Extracting information in spike time patterns with wavelets and information theory

Vítor Lopes-dos-Santos,1,2 Stefano Panzeri,3,4 Christoph Kayser,5,6 Mathew E. Diamond,7 and Rodrigo Quian Quiroga2

1Brain Institute, Federal University of Rio Grande do Norte, Natal, Brazil; 2Centre for Systems Neuroscience, University of Leicester, Leicester, United Kingdom; 3Center for Neuroscience and Cognitive Systems @UniTn, Istituto Italiano di Tecnologia, Rovereto, Italy; 4Max Planck Institute for Biological Cybernetics, Tübingen, Germany; 5Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom; 6Bernstein Center for Computational Neuroscience, Tübingen, Germany; and 7Tactile Perception and Learning Laboratory, International School for Advanced Studies, Trieste, Italy

Submitted 23 May 2014; accepted in final form 12 November 2014

Lopes-dos-Santos V, Panzeri S, Kayser C, Diamond ME, Quian Quiroga R. Extracting information in spike time patterns with wavelets and information theory. J Neurophysiol 113: 1015–1033, 2015. First published November 12, 2014; doi:10.1152/jn.00380.2014.—We present a new method to assess the significance of spike timing in a broad range of scenarios. We show that the Wavelet-Information (WI) method performs better and is more robust to spike time-jitter, background noise, and sample size than well-established approaches, such as principal component analysis, direct estimates of information from digitized spike trains, or a metric-based method. Furthermore, when applied to real spike trains from monkey auditory cortex and from rat barrel cortex, the WI method allows extracting larger amounts of spike timing information. Importantly, the fact that the WI method incorporates multiple time scales makes it robust to the choice of partly arbitrary parameters such as temporal resolution, response window length, number of response features considered, and the number of available trials. These results highlight the potential of the proposed method for accurate and objective assessments of how spike timing encodes information.

The importance of precise spike timing in carrying meaningful information has attracted much attention (Quian Quiroga and Panzeri 2009; Rieke et al. 1999). Does the temporal structure of spike trains provide information beyond the total spike count, or does it merely reflect noise? According to the “rate coding” view, neurons represent stimuli solely by the rate of firing within an encoding time window (Adrian and Zotterman 1926; Shadlen and Newsome 1994). In contrast, according to the “temporal coding” view, the time structure of the responses conveys additional information not provided by the total spike count (de Ruyter van Steveninck et al. 1997; Optican and Richmond 1987; Richmond and Optican 1987; Victor and Purpura 1996).

Experimental evidence accumulated over the last 3 decades has suggested that precise spike patterns, on the scale of milliseconds, do indeed convey information not available in rate codes (Arabzadeh et al. 2006; de Ruyter van Steveninck et al. 1997; Di Lorenzo et al. 2009; Eckhorn and Popel 1975; Foffani et al. 2009; Fontanini and Katz 2006; Kayser et al. 2010; Laurent et al. 1996; Montemurro et al. 2007; Panzeri et al. 2001, 2010; Quián Quiroga and Panzeri 2009; Richmond and Optican 1987; Victor 2000). For this, a straightforward way to assess the significance of spike timing has been to represent spike trains as sequences of “0s” and “1s,” denoting the absence or the presence of a spike in poststimulus time bins and then, using the formalism of information theory, evaluate whether the information about stimulus identity carried by such patterns is significantly larger than the information carried by spike counts alone (de Ruyter van Steveninck et al. 1997; Kayser et al. 2009; Panzeri et al. 2001; Strong et al. 1998). However, this approach leads to a combinatorial explosion (the “curse of dimensionality”) because the number of possible response patterns increases exponentially with the number of bins (Panzeri et al. 2007). Thus, for an experimentally feasible number of trials, this limits the precision of the temporal patterns to be studied (i.e., the size of the time bin) and the length of the response considered.

A solution to the combinatorial explosion problem is to reduce the dimensionality of the spike trains. To this end, a well-known approach is to compress the neural responses into a small number of features using principal component analysis (PCA). By using this method, Richmond and Optican (1987) showed that time patterns in responses from neurons in the macaque inferior temporal cortex could disambiguate visual stimuli that could not be distinguished by firing rate alone (Optican and Richmond 1987; Richmond and Optican 1987). Despite the value of this application, the PCA-based decomposition has two main caveats. First, PCA represents directions of maximum variance, which are not necessarily the directions with the largest information. Second, PCA coefficients are not localized in time and may not capture sources of information that are precisely localized at one or a few poststimulus time...
points (Panzeri et al. 2001) or even encoded at multiple temporal scales (Fotowat et al. 2011; Harvey et al. 2013; Kayser et al. 2009; Panzeri et al. 2010). Here, we propose another type of dimensionality reduction that is able to capture time-localized information encoded at multiple time scales (Fig. 1). The method combines wavelet decomposition and information theory to identify first the features in the spike patterns carrying relevant information and then use these features to quantify the amount of sensory encoding carried by these responses using a decoding approach. We validate the method on simulated spike trains and compare its performance with that obtained with PCA, direct estimations of information from digitized neuronal responses, and a widely used metric-space (MS) method (Victor and Purpura 1996). Results on simulated data demonstrate that the Wavelet-Information (WI) method is more robust and extracts more spike timing information than previous methods for a wide range of background firing rates and intertrial jitters. The advantages of the WI method are confirmed by evaluating its performance with experimental data from the monkey auditory cortex and the rat somatosensory cortex. Additionally, we show that the same approach can be used: 1) to denoise spike trains, providing a more robust quantification of the stimulus selectivity; and 2) to assess and visualize the information carried by the synchronous firing of neurons.

MATERIALS AND METHODS

Wavelet Decomposition

The wavelet transform is the inner product of a signal with dilated and translated versions of a wavelet function (Mallat 2008; Strang and Nguyen 1996). Formally, given a signal \( x(t) \) and a wavelet function \( \psi_{a,b}(t) \), the continuous wavelet transform (CWT) is defined as:

\[
W_{\psi}X(a, b) = \langle x, \psi_{a,b} \rangle, \quad \text{with} \quad \psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right),
\]

where \( a, b \in \mathbb{R} \) are the scale and translation parameters, respectively. The translation parameter changes the location of the wavelet function, whereas the scale parameter dilates or compresses it. The correlation of the signal \( x(t) \) with the dilated (contracted) versions of the wavelet \( \psi_{a,b}(t) \) gives the low- (high-) frequency components. The CWT is redundant, and, without any loss of information, it is practical to define the wavelet transform only at discrete scales \( a_j = 2^j \) and times \( b_{j,k} = 2^j k \), which is called the dyadic wavelet transform (DWT).

The DWT is nonredundant in the sense that from \( N \) data points we obtain \( N \) wavelet coefficients, each of them representing the amount of activity of the original signal at a specific time and scale. Furthermore, patterns in the signal with different frequency and time localizations are represented by specific wavelet coefficients. The DWT can be computed using a hierarchical and very efficient algorithm called multiresolution decomposition (Mallat 1999). This algorithm successively divides the signal into coarse approximations and details at different scales. The end result is the decomposition of the original signal into a series of detail scales and a final approximation, corresponding to the time-localized activity in different frequency bands.

Starting from the binned spike trains, in this study we implemented a five-scale dyadic wavelet decomposition using Haar wavelets, which is a square function that is ideally suited to identify local contrasts at different scales. The spike trains were always binned with 1-ms windows unless stated otherwise.

Selection of Wavelet Coefficients

From the total set of wavelet coefficients, equal to the number of bins in the spike trains, we selected a subset of coefficients based on their mutual information with the stimuli, defined as (Shannon 1948):

\[
I_{S,w} = \sum_{S,w} P(S, w) \log_2 \frac{P(S, w)}{P(S)P(w)},
\]

where \( P(S) \) denotes the probability of the stimulus \( S \), and \( P(w) \) denotes the probability of the wavelet coefficient \( w \), and \( P(S, w) \) denotes the joint probability of the stimulus \( S \) and the wavelet coefficient \( w \). The total spike count (left) or the informative wavelet coefficients (right) as inputs to the decoder.
where $S$ is the set of stimuli and $w_c$ is the set of values of a wavelet coefficient $c$. The significance of the information given by each coefficient $I_{wc}$ was established based on surrogate testing: for each coefficient, we calculated a distribution of information values obtained by shuffling trials (i.e., randomizing trial-stimulus relations) 20 times. Surrogate distributions were calculated separately for each decomposition level (information values obtained from coefficients of the same levels were combined), and the 95th percentiles of each distribution were used as statistical thresholds (horizontal dashed lines in Fig. 1C). To avoid having too many features with significant information, if $>25$ coefficients were significant, we used the 25 with the largest unbiased information (the same restriction was applied to the other methods). In this context, unbiased information was defined as the difference between the direct measure of information and its corresponding statistical threshold computed from shuffling the trials. Additionally, if none of the coefficients crossed the statistical threshold, we used the 2 with the largest information.

**Stimulus Decoding and Information Estimation from Confusion Matrices**

To estimate the information in a set of features, we used a cross-validated (leave-1-out) naive Bayesian decoder to assign the response on each trial in the testing set to a given stimulus, which gives a lower bound of the information available in the spike trains (Quian Quiroga and Panzeri 2009). Decoding performance was computed as the proportion of correct predictions. The decoder optimization and the selection of features were based solely on the training trials. For comparison, we also used linear discriminant analysis (Fisher 1936; Quiroga et al. 2007) and nearest neighbors classifiers (in which case, features were assigned $z$-scores to avoid scaling problems) and obtained virtually the same results.

In cases where the linear decoder introduced systematic errors (Fig. 7), we computed the mutual information between the actual and the predicted stimuli from the confusion matrices:

$$I(S, S') = \sum_{S, S'} P(S, S') \log \frac{P(S, S')}{P(S)P(S')}$$

where $S$ is the set of actual stimuli presented to the decoder and $S'$ is the set of predicted stimuli by the decoder. To correct for the upward limited-sampling bias in the information estimate, we used the quadratic extrapolation procedure described elsewhere (Panzeri et al. 2001; Strong et al. 1998) and implemented in Magri et al. (2009).

**MS Method**

We also compared the WI method with the MS approach (Victor and Purpura 1996), which clusters responses based on a distance metric between spike trains. This distance is defined as the minimum "cost" of converting a spike train into another one by deleting, inserting, or moving spikes. The cost of deleting or inserting a spike is increased, more weight is given to the precise timing of the spikes. The cost of deleting or inserting a spike defines the temporal precision of the analysis. With the MS method, we classified trials using a nearest neighbor decoder; more specifically, we assigned each tested trial to the class of its nearest neighbor in the training set. We systematically varied $q$ from 0.001 to 0.524/ms in half-octave intervals. Only representative $q$-values are reported. For computing the spike train distances, we used a MATLAB function available at http://www-users.med.cornell.edu/~jdvicto/spkdm.html.

**Spice Train Denoising**

To visualize spike patterns containing information, we adapted the WI method to denoise the spike trains by: 1) computing the wavelet decomposition of the mean peristimulus time histogram (PSTH) of each stimulus; 2) denoising the mean PSTHs by reconstructing them using only the wavelet coefficients with significant information; 3) setting to 0 the denoised PSTHs values below a threshold at 1 SD of the absolute values (taken from the denoised PSTHs of all stimuli); and 4) using the denoised and thresholded PSTHs as masks, preserving only spikes in bins passing this threshold. This effectively preserves spikes conveying stimulus information and deletes the others. We remark that although the selection of informative wavelet coefficients was the same for all stimuli, the masks were different for each stimulus.

**Estimation of Information in Correlated Spike Patterns**

To estimate the information given by the correlated firing of pairs of neurons $(i,j)$, we computed the wavelet decomposition for each trial $n$ and calculated the normalized distance between the values of a corresponding wavelet coefficient (i.e., considering a specific time location and scale) in both neurons as:

$$D_{ij}^n = \left| \frac{w_{i,c,n} - w_{j,c,n}}{w_{i,c,n} + w_{j,c,n}} \right|,$$

where $w_{i,c,n}$ is the value of wavelet coefficient $c$ of neuron $i$ at trial $n$. Analogous to the procedure described in Fig. 1C, we then selected the distances $D_{ij}^n$ that had significant information about the stimuli and used these distance values for decoding. In other words, we implemented the same procedure as before but using the distances $D_{ij}^n$ between the wavelet coefficients of each neuron instead of the value of the coefficients themselves.

**Experimental Data**

All procedures with monkeys reported here were approved by the local authorities (Regierungspräsidium Tübingen) and were in full compliance with the guidelines of the European Community (EUVD 86/609/EEC). The animals were socially (group-) housed in an enriched environment under daily veterinary supervision. Experiments with rats were conducted in accordance with National Institutes of Health and international standards for the care and use of animals in research and were approved by the International School for Advanced Studies Ethics Committee. Procedures were supervised by a consulting veterinarian.

**Monkey A1 data.** As described in a previous work (Kayser et al. 2010), neural activity was recorded from caudal auditory cortex (mainly areas A1 and caudal belt) of three alert animals using multiple microelectrodes. The data were high-pass filtered (4 Hz), amplified (AlphaOmega, Nazareth, Israel), and digitalized at 20.83 kHz. Recordings were performed in a dark and anechoic booth while animals passively listened to acoustic stimuli. The sound stimulus consisted of a 40-s sequence of pseudorandom tones ("random chords"). This sequence was generated by presenting multiple tones (125-ms duration) in different sequences (12 fixed-frequency bins per octave) with each tone frequency appearing (independently of the others) with an exponentially distributed intertone interval (range 30–1,000 ms, median 250 ms). To estimate spectrograms of the acoustic stimulus, the signal was convolved with complex Morlet wavelets with central frequencies ranging from 20 to 1,600 Hz. Then, the instantaneous amplitude of each frequency was computed as the norm of the complex values. The $z$-scored instantaneous amplitudes were used for computing spike-triggered averages (STAs) for reverse correlation.

**Rat primary somatosensory cortex (SI) data.** As described in previous works (Lebedev et al. 2000; Panzeri et al. 2001), recordings...
 Innovative Methodology

in the somatosensory cortex of adult Wistar rats were performed with an array of six tungsten microelectrodes. Neurons in barrel columns C1–3, D1–3, and E1–3 were recorded while their corresponding whiskers were stimulated individually. The stimulus was an up-down step function of 80-μm amplitude and 100-ms duration delivered 48 times for each vibrissa with a 1-s interstimulus interval. Neuronal activity was amplified and band-pass filtered in the range 300–7,500 Hz. Spike waveforms were digitized at 25 kHz (Discovery; DataWave Technologies, Boulder, CO).

RESULTS

Outline of the WI Framework

The first step of the WI method is to convolve the spike train responses (to repeated presentations of a set of stimuli; Fig. 1A) with Haar wavelets (Fig. 1B; see MATERIALS AND METHODS). Thus each trial is decomposed into a set of wavelet coefficients representing local spike patterns at different time scales. To identify wavelets carrying meaningful information, we then compute the mutual information between each coefficient and the stimuli (Fig. 1C). Information values are compared with distributions constructed by stimulus label permutations (see MATERIALS AND METHODS). Then, the wavelet coefficients with significant information are used to represent the data. We used a decoding approach to quantify the performance of the WI method (and other methods for comparison) in extracting stimulus information. Figure 1D shows confusion matrices of naïve Bayesian decoders trained to classify responses with time patterns as in Fig. 1A with either spike counts or the selected wavelet coefficients. As expected, by construction, the performance with the WI method clearly outperformed the one obtained with spike counts.

Performance with Simulated Data

We used simulated data to quantify the performance of the WI method and compared it with other approaches. The simulated data consisted of a set of 200-ms responses to four hypothetical stimuli created with a two-step procedure: 1) a specific spike time pattern (a sequence of predefined spike times) was assigned to each stimulus and inserted into the response of a given trial with a random shift within a window centered at their original time; and 2) background activity was generated independently for each trial following a Poisson process with a given mean rate and then added to these patterns.

The example in Fig. 2A illustrates the ability of the WI method to extract information at different time scales. A relatively precise spike timing distinguishes the first two stimuli (at the bottom; generated using a jitter of 0.5 ms), and a pattern at a coarser scale distinguishes the remaining two (at the top; generated using a jitter of 8 ms). The mean background rate was eight spikes per second. Figure 2B, right, shows the outcome of the WI method using a fine (1 ms) binning of the data. The mean decoding performance was close to perfect (0.975), thus indicating the ability of the method to capture time patterns at different scales. To understand this result further, in Fig. 2C we show the decoding outcomes obtained when considering coefficients of each wavelet scale separately. The coefficients from the coarser scales (scales 4 and 5 and last approximation) distinguished the coarse time patterns of stimuli 3 and 4, and the high-frequency coefficients (scales 1–3) distinguished the more precise patterns of stimuli 1 and 2. For scale 1, no coefficient crossed the statistical threshold, and we therefore used the two with the largest information.

For comparison with another dimensionality reduction method, we applied the PCA-based developed by Richmond and Optican (1987). For this, we computed the principal components (PCs) from the spike trains binned with either 1- or 8-ms windows (vertical lines in Fig. 2A). As with the WI method, we estimated the time-pattern information by decoding stimuli based on the scores of the four PCs with the largest variance. As shown in Fig. 2B, neither the 1- nor 8-ms bins could capture the information at both time scales. In particular, with the 1-ms binning, the decoder could distinguish between stimuli 1 and 2, but not 3 and 4, given that the pattern of these 2 stimuli were scattered across several bins. Likewise, when using the 8-ms bins, the decoder could distinguish between stimuli 3 and 4, but this gridding was too coarse to distinguish between stimuli 1 and 2.

To test the WI method in scenarios mimicking different recording conditions, we generated three examples (Fig. 3) including patterns with different time localizations, precision, and complexity, and we systematically varied the background firing rate (2–64 Hz) and the time jitter (2–64 ms). Left panels show simulations with low background rate and jitter (4 Hz and 2 ms, respectively), whereas right panels show simulations of the same patterns but with larger baseline firing and jitter values (16 Hz and 8 ms).

We further implemented a similar PCA-based approach but selecting the PCs with the highest information. Also, we calculated performance using the whole binned responses (i.e., with no reduction of dimensionality). We then compared these time pattern strategies to a total spike count decoding, simply summing the total number of spikes of each trial. The selection of features was always performed on a set of trials used to train the decoder (training set), and then performance was evaluated in a different set of trials (test set).

Figure 4A displays results for different jitters and background rates for example 1 of Fig. 3 (using 15 trials per stimulus for training and 20 trials for testing). The results displayed are the averages of 20 simulations for each combination of parameters. As expected, there was an overall decay of performance when increasing the background firing rate and jitter due to the increasing difficulty in extracting time patterns. Still, the WI method provided the best decoding accuracy in nearly all cases except when very large jitters were used (of the order of or larger than the time patterns themselves). In this case, all information in the time patterns was destroyed, and only spike count carried information. Similar results were obtained for examples 2 and 3.

We then investigated how the number of trials used for training the decoder and for selecting the set of response features used for decoding affected performance. To do so, we repeated the analysis of Fig. 4A but systematically varied the training set size (from 5 to 65 trials per stimulus in 5-trial steps). We used 20 simulations for each method and each set of parameters (jitter and background rate). Figure 4B reports results for all 3 examples averaging across all baseline firing and jitter values of Fig. 4A. The performance of the WI, PC coefficients chosen based on information (PCinfo), and no-reduction method increased monotonically with the number of trials used for training. In contrast to the PCinfo and no-

J Neurophysiol • doi:10.1152/jn.00380.2014 • www.jn.org
Fig. 2. Decoding time patterns at different time scales. A: simulated responses to 4 stimuli. Marks represent spikes at a given time (horizontal axis) in a given trial (vertical axis). Twenty trials per stimulus are displayed (dashed lines separate trials of different stimuli), and only the 1st 152 ms of each trial are displayed for clearer visualization. Vertical axes denote the size of the bins used (light gray: 1-ms; dark gray: 8-ms bins). B: confusion matrices of naïve Bayesian decoders trained with principal components (PCs) extracted from responses binned with 1- or 8-ms bins (left and middle, respectively) and based on informative wavelet coefficients (right). Decoding performances are shown in white. PCA, PC analysis. C: confusion matrices obtained when using wavelet coefficients from specific decomposition levels as labeled.
reduction methods, the WI method reached a value close to its maximum performance within \( \frac{1}{2} \) trials, stressing its robustness to undersampling. To evaluate the efficiency of each method in reducing the dimensionality of the responses without loss of information, we computed the performance of the naïve Bayesian classifiers as a function of the number of response features used for classification. For this, we ranked the features either by variance, in the case of PCs (PCAvar), or by information, in the case of wavelets and PCinfo. Additionally, we performed a similar analysis by ranking the response time bins by their amount of information and then selecting only the \( n \) most informative ones (referred hereafter as binned responses). Figure 5A reports the results for each of the 3 examples of Fig. 3 when using a training set of 15 trials per stimulus. Whereas the WI method needed 10 or fewer response features to reach maximal performance, all other methods needed a larger number of response features, which also varied substantially across examples. Thus the performance for these latter methods was very sensitive to the number of response features used. In sum, the WI method reduced the dimensionality of responses in a more efficient and robust way.

Interestingly, the performance using PCs with the largest variance (PCvar) had a much steeper increase of information.

Fig. 3. Results for different examples, background firing, and jitter. The panels show 3 examples of simulated responses to 4 different stimuli. Ten trials per stimulus are displayed. The background firing was modeled as a Poisson process with mean rate of 4 (left panels) or 16 spikes per second (right panels). Stimulus-specific time patterns were added to the background firing with a 2- (left panels) or 8-ms (right panels) jitter.
with the number of features compared with the performance obtained with PCinfo. This seemingly counterintuitive result can be attributed to the low number of trials (15 per stimulus) used for training in this case, which gave a relatively poor estimation of information carried by each PC. To verify this, we ran the same analysis as in Fig. 5A but using 50 trials per stimulus (Fig. 5B). As expected, with the larger training set, PCinfo showed a clear increase in performance and was much more efficient than PCvar for small numbers of features. Consistent with the results shown in Fig. 4B, increasing the number of trials had little impact on the performance of the WI method, thus highlighting its robustness to sample size.

Performance with Data from the Monkey Auditory Cortex

Single neuron recordings were performed in primary auditory (A1) cortex in response to a 40-s-long sequence of pseudorandom tones (see MATERIALS AND METHODS for details). We divided the sequence into 500-ms time intervals and denoted each time interval as a different discrete stimulus. For this data set, we trained naïve Bayesian decoders to predict
In total, 34 responsive neurons (with >1-Hz mean rate) recorded in 12 sessions were included in this analysis. Each session consisted of 50–60 presentations of the stimuli, which we separated into 2 nonoverlapping sets of training and test trials.

For each neuron, we first evaluated the performance of the various methods with a time resolution of 1 ms using 15 trials for training. Figure 6A shows the decoding performance of each neuron using the different approaches described above (y-axis) against the performance achieved with wavelets (x-axis). Note that the WI method outperformed the other methods for virtually all neurons.

Figure 6B shows decoding performance vs. training set size and reveals that the WI method outperformed the other methods for virtually all neurons. We then quantified the impact of temporal precision used to quantify the neural responses. Figure 6C shows that decoding performance was maximal when using a bin size of about 5–10 ms (and more toward 10 ms for the PCA-based methods). This result is comparable with the optimal resolution previously reported for these data using a direct information estimate (Kayser et al. 2010). In all cases, the performance decreased as the bin size increased, meaning that larger bin sizes missed relevant information arising from the precise pattern of firing of these neurons. Moreover, the performance of all methods also decreased for bin sizes smaller than 5–10 ms. This arises because the outcome of increasing temporal precision is the trade-off of two opposing effects. On one hand, a finer resolution leads to potentially higher information content in the neural responses. Because of the data processing inequality (Quiroga and Panzeri 2009), increasing resolution can only increase or leave invariant the information available in the responses. On the other hand, a finer resolution increases the dimensionality of the responses, thus making it more difficult for a decoder to extract the available information. A drop in decoding performance when increasing the resolution thus means that from that resolution onward the additional infor-
information available at finer resolutions is not sufficient to overcome the added difficulty in decoding many extra and weakly informative dimensions. Thus the ability to extract more information at finer resolutions with WI compared with other methods is due to the optimal dimensionality reduction implemented in the WI method. For example, with a bin size of 1 ms (i.e., an order of magnitude increase of the dimensionality of the response space used for decoding), the performance with wavelets decreased only $\frac{1}{110}$ with respect to the 5- to 10-ms resolution, whereas decreases of $\frac{1}{110}$ and $\frac{1}{50}$ were observed for the no-reduction and the PCA-based methods, respectively.

Next, we evaluated the ability of all methods to extract information from populations of simultaneously recorded cells. For this, we used the data of 10 (out of 12) sessions where 2 or more responsive neurons were recorded simultaneously. For each session, all possible combinations of a varying number of cells were used. In this case, we did not set a minimum number of wavelet coefficients or PCs for each neuron; i.e., if a neuron had no significantly informative features, then this neuron would provide no features to the decoder. Results (Fig. 6D) show that the decoding performance with wavelets was significantly larger than the one achieved with the other methods.

Performance with Data from the Rat Barrel Cortex

To evaluate the potential of the WI method to extract time-localized information optimally, we analyzed neuronal responses in the rat barrel cortex. In these data, a precise onset time is given by the time of whisker stimulation. We used a naïve Bayesian decoder, as before, to classify which vibrissa (from the set of C1–3, D1–3, and E1–3) was stimulated in each trial. Forty-eight trials were available for each vibrissa, and we randomly assigned half of these trials for training and the other half for testing. Figure 7A shows the responses of a representative neuron. Responses were binned in 1-ms windows, as before, and we considered a 200-ms response window starting at stimulus onset. Figure 7B compares the WI method with the other approaches for 84 responsive neurons (with $>1$ Hz mean firing rate). The WI method again outperformed all other methods for the large majority of the cells. Figure 7C, left, displays the decoding performance as a function of the poststimulus window (i.e.,...
a window of 200 ms means taking the whole response shown in Fig. 7A). Consistent with the previous results, the WI method gave the best results for nearly all time windows considered. In contrast to the auditory data above (Fig. 6), in this case the no-reduction approach did not give a good performance because only a fraction of the bins provided relevant information. Moreover, whereas for wavelets the performance kept increasing while increasing the poststimulus window, for the no-reduction and the PCA-based methods, the performance decayed or remained at the same level. This again indicates that these methods do not perform as well as wavelets when the dimensionality of the response...
increases. We also observe that for these data the spike count decoding provided better results than both PCA-based (and the no-reduction) approaches. However, this was in part due to systematic errors in the decoder. For example, the PCAvar decoder classified D1 responses as D1, D2, or D3 (but still ruling out the possibility that other vibrissae were stimulated). To verify this, we computed the information between the predicted and the actual stimuli from the confusion matrices shown in Fig. 7C, right (see MATERIALS AND METHODS for details). In this case, PCAvar provided more information than the spike counts. As before, in nearly all cases, the wavelet-based decoding provided the best results.

Figure 7D shows both the decoding performance (left) and the information extracted from the confusion matrices (right) as a function of the bin size used. The peak of performance and information was at a bin size of ~25 ms for all methods except wavelets. In particular, all methods showed decay in information for larger bin sizes because with larger bin sizes the information given by precise time patterns is lost. Interestingly, however, all methods except wavelets showed also decay in performance (and information) for bin sizes <25 ms. As with the monkey data, this is due to the increase in the response space dimensionality accompanying the increase in resolution. With these same data, a previous work (Panzeri et al. 2001) reported an optimal bin size <25 ms using a method analogous to the no-reduction shown here, i.e., calculating the mutual information from the binned responses. However, in that case, a high-dimensional space was avoided by considering a much smaller response window of 20 ms. In this regard, the advantage of wavelets is crucial whenever the optimal response window is not known a priori.

Finally, we studied the information carried by populations of neurons. For this, we assumed that all neurons were recorded simultaneously (an approach that does not take into consideration the effects of correlations) and repeated the procedure used in Fig. 6C. Since 100 neurons were available, we averaged across 30 randomly chosen combinations for each number of neurons. Additionally, to avoid an excessive number of features, we only allowed a maximum of 5 wavelet coefficients or PCs per neuron (instead of 25 as before). Results are shown in Fig. 7E. The performance achieved with wavelets surpassed the 1 achieved with the other methods. Except for the no-reduction and PCinfo cases (due to the above-mentioned limitation in dealing with high-dimensional response spaces), the performance increased monotonically with the number of neurons.

Comparison with the MS Method

Next, we compared the WI method with the MS approach for different q-values (see MATERIALS AND METHODS) with both the simulated and real data reported above. Figure 8A shows the results of such comparison with the same 20 simulations illustrated in Fig. 2A using different background firing rates. Note that WI is more robust to increases of noise levels for this example. Figure 8B shows the results for the simulations shown in Fig. 4A (to save space, results with 32- and 64-ms jitter are not shown since in these cases the performance with all methods was close to chance), and Fig. 8B the result for the 3 examples presented in Fig. 4B. Altogether, we observe an overall better performance with the WI method proposed here. Note also that the performance of the MS method is dependent on the choice of the parameter q. For instance, a q-value of 0.128 gave the best performance with a jitter of 16 ms but also the worst performance with a jitter of 2 ms. Conversely, a q-value of 0.362 gave the best performance for the 2-ms jitter case but a relatively poor performance for larger jitters. In this respect, the advantage of the WI method is that it does not require the tuning of any parameter and automatically gives a performance that in most cases surpassed the one obtained with the MS method even when choosing the optimal q-value.

Figure 8C displays results as function of training set size (as in Fig. 4B). For these examples, a q-value of 0.128 gave the best overall results for the MS method. These results were, however, not as good as the ones obtained with the WI method. We note that the MS method was remarkably robust to undersampling as can be seen in the results of example 2 where it outperformed wavelets when <10 trials per stimulus was used for training, likely due to the fact that this low number of trials was not sufficient for a good selection of wavelet coefficients.

Results of the comparison between the WI and MS method for the real data are shown in Fig. 9. In particular, Fig. 9A (left panels) displays the results obtained for the A1 monkey neurons (as in Fig. 6) using four representative q-values for the MS method. For this data set, the average performance obtained with the WI method was significantly better than that obtained with the MS method for all q-values (paired t-test, P < 0.05 in all cases; see Fig. 9A, right). Figure 9B shows the results obtained for the barrel cortex neurons. As with the monkey data, performance with the WI method was significantly better than that obtained with the MS approach for all q-values (paired t-test, P < 10^-23 in all cases; see Fig. 9B, right). The relatively poor performance of the MS method for the rat data set is likely due to the compact time localization of the informative spikes.

Denosing Spike Patterns

We next investigated whether the WI method could be used to denoise single-trial spike trains. Denoising entails, in brief, identifying spikes correlated with the informative wavelet

Fig. 7. Time patterns in rat barrel cortex neurons. A: responses of a representative neuron to mechanical stimulation of different whiskers (C1–3, D1–3, and E1–3). Each dot denotes a spike in a given trial (vertical axis) at a given time (horizontal axis). Forty-eight trials were recorded for each whisker. Stimuli were delivered at 0 ms and lasted 100 ms, which corresponded to the OFF response. B: decoding performance for 84 responsive neurons with wavelets (x-axis) compared with spike count, PCainfo, PCAvar, and no dimensionality reduction, respectively. Most entries are below the diagonal (dashed line), thus indicating a better performance with wavelets. C, left: mean decoding performance for different methods as a function of the poststimulus time window. Right: mean information for the different methods estimated from the confusion matrices (see main text for details). Solid lines denote averages, and shaded areas SE. D, left: decoding performance for different methods as function of bin size. Right: stimulus information for different methods estimated from confusion matrices as function of bin size. E: mean decoding performance for the different methods as a function of number of neurons used. Solid lines show averages across a set of 30 randomly chosen neurons (except for 100 neurons, where only 1 combination is possible), and shaded areas denote SE barely visible due to the very small error). As before, note the better performance of wavelets for nearly all poststimulus time windows and any number of neurons considered.
coefficients and discarding the remaining, noninformative ones (see MATERIALS AND METHODS). Figure 10A shows the denoising of the time pattern of example 3 (Fig. 3) with a jitter of 4 ms and a background firing rate of 64 Hz. We observe that a large amount of the background noninformative spikes were removed, and the remaining spike rasters after denoising (Fig. 10A, bottom left) were very similar to the spike patterns embedded in the data (see Fig. 3). A similar reduction of background “noisy” activity is evident when comparing the PSTHs before and after denoising (Fig. 10A, right panels).

To quantify these observations, we repeated this procedure varying systematically the jitter and background firing rate (using 15 trials per stimulus). Figure 10B shows the number of errors obtained by the denoising procedure for different jitters as a function of background rate. Here, we defined errors as the sum of false positives (not deleting a spike corresponding to background activity) and false negatives (incorrectly discarding a spike that belonged to the informative time pattern). For comparison, we also calculated the number of errors obtained when thresholding the original PSTHs but without a prior wavelet denoising. This was done to assess whether denoising could be achieved by a simple PSTH thresholding. Furthermore, to show that results were not just due to the smaller number of spikes obtained after denoising, we also calculated the number of errors obtained when randomly erasing the same number of spikes. In general, the wavelet-based denoising approach gave the lowest number of errors. Results obtained for the other examples of Fig. 3 were similar.

Next, we applied this denoising approach to the spikes of a representative neuron taken from the auditory data set. Figure 11A

Fig. 8. Comparison of the WI and metric-space (MS) methods with simulated data. A: performance of WI and MS (for 3 different values of \( q \) as labeled) methods for multiscale responses of Fig. 2A with varying background rates. Shown are mean and SE across 20 simulations for each noise level. B: performance of WI and MS for the same examples in Fig. 4A. For space reasons, we do not show results with 32- and 64-ms jitters (performance was close to chance levels in all of these cases). C: same as Fig. 4B comparing the WI and MS approaches.

Innovative Methodology
NEW METHOD TO EXTRACT INFORMATION IN SPIKE TIME PATTERNS
J Neurophysiol • doi:10.1152/jn.00380.2014 • www.jn.org
shows the raster plots before (top) and after (bottom) denoising, where it is clear that time patterns (spikes consistent across trials for each stimulus) are better visualized after denoising.

Given this encouraging result, we then examined whether wavelet denoising could lead to obtaining cleaner and sharper STAs of the stimulus. STAs are commonly used representations to assess which stimulus features (out of the many in a complex dynamic stimulus) drive the responses of the neurons (Dayan and Abbott 2005). For this, we used a time-frequency representation of the acoustic stimulus (see MATERIALS AND METHODS) and computed the STA in the form of the spectrotemporal receptive field, i.e., the average frequency spectrum of the stimulus around the time of spiking. To test the effectiveness of denoising in removing the detrimental effect of
nonstimulus-driven spikes, we first computed the STA for the original response, then added background Poisson noise with a mean firing rate of 100 Hz, and computed the STA in this noisy condition. Finally, we denoised the data with the Poisson noise and recomputed the STA. Results are shown in Fig. 11B. The original STA showed a clear tuning to stimuli with energy in the 200- to 300-Hz frequency range, ~50 ms preceding the spike occurrences. This stimulus selectivity was dramatically

Fig. 10. Denoising time patterns with simulated data. A, left: a 100-trial realization (25 trials per stimulus) of example 3 from Fig. 2 before (top) and after (bottom) denoising. Right: peristimulus time histograms (PSTHs) for each of the 4 stimuli before (top) and after (bottom) denoising. B: number of errors (see main text) after denoising (in red) obtained for example 3 with different jitters (same values used in Fig. 3) as function of background firing rate. Results are averages across 40 simulations (error bars denote SE but in most cases are too small to be noticeable). For comparison, results obtained when thresholding the original PSTHs (without prior wavelet denoising; see main text for details) are shown in green. Results obtained by randomly erasing the same number of spikes as with wavelet denoising are displayed in black. In most cases, the lowest number of errors was obtained after wavelet denoising.
We ran several simulations like the one illustrated in Fig. 12A, left, varying systematically the duration of the synchrony window and adding time jitters in the coactivation. We used 20 trials per stimulus for training the decoders and 20 for testing. Figure 12B shows the decoding performances for different jitters as function of the length of the synchrony duration. Clearly, the WI method outperformed the coincidence count for nearly all window sizes and jitters. The overall decoding performance for all methods is shown in Fig. 12C.

We then extended the spike train denoising procedure described above to visualize better the informative patterns of synchronous firing. For this, we used the distances of the selected wavelets. The mask for a given stimulus (analogous to the denoised PSTHs in the previous case) was constructed by averaging the distances of each wavelet coefficient across trials and then adding together the mean distance of each coefficient multiplied by its time support (with a value of 1 within the time range spanned by the wavelet and 0 elsewhere). Figure 12A, right panels, displays the raster plots of the neurons shown on the left panels after denoising where it is clear that the coincident activations are highlighted and noninformative spikes deleted.

**DISCUSSION**

A key problem in assessing the contribution of precise time patterns to sensory coding is the high-dimensionality of the data sets. Although several statistical methods have been developed to correct for sampling biases arising in these cases (Montemurro et al. 2007; Nemenman et al. 2004; Paninski 2003; Panzeri et al. 2007; Panzeri and Treves 1996; Strong et al. 1998), these methods are still of limited value when dealing with long response spaces with high temporal precision. Other approaches to tackle this issue have been proposed, for example, based on simplifying the structure of interactions among possible information-carrying symbols with minimum information loss (Ganmor et al. 2011; Panzeri and Schultz 2001; Shew et al. 2011), developing binless estimations (Victor and Purpura 2002) or defining spike train distances to quantify information (Victor and Purpura 1996).
not match the dimensions with relevant information. Here, following the general idea of dealing with high-dimensional spaces by implementing a dimensionality reduction that captures relevant information, we proposed a new computational approach to assess information carried by time patterns in single and multiple neurons. This approach is based on: 1) extracting features of the spike trains with the wavelet transform; 2) a dimensionality reduction by which a subset of...
wavelet coefficients are selected using information theory; and 3) a quantification of time-pattern information by using a decoding approach.

With both simulated and real data, we demonstrated a robust performance of the WI method in capturing meaningful information in the spike trains without committing to specific assumptions about the time scales at which information is encoded and even capturing information in localized patterns at multiple time scales. This feature is of utmost importance considering that recent studies have shown that neural responses carry complementary information at a number of different time scales, ranging from millisecond-precise spike patterns to slow rate variations or slow network oscillations on the scale of hundreds of milliseconds (Bullock 1997; Fairhall et al. 2001; Lisman 2005; Nadasy 2009; Panzeri et al. 2010; Victor 2000). However, it has been challenging to characterize how these neural responses work together to represent information because most spike train analysis methods are committed from the beginning by the specific choice of an optimal time scale.

The performance of the WI method was minimally affected by increases in the dimensionality of the responses (obtained by increasing the time resolution, the length of the response considered, or the number of neurons) compared with other methods. In fact, compared with methods like PCA, using the whole (binned) response space (i.e., without any dimensionality reduction) or using spike counts, the WI method was able to extract more information from the spike trains, as quantified by decoding performance. In addition, the information obtained with wavelets was more robust to varying degrees of background activity and jitter in the precise timing of the spikes. For the cortical data analyzed here, high information values were found with PCA but only when the optimal time scale (i.e., response length and resolution bin) was considered. Crucially, information values obtained with wavelets were much less sensitive to the choice of the time scale used to study the neural responses. In our view, these advantages arise because the convolution with Haar functions implemented with the DWT identifies local contrasts at different time locations and at different time scales. In other words, the wavelet transform offers a time-resolved, multiscale representation that automatically and efficiently represents time patterns of different lengths and resolutions, appearing at different times. The advantages of wavelets for spike train analysis reported here are in line with the reported advantages of wavelets for spike sorting (Quiroga 2012; Quiroga et al. 2004) and for denoising evoked potentials (Ahmadi and Quian Quiroga 2013).

We also compared the WI method with the MS method (Victor and Purpura 1996), a widely used metric-based approach to estimate information in spike trains. We found that overall the WI method performed better and more robustly than the MS method. In particular, MS results varied substantially depending on the choice of \( q \) (defining the weight given to precise timing vs. number of spikes), and no \( q \)-value gave good results in all conditions, i.e., for different firing rates and jitters. Therefore, it may not be possible to find a single \( q \)-value that is suitable for different neurons in a data set (with different firing rates, time pattern resolution, degrees of information, etc.). On the contrary, WI is parameter-free and performed well in all tested cases. With the real data, the performance of the MS method was lower than WI, likely because of the compact time localization of the informative spikes, something that is captured by the WI method when doing the selection of informative wavelet coefficients. Another aspect of practical importance is that the MS method took significantly longer to compute than WI. For the monkey data, whereas the WI method required consistently between 100 and 150 s to compute the results for each neuron, the MS method processing time was strongly correlated with the mean rate of the responses: computations for neurons firing \( \approx 2.5 \) Hz on average took \( \approx 200 \) s or more per \( q \)-value, whereas computations for neurons firing \( \approx 5 \) Hz took \( \approx 400 \) s. Since, as in previous works (Roussin et al. 2012; Victor and Purpura 1996), results were calculated for dozens of different \( q \)-values, computations with the MS method took about two orders of magnitude longer compared with wavelets.

A key strength of the WI approach is its data robustness. In particular, the WI method required fewer trials to achieve optimal results and needed fewer features to represent the relevant information in the spike trains compared with other methods. This efficient compression of the responses mitigates the curse of dimensionality and allows the analysis of larger responses, the use of higher resolutions, and the possibility of population coding analyses where features of several neurons are considered together. This represents a significant advance compared with the other dimensionality reduction-based approaches examined in our study, which in some cases tended to underestimate the time resolution or the amount of information in the spike trains because actual increases in information were counterbalanced by the limitations of these methods to deal with higher-dimensional responses.

It is of interest to discuss briefly how the data robustness of the WI method (due to its highly efficient reduction of dimensionality) relates to the biases in extracting information from neural responses often discussed in the literature (Panzeri et al. 2007). First, the success of the WI method in effectively compressing the responses to a very small number of informative dimensions leads to a strong reduction in the downward bias in decoding performance, which is given by a limited number of data for training the decoder (Jacobs et al. 2009; Quiñon Quiroga and Panzeri 2009). This is demonstrated by the success of the WI method to extract high information values even for very fine temporal resolutions in real data and the nearly optimal performance obtained with relatively few trials with the simulated data. As a rule of thumb, 10 or more trials per stimulus were found to be enough to avoid a major downward bias problem (see Fig. 4B). Second, the higher information values achieved with the WI method allow reducing (and correcting more efficiently) the upward bias in estimation of information from the confusion matrix due to the limited number of experimentally available test data. As shown in Panzeri and Treves (1996), this upward bias is roughly proportional to the number of different stimuli that are predicted by the decoder when a given stimulus is presented. As a rule of thumb, the corrections for the upward bias in the information calculation from the confusion matrix work well if the number of different stimuli that are predicted by the decoder is two to four times smaller than the number of trials per stimulus (Panzeri et al. 2007). The WI method gives less decoding errors and therefore a smaller upward bias in confusion matrix information calculations.
Innovative Methodology

NEW METHOD TO EXTRACT INFORMATION IN SPIKE TIME PATTERNS

We stress that with the WI method we do not estimate directly the mutual information contained in the stimulus response probabilities in the form of either binary words, as in Strong et al. (1998), or PC scores, as in Optican and Richmond (1987). Instead, we limited ourselves to the calculation of cross-validated confusion matrices. These estimations give a lower bound to the information available in the data (Quiroga and Panzeri 2009). However, it is in principle possible to extend the WI method to perform also direct calculations of information from neural activity (Strong et al. 1998). It is well-known that such direct computations of information from binary response words are in practice possible only for short response windows due to the curse of dimensionality (Kayser et al. 2009). In this respect, the WI method may be further developed to extend the applicability of direct calculations of information to longer windows by providing a relatively low-dimensional representation of the binary word response, thus providing more data-robust, direct calculations of information. Such implementations are beyond the scope of the present paper but are ripe for future work.

We have also shown that the WI approach can be adapted to denoise spike trains and to estimate and visualize correlations across neurons. In particular, we found an improved performance in estimating information in the correlated firing of neurons compared with a MS approach (Victor and Purpura 1996) or the standard technique of analyzing coincidence counts, what has been called “unitary events” (Grun et al. 2002). As in the case of other methods to assess time-pattern information, the caveat of the unitary event approach is that it is highly dependent on the time resolution used to bin the spike trains. For instance, coactivations can be missed if the time scale at which they occur does not match the window used for binning the spiking activity (Lopes-dos-Santos et al. 2013). Likewise, results with the MS approach rely on the choice of an optimal $q$-value. The advantage of wavelets in this respect is the fact that it allows evaluating correlated firing at different time scales and for specific time localizations.

Finally, the possibility of denoising spike trains allows a much clearer visualization of informative time patterns and a better characterization of the selectivity of the neuron (through reverse correlations) as we illustrated with data from monkey auditory cortex. The spike train denoising based on WI could in principle be used for a number of other novel applications. For example, it can be used to characterize better the relationships between spiking activity and local field potentials (particularly their phase at selected frequency bands) carrying out information about the stimuli. The better visualization of the stimulus-driven spikes can be also useful to study relationships between stimulus-driven and “internal state” components of neural activity, a topic of current interest in systems neuroscience (Harris and Thiele 2011). In general, the more accurate characterization of the tuning of the neurons offered by WI denoising will likely offer important practical advantages for the discovery of the “thesaurus” that relates sensory stimulus to neural responses and could further facilitate the understanding of what features encoded by the neurons do contribute to behavior.

A MATLAB implementation of the WI is available from http://www.le.ac.uk/csn/WI.

GRANTS

V. Lopes-dos-Santos was funded by the Science Without Borders program from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Foundation, Ministry of Education of Brazil. This work was supported by grants from the Engineering and Physical Sciences Research Council and the Human Frontier Science Program. S. Panzeri acknowledges support from the SI-CODE European Union Future Emerging Technology (FET)-Open FP7-284533 Project and from the Autonomous Province of Trento (Call “Grandi Progetti 2012,” Project “Characterizing and Improving Brain Mechanisms of Attention-ATTEND”).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS


REFERENCES


Lisman J. The theta/gamma discrete phase code occurring during the hippocampal phase precession may be a more general brain coding scheme. Hippocampus 15: 913–922, 2005.


