Chapter 2

Extracting Information From Spike Trains

A fundamental task of any information theoretic analysis of the neural code is to estimate the mutual information (MI) that neural responses convey about a set of stimuli. This estimation task is then used as a building block for more advanced questions such as *"What aspects of the stimuli do neurons code?"* or *"How do neurons interact to transmit information together?"*.

This information estimation task involves both methodological aspects - the degree of accuracy and robustness of the estimation, and scientific implications - identifying the components of spike trains that carry the information, and the stimulus components about which neurons are informative. These two aspects are the focus of the current chapter.

The current chapter therefore focuses on the methodology of extracting information from spike trains, and as a by-product, characterizes the relative importance of certain components of the neural code. This is achieved by comparing different MI estimation methods, each focusing on different aspects of neural activity. The chapter is organized as follows. The next section introduces the motivation for dimensionality reduction for MI estimation. Section 2.2 discusses the issue of estimating the joint distribution of stimuli and responses, as well as estimating the MI encapsulated in this joint distribution from finite samples. Then, section 2.3 systematically reviews and applies a series of dimensionality reduction methods to our data that focus on various aspects of spike trains. The performance of these methods is compared in section 2.4, together with a discussion of the results.

2.1 Preliminaries

The challenge of obtaining a reliable MI estimation

Estimating mutual information from empirical distributions is a difficult task, in particular with the small sample sizes of typical electrophysiological data. A naive approach to this problem would be to estimate the joint distribution of stimuli vs. all possible neural responses, and then to estimate the mutual information of this high-dimensional distribution. Unfortunately this approach is almost always bound to fail due to the potential richness of neural responses. For example, a typical pyramidal neuron in the cortex fires spikes that should be measured with a relevant temporal resolution of 1-4 milliseconds [Singer and Gray, 1995], and can thus produce in theory at least 2^{250} different spike trains in a single second. Since a robust estimation of a probability density function requires obtaining many samples relative to the number of possible responses (see e.g. [Devroye and Lugosi, 2001]), this approach is doomed to fail[1].

The crucial observation is that MI estimation does not in fact require estimating the full joint distribution of stimuli and responses. There are two important reasons for this. First, the set of functionally distinct neural responses is much smaller. Many spike trains are considered equivalent by the physiological decoding mechanisms. This is caused by the noisy nature and bounded complexity of neural decoders, and should allow us to reduce the complexity of our statistical decoding procedures. Secondly, the MI is a scalar function of the distribution, that actually averages the log likelihood ratio $\log \frac{p(x,y)}{p(x)p(y)}$ over all x's and y's. Its estimation is therefore expected to be more robust than the estimation of the distribution itself [Nemenman *et al.*, 2002], even though the log function in principle requires estimating an infinite number of moments [Paninski, 2003].

[1]Except when the neural responses are limited to a relatively small typical set, and very stable recordings can be made(e.g. in the visual system of the fly [Bialek *et al.*, 1991, deRuyter van Steveninck *et al.*, 1997]).

The estimation of MI from a finite sample involves an important **trade-off between model complexity and the reliability of estimation**. To understand this issue we may view the MI estimation task in the context of classical supervised learning, as a problem of estimating a nonlinear (scalar) function of empirical data in a way that resembles nonlinear regression. In supervised learning, there is a widely discussed tradeoff between the complexity of the models used for learning and the resulting generalization error (see e.g. [Vapnik, 1995]). This tradeoff emerges since during learning, complex models get tuned to spurious structures in the data that do not reflect true regularities but rather finite sample artifacts[2]. There is extensive literature that tries to quantify correct complexity measures, and use them to build optimally-complex models for a given size of empirical data (see e.g. [Rissanen, 1978] and chap. 7 in [Hastie *et al.*, 2001]).

As an example of this effect in the MI estimation problem, consider a simple non parametric model for the joint distribution of a discrete stimulus set S and a response set R that consists of a list of probabilities to see a stimulus and response pair (s, r). In this model, for any finite sample size n, the reliability of the density estimation $\hat{p}(s, r)$ drops with the dimension of the joint probability matrix $|S| \times |R|$. Consequently, more reliable MI estimates can be obtained if instead of estimating the joint distribution $\hat{p}(s, r)$, one looks at low dimensional functions $T(R)$ of the responses R, and estimates the distribution of $\hat{p}(s, T(r))$. On the other hand, as explained in detail in the next section, such low dimension functions tend to reduce the mutual information $I(T(R); S)$

The challenge in MI estimation is therefore to find low complexity representations of spike trains that are still highly informative. This makes it possible to obtain both a high level and a reliable estimation of the information they convey. We therefore turn to describe the effect of dimensionality reduction on the mutual information.

[2]This tradeoff is sometimes called the *bias variance* tradeoff, since complex models are more prone to over fitting which increases the variance of the learning machine, and oversimplified models lead to consistent deviations from the real values that the learning machine has to learn. This should not be confused with the bias and variance of the MI estimator for matrices, that we discussed in details in the next sections.

Dimensionality reduction and data processing inequality

The effects of projecting our data to simpler representations are formally analyzed using the *Data processing inequality.* This states that any such dimensionality reduction $T(R)$ is bound to reduce the mutual information between stimuli and responses. More formally,

Lemma 2.1.1 : Data processing inequality
If $X \to Y \to Z$ form a Markov chain (X and Z are independent given Y), then $I(X;Y) \geq I(X;Z)$.

Proof: The mutual information $I(X;Y,Z)$ can be written in two ways

$$I(X;Z) + I(X;Y|Z) = I(X;Y,Z) = I(X;Y) + I(X;Z|Y) \tag{2.1}$$

Since X and Y are conditionally independent given Y we have $I(X;Z|Y) = 0$. From the positivity of the information $I(X;Y|Z) \geq 0$ we have $I(X;Y) \geq I(X;Z)$. $\square$

Corollary 2.1.2 : *For a discrete set of stimuli S, a discrete set of neural responses R and a function of the responses $T(R)$*

$$I(S;R) \geq I(S;T(R)) \ . \tag{2.2}$$

Proof: $S \to R \to T(R)$ form a Markov chain, since $T(R)$ is a function of R alone. $\square$

Since projecting the data is bound to reduce the information, we would prefer projections that maximally preserve information, since these yield better estimates of the true MI. Therefore, the goal is to find functions $T(R)$ over the responses R that maximize the mutual information with the stimuli S

$$\max_T I(S;T(R)) \quad . \tag{2.3}$$

As an example, let $R \in \{0,1\}^{100}$ be a binary string that represents the occurrence of spikes during a time window of one hundred ms at a 1-ms resolution, and let $T : \{0,1\}^{100} \to \mathcal{N}$ be the spike count during this window, which in practice takes values between 0 and 100. As another example, $T(R) : \{0,1\}^{100} \to \{r_1, .., r_{10}\}$ can map each spike train to one representative spike train r_i to which it is the most similar. These two examples represent two distinct types of dimensionality reduction approaches. The first is the projection of the spike train to a low dimensional (often scalar) statistic. The second exploits the fact that the typical set of neural responses is limited and does not span the whole space of possible responses. Its density can therefore be well estimated in the more densely populated regions of the response space space.

In practice, another complicating factor must be considered. We can only **estimate** the joint probability $\hat{p}(S, R)$ and thus cannot calculate the true information $I(S; R)$, but rather are limited to its estimate $\hat{I}(S; R)$. In this case it is no longer true for every estimation method of I that $\hat{I}(S; T(R)) \leq \hat{I}(S; R)$ or that $\hat{I}(S; T(R)) \leq I(S; R)$. Thus even though we seek functions T that maximize the estimated information $\max_T \hat{I}(S; T(R))$, it is necessary to avoid overfitting of T which leads to overestimation of $\hat{I}(S; T(R))$. These considerations are discussed in Section 2.2.

Sufficient statistics

A common approach to modeling neural responses is to use a parametric model whose parameters are stimulus dependent. For example, spike trains are often modeled as Poisson processes, whose underlying rates are determined by the stimulus. In such a model the following relation holds

$$S \to \theta \to R \tag{2.4}$$

where S are the stimuli, θ are the parameters (e.g. the rate) and R are the neural responses (e.g. spike trains). Although we are interested in $I(S; R)$, this MI is bounded from above by $I(\theta; R)$. When the mapping between the stimulus and parameter is reliable, that is, the information loss in $S \to \theta$ is small, we have $I(\theta; R) \approx I(S; R)$. We therefore wish to find ways to reduce the dimensionality of the responses R, using some simple statistics of the spike trains, while maintaining $I(\theta; R)$ as large as possible.

The theoretical basis for choosing such statistics lies in the notion of *sufficient statistics* [Fisher, 1922, Degroot, 1989] and its application to point processes [Kingman, 1993]. Consider the case where we are given a sample $R^n = \{r_1, ..., r_n\}$ from a known parametric distribution $f(R|\theta)$ (these can be for example spikes in a train whose rate is θ). A sufficient statistic is a function of the sample $T(r_1, ..., r_n)$, that obeys

$$Pr(R^n|\theta, T(R^n)) = Pr(R^n|T(R^n)). \tag{2.5}$$

Therefore, given the sufficient statistic T, the probability of observing the sample is independent of the distributions parameter's θ. In other words, the sufficient statistic summarizes all the information about θ that exists in the sample. Indeed if T is a sufficient statistic then

Lemma 2.1.3 : *$T(R^n)$ is a sufficient statistic for the parameter θ if and only if it achieves an equality in the data processing inequality*

$$I(R^n; \theta) = I(T(R^n); \theta). \tag{2.6}$$

Proof: Consider two opposite weak inequalities. First, note that T is a function of X^n and therefore independent of θ given X^n. Therefore the following Markov relation holds $\theta \to X^n \to T$, and according to the information inequality $I(X^n; \theta) \geq I(T; \theta)$. Conversely, because T is a sufficient statistic, X^n is independent of θ given T and therefore the following Markov relation holds $\theta \to T \to X^n$ and consequently $I(X^n; \theta) \leq I(T; \theta)$. Together with the first inequality this requires $I(X^n; \theta) = I(T; \theta)$ which completes the proof. $\square$

How are these notions used for estimating the information in spike trains? If spike trains can be accurately described given a parametric model with stimulus dependent parameters, using their sufficient statistic allows us to reduce the dimensionality of the responses R while preserving the information it carries about θ and hence about S. Therefore, if T is a sufficient statistic of the neuronal responses R we only need to estimate $I(S; T(R))$ instead of the more difficult problem of estimating $I(S; R)$. When we cannot find a low dimensional statistic which is sufficient, we aim to find statistics that largely preserve the information about the parameters.

Summary

Reliable estimation of the MI between stimuli and responses requires reducing the dimensionality of the responses, by considering statistics of the responses. Such an operation reduces the information in the responses, unless the data can be accurately described using a parametric model and the statistics used are sufficient. The goal is therefore to reduce the dimensionality of spike trains while preserving maximal information about the stimuli.

A road map

In practice, a plethora of techniques have been developed in the literature to achieve dimensionality reduction of neural responses, each focusing on a different aspect of the spike trains. Applying these methods involves two interconnected issues, which are the subject of the current section.

- **What aspects of the spike trains should we look at?**
  Different aspects of spike trains may carry different information about the stimuli, and may reach different overall MI levels.

- **How do we estimate the MI carried by a specific aspect of the spike train?**
  The task here is to develop estimators that are non biased and reliable. MI estimation is commonly based on two steps:

  - First, estimating the joint distribution of stimuli and reduced spike trains $p(S, T(R))$. This is often done by binning the responses $T(R)$, but binless estimators have also been developed.
  - Secondly, estimating the MI of this distribution. The bias and variance of MI estimators based on binned density estimations are discussed in section 2.2.1. Section 2.2.3 discusses binless MI estimates.

These issues are inter-dependent. On one hand, choosing the aspect of the spike trains we are interested in may affect the methods we choose to estimate MI. For example, the statistics we are interested in may be continuous (as with first spike latency), ordinal but discrete (as with spike counts), or even non ordinal (as with spike patterns represented as binary words). Each of these may allow different estimation methods. On the other hand, the effectiveness of estimation methods affects the statistics we choose to use.

To simplify the structure of the current chapter we start the discussion with a family of estimators that is based on simple statistics of the spike trains. These include, for example, spike counts and first spike latencies. We describe MI estimators based on these statistics that use a binning procedure for density estimation and discuss the bias and variance properties of these estimators, as well as binless MI estimators (section 2.2). We then turn to review a series of methods developed for spike train dimensionality reduction (2.3). Finally the results of applying these methods to our auditory datasets are described in section 2.4.

2.2 Methods I: Estimating MI from empirical distributions

In this section we discuss the case where a simple (usually one dimensional) statistic of the spike train is used to represent neural responses, and its joint distribution with the stimuli is estimated. The discussion of this scenario generalizes over several possible statistics that will be discussed in the next section.

2.2.1 Density estimation using binning procedures

The common method for estimating the distribution of a random variable is to discretize its values with some predefined resolution and calculate the histogram, or empirical counts. Similarly, to estimate the joint distribution of the stimulus S and some statistics of the responses $T(R)$ the corresponding contingency table is calculated from the count $n(S = s, T(R) = t)$. Then, the empirical distribution $\hat{p}(s,t) = \frac{n(s,t)}{n}$ can be used to calculate an estimator of the MI

$$\hat{I}(S;T(R)) = D_{KL}[\hat{p}(S,T)||\hat{p}(S)\hat{p}(T)] \tag{2.7}$$

This approach can be used for any statistic, and we focus here on the example of spike counts.

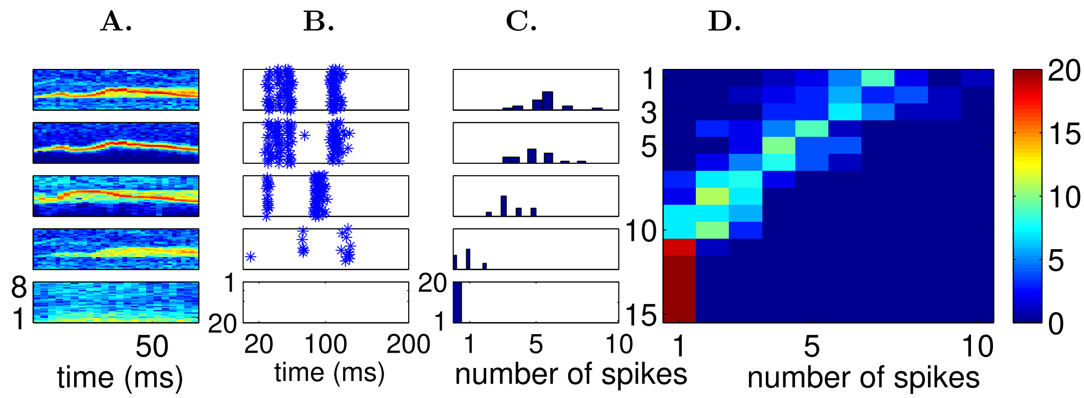

Figure 2.1: An illustrative example of estimating the mutual information in spike counts using naive binning. **A.** Spectrogram of five stimuli. **B.** Raster plots of neural activity in response to 20 presentations of each of the five stimuli. **C.** Distribution of spike counts for each stimulus. **D.** Joint distribution of 15 stimuli and spike counts. The five stimuli on the left correspond to rows 1,3,5,10 and 15.

Figure 2.1 demonstrates this method for estimating the MI carried by the spike counts of a single MGB cell. In this experiment, 15 stimuli were presented twenty times each. For purposes of demonstration, the spectrograms of five of these stimuli are plotted (2.1A) together with the raster plots of the responses they elicited in the cell (2.1B). Figure 2.1C plots the distribution of spike counts following the presentation of each of the stimuli. The distribution of counts for all 15 stimuli is plotted in Fig. 2.1D, where the stimuli were ordered by decreasing average spike count. This joint distribution suggests that there is a strong relation between the identity of the presented stimulus and the distribution of spike counts.

When using bins to estimate the density, the complexity-generalization tradeoff discussed above can easily be illustrated. When the number of bins is small, different R values are merged into a single bin, reducing the resolution in the representation of R and causing a loss of information (thus increasing the deviation[3] of $\hat{I}$ from the true MI). When the number of bins is large, the number of samples in each bin decreases, leading to a more variable estimation of the probability in each bin, and correspondingly, increasing the variance of the estimator $\hat{I}$.

[3] In the bias-variance tradeoff formulation, this deviation is referred to as the bias.

Unified Bins

Input:
A joint count $n(x, y)$.

Output:
I, An estimation of the MI in $n(x, y)$

Initialization:
$i = 0$
$n_i(x, y) \leftarrow n(x, y)$

Main loop

repeat
$i = i + 1$
calculate $I_i = I[n_i(x, y)]$, (bias corrected)
find column or row with the smallest marginal
unite it with its neighbor with smallest marginal, yielding $n_{i+1}(x, y)$
until (#rows$<$ 2 or #columns$<$ 2)

$I = \max_i(I_i)$

Figure 2.2: Pseudo-code of the "unified bins" procedure. $I[n(x, y)]$ is the naive mutual information estimator calculated over the empirical distribution $\hat{p}(x, y) = \frac{1}{n}n(x, y)$, and corrected for bias using the method of [Panzeri, 1995].

Instead of choosing the bins linearly, a better estimator of the distribution can be obtained by choosing the bins in a data dependent manner, such that the distribution across the bins is as homogeneous as possible ([Degroot, 1989] chap 9). For discrete variables this can be achieved by starting with a large number of bins, and then iteratively unifying the bin with the smallest probability to its neighbor with the smallest probability. The pseudo code of this procedure, that we call *Unified bins*, appears in Fig. 2.2.

Figure 2.3 plots the mutual information obtained using a linear binning method and the above binning method for spike counts in three brain regions. The number of bins in the naive method was enumerated over for each cell separately. All the MI estimates were corrected for bias by the method of Treves (1995).

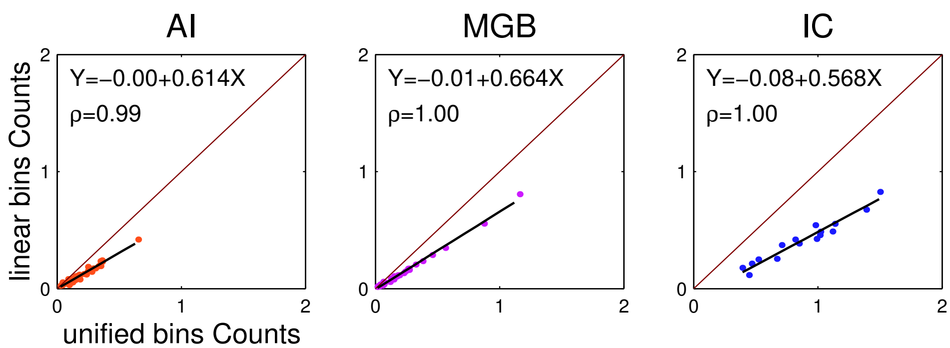

Figure 2.3: Comparison between naive (linear) binning and data dependent binning that operates to preserve homogeneous marginals. Both methods use spike counts. Each point corresponds to a different cell. The red line is the $y = x$ curve. The black line is the regression curve (adjusted for sample size), whose equation is printed within each sub-plot. ρ is the correlation coefficient.

These results show that using the above adaptive binning procedures succeeds in extracting around 50 percent more information from spike counts than with nonadaptive bins. Similar comparisons for other statistics, such as the first spike latency, also yielded higher information with "unified-bins". On the other hand, we tested this procedure using simulations with synthetic data and found that it does not overestimate the MI due to overfitting [Nelken *et al.*, 2003]. In the remainder of this work we therefore use the adaptive binning procedure "unified-bins" for estimating MI in binned joint distributions.

2.2.2 MI estimators based on binned density estimation

After deciding on a binning procedure for estimating the density, we are in the following situation: Given an empirical joint count matrix $n(r,s)$, that was created according to some unknown distribution p^*, we wish to calculate the mutual information of the underlying distribution $I[p^*]$. The problem is of course that we only observe the empirical sample $n(r,s)$. Typically, MI is estimated by calculating MI of the empirical distribution observed in the joint count.

$$\hat{I}(R,S) = \sum_{s \in S} \sum_{r \in R} \hat{p}(r,s) \log \left(\frac{\hat{p}(r,s)}{\hat{p}(r)\hat{p}(s)} \right) \tag{2.8}$$

where $\hat{p}(r,s) = \frac{1}{n} n(r,s)$ is the empirical distribution of the pairs (r,s), and $\hat{p}(r)$ and $\hat{p}(s)$ are the empirical distributions of r and s.

In the current section, we discuss the properties of this estimator $\hat{I}$, focusing on its bias as compared to the "true" MI, $I[p^*]$. We start by a simple approximation of MI to the Chi-square statistic that provides good intuition about the case of independent variables, continue with the characterization of sample size regimes, and conclude with a comparison of various bias estimation methods.

Chi square approximation

When r and s are independent and n is sufficiently large, the empirical MI of Eq. 2.8 (sometimes termed the *Likelihood Ratio Chi-square* statistic) can be approximated by the *Pearson's Chi-square* statistic for independence given the marginals

$$\chi^2(R,S) = \sum_{s \in S} \sum_{r \in R} \frac{\left[n(r,s) - \frac{1}{n} n(r) n(s) \right]^2}{n(r) n(s)/n} \tag{2.9}$$

where $n(x)$ is the number of observations of x, and n is the total number of samples.

To see this, we first consider two distributions p and q, and develop the relation between their Kullback Liebler divergence D_{KL} and their Chi-square statistic

$$\begin{aligned} D_{KL}[p||q] &= \sum_i p_i \log_2\left(\frac{p_i}{q_i}\right) \\ &= -\frac{1}{\log(2)}\sum_i p_i \log\left(1+\frac{q_i-p_i}{p_i}\right) \\ &\approx -\frac{1}{\log(2)}\sum_i p_i\left(\frac{q_i-p_i}{p_i}-\frac{(q_i-p_i)^2}{2p_i^2}\right) \\ &= \frac{1}{2\log(2)}\sum_i \frac{(q_i-p_i)^2}{p_i} \end{aligned} \tag{2.10}$$

where the approximation holds when all p_i are close to q_i. Now, substituting $p(r,s)$ for q and $p(r)p(s)$ for p we obtain

$$\begin{aligned} I(R;S) &= D_{KL}[p(r,s)||p(r)p(s)] \\ &\approx D_{KL}[p(r)p(s)||p(r,s)] \\ &\approx \frac{1}{2\log(2)}\sum_i \frac{(p(r,s)-p(r)p(s))^2}{p(r)p(s)} \\ &= \frac{1}{2\log(2)n}\sum_i \frac{(n(r,s)-n(r)n(s)/n)^2}{n(r)n(s)/n} \\ &= \frac{1}{2\log(2)n}\chi^2((p(r)p(s)||p(r,s)) \end{aligned} \tag{2.11}$$

where the first approximation only holds when R and S are nearly independent. As a result, the mutual information statistic can be approximated by the Chi-square statistic under the hypothesis that R and S are independent.

This approximation provides a characterization of the distribution of the mutual information statistic under the null hypothesis. In the limit of large sample size n, the Chi-square statistic has a Chi-square distribution. The expectation of this distribution equals the number of degrees of freedom, which in the case of the Chi-square statistic for independence equals $(|R|-1)(|S|-1)$. The expected value of the mutual information statistic is therefore

$$E(I(R;S)) \approx \frac{(|R|-1)(|S|-1)}{2\log(2)n} \tag{2.12}$$

where $|R|$ and $|S|$ are the number of bins used to represent the R and S values respectively, and n is the number of samples. The statistical rule of thumb asserts that at least 5 samples in each bin are required for these approximations to hold.

Furthermore, the variance of the Chi-square distribution also equals to the number of degrees of freedom, and thus the variance of the mutual information estimator (when stimuli and responses are independent) can be approximated by

$$Var(I(R;S)) \approx \frac{(|R|-1)(|S|-1)}{2\log(2)n} \qquad (2.13)$$

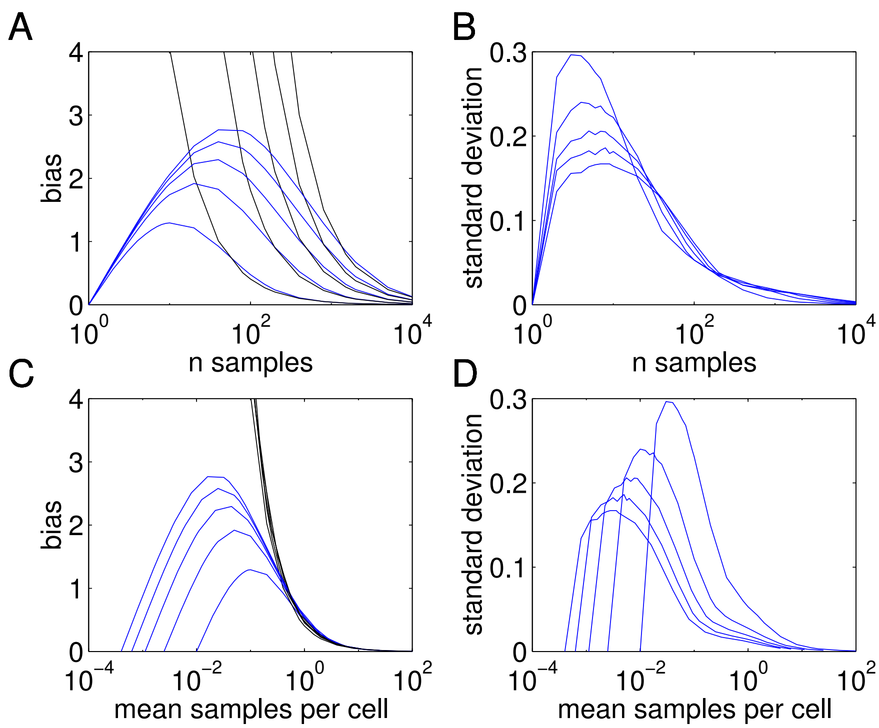

Figure 2.4: **A.** Bias and **B.** variance of the MI estimator as a function of sample size. **C.** Bias and **D.** variance of the MI estimator as a function of mean number of samples per cell. Blue curves correspond to different matrix sizes $k = 10$, 20, 30, 40 and 50. Black curves depict the bias expected from Eq. 2.12.

Small samples regime

When the sample size is small, the above equations of bias and variance terms do not describe the behavior of the MI estimator well. Interestingly, a regime of small samples exists in which the variance is very small, but the estimator is very poor. Moreover, both the variance and the bias increase when the sample size is increased. This effect is demonstrated in Fig. 2.4. 5000 samples of various size n were created from an independent distribution of two variables, each having k possible values. The expected number of samples per cell was therefore n/k^2. For each sample, the MI was calculated and the mean and standard deviation of the MI estimator are plotted as a function of the sample size. Figure 2.4A plots the standard deviation as a function of sample size. Different curves correspond to $k = 5, 10, 20, 30, 40$. Figure 2.4A plots the bias calculated numerically (blue curves), together with the bias expected from Eq. 2.12 (black curves). Figures 2.4C and 2.4D display the bias and variance as a function of the mean number of samples per cell. Paradoxically, for very small samples the variance of the estimator is small, while its bias may be large. Bias estimation according to Eq. 2.12, becomes a good approximation of the real bias for mean density of 1 per cell or higher. Below this density, Eq. 2.12 largely overestimates the bias. This is demonstrated in Fig. 2.4D in which the black curves are a good approximation of the blue curve only for a mean density larger than 1.

Two conclusions should be drawn from this demonstration. First, a low variance is not a sufficient condition for a good estimation of the MI, and for small samples the bias may be high despite the low variance. This precludes the use of cross validation techniques for bias estimation of the MI estimator. Secondly, the above approximations hold when the mean number of sample per cell is higher than 1. Recently [Paninski, 2003] formally established these results. All the neurophysiological results presented in the current work are in this regime where the bias correction method is applicable.

One should also note that when using the unified-bins procedure for estimating the MI, the bias of this estimator is no longer described well by Eq. 2.12, since the adaptive procedure performs a maximization step that reduces the bias.

Bias and variance of MI of Dependent variables

What about the case where the variables are not independent? It was shown in [Miller, 1955, Treves and Panzeri, 1995, Panzeri and Treves, 1996], that the bias of the mutual information statistic is to a first order equal to the bias in the independent case. The bias in the dependent case has a series of additional terms which explicitly depend on the underlying distribution $p(r,s)$, and are weighted by inverse powers of the number of samples n

$$\begin{aligned} E(I_n) &= I + \sum_{i=1}^{\infty} C_i \qquad (2.14)\\ \text{with} & \\ C_1 &= \frac{(|S|-1)(|R|-1)}{2n\log(2)} \\ C_2 &= \frac{\sum_s (p_n(s))^{-1} \sum_r \left[(p(r|s))^{-1}-1\right] - \sum_r (p(r))^{-2}+1}{12n^2\log(2)} \\ C_3 &= \frac{\sum_s (p_n(s))^{-2} \left[\sum_r (p(r|s))^{-2} - (p(r|s))^{-1}\right]}{12n^3\log(2)} \\ &\quad - \frac{\sum_r \left[(p(r))^{-2} - (p(r))^{-1}\right]}{12n^3\log(2)} \end{aligned}$$

where $E(I_n)$ is the expected information from a sample of size n, and I is the true MI of $p(r,s)$ [Treves and Panzeri, 1995]. The major caveat of this expansion is that it is not guaranteed to converge, and may strongly fluctuate with small changes in the underlying probability. [Treves and Panzeri, 1995] have used numerical simulations and showed several cases in which the expansion converges to the correct bias value. In these cases, the first correction term alone was found to be a good approximation of the real bias.

Unfortunately, this method largely overestimates the bias, when applied to contingency tables of spike counts and stimuli. The reason is that many of the bins are structurally empty, because different stimuli span different regions of the spike count distribution. The phenomenon is clearly demonstrated in Fig. 2.1, where some stimuli do not elicit any spikes. Thus, some of the empty bins should be treated as structural zeros. The number of effective degrees of freedom should therefore be smaller than $(S-1)(R-1)$, and larger than the actual number of occupied bins. On the other hand, The number of observed non-empty bins is of course a lower bound on the number of potentially non-empty bins, since with finite sample, it is possible that some empty bins are not structurally zero but empty due to the finite sample. The number of structurally non empty bins therefore lies between the total number of bins and the observed number of non empty ones. [Panzeri and Treves, 1996] thus suggested a maximum likelihood approach for estimating the true number of non empty bins.

To choose the best unbiased estimators of the MI, a test dataset was created based on the typical characteristics of auditory cortical responses, and the MI was estimated using the methods of [Treves and Panzeri, 1995], [Panzeri and Treves, 1996], and unified-bins. The unified-bins method was found superior and we therefore adopted it as our MI estimator of choice [Nelken *et al.*, 2003].

2.2.3 Binless MI estimation

The previous section discussed mutual information estimators based on joint distributions calculated through a binning procedure. The alternative approach that we discuss now avoids the use of bins and estimates densities according to distances between samples. The potential advantage of this approach lies in the fact that the resolution with which the density is estimated is not predefined as in histogram based methods, but changes continuously in a way that adapts to the distribution we estimate. Highly probable regions have more dense samples that effectively yield fine resolution, while low probability regions have low effective resolutions as though large bins were used. Victor [Victor, 2002] described how to use this approach for estimating mutual information.

Let f be a probability density function over a metric space, and suppose we have n samples $x_1, ..., x_n$ drawn according to f. Let λ_j be the distance from a point x_j to its nearest neighbor as illustrated in Fig. 2.5.

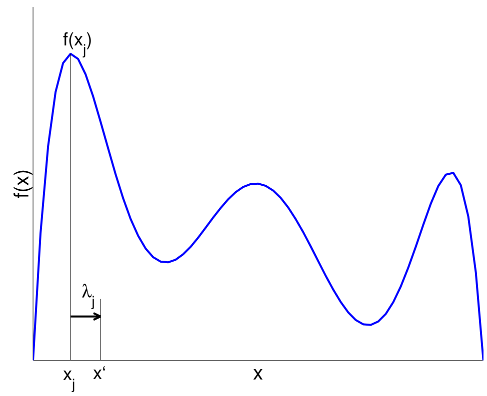

Figure 2.5: Illustration of the relation between the density $f(x_j)$ at some point x_j and the distance to its nearest neighbor x'.

Naturally, the probability density at each point $f(x_j)$ determines the expectation of the distance λ_j, as in more dense regions the expected distance between neighbors is smaller. Conversely, the distance λ_j from a sample point x_j to its nearest neighbor can be used to estimate the probability density at this point $f(x_j)$. The estimated density can be plugged into the expression of the entropy, yielding

$$\begin{aligned} H &= -\frac{1}{n}\sum_{j=1}^{n}\log_2(p(x_j)) \qquad (2.15) \\ &\approx -\frac{1}{n}\sum_{j=1}^{n}\log_2(\lambda_j) + \log_2(2(n-1)) + \frac{\gamma}{\log(2)} \end{aligned}$$

where γ is a constant ($\gamma \approx 0.577$). Information is then estimated from the difference between total entropy and stimulus conditioned entropies yielding

$$I \approx \frac{D}{n}\sum_{j=1}^{n}\log_2\left(\frac{\lambda_j^{total}}{\lambda_j^{stim}}\right) - \sum_{s=1}^{|S|}\frac{n_s}{n}\log_2\left(\frac{n_s-1}{n-1}\right) \qquad (2.16)$$

where $|S|$ is the number of stimuli, n_s is the number of samples that belong to the stimulus s, and D is the dimension of the metric space. This method is asymptotically unbiased, and is claimed to provide better information estimation than binning strategies when sufficient samples are available.

This method can be applied to continuous statistics for which the probability of observing identical samples vanishes. For example, the first spike latency measured in 1 micro second resolution can be viewed as a continuous variable. With spike train analysis however, one must handle discrete properties, as in the case of stimuli that elicit no responses. Clearly, all such "empty" spike trains have zero distance. We handle this issue by considering trials with no spikes separately. Define R to be the spike count, S the stimulus, and E to be a binary variable that denotes if there were spikes ($R > 0$, $E = e_+$) or not ($E = e_0$). Now the MI can be estimated using

$$\begin{aligned}
I(R;S) &= H(R) - H(R|S) \qquad (2.17)\\
&= H(R|E) + H(E) - H(R|S,E) - H(E|S)\\
&= p(e_+)H(R|E=e_+) + p(e_0)H(R|E=e_0)\\
&\quad -\sum_s H(R|E,S=s)p(s) + I(E;S)\\
&= \sum_s p(e_+,s)\left[H(R|E=e_+) - H(R|E=e_+,S=s)\right]\\
&\quad +\sum_s p(e_0,s)\left[H(R|E=e_0) - H(R|E=e_0,S=s)\right] + I(E;S)\\
&= \sum_s p(e_+,s)I(R;S=s|E=e_+)\\
&\quad +\sum_s p(e_0,s)I(R;S=s|E=e_0) + I(E;S)\\
&= \sum_s p(e_+,s)I(R;S=s|E=e_+) + 0 + I(E;S)\\
&= p(e_+)I(R;S|E=e_+) + I(E;S)
\end{aligned}$$

To evaluate the performance of this method we compare it with our method of MI estimation using adaptive binning: "unified-bins". Figure 2.6 compares histogram based and metric based MI estimation for first spike latency. Both methods yielded essentially the same estimates. For units with high MI value, the metric based estimation yielded slightly higher values than the unified bins procedure, but on the other hand yielded slightly lower MI values for the less informative units. These difference were found to be non significant using a paired t test (with $p > 0.25$ in AI and $p > 0.5$ in MGB and IC). We therefore chose to use the unified-bins procedure, since it can easily be applied to both continuous and discrete statistics.

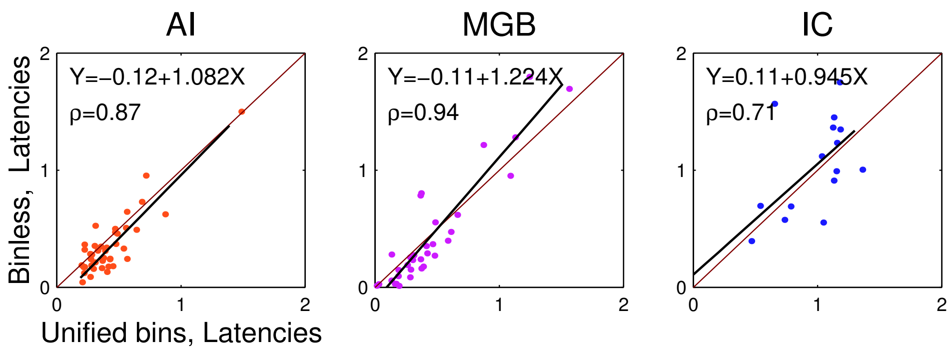

Figure 2.6: Comparison between metric based and histogram based estimation of MI in first spike latency in data from three different brain regions.

2.3 Methods II: Statistics of spike trains

The previous section focused on obtaining a reliable and efficient estimation of information in simple statistics of spike trains. We now turn to discuss six specific methods for transforming spike trains into low dimensional simple representations. Each of these methods focuses on other aspects of the spike trains, and comparing them reveals the relative coding importance of the different spike trains components.

1. Spike counts
2. Spike counts weighted by Inter-spike-intervals.
3. First spike latency
4. Spike patterns as binary words (the direct method).
5. Legendre polynomials embedding.
6. Second order correlations between spikes

Comparing the information that is obtained with these different methods can reveal the informative value of neural activity components. For example, the information value of temporal structres in spike trains is quantified by comparing a code that ignores such structures (spike counts) to codes that take them into account (e.g. the direct method).

The current section describes these methods in detail and their expected advantages. Their application to our neurophysiological data is presented in section 2.4.

2.3.1 Spike counts

Spike counts are probably the most widely used method of quantifying neural responses. A theoretical justification for using it as an informative statistic of spike trains is based on a widely used model for spike trains (see e.g. chap. 1 in [Dayan and Abbot, 2002]): the homogeneous Poisson process [Kingman, 1993]. In this model there is a single stimulus-dependent parameter, the firing rate, whose sufficient statistic is the spike count. As explained in section 2.1, if our spike trains are indeed created by a stimulus-dependent Poisson process, then using the spike count should extract all the information carried about the stimulus.

2.3.2 ISI weighted spike counts

Another type of point processes which are of special interest in neural modeling are *renewal processes*. These are point processes in which the distribution of inter-event-intervals is independent of past events, but is not necessarily exponential as in a Poisson process. Renewal processes are a useful model for spike trains, due to the fact that spiking largely resets numerous biophysical processes in the cell such as membrane voltage and ion channel configuration in the soma and proximal dendrites. The past spiking activity of a cell is thus partially decoupled from the future one. The Poisson process is a special case of a renewal process in which the inter-event-interval distribution is exponential. A useful model of renewal processes is the inverse-Gaussian process [Chhikara and Folks, 1989, Seshardri, 1993], in which the inter-interval distribution is

$$f(x|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} exp(-\lambda \frac{(x-\mu)^2}{2\mu^2 x}) \tag{2.18}$$

For this process, it was shown by [van Vreeswijk, 2001], that the sufficient statistic can be calculated on line by the following formula

$$T(X) = \sum_{t_i} 1 - \frac{K}{t_i - t_{i-1}} \tag{2.19}$$

where K is a parameter that corresponds to temporal correlations in the data and can be estimated from the data.

In contrast with non weighted spike counts, this statistic not only takes into account ont only the number of spikes but also some of their temporal structure. This structure is however limited to the inter-spikes-interval and does not take into consideration more complex patterns of spikes.

2.3.3 First spike latency

The timing of the first spike after stimulus onset is a simple statistic of spike trains that can carry considerable information about the stimuli. It was specifically proposed as a candidate for coding in the auditory system [Heil and Irvine, 1996, Heil, 1997, Jenison, 2001, Krishna and Semple, 2001].

Thorpe and colleagues [Thorpe *et al.*, 1996, VanRullen and Thorpe, 2001, Fabre-Thorpe *et al.*, 2001] have argued convincingly that first spike latencies provide a fast and efficient code of sensory stimuli, and showed how excellent reconstructions of natural visual scenes can be obtained when computation times are limited to short durations at which only the first spikes of 5-10 percent of the coding neurons are used. A simple intuition explaining the success of this method is that for neurons whose firing fits a homogeneous Poisson model, the inter spike interval is exponentially related to the firing rate

$$Pr(ISI = l) \propto exp(-\lambda l) \ . \tag{2.20}$$

This allows us to obtain an estimate of the rate in a short time: While estimating spike counts requires averaging over a relatively long time window, the time interval between stimulus onset and the first spike conveys the same information but only requires us to observe a single spike [4]. These observations suggest that the first spike latency is a natural candidate for carrying information in auditory spike trains.

[4] The variance of the latency estimator is however larger than the counts estimator.

2.3.4 The direct method

The direct method, presented in [deRuyter van Steveninck *et al.*, 1997], is aimed at capturing potential information in temporal patterns of spikes. This is achieved by representing each spike train as a binary string R, by discretizing time with some resolution Δt, and calculating the MI from the joint distribution of stimuli and binary strings.

This method differs from the ones we discussed above in that with asymptotically infinite data and using infinitesimal resolution, it should capture all the information that the spike trains convey regardless of their underlying distribution. This is inherently different for example from spike counts, as these will only capture all the information in the case of homogeneous Poisson processes.

In practice, with finite data, the direct method should be viewed as any other dimensionality reduction method since it requires to discretize the spike trains with some predefined temporal resolution (usually on the order of a few milliseconds), and this coarse resolution reduces the information carried by the spike train.

With coarse resolution the direct method therefore extracts the coarse temporal structure of the responses, similar to the one captured by inhomogeneous rate modulation models.

2.3.5 Taylor expansion

Panzeri and Schultz [Panzeri and Schultz, 2001] developed a second order Taylor approximation of the mutual information $I(\{t_i\};S)$ between stimuli S and spike trains t_i. This method uses mean firing rates and second order correlations between spike times, and can be viewed as a reduction method, since it ignores higher order correlations.

Denoting with t_i^a the time of the i-th spike of the cell a, the information between the spike trains and the stimulus is represented as a Taylor expansion in the time window T

$$I\left(\{t_i\};S\right) = I_t\left(\{t_i\};S\right)T + I_{tt}\left(\{t_i\};S\right)\frac{T^2}{2} + \ldots \tag{2.21}$$

where $I_t(\cdot)$ and $I_{tt}(\cdot)$ are first and second order derivatives of the information with respect to the length of the time window T.

The first order term is [Bialek *et al.*, 1991]

$$I_t\left(\{t_i\};S\right)T = \sum_{a=1}^{n}\int_{dt_1^a}\left\langle \bar{r}_a(t^a;s)\log_2\left(\frac{\bar{r}_a(t^a;s)}{\langle \bar{r}_a(t^a;s')\rangle_{s'}}\right)\right\rangle_s \tag{2.22}$$

where $\bar{r}_a(t;s)$ is the average firing rate of cell a at time t when presented with the stimulus s.

The second order term consists of three components

$$\begin{aligned} I_{tt}\left(\{t_i\};S\right)\frac{T^2}{2} \;\propto\; & \sum_{a=1}^{n}\sum_{b=1}^{n}\int_{dt_1^a}\int_{dt_2^b}\langle \bar{r}_a(t_1^a;s)\rangle_s\left\langle \bar{r}_b(t_2^b;s)\right\rangle_s \\ & \times\left\{\nu_{ab}(t_1^a,t_2^b)-(1+\nu_{ab}(t_1^a,t_2^b))\log_2\left(1+\nu_{ab}(t_1^a,t_2^b)\right)\right\} \\ + \; & \sum_{a=1}^{n}\sum_{b=1}^{n}\int_{dt_1^a}\int_{dt_2^b}\left\langle \bar{r}_a(t_1^a;s)\bar{r}_b(t_2^b;s)\gamma_{ab}(t_1^a,t_2^b;S)\right\rangle_s \\ & \times\log_2\left(\frac{1}{1+\nu_{ab}(t_1^a,t_2^b)}\right) \\ + \; & \sum_{a=1}^{n}\sum_{b=1}^{n}\int_{dt_1^a}\int_{dt_2^b} \\ & \left\langle \bar{r}_a(t_1^a;s)\bar{r}_b(t_2^b;s)\left[1+\gamma_{ab}(t_1^a,t_2^b;S)\right]\times\log_2\left(\mu\right)\right\rangle_s \end{aligned} \tag{2.23}$$

where $\nu_{ab}(t_1^a,t_2^b)$ is a scaled correlation density between cells a and b

$$\nu_{ab}(t_1^a,t_2^b) = \frac{\left\langle \bar{r}_a(t_1^a;s)\bar{r}_b(t_2^b;s)\right\rangle_s}{\langle \bar{r}_a(t_1^a;s)\rangle_s\left\langle \bar{r}_b(t_2^b;s)\right\rangle_s} - 1 \quad , \tag{2.24}$$

γ is the scaled noise correlation (stimulus conditioned correlations)

$$\gamma(t_1^a,t_2^b;s) = \frac{\overline{r_a(t_1^a;s)r_b(t_2^b;s)}}{\bar{r}_a(t_1^a;s)\bar{r}_b(t_2^b;s)} - 1 \tag{2.25}$$

and μ is

$$\mu = \frac{\left\langle \bar{r}_a(t_1^a;s)\bar{r}_b(t_2^b;s)\right\rangle_s(1+\gamma(t_1^a,t_2^b;s))}{\langle \bar{r}_a(t_1^a;s)\bar{r}_b(t_2^b;s)(1+\gamma(t_1^a,t_2^b;s))\rangle_s}. \tag{2.26}$$

The first term is always non positive. When the spike trains are samples of an inhomogeneous Poisson, the second and third terms asymptotically vanish. When spikes are not independent they can contribute to the MI through non-zero autocorrelations $\gamma_{kk}(t_1^k, t_2^k)$ or cross correlations $\gamma_{kl}(t_1^k, t_2^l)$. The second term measures the stimulus independent correlations, and the third term corresponds to stimulus dependent correlations.

2.3.6 Legendre polynomials embedding in Euclidean spaces

A method for embedding spike trains in a low dimensional Euclidean space was suggested by [Victor, 2002]. After this embedding is performed, MI can be estimated through binless estimation using the distances between the embedded points (Section 2.2.3). The embedding procedure consists of two steps. First, a monotonic time warping is applied to the spike trains, such that all spikes are equally spaced in the range $[-1, 1]$. This is achieved by first sorting all spikes and then assigning the j^{th} spike to $\tau_j = \frac{2}{M}(j - \frac{1}{2}) - 1$, where M is the total number of spikes.

Then, the embedding coordinates are calculated by the Legendre polynomials P_h, which are orthogonal on $[-1, 1]$. Consider a spike train x_i that contains n_i spikes that were warped to $\tau_1, ..., \tau_{n_i}$. The coordinate of the embedded spike train x_i that is calculated using the Legendre polynomial P_k to be

$$c_k = \sqrt{2k+1} \sum_{j=1}^{n_i} P_k(\tau_j) \tag{2.27}$$

Therefore, choosing a set of d Legendre polynomials $P_1, ..., P_d$ allows us to project each spike train x_i to a d dimensional vector, with the coordinates $c_1, ..., c_d$. the Legendre statistic is therefore defined as $T_{legendre} = (c_1, ..., c_d)$.

It is difficult to assign a complete intuitive interpretation to the aspects of the spike trains to which this method is sensitive. Clearly it is sensitive to temporal structures through the relative ordering of spikes across all stimuli. Moreover, since only the relative order of spikes matters, the relevant temporal resolution is data dependent, and the statistic is insensitive to the fine temporal structures in time windows where one stimulus elicits spikes. It thus invests higher effective resolution in periods where many stimuli cause the neurons to respond.

2.4 Results

We estimated the MI in each of the methods described in the previous section. The results are detailed in the next sections. Sections 2.4.1-2.4.6 describe MI results obtained with specific methods, and a comparison of these results is given in section 2.4.7.

2.4.1 Spike counts

We estimated the MI conveyed by the spike counts $T_{counts}(R)$ in some time window using the joint count matrix $n(s, T_{counts}(r))$ and calculating $\hat{I}(T_{counts}(R); S)$ using the unified-bins procedure. Figure 2.7 plots the mean mutual information obtained from the spike counts of 45 AI cells in our data, for several time windows.

The MI in spike counts is fairly insensitive to the exact location of the window, as long as it covers the period that begins at 30 ms and ends at 80 ms after stimulus onset. There is a slight decrease in MI for longer time windows due to spontaneous activity of some of the cells that increases the noise. Similar analysis for IC and MGB cells show that the relevant windows in these areas start 20 ms after stimulus onset.

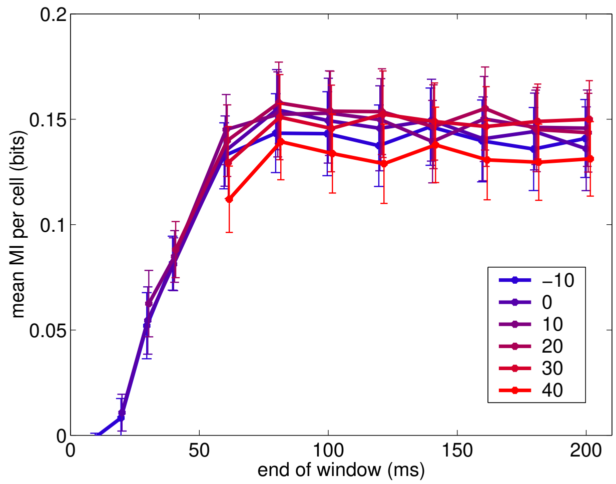

Figure 2.7: Mutual information between the stimulus identity and the spike counts of 45 AI cells, for different windows in which spikes are considered. Different curves correspond to different start point of the window after stimulus onset. MI was estimated using the unified-bins procedures. Similar results with respect to optimality of time window were obtained with linear bins as well. The error bars designate standard error of the mean over the population of cells.

Figure 2.8 plots the information obtained using the spike counts of each cell in the three brain regions, against its firing rate. Large markers denote the mean of the population with regard to both the x and y axes.

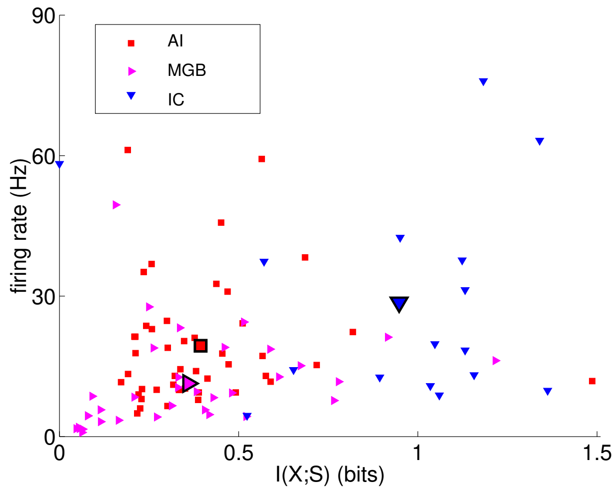

Figure 2.8: The firing rate vs the MI estimated with spike counts in three braing regions. Large markers denote the mean of each population. Firing rates and MI were calculated over the most informative time window of each population.

Even though the mean firing rates are relatively similar in all three brain regions, the information carried by IC spike counts is about double the information carried by MGB and AI spike counts. The intuition behind this observation is that higher firing rates are not necessarily more informative, since it is the variability in spike counts across stimuli that carries information about the identity of the stimulus.

2.4.2 Weighted spike counts

Figure 2.9 plots the MI obtained from weighted spike counts, as a function of the beginning point of the window size and the values of the parameter K. The end point of the window was optimized. Values are the means over a population of 45 AI cells.

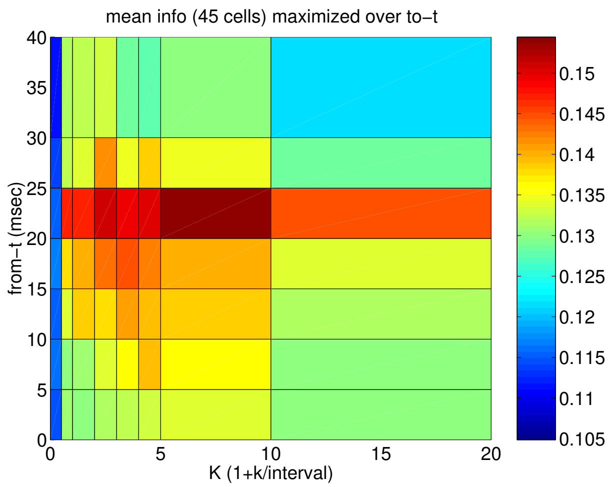

Figure 2.9: Mutual information conveyed by the value of the sufficient statistic for the inverse-Gaussian process, as a function of the parameter K and the window starting point after stimulus onset. MI was estimated using the unified-bins procedure. Results are the mean over 45 AI cells.

This figure shows that this method achieves the same MI values as the ones obtained with uniformly weighted spike counts, both reaching a maximum of about 0.15 bit per cell on average. The optimal time window was also found to be similar.

2.4.3 First spike latency

Figure 2.10 plots the mean mutual information obtained from a population of 45 AI cells, estimated with different temporal windows as in Fig. 2.7.

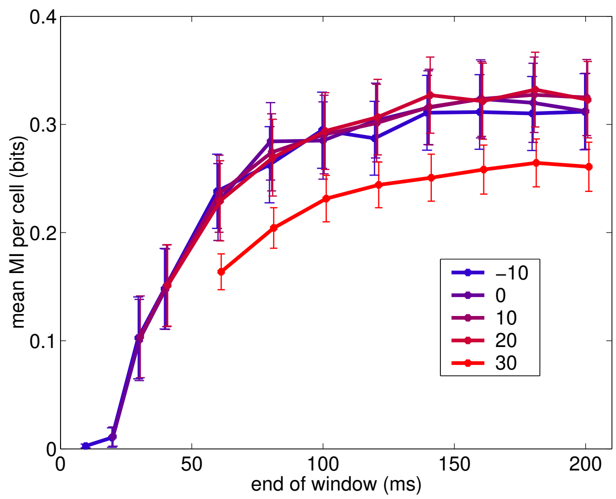

Figure 2.10: Mutual information conveyed by the first spike latencies of 45 AI cells, for different windows in which spikes are considered. Different curves stand for different start points of the window after stimulus onset. MI was estimated using the unified-bins procedure. Trials with no spikes were assigned to a separate bin. Error bars designate standard error of the mean over the population of cells.

Most importantly, the MI in the first spike latency yields almost double the information obtained with spike counts.

These MI levels are again not sensitive to the exact location of the time window, as long as it covers at least ~ 120 ms after stimulus onset, and as expected the MI in the first spike latency is monotonic in the length of the window. MI is slightly higher when ignoring spontaneous spikes that occur in the time window 0-20 ms after stimulus onset.

2.4.4 The direct method

We applied the direct method to our data in the following way. For a given temporal resolution Δ, all spike trains were discretized and converted to a binary string of length T/Δ, where T is the length of the time window considered. The i^{th} bit in the string was 1 if there was at least one spike in the corresponding bin, and was zero otherwise. Note that we only had a single "binary word" per each spike train, but since each stimulus was repeated n times (usually $n = 20$), we had a distribution of binary words for every stimulus. Given these joint counts of stimulus and binary words we applied the unified-bins procedure to estimate the MI.

As with previous methods we enumerated over the time window and also over resolutions within the set $1, 2, 4, 8, 16, 32$ ms. The majority of the cells achieved maximal MI with temporal resolutions of $2-4$ ms.

Figure 2.11 compares the MI obtained with the direct method with the MI obtained with first-spike latencies. The two methods achieve similar MI levels in AI and MGB, but the direct method achieves higher MI levels in the IC. This is probably since many IC neurons are sensitive to features near the stimulus onset and are common to several stimuli. The direct method successfully extracts more information since it is not limited to the first spike but rather sensitive to the temporal structure of the responses after the onset.

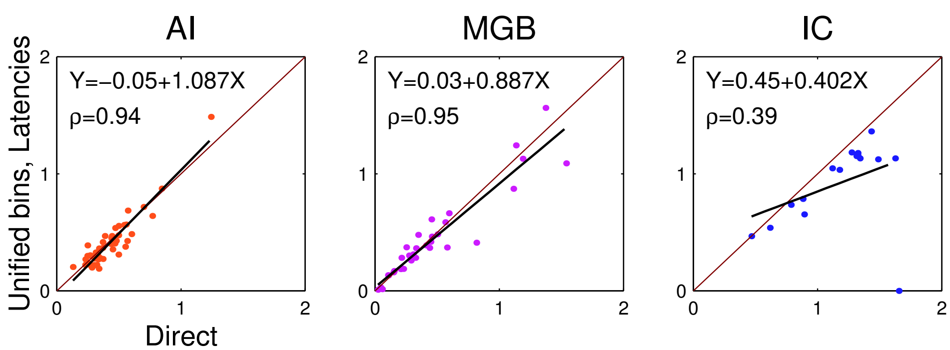

Figure 2.11: MI estimated using the direct method and using first spike latency in three brain regions.

2.4.5 Taylor expansion

To apply this method to our data, we enumerated over temporal resolutions in the set $\{1, 2, 4, 8, 16, 32\}$ milliseconds. Correction for bias in the MI estimation was done by estimating the MI after shuffling the trials, and subtracting the MI of the shuffled data. With shuffled trials, the MI should be asymptoticaly zero, thus (posisitive) MI values estimate the bias under the independent case. This method tends to underestimate the bias of the dependent case, and thus one should take into account that the MI in this method may be overestimated.

Figure 2.12 compares the MI obtained with this method with the MI obtained with the direct method. With our data, this method performed poorly, in the sense that it yielded lower MI levels than the rest of the methods.

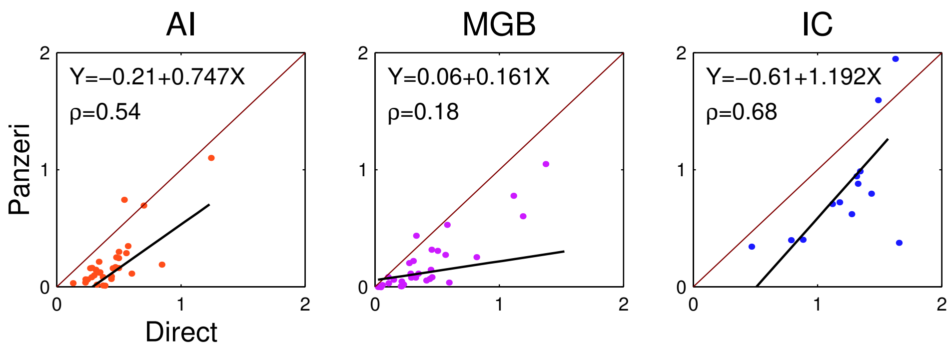

Figure 2.12: MI estimated using second order correlations vs the direct method in three brain regions.

2.4.6 Legendre polynomials embedding

Figure 2.13 compares the MI obtained with this method with the MI obtained with the direct method. We enumerated over embedding dimensions in the range $1-5$ and the optimal maximal information was obtained with a dimension $d=5$. With our data, the embedding method was able to extract the same level of information as the direct method, and no significant difference was found using a paired t test ($p > 0.2$ in AI, $p > 0.5$ in MGB and IC).

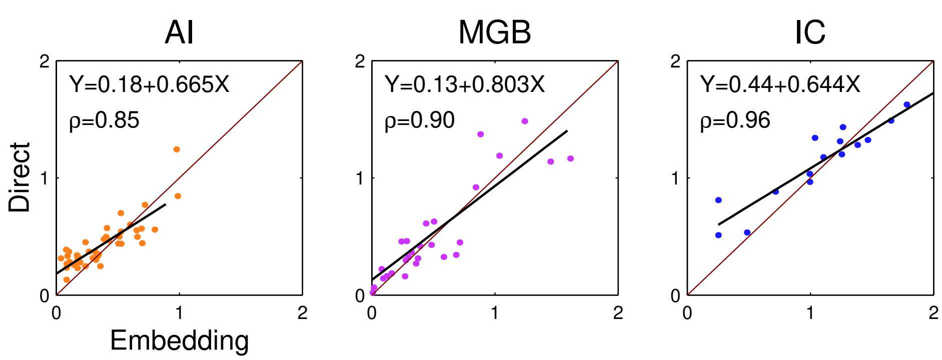

Figure 2.13: MI estimated using the direct method vs the Legendre polynomials embedding in three brain regions.

2.4.7 Comparisons

Distribution of MI across cell populations

The distribution of MI values across cells in the different brain regions is given in the following figure.

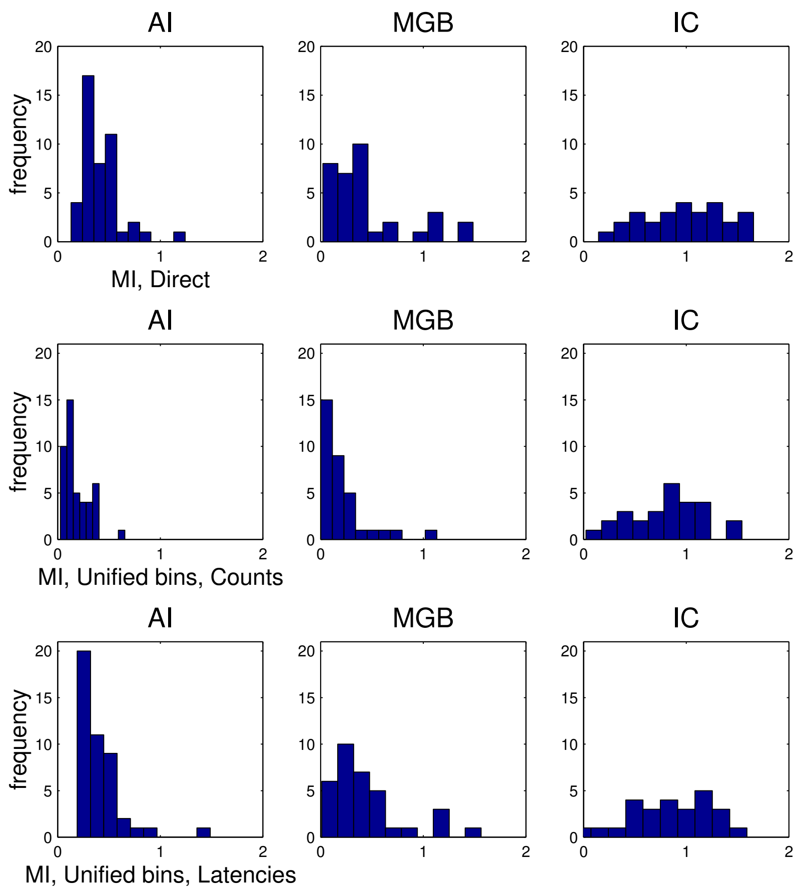

Figure 2.14: Comparison of mean MI obtained in three brain regions using four different methods. Error bars are standard errors of the mean over the population of cells.

In AI and MGB the distribution is largely skewed: most of the cells convey only little information, and few cells are highly informative. The distribution is more symmetric in the IC.

Comparison of MI levels

Figure 2.15 summarizes the mean mutual information in bits per AI cell that is obtained using each of these methods. Interestingly, spike counts provide only about half the information that can be extracted using the direct method.

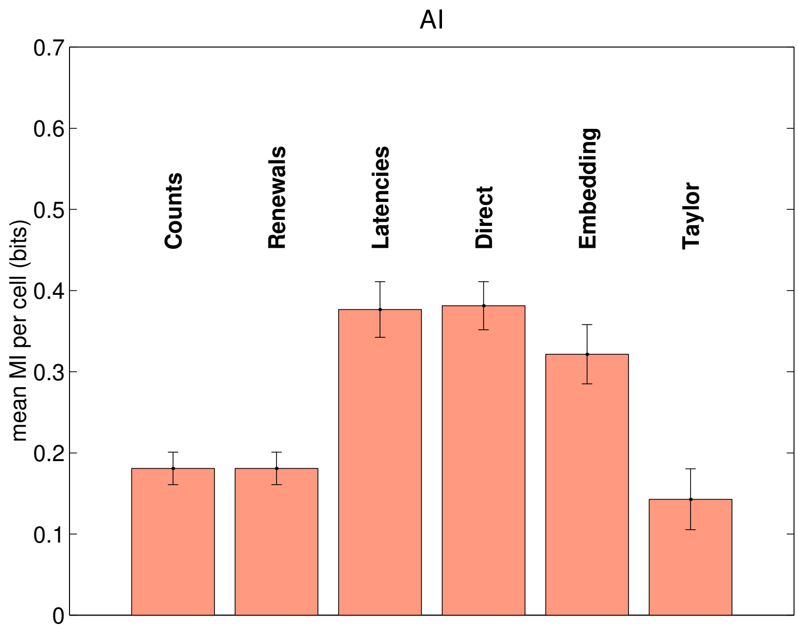

Figure 2.15: Mean MI obtained from all 45 cells using six different methods.

An interesting result of this comparison is that similar MI level are obtained by three different methods: using first spike latency, using the direct method and using the Legendre polynomials embedding method. This similarity suggests that these (very different) methods saturate the full information in the spike trains. Theoretically the MI is bounded by the entropies $I(S;T(R)) \leq H(S)$ and $I(S;T(R)) \leq I(S;R) \leq H(R)$. While we cannot estimate $H(R)$, the stimulus entropy in our experiments (—S—=15) yields an upper bound of $log_s(15) = 3.9$ bits, to be compared with the mean MI level of about 0.35 bits per cell obtained in AI. These results suggest that allthugh AI cells convey completely independent information , only about

10-11 cells are needed to fully discriminate between the 15 stimuli. The nature of interactions between cells is the subject of the next two chapters of thesis.

Figure 2.16 summarizes the mean mutual information obtained using the four main methods, in all three brain regions.

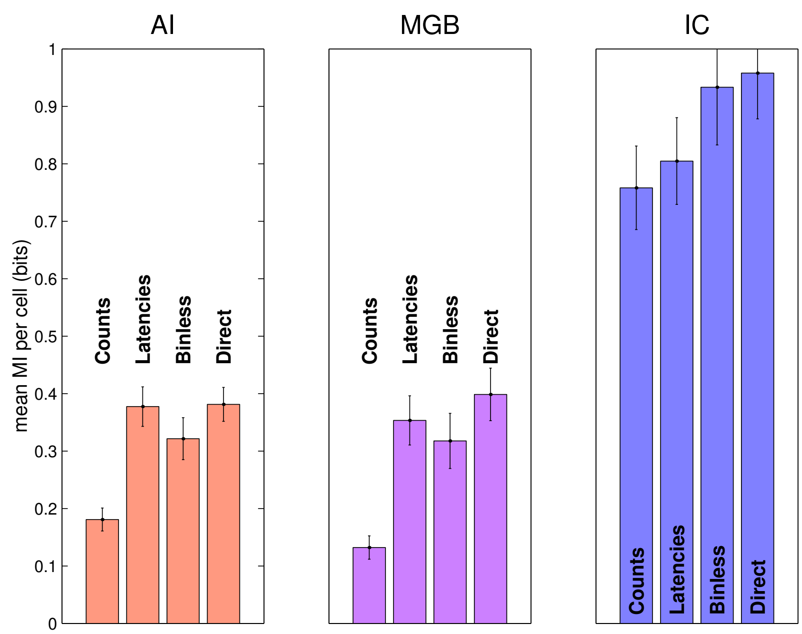

Figure 2.16: Comparison of mean MI obtained in three brain regions using four different methods. Error bars are standard errors of the mean over the population of cells.

We find that IC cells convey considerably more information about the stimulus identity regardless of the specific method used for MI extraction.

2.5 Conclusions

Estimating information from spike trains is difficult with the typical amount of neurophysiological data, because the dimensionality of the spike trains is too large for estimating the full joint distribution of spike trains and stimuli. For this reason, one has to devise methods to reduce the dimensionality of the spike trains, while loosing as little information as possible. We compared several methods for such dimensionality reduction, and showed that several of them converge to highly similar values in our data. These include the first spike latency, Legendre polynomials embedding and binary words representation of spike patterns (the direct method).

Using spike counts alone extracted about half of the maximum information, while the first spike latency and the distribution of binary words achieved the maximum information. The fact that spike counts are commonly found to achieve only one to two thirds of the total information is often used as an argument in favor of temporal coding, i.e. that precise temporal patterns carry information that does not exist in rate coding. The above analysis shows that this argument is not entirely correct: most of the information can be extracted using relatively simple statistics of the responses, although spike counts by themselves are clearly insufficient.

Bibliography

[Bialek *et al.*, 1991] W. Bialek, F. Rieke, R.R. deRuyter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1854–1857, 1991.

[Chhikara and Folks, 1989] R.S. Chhikara and J.L. Folks. *The inverse Gaussian distribution: Theory, methodology and applications.* Marcel Dekker, New York, 1989.

[Dayan and Abbot, 2002] P. Dayan and L.F. Abbot. *Theoretical Neuroscience.* MIT Press, Cambridge, MA, 2002.

[Degroot, 1989] M. Degroot. *Probability and Statistics.* Addison Wesley, Reading MA, 2nd edition, 1989.

[deRuyter van Steveninck *et al.*, 1997] R. deRuyter van Steveninck, G.D. lewen, S.P. Strong, R. Koberele, and W. Bialek. Reproduciblity and variability in neural spike trains. *Science*, 275:1805–1808, 1997.

[Devroye and Lugosi, 2001] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York, 2001.

[Fabre-Thorpe *et al.*, 2001] M. Fabre-Thorpe, A. Delorme, C. Marlot, and S. Thorpe. A limit to the speed of processing in ultra-rapid visual categorization of novel and natural scenes. *Journal of Cognitive Neuroscience*, 13(2):171–180, 2001.

[Fisher, 1922] R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, 222:309–368, 1922.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer series in statistics. Springer, 2001.

[Heil and Irvine, 1996] P. Heil and D.R.F Irvine. On determinants of first-spike latency in auditory cortex. *NeuroReport*, 7:3073–3076, 1996.

[Heil, 1997] P. Heil. Auditory cortical onset responses revisited I. first-spike timing. *The Journal of Neurophysiology*, 77(5):2616–41, May 1997.

[Jenison, 2001] R.L. Jenison. Eliminating nuisance parameters for cortical first-spike latency decoding of sound source direction. NIPS*2001 Workshop: Information and Statistical Structure in Spike Trains, 2001.

[Kingman, 1993] J.F.C. Kingman. *Poisson Processes*. Oxford studies in probability. Oxford University Press, 1993.

[Krishna and Semple, 2001] B.S. Krishna and M.N. Semple. A computational model for first-spike latency and variability in the auditory nerve. The Association for Research in Otolaryngology MidWinter Meeting 2001, 2001.

[Miller, 1955] G.A. Miller. Note on the bias in information estimates. In H. Quastler, editor, *Information theory in Psychology, Problems and methods II-B*, pages 95–100. The Free Press, 1955.

[Nelken *et al.*, 2003] I. Nelken, G. Chechik, A. King, and J. Schnupp. The neural code in primary auditory cortex. In preparation, 2003.

[Nemenman *et al.*, 2002] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[Paninski, 2003] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

[Panzeri and Schultz, 2001] S. Panzeri and S.R. Schultz. A unified approach to the study of temporal correlational and rate coding. *Neural Computation*, 13(6):1311–1349, 2001.

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

[Rissanen, 1978] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[Seshardri, 1993] V. Seshardri. *The inverse Gaussian distribution*. Oxford, 1993.

[Singer and Gray, 1995] W. Singer and C.M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annual Reviews in Neuroscience*, 18:555–586, 1995.

[Thorpe *et al.*, 1996] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

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

[van Vreeswijk, 2001] C. van Vreeswijk. Information trasmission with renewal neurons. In J.M. Bower, editor, *Computational Neuroscience: Trends in Research*. Elsevier Press, 2001.

[VanRullen and Thorpe, 2001] R. VanRullen and S. Thorpe. Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex. *Neural Computation*, 13(6):1255–1283, 2001.

[Vapnik, 1995] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, Berlin, 1995.

[Victor, 2002] J.D. Victor. Binless strategies for estimation of information from neural data. *Physical Reviews E*, 66(51903):1–15, 2002.