estimated by a neural network with six hidden units and a learning rate of 0.003. I (S; PC123) is shown also in Figures 1C and 1F.

is especially strong for the magnocellular cell, as the first PC weighting function suppresses contributions from the spikes after the first 100 ms, which are mainly noise.

Figure 1 shows that the raw binning is strongly biased upward. The difference between the estimates made with raw and corrected binning almost does not vary with N_s, for PC1 and PC123. This is because the first-order correction term (see equation 2.8) is approximately proportional to R/N_s, which we choose to keep roughly constant in our calculation. For NOS there is no point choosing R above $1+\mathrm{NOS}_{max}$; hence the correction term, and with it the raw estimate, decreases as N_s is increased.

Both the corrected binning method and the network method tend to underestimate the information in PC1 and PC123 (the only counterexample is shown in Figure 1E, where the binning method overestimated it). This is because both methods involve a regularization of the responses (explicit in the binning, and done implicitly by the network), and regularization always decreases the amount of information present in the raw response. For example, the corrected binning method underestimates the information whenever the number of bins is too small to capture important features of the probability distribution of the responses. Therefore, the effect is strong for PC123 when the number of bins in the direction of the first PC is not large enough, and the bias downward decreases with increasing N_s because the number of bins increases too. The underestimation does not occur with NOS

because the maximal number of spikes in our examples is around 30, and there is no meaning to using finer binning. In general, the underestimation due to regularization is more prominent for the higher-dimensional code (PC123) because the effect of adding more bins in each dimension is stronger when the number of bins is small.

The neural network tends to find a larger amount of information carried by the first three principal components together with the spike count than by the first three principal components alone (Heller et al. 1995). For the parvocellular cell, $I(S; \text{PC123s})$ exceeds $I(S; \text{PC123})$ by 0.025 bit (2.6%) for $N_s = 64$ and by 0.008 bit (0.8%) for $N_s = 128$; for the magnocellular cell, the extra information is 0.03 (7%) for $N_s = 64$ and 0.04 bit (9%) for $N_s = 128$; compare the solid lines in Figures 1 and 2. This increase occurs despite the high correlation between the first principal component and the number of spikes, especially for the parvocellular cell.

In Figure 3 we present the differences $I(S; \text{PC1}) - I(S; \text{NOS})$ between the information carried by PC1 and NOS, and $I(S; \text{PC123}) - I(S; \text{NOS})$, between PC123 and NOS. For all the cases considered here, the network yields a value for the extra information that is biased downward. This shows that the network automatically regularizes responses, and apparently the regularization is stronger for the higher-dimensional code. The corrected binning technique, on the other hand, gives both downward- and upward-biased values for the extra information—in this instance, downward for the parvocellular cell.

Since the choice of number of bins along each dimension is somewhat arbitrary for a multidimensional code, we checked the effect of using different binning schemes for $N_s = 128$. The results are summarized in Table 2A. The information differences for the parvocellular cells are in a 30% range; the information differences for the magnocellular cells are all nearly the same, no matter which method is used. Thus, even in the least favorable case, information differences using different binning schemes remain in the range of the remaining (downward) systematic error—about 30%. In a similar way, we varied the parameters of the network: the number of hidden units and the learning rate (see Table 2B). The difference in information varies within 10% for both cells, indicating that changes in these parameters are less important than the downward bias due to the regularization. The test error for the various network parameters is quite similar, with differences within 0.4% for both cell types. Thus, it is difficult to determine the best result of the network from just this number.

4 Discussion

The main result of this work is that information carried in the firing of a single neuron during a certain time interval about a set of stimuli, and quantified by a certain code, can be estimated with a reasonable accuracy (within 0.05–0.1 bit, or 10%) using either of the two procedures described

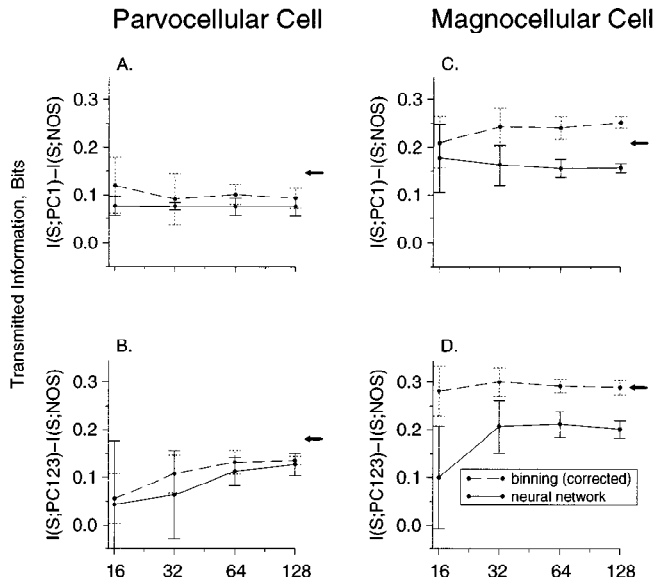

Figure 3: Differences between measures used for quantifying the response of a parvocellular cell (A,B) and a magnocellular cell (C,D). We show in A and C the difference between the information carried by the first PC and that carried by the number of spikes, and in B and D the difference between the information in the first three PCs and that in the number of spikes. Dashed lines indicate results obtained from normalized equipopulated binning; solid lines indicate those from the neural network.

In A and C, the number of bins is 16 for $N = 16$ and 36 for $N \geq 32$. In B and D the number of bins in the first three principal components is as given in Table 1, and the number of bins for NOS is equal to R(1) in that table, that is, to the number of bins for the first PC (4, 6, 7, and 8, respectively). The arrows at the right indicate an asymptotic value obtained with equispaced binning with $36 \times 20 \times 20$ bins and 10^6 trials per stimulus.

above. This result is achieved with the temporal response quantified by a three-dimensional code (PC123). With a one-dimensional code (NOS or PC1) the accuracy is even better. The relatively good accuracy obtained with our methods enables one to compare various codes and determine which one carries the most information (e.g., Heller et al. 1995).

To understand the results in more detail, it is important to remember that information measures are specific to the stimulus set considered, but also specific to the quantities chosen to quantify neuronal responses; also,

Table 2: Difference in Mutual Information $I(S; \text{PC123}) - I(S; \text{NOS})$ (Bits) for $N_s = 128$

A. Several Binning Schemes

$R = R(1) \times R(2) \times R(3)$	$7 \times 6 \times 3$	$8 \times 4 \times 4$	$10 \times 4 \times 3$	$15 \times 3 \times 3$
Parvocellular cell	0.145	0.135	0.125	0.114
Magnocellular cell	0.289	0.289	0.291	0.289

B. Several Network Schemes

Hidden units, learning rate	6, 0.0003	6, 0.001	10, 0.001	6 , 0.003
Parvocellular cell	0.121	0.127	0.123	0.118
Magnocellular cell	0.219	0.201	0.219	0.206

Note: The standard deviations of the difference are about 0.01.

information measures are affected by limited sampling, which results typically in an upward bias, but also, since it is always necessary to regularize continuous responses, they may be affected by the regularization, which results in a bias downward.

When a simple binning of the responses was used, raw information measures were strongly biased upward, and thus it was important to apply a correction for limited sampling. If one follows this procedure, the only parameter that has to be set is R, the number of response bins. If R is chosen too large, subtracting the term C_1, equation 2.8, will not be enough to correct for limited sampling (see also Treves & Panzeri 1995); if it is chosen too small, a strong regularization will be imposed, and information will be underestimated. The results indicate that it is sensible to set the number of response bins at roughly the number of trials per stimulus available, $R \simeq N_s$. C_1 is inversely proportional to the number of trials available and roughly directly proportional to the number of response bins, and this choice approximately balances the upward bias due to finite sampling with the downward one due to the regularization. Choosing the number of bins for each of the first three principal components in PC123 was more delicate that for NOS or PC1 and tended to yield a stronger downward bias in information values. This suggests that the use of the binning procedure alone becomes insufficient for higher-dimensional codes, as when the spike trains of several cells are considered together.[1]

[1] One useful procedure for the computation of information from multiple neurons is to use decoding to extract the relative probabilities of the stimuli from the responses and thus to reduce the original set of responses to the size of the stimulus set (Gochin et al. 1994; Rolls et al. 1996a, 1996b).

For all the cases considered, the network *underestimated* both the mutual information and the extra information in the temporal response in comparison to the number of spikes. This indicates that the regularization induced by the network is enough to dispose of the finite sampling bias, at the price of underestimating information values, especially for higher-dimensional codes, which are more strongly regularized.[2] Information values generally increase weakly with N, which indicates that the regularization induced has decreasing effects as N becomes large. Since the underestimation is stronger for $I(S;\mathrm{PC123})$ than in the case of unidimensional codes, estimates of the extra information in the second and third principal components (see Figure 3) are also biased downward. Several parameters need to be set when the network is used. Some (e.g., the number of iterations) can be set by cross-validation. Others (e.g., the learning rate) have little effect on the results across a broad range of values, as indicated in Table 2. Another virtue of the network procedure is that it effectively incorporates a decoding step and as such can be immediately applied to high-dimensional (e.g., multiple single-unit) data.

Although our results were obtained with simulated data, we believe that the conclusions apply to real data as well. In particular, our network results are consistent with those of Kjaer et al. (1994, Figure 10), obtained from complex V1 cells. They also found that the estimated mutual information increases with N_s and almost reaches a plateau for $N_s \approx 15$ for the first principal component but not for the first three principal components.

In this work, principal components are used for quantifying the data with a low-dimensional code, because the first principal components carry most of the difference among the responses to different stimuli. However, principal component analysis is not essential to our procedures for handling finite sampling problems; any kind of N-dimensional response extracted from the neuronal firing patterns can be used (e.g., PC123s, the NOS of several neurons, or neuronal spiking times; Heller et al. 1995). Estimating the accuracy of the methods can be done only for $n \leq 3$, because for higher n, calculating the asymptotic value of the transmitted information using straightforward binning is too consuming of computer time. Our procedures of finite sampling corrections were demonstrated here on stimuli with a sharp onset in time. They are also applicable, however, to continuously changing stimuli (after a suitable discretization), as long as the response to each stimulus is measured during a fixed time interval T, and stimuli are either discrete or have been discretized.

Important results have been obtained by Bialek and collaborators by extracting information measures from neuronal responses (Bialek 1991; Bialek et al. 1991; de Ruyter van Steveninck & Laughlin 1996; Rieke et al. 1993). These results are based on invertebrate (insect) data, which are usually easy

[2] One could even compute the corresponding C_1 term (Panzeri & Treves 1996).

to collect in large quantities, and thus the limited sampling artifacts are less severe. Our analysis is particularly relevant to primate data, which are typically much less abundant. Note, however, that a recent paper (Strong et al. 1996) analyzes finite sampling effects following the method of binning with finite size correction (Treves & Panzeri 1995). A second, less crucial difference between the type of experimental data considered here and that used by those investigators is that the stimulus in their case is a continuous quantity, which opens up the possibility of using notions such as the linearized limit and the gaussian approximation that do not apply to our situation with a discrete nonmetric set of stimuli and that when applicable can alleviate finite sampling effects further.

The network procedure needs a long computational time. A typical calculation for $N_s = 128$ and 32 stimuli runs for about 7 CPU hours on an SGI-ONYX computer. The binning-and-correcting technique is much faster, and most of the CPU time is taken up by sorting responses in order to construct equipopulated bins. For $I(S;\mathrm{NOS})$, which involves no sorting, a calculation with $N_s = 128$ runs for less than 1.8 seconds on an HP-Apollo computer.

What is the minimum number of trials to plan for in an experiment? In our instance, results seemed reasonable when $N_s = 32$. However, this number depends on the stimuli, response representation, and neurons used. The argument sketched above indicates how to estimate the minimum number of trials in other experiments. The binning-and-correcting procedure functions reasonably up to $N_s \simeq R$, and the minimum R that may, if the appropriate type of response is chosen, not throw away much information,[3] is $R = S$. Therefore a minimum of $N_s = S$ trials per stimulus is a fair demand to be made on the design of experiments from which information estimates are going to be derived.

Acknowledgments

AT and SP have developed their procedure using data from the labs of Edmund Rolls, Bruce McNaughton, and Gabriele Biella, all of which was extremely useful. Partial support was from CNR and INFN (Italy).

References

Abbott, L. F., Rolls, E. T., & Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex*, *6*, 498–505.

Bialek, W. (1991). Optimal signal processing in the nervous system. In W. Bialek (Ed.), *Princeton lectures on biophysics*. London: World Scientific.

[3] $R = S$ is the effective R used when applying a decoding procedure, as with multiunit data by Rolls et al. (1996b).

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*, 1854–1857.

Carlton, A. G. (1969). On the bias of information estimate. *Psych. Bull., 71*, 108–109.

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

de Ruyter van Steveninck, R. R., & Laughlin, S. B. (1996). The rates of information transfer at graded-potential synapses. *Nature, 379*, 642–645.

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

Gochin, P. M., Colombo, M., Dorfman, G. A., Gerstein, G. L., & Gross, C. G. (1994). Neural ensemble encoding in inferior temporal cortex. *J. Neurophysiol., 71*, 2325–2337.

Golomb, D., Kleinfeld, D., Reid, R. C., Shapley, R. M., & Shraiman, B. I. (1994). On temporal codes and the spatiotemporal response of neurons in the lateral geniculate nucleus. *J. Neurophysiol., 72*, 2990–3003.

Heller, J., Hertz, J. A., Kjaer, T. W., & Richmond, B. J. (1995). Information flow and temporal coding in primate pattern vision. *J. Comp. Neurosci., 2*, 175–193.

Hertz, J. A., Kjaer, T. W., Eskander, E. N., & Richmond, B. J. (1992). Measuring natural neural processing with artificial neural networks. *Int. J. Neural Syst., 3* (suppl.), 91–103.

Kjaer, T. W., Hertz, J. A., & Richmond, B. J. (1994). Decoding cortical neuronal signals: Networks models, information estimation and spatial tuning. *J. Comp. Neurosci., 1*, 109–139.

McClurkin, J. W., Optican, L. M., Richmond, B. J., & Gawne, T. J. (1991). Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science, 253*, 675–677.

Miller, G. A. (1955). Note on the bias of information estimates. *Info. Theory Psychol. Prob. Methods, II-B*, 95–100.

Optican, L. M., Gawne, T. J., Richmond, B. J., & Joseph, P. J. (1991). Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biol. Cybernet., 65*, 305–310.

Optican, L. M., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: II. Quantification of response waveform. *J. Neurophysiol., 57*, 147–161.

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

Reid, R. C., & Shapley, R. M. (1992). Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus. *Nature, 356*, 716–718.

Richmond, B. J., & Optican, L. M. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: III. Information transmission. *J. Neurophysiol.*, **57**, 162–178.

Rieke, F., Warland, D., & Bialek, W. (1993). Coding efficiency and information rates in sensory neurons. *Europhys. Lett., 22*, 151–156.

Rolls, E. T., Treves, A., Tovée, M. J., & Panzeri, S. (1996a). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. Submitted.

Rolls, E. T., Treves, A., & Tovée, M. J. (1996b). The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Experimental Brain Research*. In press.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1996). Entropy and information in neural spike trains. *Los Alamos archives cond-mat 9603127*. Submitted.

Tovée, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the response of single neurons in the primate temporal visual cortex. *J. Neurophysiol.*, *70*, 640–654.

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

Received February 5, 1996; accepted June 24, 1996.

This article has been cited by:

2. Hiroyuki Nakahara , Shun-ichi Amari . 2002. Information-Geometric Measure for Neural Spikes. *Neural Computation* **14**:10, 2269-2316. [Abstract] [PDF] [PDF Plus]
3. Inés Samengo . 2002. Information Loss in an Optimal Maximum Likelihood Decoding. *Neural Computation* **14**:4, 771-779. [Abstract] [PDF] [PDF Plus]
4. Stefano Panzeri, Alessandro Treves, Simon Schultz, Edmund T. Rolls. 1999. On Decoding the Responses of a Population of Neurons from Short Time Windows. *Neural Computation* **11**:7, 1553-1577. [Abstract] [PDF] [PDF Plus]
5. Edmund T. Rolls, Martin J. Tovée, Stefano Panzeri. 1999. The Neurophysiology of Backward Visual Masking: Information Analysis. *Journal of Cognitive Neuroscience* **11**:3, 300-311. [Abstract] [PDF] [PDF Plus]