An Information Theoretic Approach to the Study of Auditory Coding

Thesis submitted for the degree
"Doctor of Philosophy"
by

Gal Chechik

Submitted to the Senate of the Hebrew University

July 2003

This work was carried out under the supervision of Prof. Naftali Tishby and Dr. Israel Nelken.

Acknowldegements

Someone once said that every manuscript is just a graphomaniac appendix to the acknowledgment page. The current page provides this hypothesis with hard empirical evidence.

This thesis summarizes a wonderful period I spent in the Hebrew University. I was lucky to study and work with many gifted people who enriched me in numerous and often unexpected ways. First of all, I am in great debt to Naftali Tishby and Israel Nelken, my supervisors. I learned from them very different lessons, typical to the experimental and theoretical approaches to scientific work. Tali instructed me how to seek principled approaches, and to base methodologies on fundamental first-principles. Eli guided me how to explore experimental data in order to reveal its organization principles, and how to turn empirical findings into well established insightful observations. They both taught me not only how deep and creative thinking can be combined with careful and detailed investigation, but also the importance of polishing to perfection the clear presentation of findings and ideas. Their patient guidance, wide knowledge, generosity and friendship turned our joint work into pleasure, and made my PhD period an experience I will always be happy to recall.

My work was based on data collected by a series of dedicated people in several labs, all connected through a joint project under a Human Frontiers Science Project (HFSP) grant. Prof. Eric Young and Dr. Mike Anderson from Johns Hopkins University, Baltimore, not only provided me with their recordings in the inferior colliculus, but also with help and guidance in investigating the problems of auditory neural coding. Prof. Ad Aertsen from university of Freiburg together with Alexandre Kuhn were helpful in discussing various aspects of the work.

I received great help from the members of my Ph.D. committee, both within our formal meetings and beyond. Moshe Abeles advised me whenever I needed, and helped to direct my thinking to useful alleys. Yaacov Ritov was extremely helpful when I came to him with my weird ideas about the distribution of the mutual information statistic and guided me in that research (not included in this thesis). Finally, I owe special thanks to Eytan Ruppin, who introduced me to scientific work as my M.Sc. advisor, and continued to provide me his guidance ever since. His true friendship and continuous support throughout the course of my studies are invaluable.

Several faculty members of the ICNC, the department of computer sciences and the Tel Aviv University helped me in various ways. Idan Segev was helpful and inspiring, Eilon Vaadia was always willing to indulge in deep long discussions, and Haim Sompolinsky taught me to search for the deepest and clearest description of any phenomenon. David Horn from Tel-Aviv University has helped me throughout my academic studies, providing scientific intuitions and good advice and making our collaboration a true pleasure. Isaac Meilijson, my M.Sc. advisor, continued to help me whenever I needed, and educate me with delight about probabilistic modeling. Yair Weiss, a dear friend, provided his deep intuition that made our collaborations a pleasure, and wise advices that were of great help. Yoram Singer and Nir Friedman were always available and helpful, and Nir is the one who introduced me to the exciting field of computational molecular biology.

Amir Globerson has been my partner both to four papers and to long rides on the Tel-Aviv - Jerusalem road. He is the living proof that friendship and work can and should be mixed together. Elad Schneidman, a lab's veteran, has helped me a lot during my first days at the lab and has continued to help ever since. My roommates Rani Gilad-Bachrach and Amir Navot were always happy to assists in any weird question I had, and initiated the most bizarre discussions ever (including, but not limited to, the types of oral surgical operations required for being a performing magician). Rony Paz reviewed my paper on redundancy reduction and Ranit Aharonov and Tuvik Beker commented on my papers on synaptic learning rules (not included in this thesis), as well as on my dubious yachting skills. They all cheered me up when I needed. Gill Bejerano helped me thinking about distribution of mutual information statistic. Gal Elidan, Kobi Crammer, and Iftach Nachman were happy to comment on various ideas and Noam Slonim kindly shared with me his knowledge, data and IB code. Finally, all members of the auditory neuroscience laboratory in Haddasah medical school have participated in the challenging task of collecting the data analyzed here, and helping me to try and understand it: Omer Bar Yosef, Nachum Ulanovsky, Gilad Jacobson, Liora Las and Dina Farkash. Omer also created the set of natural and complex stimuli analyzed here, and Nachum was always eager to help, a merit I was happy to use.

A special thank you is due to Alisa Shadmi, who was there for me whenever an administrative issue came up, and saved me the agonies of academic bureaucracy. I got plenty of help in administrative matters from Ruthi Succi, Ziva Rechani and Regina Krizhanovsky, to all of whom I owe great gratitude.

The research described in this thesis was supported by several external funding sources. The Ministry of science, the Eshkol foundation, provided an ongoing full scholarship throughout my studies. Intel Corp. and the Wolf foundation provided additional generous support.

Finally, my deepest thanks are to my family: my parents Rachel and Aharon, my wife Michal, and my kids Itay and Maayan. Their infinite love and support is what make it all happen.

To Itay, The happiest learning machine I ever created. *shigen.*

Abstract

This dissertation develops information theoretic tools to study properties of the neural code used by the auditory system, and applies them to electro-physiological recordings in three auditory processing stations: auditory cortex (AI), thalamus (MGB) and inferior colliculus (IC). It focuses on several aspects of the neural code: First, robust estimation of the information carried by spike trains is developed, using a variety of dimensionality reduction techniques Secondly, measures of informational redundancy in small groups of neurons are developed. These are applied to neural activity in a series of brain regions, demonstrating a process of redundancy reduction in the ascending processing pathway. Finally, a method to identify relevant features, by filtering out the effects of lower processing stations is developed. This approach is shown to have numerous applications in domains extending far beyond neural coding. These three components are summarized below.

The problem of the *neural code* of sensory systems combines two interdependent tasks. The First is identifying the *code words*: i.e., the components of neural activities from which a model of the outside world can be inferred. Common suggestions for these components are the firing rates of single neurons, correlated and synchronized firing in groups of neurons, or specific temporal firing patterns across groups of neurons. The second task is to identify the components of the sensory inputs about which neurons in various brain regions carry information. In the auditory realm, these can be "physical" properties of acoustic stimuli, such as frequency content or intensity level, or more abstract properties such as pitch or even semantic content of spoken words.

We first address the first task, that of identifying code words that are informative about a predefined set of natural and complex acoustic stimuli. The difficulty of this problem is due to the huge dimension of both neural responses and the acoustic stimuli, which forces us to develop techniques to reduce the dimensionality of spike trains, while losing as little information as possible. Great care is taken to develop and use unbiased and robust estimators of the mutual information. Six different dimensionality reduction approaches are tested, and the level of information they achieve is compared. The findings show that the maximal information can almost always be extracted by considering the distribution of temporal patterns of spikes. Surprisingly, the first spike

latency carries almost the same level of information. In contrast, spike counts convey only half of the maximal information level. In all of these methods, IC neurons conveyed about twice more information about the identity of the presented stimulus than AI and MGB neurons.

To study neural coding strategies we further examine how small groups of neurons interact to code auditory stimuli. For this purpose we develop measures of informational redundancy in groups of cells, and describe their properties. These measures can be reliably estimated in practice from empirical data using stimulus conditioned independence approximation. Since redundancy is biased by the baseline single-unit information level, we study this effect and show how it can be reduced with proper normalization. Finally, redundancy biases due to ceiling effect on maximal information are discussed.

The developed measures of redundancy are then applied to quantify redundancy in processing stations of the auditory pathway. Pairs and triplets of neurons in the lower processing station, the IC, are found to be considerably more redundant than those in MGB and AI. This demonstrates a process of redundancy reduction along the ascending auditory pathway, and puts forward *redundancy reduction as a potential generic organization principle for sensory systems*. Such a process was hypothesized 40 years ago by Barlow based on a computational motivation and is experimentally demonstrated here for the first time.

We further show that the redundancies in IC are correlated with the frequency characterization of the cells; namely, redundant pairs tend to share a similar best-frequency. This effect is much weaker in MGB and AI, suggesting that even the low redundancy in these stations is not due to similar frequency sensitivity. This result has great significance for the study of auditory coding since it cannot be explained by the standard model for cortical responses, the spectro-temporal receptive field (STRF). Finally, we measured informational redundancy in the information that single spikes convey about the spectro-temporal structure. Redundancy reduction is observed here as well. Moreover, IC cells convey an order of magnitude more information about these spectro-temporal structures than MGB and AI neurons. Since AI neurons convey half the information that IC neurons do about stimulus identity we conclude that cortical neurons code the identity of the stimuli well without characterizing their "physical" aspects. This observation hints that the cortex is sensitive to complex structures in our stimulus set, which cannot be identified with the common parametric stimuli.

In the last part of the work, we address the second task in neural coding identification. Here the goal is to develop methods to which characterize the stimulus features that cortical neurons are sensitive. One difficulty is that many features that cortical neurons are informative about result from processing at lower brain regions. For example, AI neurons are sensitive to frequency content and level, but these properties are already computed at the cochlea. This is in fact a special case of a fundamental problem in unsupervised learning. The task of identifying relevant structures in data in an unsupervised manner is ill defined since real world data often contain conflicting structures. Which of them is relevant depends on the task. For example, documents can be clustered according to content or style; speech can be classified according to the speaker or the meaning.

We provide a formal solution to this problem in the framework of the *information bottleneck* (IB) method. In IB, a source variable is compressed while preserving information about another relevance variable. We extend this approach to use side information in the form of additional data, and the task is now to compress the source (e.g. the stimuli) while preserving information about the relevance variable (e.g. cortical responses), but removing information about the side variable (e.g. 8th nerve responses). The irrelevant structures are therefore implicitly and automatically learned from the side data. We present a formal definition of the problem, as well as its analytic and algorithmic solutions. We show how this approach can be used in a variety of domains in addition to auditory processing.

Contents

Chapter 1

Introduction

1.1 Computation in the brain

1.1.1 Introduction

What do we mean when we say that the brain computes?

It is not easy to explain to the educated layman what it means that the brain computes. Often, the immediate source of confusion is that the term does not refer to a person performing calculations in his head, but rather to the operations of small circuits of neurons in his brain. The clearest way to think about it is to view computations as mappings, or (possibly high dimensional) functions. By this view, the mapping of addition is simply to map the elements *two* and *two* to an element *four*. This mapping also maps *three* and *one* to the same element *four*. In this context the theory of computation is about studying the ways in which simple mappings can be combined to create complex ones. More complicated functions can map a large set of real numbers into a smaller set that extracts important invariances; for example, by mapping arrays of gray level pixels into a small set of familiar faces, or arrays of sound-pressure levels into a set of comprehensible words. Such mappings can result from the computations performed by our sensory organs, and this dissertation centers on understanding the rules that govern them.

Can we understand how the brain computes?

The extreme difficulty in understanding such complex mappings, is only realized when the relevant quantities are stated. The influx of sensory information to a single human retina is detected by an array of millions of receptors, each capable of telling the difference between hundreds of gray levels, and having time-constants that allow them to detect dozens of new signals in a second. This input is then processed by hundreds of millions of other neurons, many of them interact with each other in complex ways that are constantly changed by the very same inputs we wish to investigate. This architecture is therefore capable of implementing extremely complex maps.

With this gigantic influx, the experimental tools available today are devastatingly weak. The current work uses electrophysiological recordings from small groups of isolated neurons. The data analyzed here were collected from about one hundred neurons only, but required several years of dedicated work done by my collaborators.

With this mismatch between the complexity of the problem and the weakness of the tools, how can we hope to obtain a well established understanding of complex neural systems? The answer lies in the hope that the system adheres to regularities and similarities that simplify the mapping it implements. For example, since neighboring neurons across the neural epithelium are exposed to similar inputs, their functions are expected to share similar properties. This suggests that averages over localized groups of neurons can improve signal to noise issues and allows for extracting coarse maps. Alternatively, developmental and evolutionary considerations can pose additional constraints on the type of maps and computations we may find.

Finally, and this is the approach taken in this thesis, there is hope that these mapping obey some generic *design principles* that guide the type of computation the neural system performs. If such principles exist, we should be able to characterize them more easily than the complex maps themselves, since they will be reflected in multiple subsystems, areas and forms. Moreover, they are expected to embody the functional properties of the neural circuits, which is our ultimate goal in understanding the neural system.

1.1.2 Design principles for sensory computation

The search for design principles that govern the processing performed by sensory systems, was boosted by the appearance of Shannon's information theory in the early 50's. Analogies between sensory systems and communication channels were suggested (Attneave, 1954; Miller, 1956), laying the ground for postulating optimization principles for neural circuits. Although several researchers discussed generic design principles that could underlie sensory processing (e.g. (Barlow, 1959b; Uttley, 1970; Linsker, 1988; Atick & Redlich, 1990; Atick, 1992; Ukrainec & Haykin, 1996; Becker, 1996), and see also chapter 10 in (Haykin, 1999)), I focus here on *Information Maximization* and *Redundancy Reduction*.

Information maximization

The information maximization principle (InfoMax) put forward by Linsker (Linsker, 1988, 1989), suggests that a neural network should tune its circuits to maximize the mutual information between its outputs and inputs. Since the network usually has some prefixed architecture, this amount to a constrained optimization problem for any given set of inputs. This approach was used in (Linsker, 1992) to devise a learning rule for single linear neurons receiving Gaussian inputs. It was extended in (Linsker, 1997) to the case of multiple output neurons utilizing local rules only in the form of lateral inhibition.

While Infomax was originally formulated such that the input-output mutual information maximization is the goal of the system, it was extended to other scenarios. Becker and Hinton (1992,1996) presented *Imax*, one of the important variants of Infomax in which the goal of the system is to maximize the information between the outputs of two neighboring neural networks. They showed how this architecture can be used to extract spatially coherent features in simulations of visual processing. Another variant was presented by (Ukrainec & Haykin, 1996), where the goal of the system was the opposite of that of *Imax*. They showed how minimization of mutual information between outputs of neighboring networks extracts spatially incoherent features, and can be usefully applied to the enhancement of radar images. In (Uttley, 1970) the *Informon* principle was described, where minimization of the input-output mutual information was used as the optimization goal. Such a system becomes discriminatory of the more frequent patterns in the set of signals.

In a paper which is not included in this dissertation (Chechik, 2003), I showed how Infomax can be extended to maximize information between the output of a network and the identity of an input pattern. This setting allows to extract *relevant information* using a simplified learning signal, instead of reproducing the networks inputs. Interestingly the resulting learning rule can be approximated by a spike time dependent plasticity rule.

Redundancy reduction

Redundancies in sensory stimuli were put forward as important for understanding perception since the very early days of information theory (Attneave, 1954). Indeed these redundancies reflect structures in the inputs that allow the brain to build "working models" of its environments. Barlow's specific hypothesis (Attneave, 1954; Barlow, 1959b, 1959a, 1961) was that one of the goals of a neural system is to obtain an efficient representation of the sensory inputs, by compressing its inputs to achieve a parsimonious code. During this compression process, statistical redundancies that are abundant in natural data and therefore also characterize the representation at the receptor level, are filtered out such that the neuronal outputs become statistically independent. This principle was hence named *Redundancy Reduction*. The redundancy reduction hypothesis inspired Atick and Redlich (1990), to postulate the principle of *minimum redundancy* as a formal goal for learning in neural networks. Under some conditions (Nadal & Parga, 1994; Nadal, Brunel, & Parga, 1998) this minimization of redundancy becomes equivalent to maximization of input-output mutual information.

Achieving compressed representations provides several predictions about the nature of the neural code after compression, namely that the number of neurons required is smaller but their firing rates should be higher. The neurophysiological evidence however does not support these predictions: For example, the number of neurons in the lower levels of the visual system is ten times smaller than in the higher ones, and the firing rates in auditory cortex are significantly lower than in the auditory nerve. This suggests that parsimony may not be the primary goal of the system.

Barlow then suggested (Barlow, 2001) that the actual goal of the system is rather *redundancy exploitation*, a process during which the statistical structures in the inputs are removed in a way that reflects the fact that the system used it to identify meaningful objects and structure in the input. These structures are later represented in higher processing levels, a process that again yields a reduction in coding redundancies of higher level elements.

1.2 Information theory

Information theory plays several different roles in the current thesis: both conceptual and methodological. At the methodological level, we use the basic quantities of information theory - such as entropy, mutual information, and redundancy - to quantify properties of the stochastic neural activity. But more importantly, information theory provides a conceptual framework for thinking about design principles of neurally implemented maps. Finally, we also use information theoretic tools to develop methods of unsupervised learning to make sense of the data.

The fundamental concepts of Information theory are reviewed in Appendix B. The reader is referred to (Cover & Thomas, 1991) for a fuller exposition.

1.3 To hear a neural code

In the studies described in this dissertation the main data source were electrophysiological recordings in the auditory system of cats. To understand the findings presented in the main body of the thesis, I now provide a short review of the architecture of this system, both in terms of its gross anatomy and its physiology.

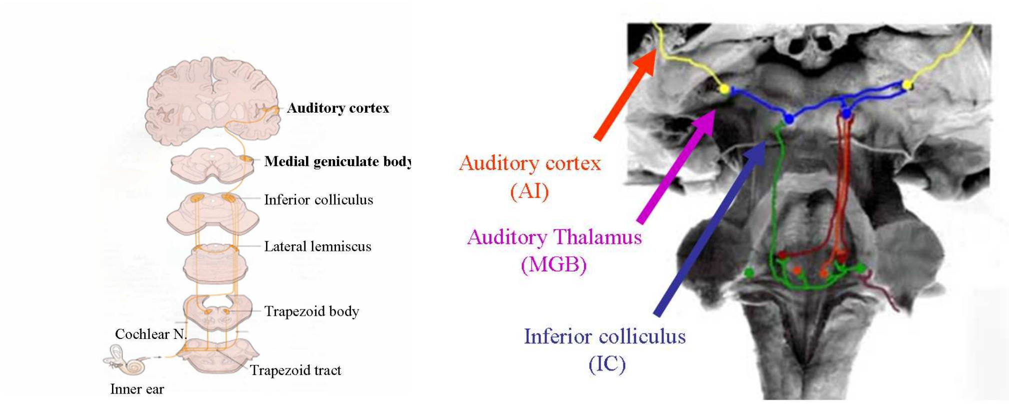

Figure 1.1: **Left.** An illustration of the anatomy of the mammalian auditory system. **Right.** A cross section of a human brain, on which the auditory pathway is marked. The three auditory processing stations analyzed in this work are designated: IC, MGB and AI.

1.3.1 Gross anatomy of the auditory system

This thesis focuses on the core pathway of the auditory system in mammals. This pathway consists of several neural processing stations: the 8th (auditory) nerve, the cochlear nucleus (CN), the superior olivary complex (SOC) and the nuclei of the lateral lemniscus (NLL), the inferior colliculus (IC), the medial geniculate body of the thalamus (MGB), and the primary auditory cortex (AI) (Popper & Fay, 1992). An illustration of the mammalian auditory system is presented in Figs. 1.1.

In addition to the ascending system, there is also a strong descending information flow, where the major descending pathway projects from the cortex to the thalamus and IC, from IC to lower centers and finally from sub-nuclei of the SOC to the cochlear nucleus and to the cochlea (Spangler & Warr, 1991).

The next subsections briefly review some of the main functional properties of the processing stations of the core pathway, and provide a few examples of raw data later used in the analysis presented in the main chapters of the thesis. Aspects of localization or binaural processing are not discussed here, and the interested reader is referred to (Middlebrooks, Xu, Furukawa, & Mickey, 2002).

1.3.2 Auditory nerve fibers

Auditory nerve fibers project information from the auditory receptors (the hair cells of the cochlea) into the cochlear nucleus, which is the first processing station of acoustic stimuli[1]. To characterize the spectral sensitivity of an auditory nerve fiber, pure tones at different frequencies and amplitudes are presented to an animal and the response of the fiber is recorded. At every frequency, the minimal sound level that elicits a

[1]The first synapse in the pathway is between the hair cells and the auditory nerve fibers, and the second synapse is in the CN.

significant response is recorded, resulting in a frequency tuning curve, an example of which is presented in Fig. 1.2. It shows that the typical frequency tuning curve consists of a fairly narrow frequency band to which the neurons are sensitive. The frequency that has the lowest threshold is called the *best frequency* (BF).

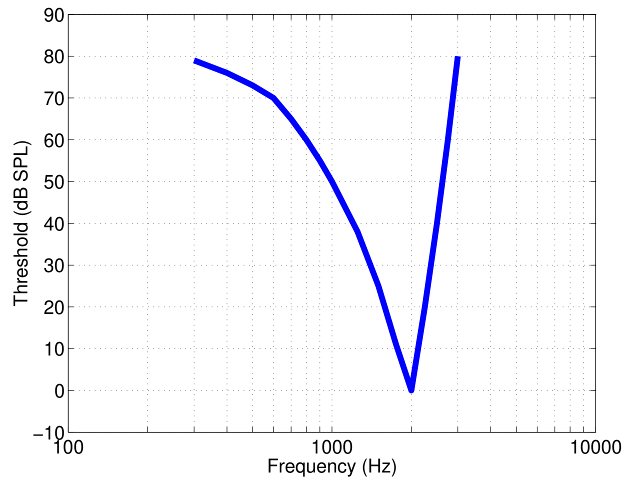

Figure 1.2: **A.** A typical frequency tuning curve of an 8^{th} nerve. It is sensitive to a band of frequencies only few kHz wide. Reproduced from publicly available data

In spite of the presence of strong non-linearities in the responses of auditory nerve fibers, the responses of the population of auditory nerve fibers can be reasonably well described to a first approximation as the output of a band-pass filter bank. A useful model of the responses of these cells is with Gamma-tone filters, where the BF's of the cells are homogeneously spaces along a logarithmically scaled frequency axis. Figure 1.3 depicts an example of such a set of filters.

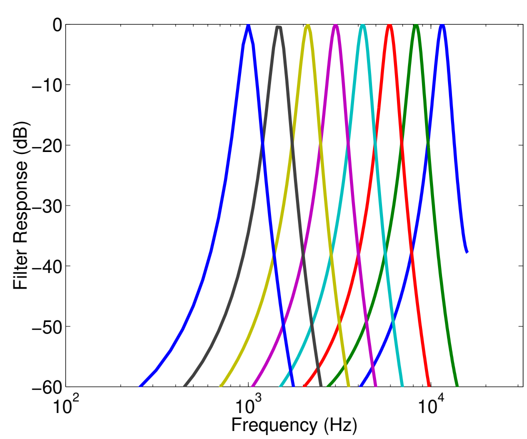

Figure 1.3: The filter coefficients for a bank of Gamma-tone filters. Taken from the auditory tool box by [Slaney, 1998]. Filters were designed by Patterson and Holdworth for simulating the cochlea.

Most interestingly, when auditory nerve fibers are probed with complex sounds such as bird chirps, their response profile can be well explained by their frequency profiles. This is demonstrated in the activity of a neuron from the ventral cochlear nucleus in Fig. 1.4. Whenever the stimulus (middle panel) contains energy in the range of frequencies within the neuron's tuning curve (left panel) as depicted with black horizontal lines, a significant rise in firing rate is observed (lower panel).

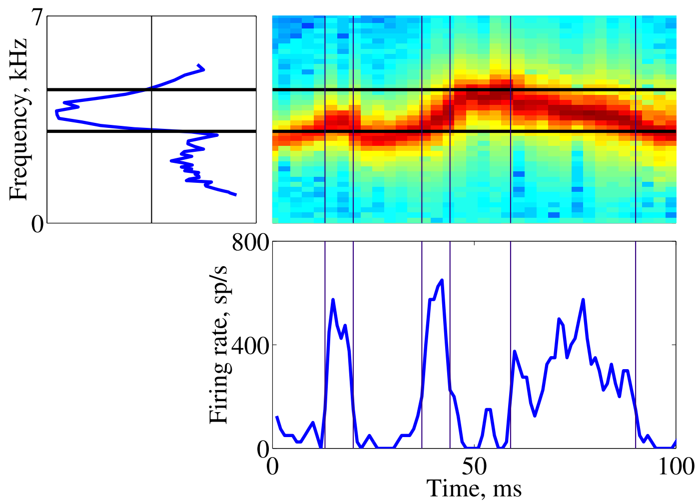

Figure 1.4: Responses of a primary like neuron in the ventral cochlear nucleus, whose behavior is also typical of an auditory nerve fiber. **Left:** A tuning curve. The blue line denotes the minimal level at which a significant response is observed. **Right:** A spectrogram of a bird chirp. Horizontal lines depict the range of frequencies for which the neuron is sensitive. **Bottom:** Firing rate in responses to the presentation of the bird chirp.

Figure 1.5 shows post stimulus time histograms (PSTH) of a model neuron in our data set, as a response to the presentation of a natural bird chirp. The behavior of this model neuron is similar to the recorded ones, in the sense that a coarse but good prediction of the responses to complex sounds can be obtained from the frequency characterization.

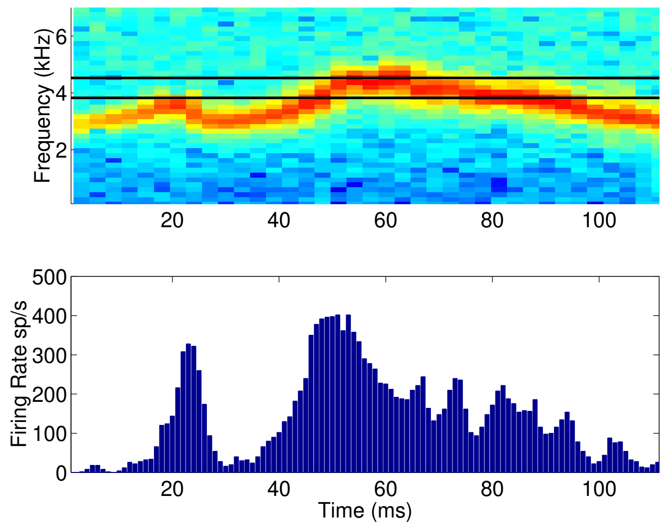

Figure 1.5: Post stimulus time histogram of the responses of a model 8th nerve neuron from our data set, created to have a best frequency at 4.5 kHz.

1.3.3 Inferior colliculus

The inferior colliculus (IC) is an obligatory station of the auditory pathway. All the separate brainstem pathways (from the CN, the SOC and the NLL) terminate in the IC. In addition, the IC receives input from the contralateral IC, descending inputs from the auditory cortex and even somatosensory inputs. As in other auditory areas, IC neurons are frequency sensitive, and exhibit a tonotopic organization within the IC. Moreover the best frequencies of neurons are arranged in an orderly manner, in a way that is fairly well preserved across mammalian species. Most interestingly, the inputs from the multiple origins converge in an arrangement that corresponds to the tonotopic organization of the IC. The IC therefore preserves the same tonotopic map for multiple converging inputs, allowing for complex integration of information within a localized area of the iso-frequency sheet. Not much is known about the organization orthogonal to the frequency gradient, although there is strong evidence for functional gradients related to temporal characteristics, such as best modulation frequency (Schreiner & Langner, 1997) and latency. IC neurons exhibit a rich spectrum of frequency sensitivities, some are sharply tuned to frequencies while some respond to broad-band noise. Some shaping of the frequency tuning is achieved by lateral inhibition (Ehret & Merzenich, 1988) , and some by other mechanisms (Palombi & Caspary, 1996). Many IC neurons are also selective to temporal structures. Temporal processing in IC include selectivity to sound durations, delays, frequency modulated sounds and more (see section 3.3 in (Casseday, Fremouw, & Covey, 2002) for detailed review). Despite all this, there is still no satisfying description of IC organization in terms of ordered functional maps.

Figure 1.6 presents responses of a typical IC neuron analysed in the current work.

When presented with bird chirps, IC response tended to be locked to some features of the stimuli, as indicated by the reliable and precise nature of spikes revealed across repeated stimulus presentations.

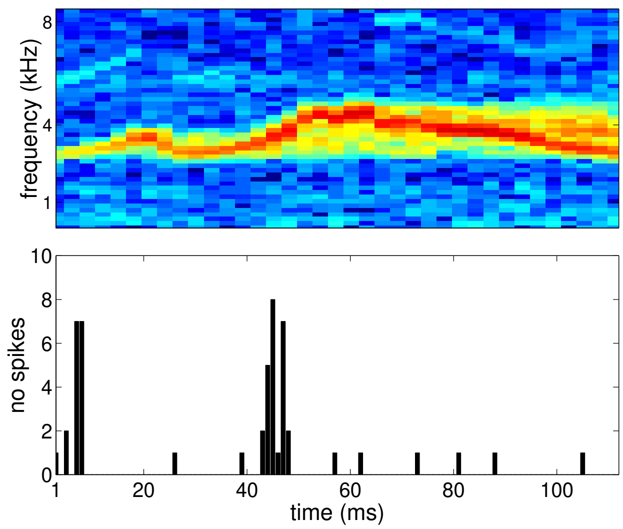

Figure 1.6: Post stimulus time histogram of the responses of an IC neuron from our data set. Notice the tight and precise locking of responses to the stimulus.

1.3.4 Spectro temporal receptive fields in auditory cortex

The auditory cortex is a focus of special interest in this work, since it is the highest processing station we investigated and presumably contains the most complex response properties.

Frequency sensitivity, as characterized with pure tones, reveals that many cortical neurons have a narrow frequency tuning curve, limited dynamic range and often are non monotonic in their responses. Cortical neurons show strong sensitivity to the shape of the tone onset, a dependence that is currently well understood (Heil & Irvine, 1996; Heil, 1997; Fishbach, Nelken, & Yeshurun, 2001).

When using more complex stimuli, the picture becomes drastically more complex. While FRA characterization could be used to obtain a good description of responses to complex sounds in ANF, this is no longer the case for cortical neurons. The FRA measured by pure tones fails to capture two important aspects of cortical processing: integration across frequencies, and sensitivity to temporal structures (Nelken, 2002).

It was suggested that a better model of cortical responses can be obtained by deriving a spectro temporal receptive field (STRF), an approach that was found useful for characterizing auditory neurons in several systems (e.g. (Aertsen & Johannesma, 1981; Eggermont, Johannesma, & Aertsen, 1983)). deCharms and colleagues (DeCharms, Blake, & Merzenick, 1998) used short random combinations of pure tones and a spike triggered averaging analysis to obtain the STRF of auditory cortical neurons in monkeys. The resulting receptive fields, demonstrated in Fig. 1.7, show complex dependencies between time and frequency, suggesting that cortical neurons are sensitive to frequency changes, as in FM sweeps. However, this type of analysis is linear in the sense that it averages the energy in spectro-temporal "pixels" while assuming independence between pixels, and it is therefore limited in its ability to capture complex interactions between frequencies and temporal structures.

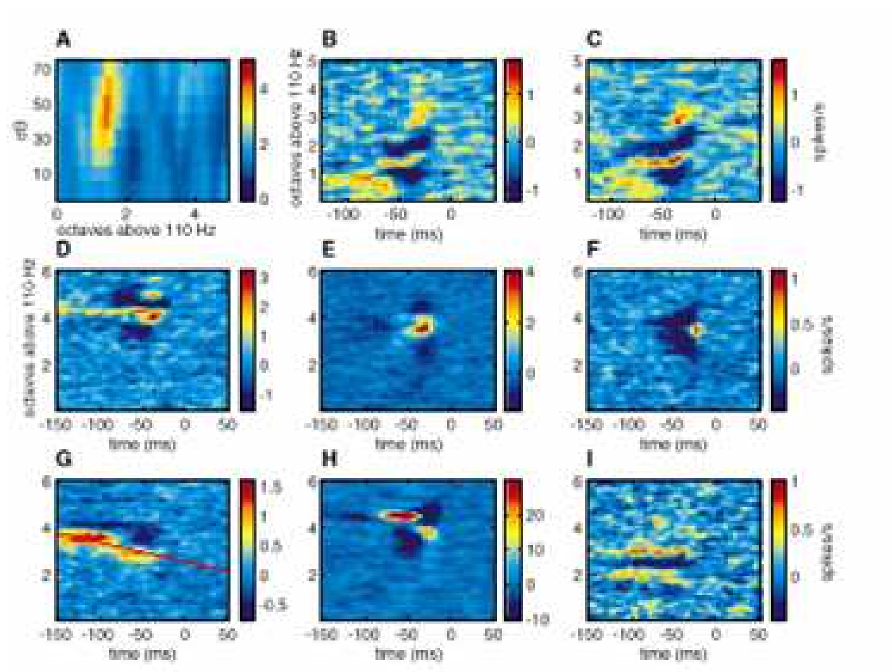

Figure 1.7: **A.** Frequency response area (FRA) of a typical cell. Notice the non monotonic response as a function of level. **B.-I.** spectro temporal receptive fields of different cells. B and C are STRFs estimated from the responses of the neuron in A using different sets of random chords. From [deCharms 1998] .

A striking demonstration of such nonlinear interactions was observed in cortical responses to natural and modified bird chirps (Bar-Yosef, Rotman, & Nelken, 2001). Bar Yosef and colleagues showed that relatively minor modifications of the stimulus, such as the removal of the background noise from a natural recording, could dramatically alter the responses of cortical neurons. This type of behavior cannot be explained using linear combinations of STRF's. These results are discussed together with the set of stimuli we used, in the next section.

1.3.5 The stimulus set: To hear a mocking bird

In order to study changes in stimulus representation along the processing pathway, one should use a set of stimuli whose processing is not limited to low level processing stations, otherwise, the properties of high level representation will only reflect low level rather than high level processing. Auditory neurons are often characterized by their spectro temporal properties, however, since the exact features for which cortical neurons are sensitive to are still unkown, we chose to use here a stimulus set, that contains several natural stimuli, which contain rich structures in terms of frequency spectrum and modulation. In addition, we added several variants of these stimuli that share some of the spectro temporal structures that appear in the natural stimuli. This set of stimuli is expected to yield redundant representations at the auditory periphery, and is therefore suitable for the investigation of informational redundancy

The stimulus set used here was created by O. Bar-Yosef and I. Nelken and is described in details in (Bar-Yosef et al., 2001). It is based on natural recordings of isolated bird chirps, whose sound wave and spectrograms are depicted in Fig. 1.8.

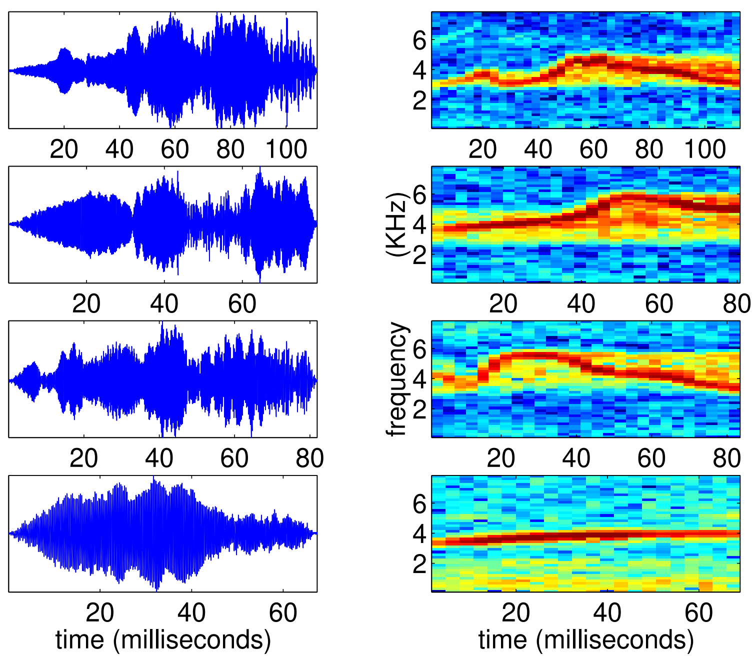

Figure 1.8: Four natural recordings of bird chirps. For each chirp, the left panel shows its sound wave and the right panel its spectrogram.

Each natural recording was then separated into two components: the *main* chirp and the background. The background was further separated into the *echo* component, and the rest of the signal, termed *noise*. These components were then combined into several configurations (*main+echo*, *main + background*). In addition, an artificial stimulus that follows the main FM sweep of the chirp was also created (termed *artificial*). The variants based on the first bird chirp are depicted in Fig. 1.9

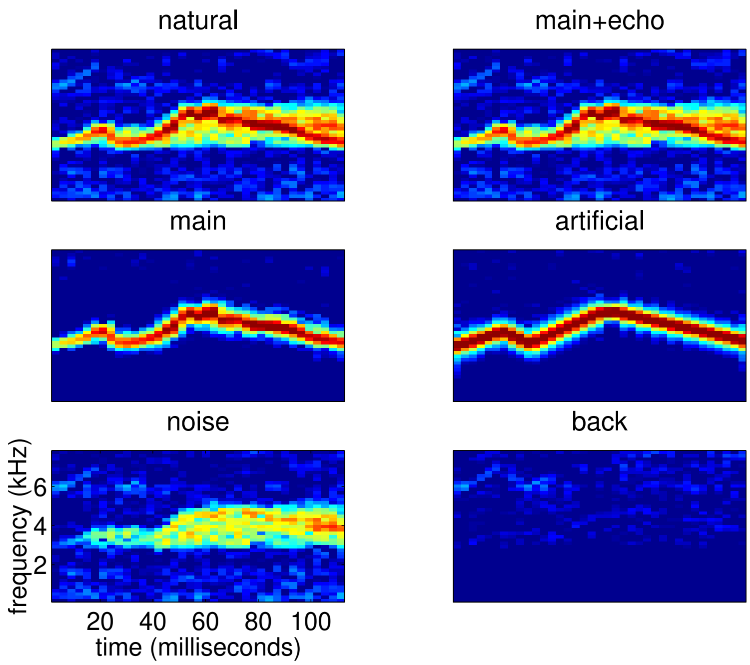

Figure 1.9: Six different variants created from a single natural bird chirp (upper panel in the previous figure) .

In some of the analyses, 32 different stimuli, based on 8 variants of 4 different bird chirps were used. In others, 15 stimuli, based on 5 variants of 3 bird chirps were used. These 15 stimuli are plot below

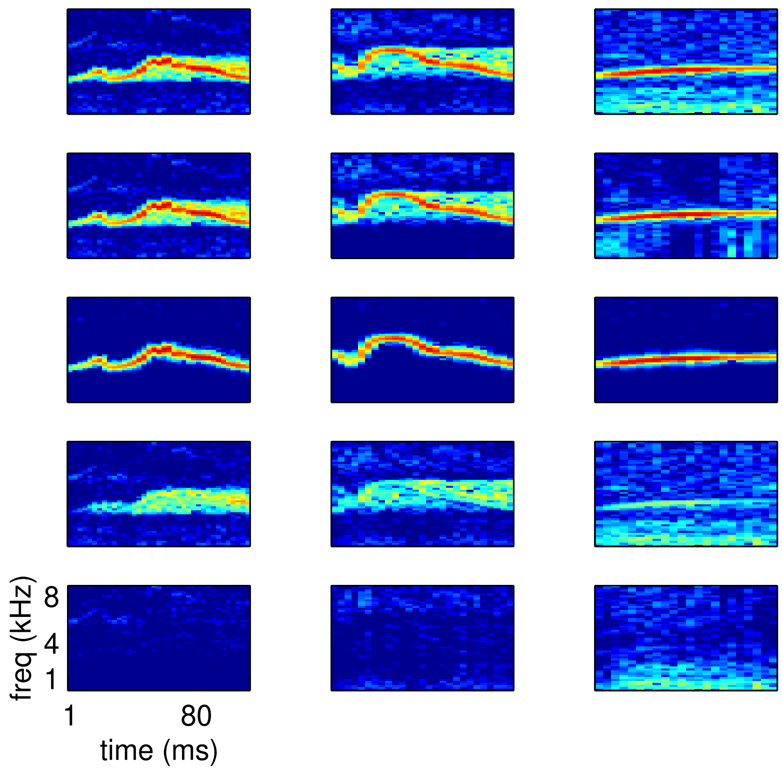

Figure 1.10: A set of 15 stimuli created from three different bird chirps..

1.3.6 The experimental setup

The electrphysiological recordings that provided the data that are analyzed in this work were performed in two laboratories. First is the laboratory of Prof. Eric Young at Johns Hopkins University, Boltimore, where electrophysiological recordings were done in the IC by Dr. Mike Anderson and Prof. Young. Secondly, the lab of Dr. Israel Nelken at Hadassah Medical School of the Hebrew University in Jerusalem, where recordings were conducted in the auditory cortex, the auditory thalamus and the inferior colliculus by Omer Bar-Yosef, Dina Farkas, Liora Las, Nachum Ulanovski and Dr. Nelken.

A detailed description of the experimental methods is given in (Bar-Yosef et al., 2001). In what follows, a brief review of these is provided.

Animal preparation

Extracellular recordings were made in primary auditory cortex of nine halothane-anesthetized cats, in the medial geniculate body of two halothane- anesthetized cats and inferior colliculus of nine isoflurane-anesthetized and two halothane-anesthetized cats. Anesthesia was induced by ketamine and xylazine and maintained with halothane (0.25-1.5 percent, all cortex and MGB cats, and 2 IC cats) or isoflurane (0.1-2 percent 9 IC cats) in 70 percent N_2O. Breathing rate, quality, and CO_2 levels were continuously monitored. In case of respiratory resistance, the cat was paralyzed with pancuronium bromide (0.05-0.2 mg given every 1-5 hr, as needed) or vecuronium bromide (0.25 mg given every 0.5-2 hr). Cats were Anesthesized using standard protocols authorized by the committee for animal care and ethics of the Hebrew University - Hadassah Medical School (AI, MGB and IC recordings) and Johns Hopkins University (IC recordings).

Electrophysiological recordings

Single neurons were recorded using one to four glass-insulated tungsten microelectrodes micro-electrodes. Each electrode was independently and remotely manipulated using a hydraulic drive (Kopf) or a four-electrode electric drive (EPS; Alpha-Omega, Nazareth, Israel). The electrical signal was amplified (MCP8000; Alpha-Omega) and filtered between 200 Hz and 10 kHz. The spikes were sorted online using a spike sorter (MSD; Alpha-Omega) or a Schmitt trigger. All neurons were well separated. The system was controlled by a master computer, which determined the stimuli, collected and displayed the data on-line, and wrote the data to files for off-line analysis. MGB neurons were further sorted off line

Most of the neurons whose analysis is described below were not recorded simultanousely.

Acoustic stimulation

The cat was placed in a soundproof room (Industrial Acoustics Company 1202). Artificial stimuli were generated digitally at a rate of 120 kHz, converted to analog voltage (DA3-4; Tucker-Davis Technologies), attenuated (PA4; Tucker-Davis Technologies), and electronically switched with a linear ramp (SW2; Tucker-Davis Technologies). Natural stimuli and their modifications were prepared as digital sound files and presented in the same way, except that the sampling rate was 44.1 kHz. Stimuli were delivered through a sealed calibrated acoustic system (Sokolich) to the tympanic membrane. Calibration was performed in situ by probe microphones (Knowles) precalibrated relative to a Brüel and Kjær microphone. The system had a flat (± 10 dB) response between 100 Hz and 30 kHz. In the relevant frequency range for this experiment (2-7 kHz), the system was even flatter (the response varied by less than ± 5 dB in all but one experiment, in which the variation was ± 8 dB). These changes consisted of relatively slow fluctuations as function of frequency, without sharp peaks or notches.

Anatomical appraoch

In AI, penetrations were performed over the whole dorso-ventral extent of the appropriate frequency slab (between about 2 and 8 kHz). In MGB, all penetrations were vertical, traversing a number of iso-frequency laminae, and most recording locations were localized in the ventral division. In IC vertical penetrations were used in all experiments except one, in which electrode penetrations were performed at a shallow angle through the cerebellum, traversing the nucleus in a caudo-rostral axis. We tried to map the full medio-lateral extent of the nucleus, but in each animal only a small number of electrode penetrations were performed. Based on the sequence of best frequencies along the track, the IC recordings are most likely in the central nucleus.

1.4 Summary of our approach

With over 100 years of neuroscience research using electrophysiological experiments, how can we hope to innovate, and gain a deeper understanding of the sensory systems?

Our approach is based on combining several ingredients. First, we use natural and complex stimuli, reflecting our belief that interesting properties of high level processing (presumably taking place in the auditory cortex) can be revealed in the responses to such stimuli. Such properties however cannot be discovered using standard linear methods.

Secondly, electrophysiological recordings from a series of auditory processing stations allows us to compare the representations of these complex stimuli, and the way they change along the processing hierarchy, thus reflecting the computational processes that the system applies. Our goal is to identify design principles that underlie the changes in these representations.

Thirdly, we use information theoretic measures to quantify how auditory cells interact to represent stimuli, and develop information theoretic methods to study what the cells represent.

Our belief is that the combination of these components can reveal novel evidence about the principles that underly auditory neural coding.

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 & 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 & 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, Shafee, & Bialek, 2002), even though the log function in principle requires estimating an infinite number of moments (Paninski, 2003).

The estimation of MI from a finite sample involves an important **tradeoff 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

[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, Rieke, Steveninck, & Warland, 1991; Steveninck, lewen, Strong, Koberele, & Bialek, 1997)).

[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.

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, Tibshirani, & Friedman, 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.

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 $\theta \rightarrow X^n \rightarrow T$, 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 $\theta \rightarrow T \rightarrow X^n$ 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.

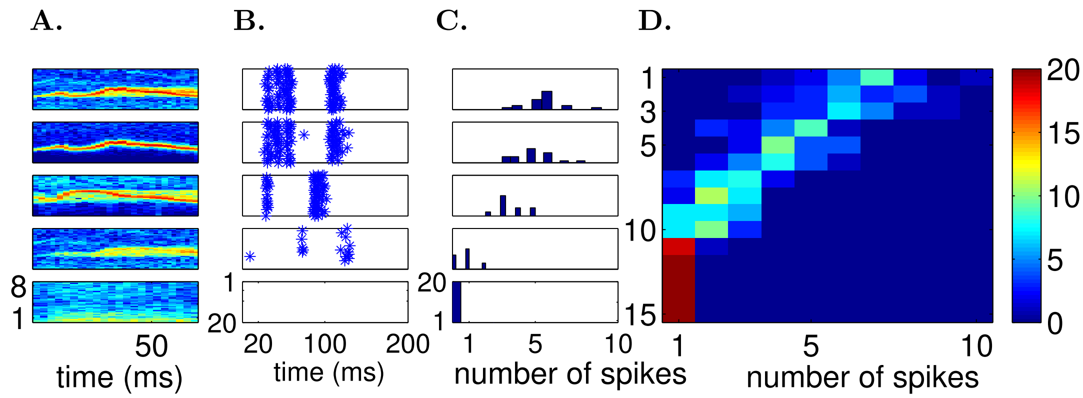

Figure 2.1: An illustrative example of estimating the mutual information in spike counts using naive binning. **A.** Spectrogram of five stimuli. **B.** Raster plots of neural activity in response to 20 presentations of each of the five stimuli. **C.** Distribution of spike counts for each stimulus. **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).

[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].

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 $\hat{I}$.

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).

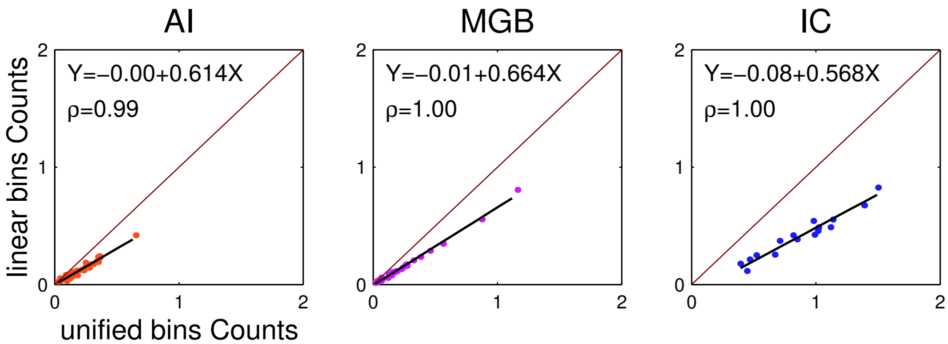

Figure 2.3: Comparison between naive (linear) binning and data dependent binning that operates to preserve homogeneous marginals. Both methods use spike counts. Each point corresponds to a different cell. The red line is the $y = x$ curve. The black line is the regression curve (adjusted for sample size), 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, Chechik, King, & Schnupp, 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

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

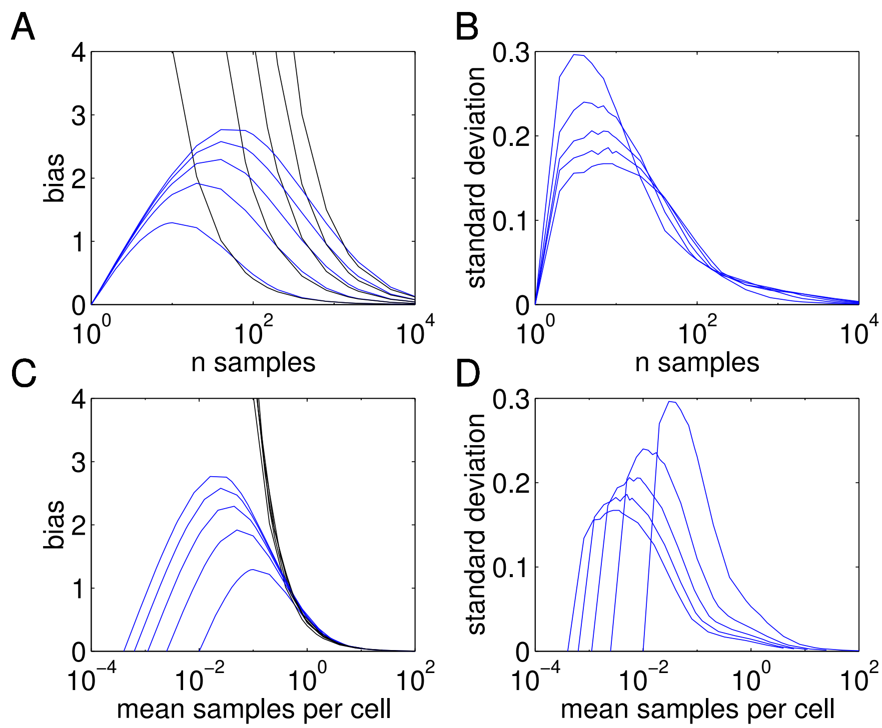

Figure 2.4: **A.** Bias and **B.** variance of the MI estimator as a function of sample size. **C.** Bias and **D.** variance of the MI estimator as a function of mean number of samples per cell. Blue curves correspond to different matrix sizes k = 10, 20, 30, 40 and 50. Black curves depict the bias 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 & Panzeri, 1995; Panzeri & 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

$$E(I_n) = I + \sum_{i=1}^{\infty} C_i \tag{2.14}$$

with

$$\begin{aligned}
C_1 &= \frac{(|S|-1)(|R|-1)}{2n\log(2)} \\
C_2 &= \frac{\sum_s (p_n(s))^{-1} \sum_r \left[(p(r|s))^{-1} - 1\right] - \sum_r (p(r))^{-2} + 1}{12n^2\log(2)} \\
C_3 &= \frac{\sum_s (p_n(s))^{-2} \left[\sum_r (p(r|s))^{-2} - (p(r|s))^{-1}\right]}{12n^3\log(2)} \\
&\quad - \frac{\sum_r \left[(p(r))^{-2} - (p(r))^{-1}\right]}{12n^3\log(2)}
\end{aligned}$$

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 & 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 & 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 & 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 & Panzeri, 1995), (Panzeri & 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.

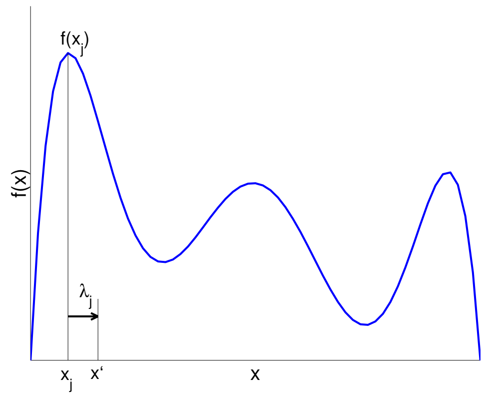

Figure 2.5: Illustration of the relation between the density $f(x_j)$ at some point x_j and the distance to its nearest neighbor x‘.

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

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

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

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

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.

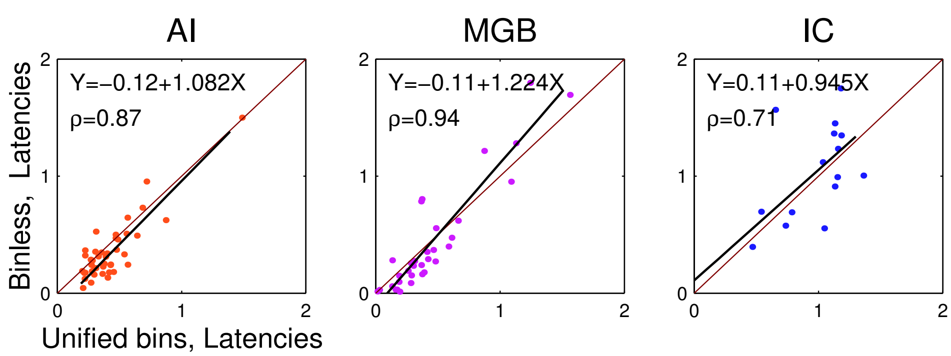

Figure 2.6: Comparison between metric based and histogram based estimation of MI in first spike latency in data from three different brain regions.

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 & 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 & 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 (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 & Irvine, 1996; Heil, 1997; Jenison, 2001; Krishna & Semple, 2001).

Thorpe and colleagues (Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001; Fabre-Thorpe, Delorme, Marlot, & Thorpe, 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.

2.3.4 The direct method

The direct method, presented in (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 & 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

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

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.

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

The second order term consists of three components

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

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

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

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

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

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.

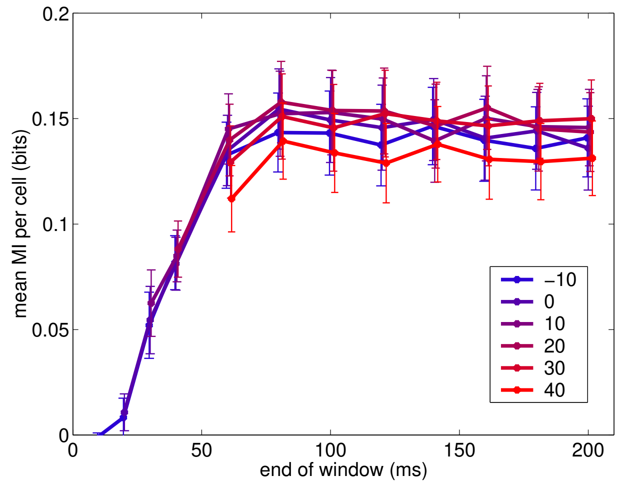

Figure 2.7: Mutual information between the stimulus identity and the spike counts of 45 AI cells, for different windows in which spikes are considered. Different curves correspond to different start point of the window after stimulus onset. MI was estimated using the unified-bins procedures. Similar 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.

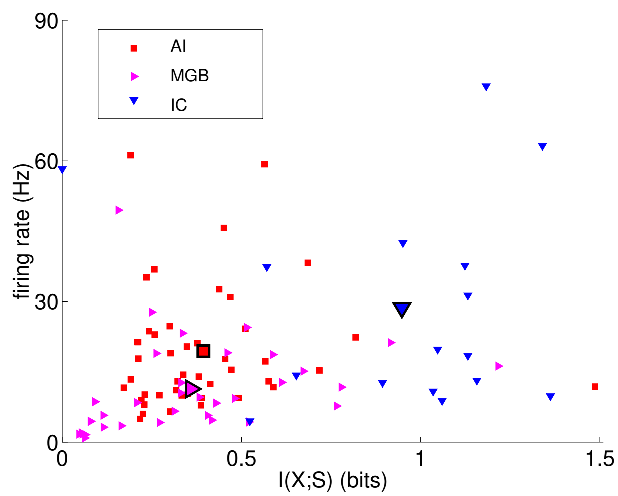

Figure 2.8: The firing rate vs the MI estimated with spike counts in three braing regions. Large markers denote the mean of each population. Firing rates and MI were calculated over the most informative time window of each population.

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.

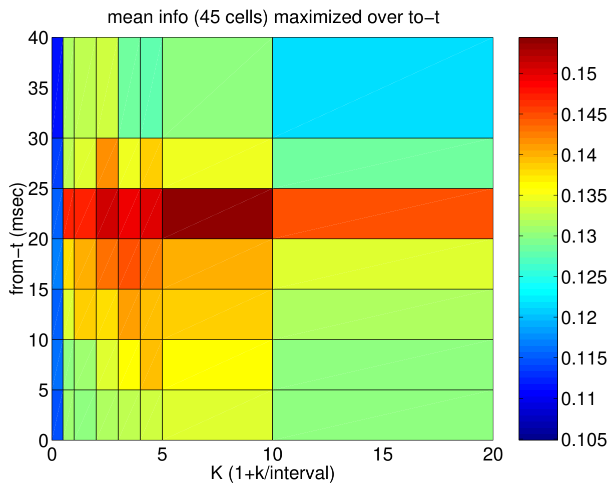

Figure 2.9: Mutual information conveyed by the value of the sufficient statistic for the inverse-Gaussian process, as a function of the parameter K and the window starting point after stimulus onset. MI was estimated using the unified-bins procedure. Results are the mean over 45 AI cells.

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.

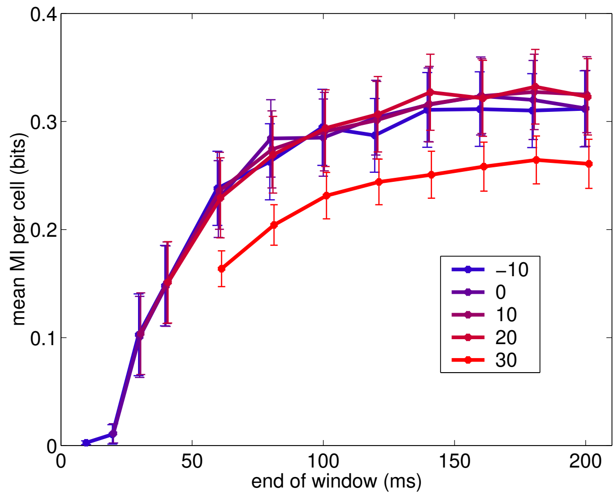

Figure 2.10: Mutual information conveyed by the first spike latencies of 45 AI cells, for different windows in which spikes are considered. Different curves stand for different start points of the window after stimulus onset. MI was estimated using the unified-bins procedure. Trials with no spikes were 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.

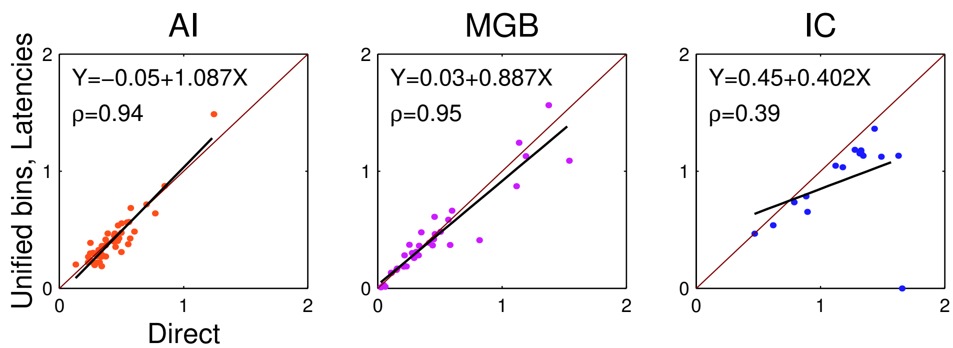

Figure 2.11: MI estimated using the direct method and using first spike latency in three brain regions.

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.

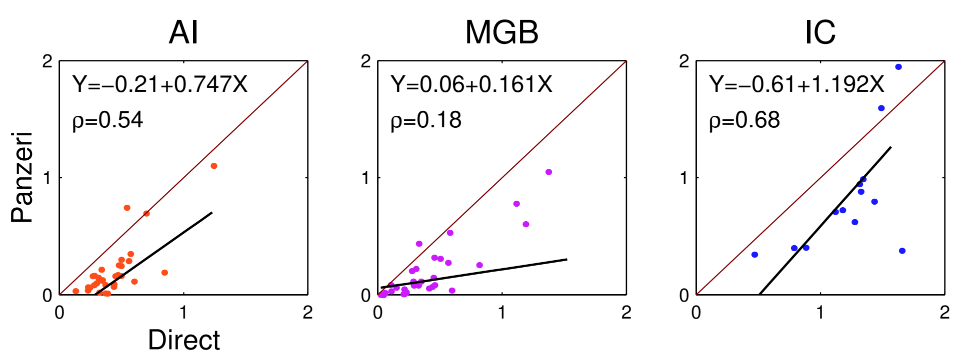

Figure 2.12: MI estimated using second order correlations vs the direct method in three brain regions.

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).

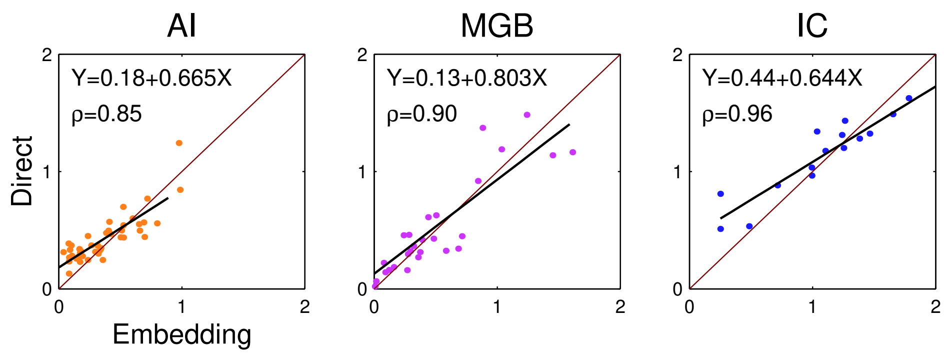

Figure 2.13: MI estimated using the direct method vs the Legendre polynomials embedding in three brain regions.

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.

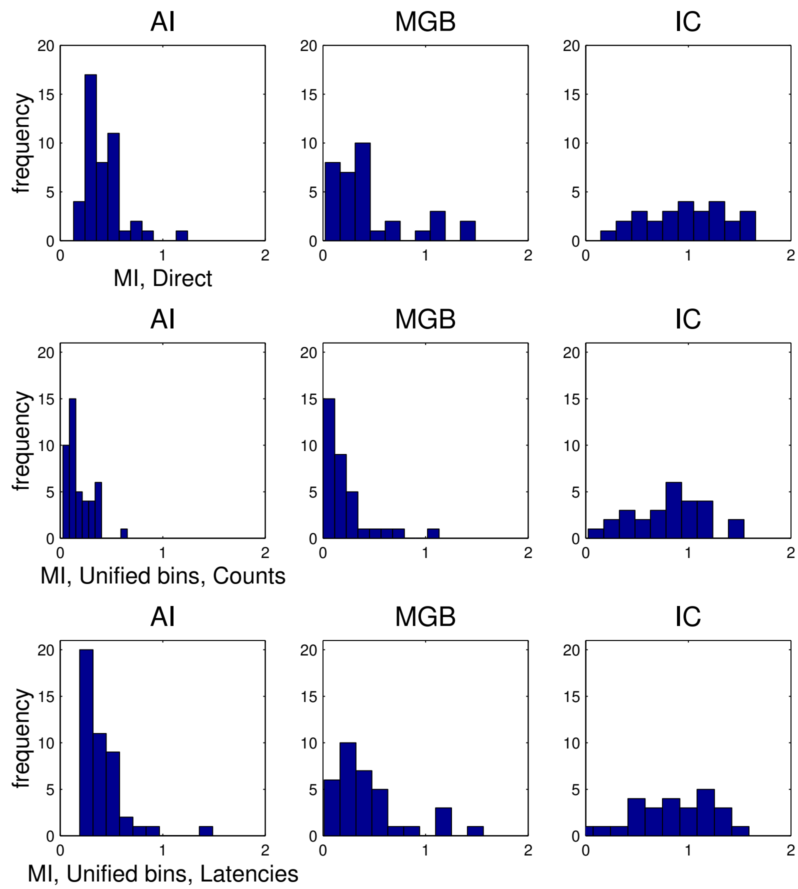

Figure 2.14: Comparison of mean MI obtained in three brain regions using four different methods. Error bars are standard errors of the mean over the population of cells.

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.

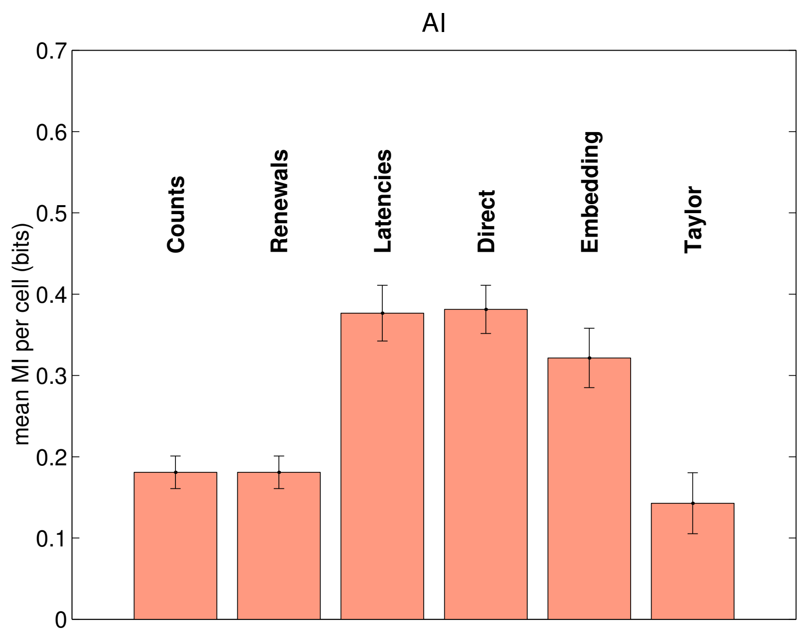

Figure 2.15: Mean MI obtained from all 45 cells using six different methods.

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.

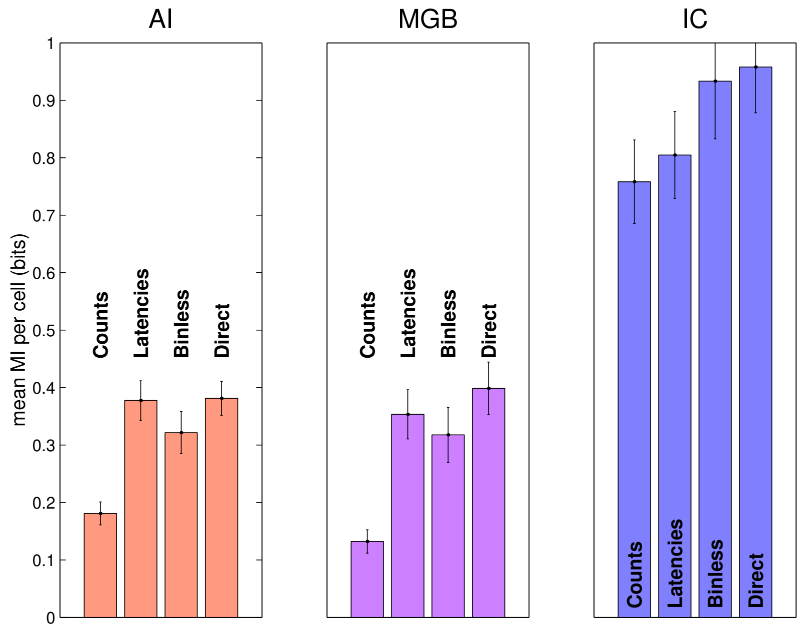

Figure 2.16: Comparison of mean MI obtained in three brain regions using four different methods. Error bars are standard errors of the mean over the population of cells.

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.

Chapter 3

Quantifying Coding Interactions

The previous chapter discussed in detail methods for estimating the information that spike trains of *single neurons* convey about stimuli. The current chapter describes an information theoretic approach to the study of high order correlations among small groups of neurons in the auditory system. Section 3.1 reviews previous studies based on the information theoretic approach to this problem, and section 3.2 discusses methodological issues involved in measuring and estimating synergy and redundancy. The application of these methods to neural coding in the auditory system are presented in the next chapter.

3.1 Previous work

Several investigators have used information theoretic measures for quantifying high order correlations in the activity of sets of neurons. These were mostly performed in the visual system, but also in high brain areas (Gawne & Richmond, 1993; Meister, 1996; Warland, Reinagel, & Meister, 1997; Rolls, Treves, & Tovee, 1997; Dan, Alonso, Usrey, & Reid, 1998; Gat & Tishby, 1999; Brenner, Strong, Koberle, Steveninck, & Bialek, 2000; Nirenberg, Carcieri, Jacobs, & Latham, 2001; Reich, Mechler, & Victor, 2001).

Gawne and Richmond (1993) defined information theoretic measures to quantify correlations in pairs of inferior temporal visual neurons. They found both independent, redundant and synergistic pairs, and concluded that IT cortical pairs were not organized in clusters of similar-properties neurons, but were heterogeneous and more independent than redundant.

Rolls, Treves and Tovee (1997), measured the information conveyed by set of neurons of varying sizes, observing an almost linear growth of information as a function of number of neurons, as expected under a distributed coding scheme.

Warland and colleagues (Warland et al., 1997) compared information extracted using the method of (Bialek et al., 1991) using two retinal ganglion cells with the information conveyed when neurons were considered independently. Their main finding was that the traditional classification into ON/OFF types determines the synergistic/independence nature of neuronal correlations.

Gat and Tishby (1999) measured synergy and redundancy in pre-motor cortical neurons, in the context of Synfire chains. Their major finding was high synergy values that were observed in cortical but not in basal ganglia neurons.

Brenner and colleagues (Brenner et al., 2000) measured information carried by compound events in spike trains of single neurons, and measured synergy and redundancy in their code. They showed how pairs of spikes can be synergistic or redundant in a way that depends on their time interval.

Recently, Nirenberg and colleagues (Nirenberg et al., 2001) measured information levels in pairs of isolated ganglion neurons from the mouse retina. The main goal of their work was to estimate the relative weight of information transmitted by temporal correlations activity in coding of natural images. They found that more than 90 percent of the information could be extracted from the neurons when their temporal correlations were ignored. One should note however that their conclusion, "retinal ganglion cells act largely as independent encoders" contradicts earlier studies (Meister, Lagnado, & Baylor, 1995; Meister, 1996). Regarding the issues discussed in the current chapter, it worth emphasizing that Nirenberg and colleagues define *excess correlation function* (ECF) as a measure of the information lost when estimating information under stimulus conditioned independence $P(r_1, r_2|s) = P(r_1|s)P(r_2|s)$, and find that the ECF is small in their preparation.

Finally, Reich and colleagues (Reich et al., 2001) measured redundancy levels in clusters of up to six simultaneously recorded neurons in primary visual cortex. Their main finding is that responses were almost independent under a *labeled lines* coding scheme, that is, when keeping track which neuron fired which spikes, but were redundant if responses were summed over clusters. This suggests that summing neuronal responses in a naive manner (for example in order to reduce variability) discards important information about the stimuli.

The common objective of this series of studies is to identify the way neurons interact to convey sensory information. This goal yields further refined questions: Are nearby neurons synergistic, independent or redundant? How should these qualitatively different behaviors be measured and quantified? Do these properties change along the ascending sensory pathways? Are they stimulus dependent? Perhaps most importantly, can such interactions be used by a readout mechanism and account for behavioral changes?

The current chapter aims at answering some of these questions. It has two main goals. First, from the methodological point of view, it will systematically define information theoretic measures of correlations between groups of cells, and discuss how these should be reliably estimated and compared. Then, from a scientific point of view, we will use these methods for the analysis of auditory neurons responses, to discover principles governing interactions between neurons in the auditory pathway. The unique nature of our data, namely neural responses from three brain regions to the same set of stimuli, allows us to compare the way typical correlations change along the auditory pathway.

3.2 Measures of synergy and redundancy

3.2.1 Preliminaries: synergy and redundancy in pairs

Let X_1 and X_2 be a pair of neurons conveying information about a stimulus S. Their synergy-redundancy measure is commonly defined as the difference between the amount of information that can be obtained when the two neurons area are considered jointly and the information obtained when they are considered individually:

$$SR_{pairs}(X_1, X_2, S) = I(X_1, X_2; S) - [I(X_1; S) + I(X_2; S)] \tag{3.1}$$

Intuitively, SR_{pairs} measures the amount of information on the stimulus S gained by observing the joint distribution of both X_1 and X_2, as compared to observing each of the two cells independently. In the extreme case where $X_1 = X_2$, the two cells are completely redundant and provide the same information about the stimulus, yielding $SR_{pairs} = I(X_1, X_2; S) - I(X_1; S) - I(X_2; S) = -I(X_1; S)$, which is always non-positive. On the other hand, positive SR_{pairs} values testify to synergistic interaction between X_1 and X_2. For example, let X_1 and X_2 be fair Bernoulli variables and S their sum modulu 2 ,$S = X_1 \oplus X_2$. In this case any isolated variable X conveys zero information about S while knowing their joint value provides all the information about S.

Although the SR_{pairs} measure is defined in an asymmetric way, it obeys the following property

Lemma 3.2.1 : *The SR_{pairs} is symmetric with respect to all three variables*

$$\begin{aligned} SR_{pairs} &= I(X, Y; Z) - I(X; Z) - I(Y; Z) \\ &= I(Z, Y; X) - I(Z; X) - I(Y; X) \\ &= I(X, Z; Y) - I(X; Y) - I(Z; Y) \end{aligned} \tag{3.2}$$

Proof: First, use the chain rule for the mutual information (section A.2.6) $I(X, Y; Z) = I(X; Z) + I(Y; Z|X)$ and write

$$I(X, Y; Z) - I(X; Z) - I(Y; Z) = I(Y; Z|X) - I(Y; Z) \ . \tag{3.3}$$

Secondly, from the symmetry in the MI definition $I(X, Y; Z) = I(Y, X; Z)$, we obtain $I(X, Y; Z) = I(X; Z|Y) - I(X; Z)$. To complete the triple symmetry, we write

$$I(X, Y; Z) \quad - \quad I(X; Z) - I(Y; Z) \tag{3.4}$$

$$
\begin{aligned}
&= H(X,Y) - H(X,Y|Z) - H(X) + H(X|Z) \\
&\quad -H(Y) + H(Y|Z) \\
&= I(X;Y|Z) - I(X;Y)
\end{aligned}
$$

□

As a conclusion from this derivation, the measure SR_{pairs} can be written as the difference of two MI terms, a stimulus conditioned term and an unconditioned one

$$
\begin{aligned}
SR_{pairs} &= I(X_1, X_2; S) - [I(X_1; S) + I(X_2; S)] \\
&= I(X_1; X_2|S) - I(X_1; X_2)
\end{aligned}
\tag{3.5}
$$

The first term in the last expression is commonly referred to in the literature as *noise correlations* (Gawne & Richmond, 1993; Panzeri, Schultz, Treves, & Rolls, 1999), measuring the level of dependence given the stimuli, while the second is known as *signal correlations*, measuring the dependencies induced by the different stimuli. This representation makes it easier to see how SR_{pairs} can assume both positive and negative values. Since the positive values of SR_{pairs} are commonly interpreted as synergy, and the negative values as redundancy, we naturally treat $I(X_1, X_2|S)$ as measuring the synergistic interaction of the pair, while $I(X_1; X_2)$ measures their redundancy interactions. Separating SR_{pairs} into the difference of the two terms allows us to measure separately the two different effects.

The case of synergistic coding, $SR_{pairs} > 0$, means that specifying the stimulus increases the MI between two units, and makes a precise definition of the idea of "stimulus dependent correlations" (Abeles, Bergmann, Margalit, & Vaadia, 1993; Vaadia et al., 1995; Hopfield, 1995; Meister et al., 1995; Meister, 1996; Singer & Gray, 1995; Brenner et al., 2000). However, care should be taken when interpreting the $I(X_1, X_2; S)$ as standing for the everyday meaning of the word synergy. For example, consider two independent binary variables whose activity is contaminated with zero mean correlated noise ξ that is independent of the stimulus, e.g. with a joint distribution

$$
P(X, Y|S = s) = \begin{pmatrix} 0.25 + \xi & 0.25 - \xi \\ 0.25 - \xi & 0.25 + \xi \end{pmatrix}. \tag{3.6}
$$

In this case the synergy term is strictly positive $I(X; Y|S) > 0$ but the correlated noise is not actually used for coding (since the noise is stimulus independent).

3.2.2 Estimation considerations

In chapter 2 we derived the bias of the mutual information estimator, showing that it is roughly proportional to the ratio between the number of free parameters estimated and the number of samples. We now use this result to derive the bias of the synergy and redundancy terms, and show that these biases are considerably different. To see this, denote the total number of samples n, the number of samples per stimulus n_s, the number of stimuli $|S|$, the number of possible responses $|R|$, and stimulus probabilities $P(S = s) = n_s/n$. Since the possible responses of the two neurons reside in a two dimensional matrix of size $|R| \times |R|$, the stimulus conditioned joint probability matrix has n_s samples in $|R|^2$ bins for each stimulus, yielding a bias of $\frac{|R|^2}{2n_s \log(2)}$ per stimulus. Thus the total bias of the synergy term equals

$$E[I(X;Y|S)] = \sum_s P(S=s)\frac{|R|^2}{2n_s \log(2)} = \frac{|S||R|^2}{2n \log(2)} \tag{3.7}$$

On the other hand, the bias of the redundancy term is S times smaller

$$E[I(X;Y)] = \frac{R^2}{2n \log(2)} = \frac{1}{|S|}E[I(X;Y|S)] \tag{3.8}$$

As a consequence, the bias of the synergy-redundancy estimator is again positive, and equals

$$E[I(X;Y|S) - I(X;Y)] \quad = \quad \frac{(|S|-1)|R|^2}{2n \log(2)} \tag{3.9}$$

If the mutual information estimators are not properly corrected for the bias, independent neurons will appear as synergistic ones. This effect is not specific to estimators that use a discretizing procedures into bins, but stems from the fact that the synergy term requires estimating a finer effect, with mutual information terms with a smaller relative number of samples than the redundancy term.

3.2.3 Extensions to group redundancy measures

We now turn to extend the measures of synergy and redundancy beyond neuronal pairs. In the multivariate case, several different definitions of synergy and redundancy naturally arise, and each measures a different effect.

Redundancies of N-tuples given singles

First, as in the case of pairs, one may be interested in the difference between information levels conveyed by the joined distribution of N variables (neurons) compared to that provided by N single independent ones. This is defined by

$$SR_{N|1} = I(X_1, ..., X_N; S) - \sum_{i=1}^{N} I(X_i; S) \tag{3.10}$$

As with SR_{pairs}, this synergy-redundancy measure may be rewritten as the difference between two multi-information terms (see definition A.3.1)

$$\begin{aligned} SR_{N|1} &= I(X_1, ..., X_N; S) - \sum_{i=1}^{N} I(X_i; S) = \\ &= H(X_1, ..., X_N) - H(X_1, ..., X_N|S) \\ &\quad - \sum_{i=1}^{N} H(X_i) - H(X_i|S) = \\ &= I(X_1; ...; X_N|S) - I(X_1; ...; X_N) \end{aligned} \tag{3.11}$$

The index $SR_{N|1}$ can thus be separated into two terms, as with SR_{pairs}. The first is always non-negative, measures *within-stimulus correlations* (noise correlations) and is again termed here *synergy*. The second is always non-positive, measures *between stimulus correlations*, (signal correlations), and again quantifies the *redundancy*.

A major difficulty in estimating $SR_{N|1}$ is that it requires estimating the joint distribution of $N + 1$ variables: the stimulus and the activity of N neurons. This often becomes prohibitive even for the small N used in electrophysiological experiments, and is discussed in section 3.3.1.

Redundancies of N-tuples given N–1-tuples

In contrast to the two-variables case, the multi-variable case enables other measures of redundancies. For example, one may be interested in the difference between information conveyed by all N variables and that conveyed by pairs, triplets, or even all the $N-1$ subsets of variables. A case of particular interest is the residual information obtained from the joint distribution of all N variables, as compared to that obtained from any subset of $N-1$ variables. As with inclusion-exclusion calculations, the SR measure for this case is defined as

Definition 3.2.2 : N-tuples given N–1-tuples synergy-redundancy

$$\begin{aligned} SR_{N|N-1} &= I(X^{(N)}; S) \\ &\quad - \sum_{X^{(N-1)} \in I^{N-1}} I(X^{(N-1)}; S) \\ &\quad + \sum_{X^{(N-2)} \in I^{N-2}} I(X^{(N-2)}; S) \\ &\quad \vdots \\ &\quad + (-1)^{N-1} \sum_{\{X_i\}} I(X_i; S) \end{aligned} \tag{3.12}$$

where I^k is the set of all subsets of size k out of the N variables.

For example, for $N = 3$ we have

$$\begin{aligned} SR_{3|2} &= I(X_1, X_2, X_3; S) \\ &\quad -I(X_1, X_2; S) - I(X_1, X_3; S) - I(X_2, X_3; S) - \\ &\quad +I(X_1; S) + I(X_2; S) + I(X_3; S) \end{aligned} \tag{3.13}$$

This definition contains a total of $2^N - 1$ terms. For $N = 2$ it coincides with $SR_{pairs} = SR_{2|1}$.

As with the previous redundancy measures, it can again be rewritten as a difference between conditional and unconditional multi-information terms

$$SR_{N|N-1} = I(X^N|S) - \sum_{X^{N-1} \in I^{N-1}} I(X^{N-1}|S) + ... \tag{3.14}$$

$$-I(X^N) + \sum_{X^{N-1} \in I^{N-1}} I(X^{N-1}) - ...$$

where in this context $I(X^N) = I(X_1; ...; X_N)$, and I^k is defined as in 3.12.

Calculating $SR_{N|N-1}$ is difficult in practice since it requires estimating $O(2^N)$ information terms. This difficulty is compounded by the difficulty encountered in estimating of $SR_{N|1}$ in which joint probabilities of exponential sizes have to be estimated. These issues are discussed in the following section.

3.3 Redundancy measurements in practice

Applying the measures described above to finite data, and comparing redundancies across different brain stations involves several delicate points. We discuss three of them here: using stimulus conditioned independence approximation, the effect of baseline single-unit information level and redundancy bias due to information ceiling effects.

3.3.1 Conditional independence approximations

The multi-variable redundancy measures $SR_{N|1}$ and $SR_{N|N-1}$ defined in the previous section are based on a joint distribution of $N + 1$ variables $X_1, \ldots, X_N$ and S. Unfortunately, given the typical recordings in electrophysiological experiments, the sample sizes are rarely sufficient for reliable estimation of these joint distributions and consequently the MI values, even for moderate N values. One approach is to approximate this joint distribution with other distributions, using some predefined assumptions on conditional independence between variables. This approach has been advocated in the recent years in the literature of graphical models (Bayes nets), where predefined (conditional) independence structures are conveniently represented in the form of graphs (Pearl, 1988; Jordan, 1998; Shafer & Pearl, 1990).

One such approximating assumption is that neural activity of different neurons is conditionally independent given the stimulus

$$\begin{aligned} p(x_1, ..., x_N | S = s) &= \prod_{i=1}^{N} p(x_i | S = s) \qquad (3.15) \\ p(x_1, ..., x_N) &= \sum_s p(s) \left(\prod_{i=1}^{N} p(x_i | S = s) \right) . \end{aligned}$$

Figure 3.1 shows a graphical model illustration of this dependence structure.

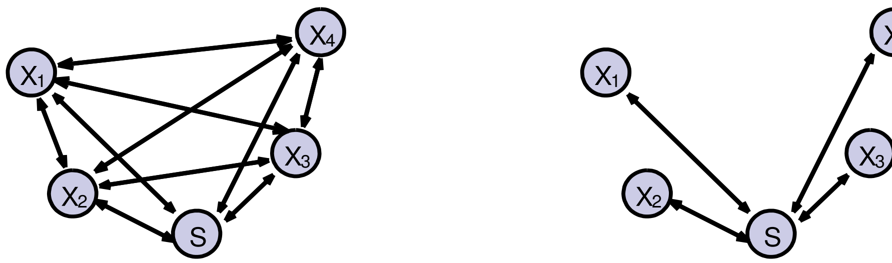

Figure 3.1: An illustration of dependence structure for two cases. **Left:** Four neurons and the stimulus, without any independence approximation. **Right:** Four neurons under the stimulus conditioned independence assumption.

Figure 3.2 depicts the joint distribution of the neural responses (number of spikes) of two IC neurons calculated under the conditional independence approximation. Each stimulus conditioned distribution was calculated under independence approximation

$$p(r_1, r_2) = \sum_s p(r_1, r_2|s)p(s) = \sum_s p(r_1|s)p(r_2|s)p(s). \tag{3.16}$$

Estimating the full joint distribution $p(R_1, R_2)$ thus requires estimating S distributions of size $|R_1|$ and S distributions of size $|R_2|$, instead of S two dimensional distributions of size $|R_1| \times |R_2|$.

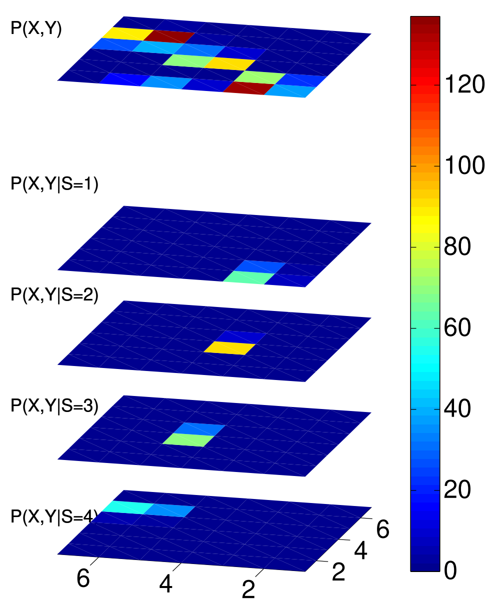

Figure 3.2: Stimulus conditioned joint distribution of spike counts for two IC cells. Each of the four lower pallets correspond to the distribution of spike counts for a different stimulus, under stimulus conditioned independence. The upper panel is the sum of conditional distributions over 15 stimuli. The color code denotes the number of spike counts events.

With stimulus conditioned independence, calculating the joint distribution used in the information terms of $SR_{N|1}$ is considerably simplified, as for each stimulus s

we only have to estimate the marginal distributions $p(X_i|S = s)$ instead of the full distribution $p(X_1, ..., X_N|S = s)$. It thus allows us to estimate an exponentially smaller number of parameters, which in the case of neuro-physiological data makes it feasible to estimate the MI in these approximated distributions. This approximation thus makes it possible to investigate redundancy among considerably larger groups of neurons than the 2-3 neuron groups often analyzed in the literature. In terms of the bias variance tradeoff discussed in Chapter 2, using this approximation decreases the variance, since it decreases the number of degrees of freedom, but increases the bias, as the estimated MI deviate from the ones calculated with a non approximated joint distribution.

Interestingly, under stimulus conditioned independence, the synergy term $SR_{N|1}$ vanishes, thus limiting neural interactions to the redundant (negative) regime. The measure we will use often in the current work is therefore

$$Redundancy_{N|1} = -I(X_1; \ldots ; X_N) \tag{3.17}$$

We will also refer at times to the quantity $I(X_1; \ldots ; X_N)$ as redundancy, but the difference should be clear from the context.

One situation in which the assumption of stimulus conditioned independence is crucial is in the analysis of non simultaneously recorded data. In this case, if we want to obtain any estimate of informational coupling between neurons, it is necessary to couple the given marginal distributions of the single neurons conditioned on the stimuli. Stimulus-conditioned independence is then the most natural way (in some sense, the maximum-entropy way) of performing this coupling.

How reasonable is the conditional-independence approximation? Naturally, its validity depends on the data at hand. In theory the approximation is expected to be good when neuronal activity is mostly determined by the presented stimulus and to a lesser extent by interactions with nearby neurons. One example is the high input regime of cortical neurons receiving thousands of inputs, where a single input has only a limited influence on the activity of the target cell. The experimental evidence in this regard is however mixed (see e.g.(Gochin, Colombo, Dorfman, Gerstein, & Gross, 1994)).

When simultaneously recorded neurons are available, it is possible to test empirically the quality of the approximation for small sets of neurons (e.g. pairs and triplets), and quantify the deviation from independence. This was done e.g. in (Nirenberg et al., 2001). We performed these tests for a limited set of simultaneously recorded neurons in AI and MGB and found that the approximation causes only a small deviation from the true distribution. These results are described in detail in section 4.1 (see Fig. s 4.7 and 4.8).

3.3.2 The effect of single-unit information on redundancy

A major goal of the current work is to compare coding characteristics across different brain regions in a sensory pathway. Clearly, we must ensure that any observed difference in redundancies between regions is not an artifact due to differences in other factors. In particular, the redundancy between neurons tends to grow when single-unit information about the stimulus grows. The differences in MI levels in different auditory stations therefore require normalizing the redundancies to a unified scale.

To demonstrate this effect, consider the following toy model (Fig 3.3), consisting of a Markov chain of three binary variables $X \to S \to Y$, where each of the pairs X, S and S, Y is a binary symmetric noisy channel with noise level $\xi \in [0, 1]$. Formally we have, $Pr(X = 1) = \frac{1}{2}$ and $Pr(X = S) = Pr(S = Y) = 1 - \xi$. Since X, S and Y form a Markov chain, X and Y are conditionally independent given the stimulus S.

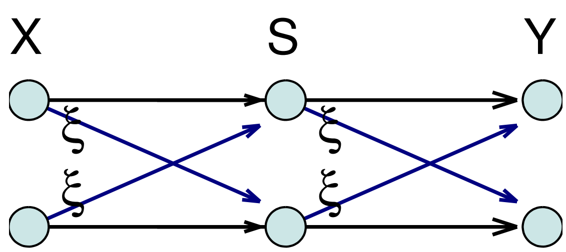

Figure 3.3: Two concatenated symmetric noisy binary channels, each with noise level ξ.

The intuitive interpretation for the notion of redundancy between X and Y, suggests that in this model redundancy should not depend on the noise level ξ. However, the noise level here effects both the single-unit information and the redundancy, as can be shown analytically, by calculating the single-unit information $I(X; S)$,$I(Y; S)$ and redundancy magnitude $I(X; Y)$ as a function of ξ

$$I(X; S) = I(S : Y) = 1 - H[\xi] \tag{3.18}$$

where $H[p]$ is the entropy of the binary distribution $(p, 1 - p)$ $H[p] = -p \log_2 p - (1 - p) \log_2(1 - p)$. In addition, the pair (X, Y) is also a symmetric binary channel, this time with a noise level of $2\xi(1 - \xi)$. The redundancy magnitude, which is the information of the concatenated channel, is therefor

$$I(X; Y) = 1 - H[2\xi(1 - \xi)] \tag{3.19}$$

Figure 3.4A plots the single-unit information $I(X; S)$ and the redundancy magnitude $I(X; Y)$ as a function of the noise level ξ. It shows that although the redundancy changes considerably with the noise level it is roughly consistent with the change in the single-unit information. This suggests that the single-unit information provides a natural scale for the problem and can be used to normalize the redundancy measure. The relation between MI and redundancy as a function of the noise level is presented as a scatter plot in Fig. 3.4B.

These data suggest that the *normalized redundancy* $I(X;Y)/I(X;S)$ (or, which is equivalent in this case $2I(X;Y)/[I(X;S)+I(S;Y)]$) provides a reasonable normalization scheme for this simple example, because it preserves redundancy values $I(X;Y)$ over a range of single-unit information values $I(X;S)$. This normalization procedure was indeed used in other studies of synergy and redundancy (Brenner et al., 2000; Reich et al., 2001).

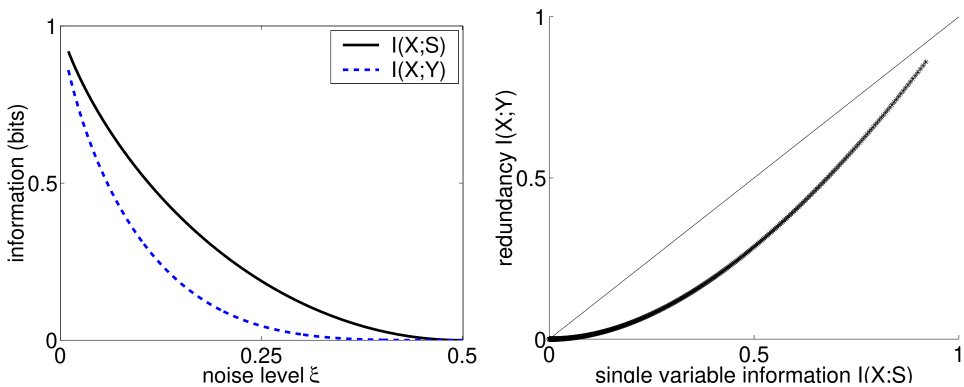

Figure 3.4: **left:** Mutual information and redundancy as a function of the noise level ξ. **right:** Scatter plot of MI vs. (non normalized) redundancy.

Consider now a second setting, where the Markov chain $X \to S \to Y$ consists of three variables, each having an alphabet of 4 letters. We compare two models, described in figure 3.5, where the intuitive notion of redundancy asserts that the model A (orthogonal noise in $p(x,s)$ and $p(y,s)$ is less redundant than the model B. For example, when the noise is maximal $\xi = \frac{1}{2}$, X and Y provide the same information on S in model B, since they serve to discriminate between the upper two alternatives, but different information on S in model A.

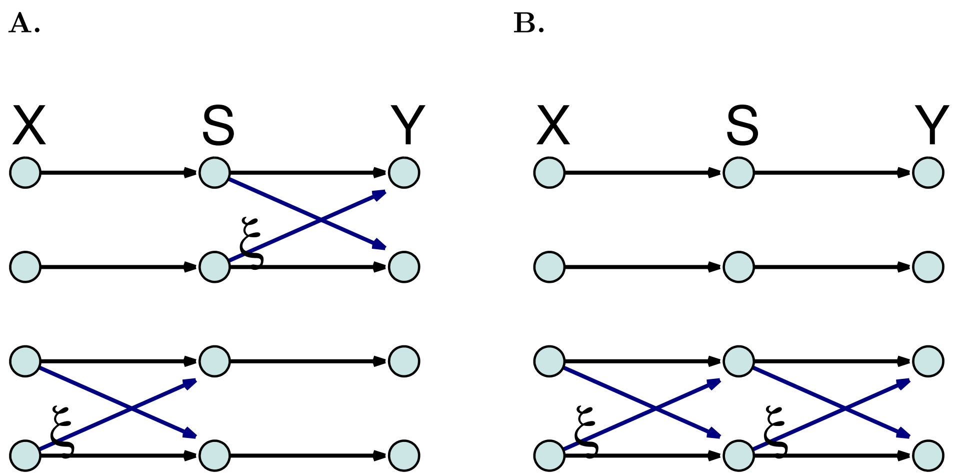

Figure 3.5: Two toy example for 4-letter channels, with different structure. Each channel consists of a concatenation of two channels, in which two of the values are transmitted with no noise, while the other two are corrupted with noise at level ξ. On the left channel, X and Y carry noiseless information about different pairs of values of S, while in the right panel both X and Y convey full information about the same pair of values of S.

Once again mutual information can be calculated analytically in these two models. For uniform $p(x)$ we have

$$\begin{aligned}
&\textbf{model A}: \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.20)\\
I(X;S) &= H(S) - H(S|X) = 2 - \frac{1}{4}\left(0 + 0 + H[\xi] + H[\xi]\right)\\
&= 2 - \frac{1}{2}H[\xi]\\
I(X;Y) &= H(Y) - H(Y|X) = 2 - \frac{1}{4}(H[\xi]+H[\xi]+H[\xi]+H[\xi])\\
&= 2 - H[\xi]\\
&\textbf{model B}:\\
I(X;S) &= H(S) - H(S|X) = 2 - \frac{1}{2}H[\xi]\\
I(X;Y) &= H(Y) - H(Y|X) = 2 - \frac{1}{2}H[2\xi(1-\xi)]
\end{aligned}$$

We now compare redundancies in these two models under different noise levels. Specifically, we wish to avoid a situation where one model seems more redundant just because its single-unit information is higher. Figure 3.6 plots the single-unit information and redundancy in the two models as a function of noise level ξ. In both models the single-unit information and unnormalized redundancies are correlated (upper panels). More importantly, The normalized redundancy of model A is usually lower than that of model B (lower panels).

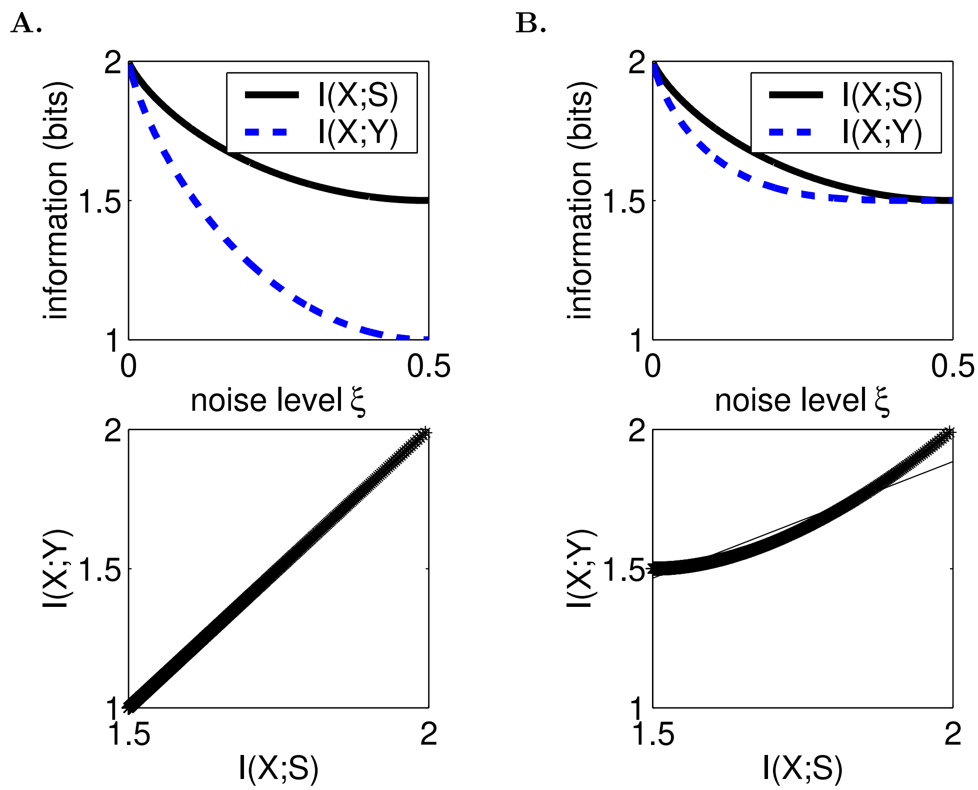

Figure 3.6: **Top:** Single information and redundancy in two synthetic models as a function of noise level. **Bottom:** Redundancy as a function of single-unit information, for the two models.

Figure 3.7 compares the normalized and non-normalized redundancies in the two models. For each of the two measures there is a regime of noise levels $\xi \in [0, \xi^{max}]$, for which any type-A model appears less redundant than any type-B model, regardless of ξ chosen independently for each model. This regime is almost three times larger for the normalized redundancy measure than the unnormalized one.

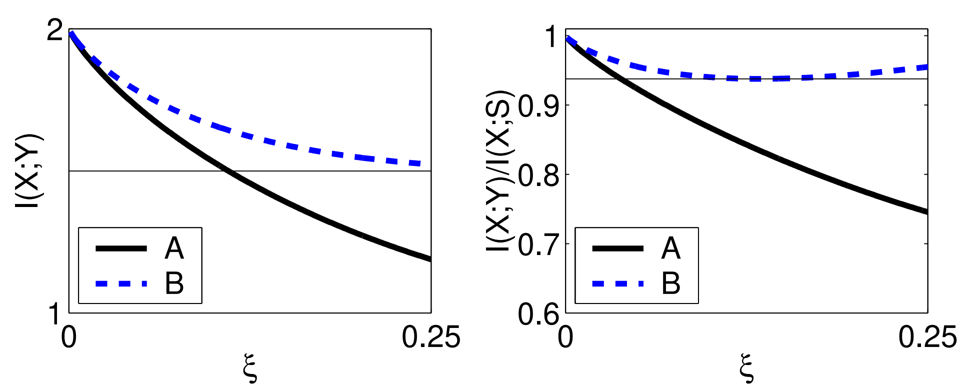

Figure 3.7: Comparing normalized and unnormalized redundancy measures in model A and B. The regime of ξ for which any type-A model appears less redundant than any type-B model is depicted with a thin line. This regime is about three times larger for the normalized redundancy measure. **A.** Non-normalized redundancy **B.** Normalized redundancy.

Clearly, these results do not guarantee that under more complex scenarios the redundancy will grow near-linearly with single neuron information. Naturally, the picture becomes even more complex for multi-neuron redundancies. The above results show however that normalizing by single neuron information captures the intuitive notion of redundancy better.

What about normalization of redundancy among larger groups of variables? Earlier we defined the redundancy in this case to be $-I(X_1;\ldots;X_N)$. This information is always bounded from above by the information between all variables and the stimulus

$$I(X_1;\ldots;X_N) \leq I(X_1;\ldots;X_N;S) \ , \tag{3.21}$$

and under stimulus conditioned independence this bound equals the sum of single unit information terms

$$\begin{aligned}
& I(X_1;\ldots;X_N;S) = \\
= & \sum_{x_1,\ldots,x_N,s} p(x_1,\ldots,x_N,s)\log\left(\frac{p(x_1,\ldots,x_N,s)}{p(x_1)\cdot\ldots\cdot p(x_N)\cdot p(s)}\right) \\
= & \sum_{x_1,\ldots,x_N,s} p(x_1|s)\cdot\ldots\cdot p(x_N|s)p(s)\log\left(\frac{p(x_1|s)}{p(x_1)}\cdot\ldots\cdot\frac{p(x_N|s)}{p(x_1)}\right) \\
= & \sum_{x_1,\ldots,x_N,s} p(x_1|s)\cdot\ldots\cdot p(x_N|s)p(s)\sum_{i=1}^{n}\log\left(\frac{p(x_i|s)}{p(x_i)}\right) \\
= & \sum_{i=1}^{n}\sum_{x_i,s} p(x_i|s)p(s)\log\left(\frac{p(x_i|s)}{p(x_i)}\right) \\
= & \sum_{i=1}^{n} I(X_i;S)
\end{aligned} \tag{3.22}$$

Interestingly, tighter bounds such as the average or minimal single unit information do not hold. To see this consider the following example. Let $X_1,\ldots,X_N$ be symmetric binary variables, whose values are determined by the value of a binary symmetric S.

$$(X_1,\ldots,X_N) = \begin{cases} (1,\ldots,1) & \text{when } S=1, \text{with } p=\frac{1}{2} \\ (0,\ldots,0) & \text{when } S=0, \text{with } p=\frac{1}{2} \end{cases} \tag{3.23}$$

In this case, each single-unit information equals one $I(X_i;S) = 1$ bit $\forall i$. Thus both the average and the minimum of the single unit information terms are also 1 bit. However, the multi-information term grows with N, and equals

$$\begin{aligned}
I(X_1;\ldots;X_N) & = \sum_{x_1,\ldots,x_N} p(x_1,\ldots,x_N)\log\left(\frac{p(x_1,\ldots,x_N)}{p(x_1)\cdot\ldots\cdot p(x_N)}\right) \\
& = \sum_{s=0}^{1} p(\vec{X}=s)\log\left(\frac{p(\vec{X}=s)}{p(x_1=s)\cdot\ldots\cdot p(x_N=s)}\right) \\
& = \sum_{s=0}^{1}\frac{1}{2}\log\left(\frac{\frac{1}{2}}{\frac{1}{2}^N}\right) \\
& = N-1
\end{aligned} \tag{3.24}$$

This example suggests that in the multivariate case redundancy should be normalized by the sum of the single unit information terms, which for a large N is a tight bound on redundancy. Therefore in this work we use the normalized multivariate redundancy measure

$$Normalized\ Redundancy_{\ N|1} = \frac{-I(X_1; \ldots; X_N)}{\sum_{i=1}^{n} I(X_i; S)} \ . \tag{3.25}$$

Note that this measure of redundancy is equal up to a constant to the normalization examples discussed earlier for two variables (figures 3.3, 3.5).

3.3.3 Bias in redundancies estimation due to information ceiling effects

To illustrate the subject of the current section, we start with a simple example. Let $X_1, ..., X_N$ be the activities of N neurons in response to the stimuli, and let us assume they were measured independently, and that all of them have the same response distribution with the stimulus $P(X_i, S)$.

According to the chain rule for mutual information, the information conveyed by such independent variables about a stimulus $I(X_1, ..., X_N; S)$ increases according to $I(X_1, ..., X_N; S) = \sum_{i=1}^{N} I(X_i; S|X_1...X_{i-1})$. On the other hand however, this information is bounded by the stimulus entropy $H(S)$, which for any finite number of stimuli must be finite as well. This shows that the information accumulate sub-additively

$$\begin{aligned} I(X_1, ..., X_N; S) &= \sum_{i=1}^{N} I(X_i; S|X_1...X_{i-1}) \\ &< \sum_{i=1}^{N} I(X_i; S) = N\ I(X_1; S). \end{aligned} \tag{3.26}$$

The reason for this sub-additivity is that given the neurons $X_1, ..., X_{k-1}$ the information that the i^{th} neuron conveys about S is smaller than the unconditional information $I(X_k; S|X_1...X_{k-1}) < I(X_k; S)$. Simply stated, after some neurons convey information about the stimulus, there isn't much left to tell, and the remaining neurons can only convey lower levels of information. This effect is important when aiming to quantify redundancies, since it creates an "artificial" source of redundancy between neurons, stemming from the experimental conditions (namely the bounded entropy of the stimulus set).

How can this effect be quantified? A first order model of this phenomenon was suggested by (Gawne & Richmond, 1993; Rolls et al., 1997). Consider an abstract information "space" of size I^{max}, and N variables each conveying $I(X_i; S)$ bits of information, thus covering a fraction $I(X_i; S)/I^{max}$ of the space. The mean information conveyed by a single neuron is therefore $I_1 = \frac{1}{N} \sum_i I(X_i; S)/I^{max}$. If the information conveyed by each variable randomly covers some fraction of this space, the spurious overlaps between the any pairs of variables will be I_1^2. The expected fraction of information covered by N variables is then

$$
\begin{aligned}
I(N) &= I^{max}\left[\binom{N}{1} I_1 - \binom{N}{2} I_1^2 + \binom{N}{3} I_1^3 - \ldots\right] \\
&= I^{max}\left(1-(1-I_1)^N\right) \ .
\end{aligned} \tag{3.27}
$$

In the limit of infinite number of neurons $N \to \infty$, $I(N)$ reaches the upper bound I^{max}.

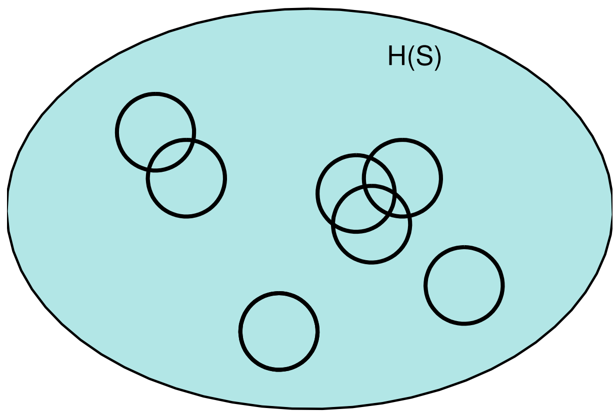

Figure 3.8: An illustration of the information space mode. The ellipsoid stands for the entropy of the stimulus $H(S)$, which is an upper bound on the maximal information an ensemble of neurons can convey about the stimulus. Each circle corresponds to a neuron, and its area corresponds to the mutual information $I(X_i; S)$. In this type of diagrams, neurons may in principle cover areas outside the ellipse, but these are not plotted here.

This model is loosely based on Venn type diagrams that correspond to entropies and mutual information through what is known as *I-measures* (Yeung, 1997). For a more extensive discussion about this type of diagrams, see e.g. (Cover & Thomas, 1991) p.20 and (Csiszar & J.Korner, 1997) p.50.

The arguments using the information plane have intuitive appeal, but the theoretical justification behind them is not formally established. However, Samengo and Treves (Samengo & Treves, 2000; Samengo, 2001) compared the prediction of the information plane approach with actual information curves for several synthetic examples and showed that the resulting equations reasonably approximate the observed information growth as a function of N. We therefore adopted this model as the null hypothesis model for estimating the baseline behavior of a set of neurons, to which we compare our experimental results when quantifying redundancy in groups of neurons.

This model can be refined to take into account the variable information values across the isolated single neurons $I(X_i; S)$. Since in this heterogeneous case the order of the series $(X_1, ..., X_N)$ determines the shape of the curve, we first define the information curve for a predefined ordered set of variables $X_1 \ldots X_N$.

$$I(N) = I^{max} \left[1 - \Pi_{i=1}^{n} \left(1 - I(X_i; S)\right)\right] \ . \tag{3.28}$$

For an non-ordered set of variables $\{X_1, ..., X_N\}$, the information curve is defined as the average over all possible orderings. It is interesting to compare this model with the homogeneous model for which I_1 is set to $I_1 = \frac{1}{N} \sum_i I(X_i; S)$. Since for any two numbers $x \times y < \left(\frac{x+y}{2}\right)^2$, the pairs I_1^2 is larger than $< I_1^i I_1^j >$. Similar consideration for higher powers shows that the homogeneous information model yields an underestimation of the curve $I(N)$, i.e. an overestimation of the redundancies.

3.4 Summary

This section has discussed methodological issues in quantifying high-order correlations among small groups of neurons, specifically redundancy and synergy. We defined information theoretic measures, and discussed ways for reliable estimation of these quantities from actual electrophysiological data. The following section describes the application of these methods to neurons in the auditory system.

Chapter 4

Redundancy Reduction in the Auditory Pathway

The previous chapter described quantitative measures of interactions among small groups of neurons. In this chapter these measures are applied to study the neural code in the ascending auditory pathway. Sections 4.1 describes coding of stimulus identity and 4.2 discusses coding of acoustic features .

The current chapter describes the main empirical result of this dissertation, namely, evidence for a process of redundancy reduction along the core ascending auditory pathway.

4.1 Coding stimulus identity

We applied the methods described in the previous chapter to electrophysiological recordings in the core auditory pathway, quantifying mean redundancies in population of neurons in three processing stations: the inferior colliculus (IC), the medial geniculate body of the thalamus (MGB) and the primary auditory cortex (AI). The current section summarizes our findings.

To first demonstrate the type of responses encountered in our data, Fig. 4.1 plots the mean firing rate of two IC neurons in response to 15 stimuli, compared with the mean firing rate of 2 AI neurons to the same stimuli. While the response profiles of the IC neurons to these stimuli are very similar, the AI neurons responded rather differently. For example, one AI neuron (in blue) responded to all stimuli at a roughly constant level, except for stimuli number 7, 8 and 9. The other neuron, on the other hand, responded to these stimuli at approximately the same level as to the other stimuli, but responded much more weakly to stimulus number 13. These results are not simply due to differences in the frequency response characteristics of the two pairs, because both had similar BF's (5.5 kHz and 6.1 kHz for the IC neurons, 5.1 for both AI neurons).

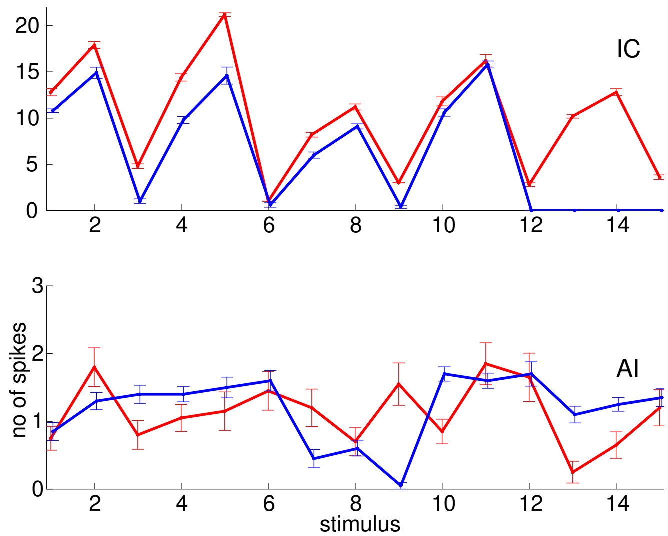

Figure 4.1: Spike counts across the stimulus ensemble for a pair of IC cells (top), (BF's 5.5 and 6.1 kHz) and a pair of AI cells (bottom), (BF's 5.1 both). Error bars denote standard error of the mean spike count, as obtained from 20 repetitions of stimulus presentations.

We quantify pairwise redundancy in neurons from the auditory processing stations. For this purpose we measured the normalized pairs redundancy under conditional independence for all pairs of neurons (as explained in section 3.3.2)

$$Normalized\ Redundancy = \frac{-I(X_1;X_2)}{I(X_1;X_2;S)} = \frac{-I(X_1;X_2)}{I(X_1;S)+I(X_2;S)} \tag{4.1}$$

and plotted its distribution. All information measures were corrected for bias as discussed in chapter 2.

Figure 4.2 plots the distribution of normalized pairwise redundancy with information obtained from spike counts. It reveals a considerable difference in redundancy level between the IC population on one hand and AI and MGB on the other.

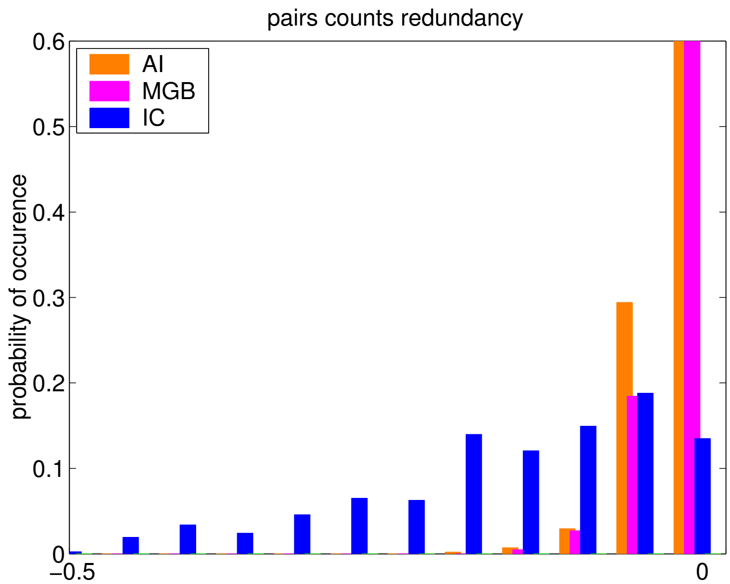

Figure 4.2: Distribution of normalized pairs redundancy $\frac{-I(X_1;X_2)}{I(X_1;S)+I(X_2;S)}$ estimated using spike counts across the population of three brain regions.

Distribution of spike counts was calculated by counting spikes within a window that was optimized for each brain region separately. The window was chosen such that it maximized the mean single-neuron information over the population. The optimal window values were AI: $20-140ms$, MGB: $20-80ms$ and IC: $0-60ms$.

Neurons in A1 and MGB are significantly less redundant than neurons in IC (Fig 3a). The median normalized redundancy in IC was -0.13 (with a median absolute deviation from the median of 0.07), whereas in MGB it was -0.02 (0.015) and in A1 -0.03 (0.015), a highly significant difference.

We further measured normalized triplets redundancies

$$\frac{-I(X_1;X_2;X_3)}{I(X_1;S)+I(X_2;S)+I(X_3;S)} \tag{4.2}$$

in the same population, which showed an even more pronounced difference (Fig. 4.3).

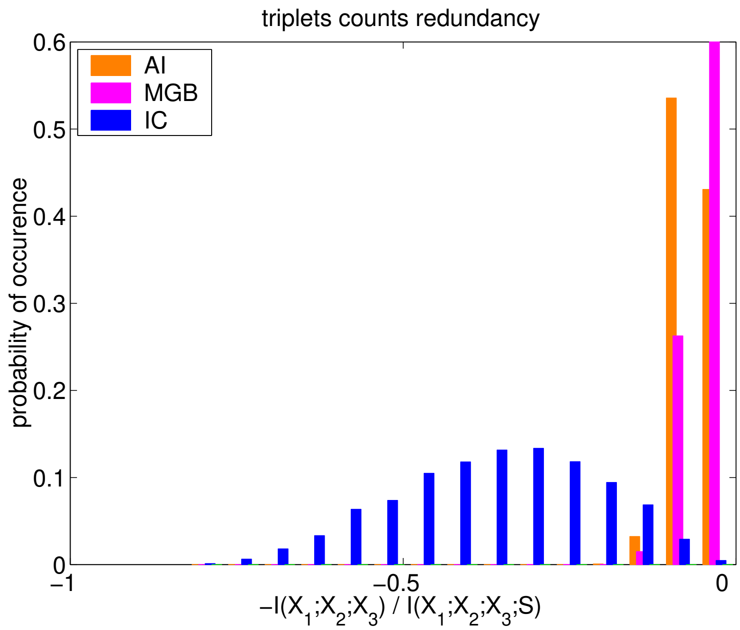

Figure 4.3: Distribution of normalized triplets redundancy across the population three brain regions.

One possible cause for low redundancy in AI compared to IC is the use of a reduced measure, the spike counts. Other statistics of the spike trains could perhaps show comparable redundancies in both areas. We therefore calculated the redundancies using first spike latency as the reduced response measure, and using the direct method (section 2.3.4). Figure 4.4 show that the general picture remained the same, with redundancies in IC substantially larger than in MGB and AI.

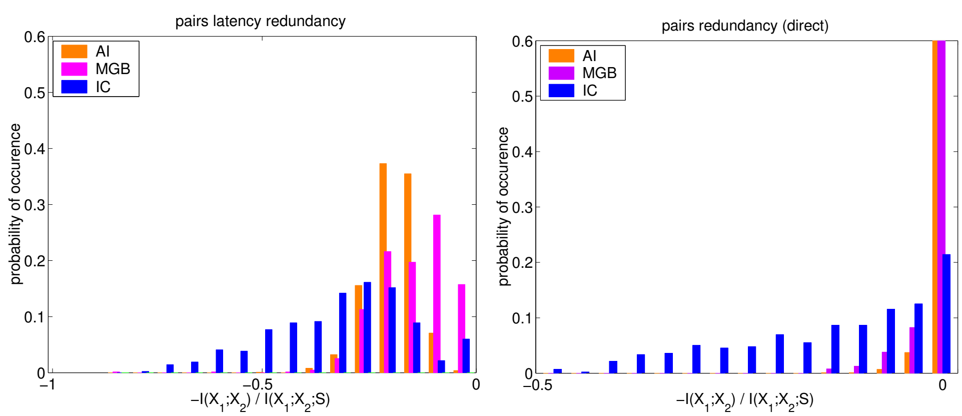

Figure 4.4: Distribution of normalized pairs redundancy across the population of three brain regions, as estimated using first spike latency and the direct method.

To demonstrate that these differences in redundancies are not caused by our normalization procedure, Figure 4.5 shows the distributions of the non normalized redundancy measure under the same conditions, showing that essentially the same results are obtained.

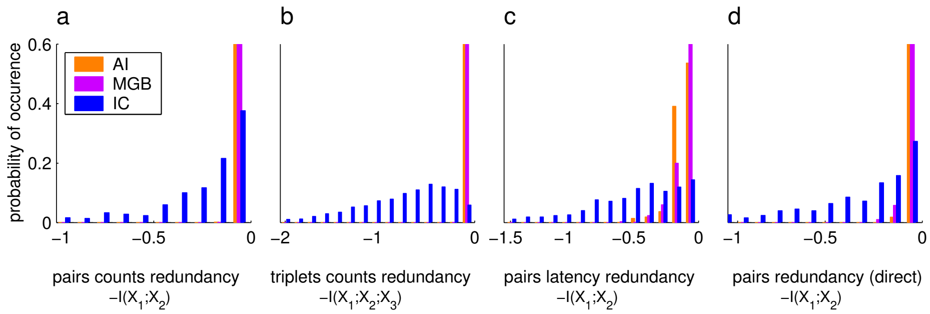

Figure 4.5: Distribution of non-normalized redundancies (A) pairs counts (B) triplets counts. (C) pairs, latencies (D) pairs, direct method .

As summarized in Table 4.6, in all these measures, the IC population of neurons had a considerably larger redundancy than lower level neurons. The following two subsections further refine the characteristics of this redundancy.

method	IC	MGB	AI
counts, pairs	-0.816±0.585	-0.038±0.026	-0.027±0.026
counts, triplets	-0.307±0.094	-0.061±0.025	-0.043±0.025
pairs, latency	-0.332±0.178	-0.186±0.055	-0.142±0.086
pairs, direct	-0.334±0.179	-0.005±0.013	-0.001±0.026

Figure 4.6: Mean and standard deviation of redundancy in four different coding schemes and three brain regions.

4.1.1 Validating the conditional independence approximation

The results of the previous section were obtained using the stimulus conditioned approximation for estimating the joint distributions of pairs and triplets of cells. As discussed in section 3.3.1, this approximation may effect the estimated level of redundancy. We now turn to estimating the validity of this approximation. We measured redundancy and synergy for a smaller number of cells that were recorded simultaneously in AI (total of 9 pairs from 15 cells) and MGB (total of 43 pairs from 29 cells).

Figure 4.7 plots the bias-corrected normalized redundancy obtained from these cells both under the conditional independence approximation and without it. It shows that using the actual coupling between the neurons, instead of the conditional independence approximation, increases the estimated redundancy by a factor of 1.74 on average in MGB cells. The corresponding increase in AI cells was 1.23 on average. It should be noted that redundancies observed in IC were about 5 to 15 times larger than those observed in MGB and AI. We conclude that the conditional independence approximation cannot be the sole cause for this difference in redundancies.

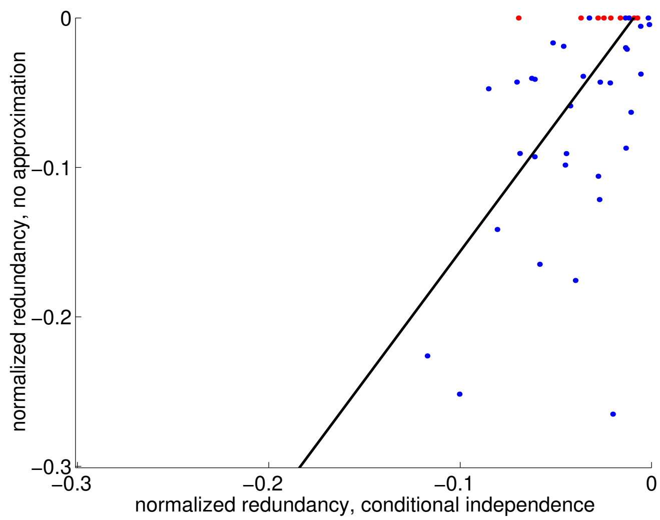

Figure 4.7: Normalized redundancy with and without stimulus conditioned independence approximation in simultaneously recorded MGB neurons. Red points are points that yielded positive redundancy estimate and were clamped to zero.

As explained in section 3.2.1, a frequently used measure of synergy and redundancy in the literature is the difference between the information conveyed by two cells together and that conveyed by the two cells considered individually, $SR_{pairs} = I(X_1, X_2; S) - I(X_1; S) - I(X_2; S)$, which can be rewritten as $I(X_1; X_2|S) - I(X_1; X_2)$.

Figure 4.8 compares this synergy-redundancy index with our redundancy index $-I(X_1; X_2)$ for MGB cells, both normalized by $[I(X; S) + I(Y; S)]$. It shows that in MGB, these two indices are correlated with a linear regression slope of 1.99, and that a linear relation of $SR = -1.0 * I(X_1; X_2)$is within the 95 percent confidence interval of the slope (b $\in$ [0.85, 3.4]). The reason for this correlation is that the synergy in our data is relatively weak (mean synergy -0.02, with a standard deviation of 0.28), and is counterbalanced by an increase in the redundancy when the conditional independence approximation is not used (Fig. 4.7). Similar results were obtained for AI cells (linear regression curve SR=-0.8 I+0.2, mean synergy 0.039 with a standard deviation of 0.12).

We are particularly interested in the redundancies induced by similar frequency sensitivity, and therefore used $-I(X_1; X_2)$ under stimulus-conditioned independence as our measure of redundancy. The result shown in Fig. 4.8 suggests that the same conclusions would be reached by using SR with the actually measured coupling between the responses. Our measure has two major advantages. First, it can be evaluated reliably for larger sets of neurons (e.g. in Fig. 4, up to 19 neurons). Secondly, it can be evaluated for neurons that have not been simultaneously measured, thus allowing for the use of a larger fraction of the recorded neurons.

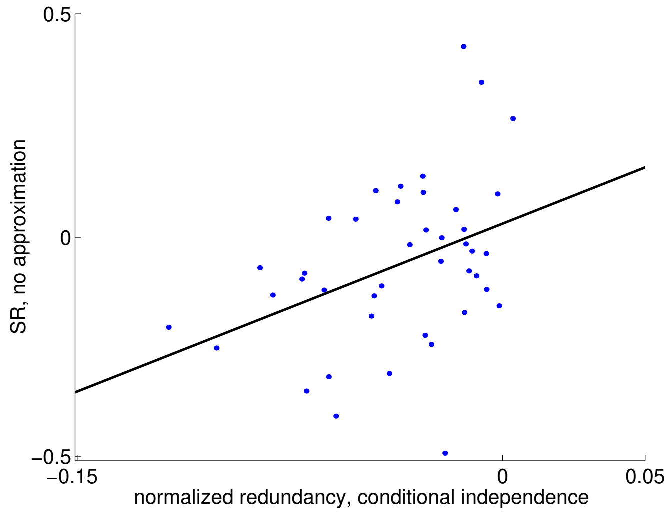

Figure 4.8: Redundancy under conditional independence approximation vs. the standard synergy-redundancy measure in simultaneously recorded MGB neurons.

4.1.2 Redundancy and spectral sensitivity

Neurons in the auditory system are traditionally characterized by their spectral sensitivity. Specifically they are often characterized by the frequency to which they are most responsive, their "best frequency" (BF). A natural question when characterizing redundancy is thus the relation between redundancy levels and spectral sensitivity.

To address this question, we studied the relation between the redundancy of each pair of neurons and their BF's. Figure 4.9 plots the normalized redundancy for each pair of AI neurons, while neurons were ordered by their BF. As a baseline for comparison we used the set of auditory nerve fibers (ANF) model neurons described in Chapter 1. The set of ANF neurons had the same set of BF's as the AI neurons. Their responses to the same set of stimuli were computed and the redundancy was estimated exactly the same way as for the AI neurons.

Figure 4.9B depicts the normalized redundancy for all pairs of ANF neurons. High redundancy (negative values) is observed along the diagonal, in particular in those frequency bands that have high energy in the stimulus set. Figure 4.9A depicts normalized redundancy values for the AI neurons

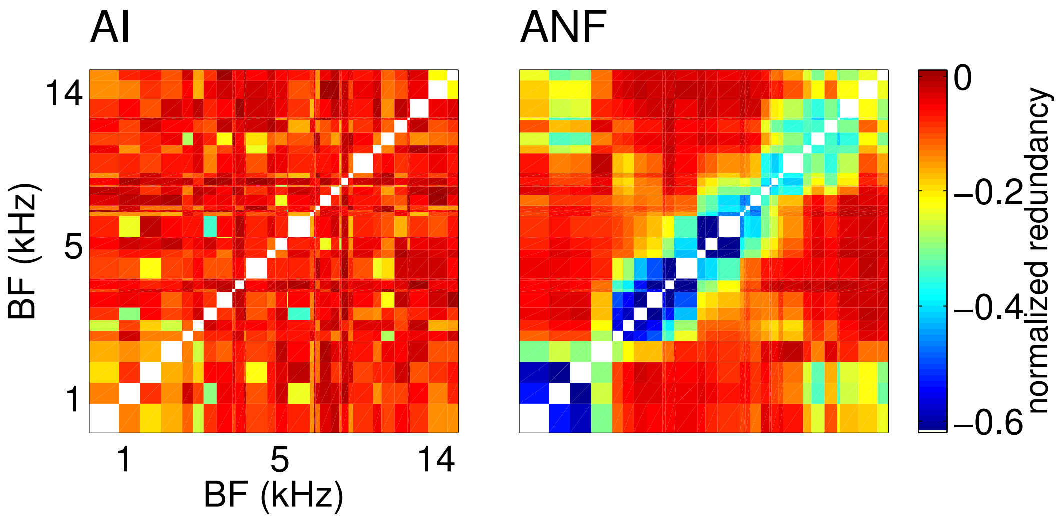

Figure 4.9: Normalized redundancy of pairs of cells ordered by their BF's.

Figure 4.10 plots the normalized redundancy for each pair as a function of the difference in BF's (in log scale), for each of the four brain regions. A strong correlation between BF difference and redundancy level is apparent for both IC and ANF-model neurons, but not for MGB and AI neurons.

In order to quantify this effect, we measured the correlation between the BF difference (in log scale) and the normalized redundancy level. In order to take into account the fact that the points in the scatter plot are the results of pair-wise comparisons, and thus are not independent, we correlated redundancy values against BF difference separately for each unit against all other units. We then tested the hypothesis that the set of correlation coefficients is significantly different from zero. After correcting for the number of samples,

both AI and MGB populations had non significant correlations of redundancy with BF (p¿0.01). On the other hand, in IC and in the ANF simulations, pairs of neurons showed significant correlations of redundancy with BF (IC: $p < 10^{-6}$;ANF: $p < 10^{-8}$.)

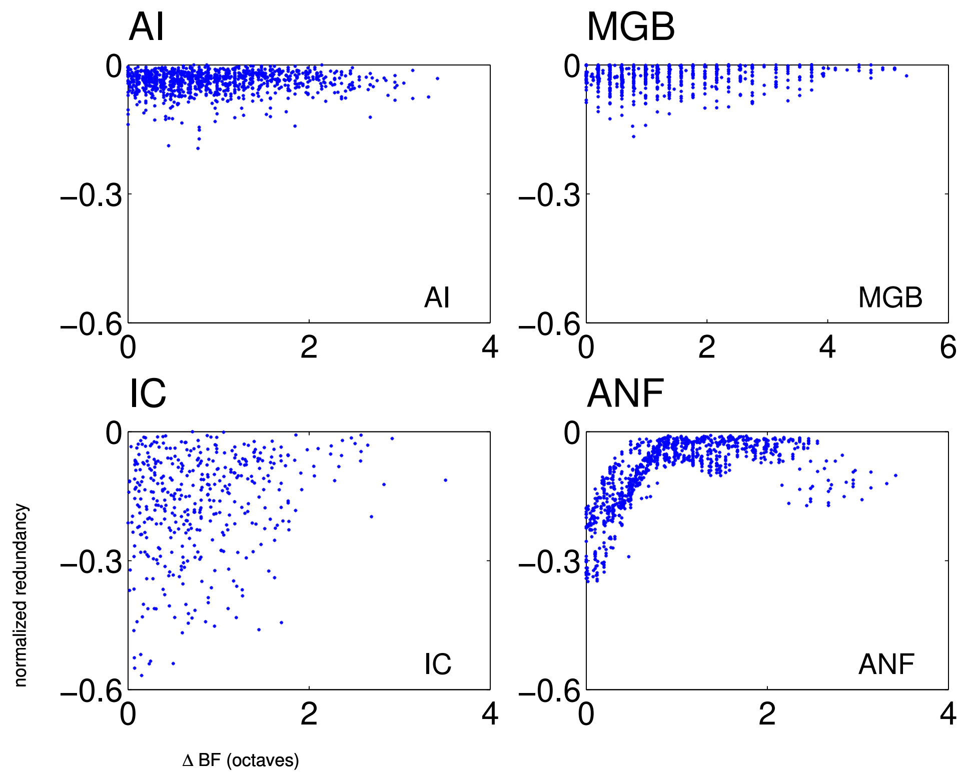

Figure 4.10: Normalized redundancy of pairs of cells as a function of their BF's difference.

The importance of these results lies in the fact that they provide strong hints about the differences in the nature of the neural code in the different stations. IC neurons are more redundant, and this redundancy is related to spectral sensitivity. More specifically, pairs of neurons that show high redundancy, and so convey similar information about the identity of the sounds tend to have similar BF. The reverse is not necessarily true, in that there are pairs of IC neurons with the same BF that have low redundancy. These are expected, since neurons in IC with the same BF may differ in other aspects of their sensitivity to the physical structure of sounds; e.g. by having different temporal sensitivities (Casseday et al., 2002).

The behavior of AI neurons is inherently different. First, AI neurons are far less redundant. In addition, even the most redundant AI pairs have larger BF difference than the most redundant pairs in IC. The complete lack of redundancy in AI neurons, even among those with similar BF, suggests that different neurons in AI tend to convey information about different aspects of the stimuli, since their redundancy cannot be accounted for by mere spectral sensitivity.

4.1.3 Redundancy and physical cell locations

The previous section characterized the relation between redundancy and cells spectral sensitivity, specifically the distance between cells' BF's. Another possible organization principle underlying redundancy may be revealed in the anatomical organization of the recorded cells. Unfortunately, the 3-dimensional locations of the recorded neurons were not available in our data. However, anatomical distances can be very coarsely estimated by separating the recorded cells into three groups according to their recording type:

- **Same penetration**: Cells that were recorded within a single penetration but on different electrodes, which must be less than 1 mm apart in the cortex. In the MGB, such neurons are less than 1 mm apart in the medio-lateral and antero-posterior axes, but could be more distant in the dorso-medial axis.

- **Same animal**: Cells that were recorded during the same experiment but from different penetration are usually up to a few mm from each other.

- **Different animals**: The distance between cells recorded from different animals is not defined. Since neural organization in different animals could be considerably different, we treat cell pairs in this group as if located at a large distance from each other.

Figure 4.11 plots the normalized redundancy in each of these four groups for MGB cells [1]. Error bars denote the standard errors of the mean of each group. For comparison the mean redundancy in IC is also plotted, showing that MGB redundancy is significantly smaller than IC in all groups. A one-way ANOVA suggests that the difference between the groups is significant ($p < 0.01$) Even more interestingly, the three groups have a monotonically decreasing order of redundancy levels as expected. This means that neurons that are physically near to each other also tend to be more redundant.

[1] Unfortunately, for AI and IC recordings there are very few pairs in the first three groups, rendering this analysis unproductive for these data.

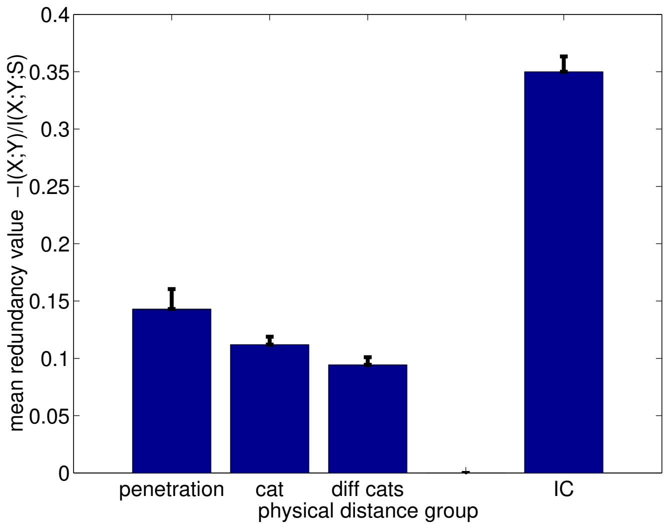

Figure 4.11: Mean normalized redundancy in MGB cell-pairs grouped according to physical distance type, and in IC. Error bars denote SEM of each group.

These results are in agreement with the idea that there is some functional segregation of neurons in the thalamus. Such a functional segregation is also expected in the cortex (Powell & Mountcastle, 1959; Hubel & Wiesel, 1962; Calvin, 1995). There is ample evidence for the presence of functional clustering and gradients in the Thalamus, both in the visual and auditory modalities. When combined with the results of the previous subsection, it suggests that MGB neurons are physically organized according to their functional sensitivity, but in addition to sensitivity to spectral content of sounds, additional acoustic components are used for thalamic organization. Furthermore, the local clustering of functional properties in the MGB is much weaker than in the IC, as suggested by the much lower redundancy values in the MGB, even among neurons recorded in the same penetration.

4.2 Coding acoustics

Neurons in the auditory system are usually analyzed in terms of their spectro-temporal selectivity (e.g. (DeCharms et al., 1998; Schnupp, Mrsic-Flogel, & King, 2001)), i.e. the patterns of stimulus energy across frequency and time to which the neuron is selective. In order to provide insight into the nature of the reduced redundancy among MGB and AI neurons, we investigated the information that single spikes convey about short-term spectro-temporal structures in our stimuli.

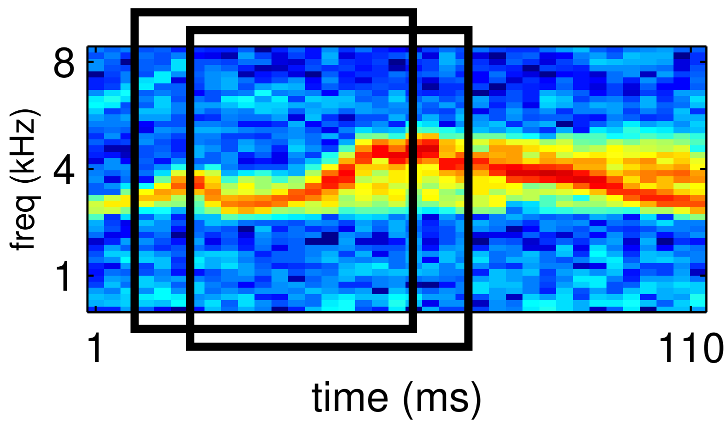

Figure 4.12: Illustration of the way overlapping segments of the bird chirp are created.

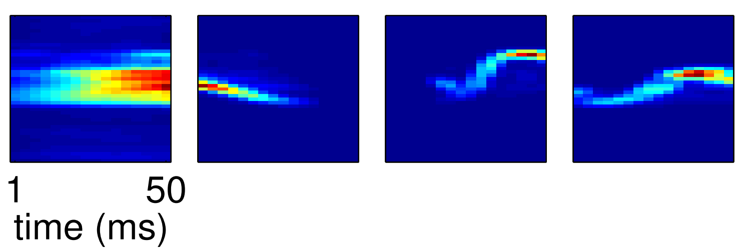

Figure 4.13: Four examples of the means of segment clusters after segments were grouped into 32 clusters.

To this end, the acoustic signal was cut into 50 ms segments with 49 ms overlaps, and the spectrogram of each segment was computed using a Hanning window with a length of 5.8 ms and a 2.8 ms overlap (Fig. 4.12). This set consisted of about 1500 segments. In order to reduce the dimensionality of the data, we aim to cluster the segments based on their physical properties, namely the spectro temporal energy pattern. The appropriate distance measure is however not known a-priori. We parametrized a family of metrics between segments by passing each segment through a sub-linear transformation $x \to x^{\alpha}$ in a pixel-wise manner. For α values smaller than 1, this transformation has the effect of emphasizing weak "pixels" as compared to pixels with high energy. The value of

α was chosen to maximize the information conveyed by the sample of AI neurons, yielding the value of 0.5. The segments were then clustered using K-means into 32 representatives using a dot product metric operated on the transformed spectrograms. The mean of each of these clusters was calculated (Fig. 4.13).

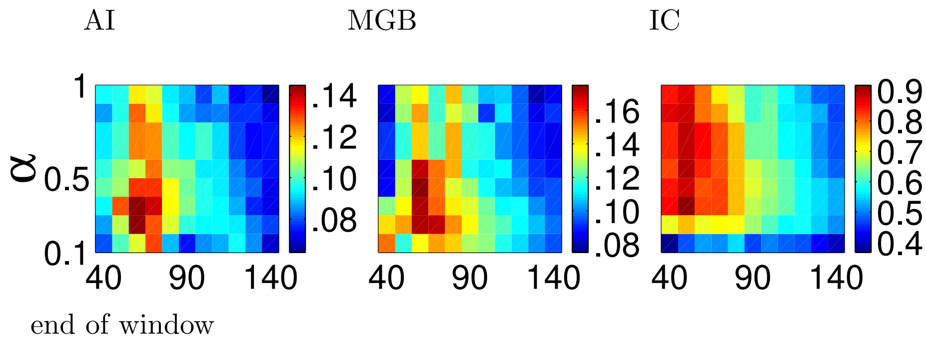

Figure 4.14: Total mutual information obtained from single spikes about acoustics, as a function of the power α and the end point of the temporal window (after stimulus onset). Start time of the window in this figure was taken to be 20 ms after stimulus onset for all three brain regions.

In order to estimate the information that neurons convey about the acoustic segments, the joint probability of spikes and representative segments was estimated. This was achieved by counting combinations of multi-cell spikes patterns, where the responses of a single neuron were considered to be 1 or 0, according to the presence or absence of a spike just following the appearance of a segment from a given cluster. Because of the simplicity of this response set, it was possible to create a joint distribution between clusters and combinations of spike patterns consisting of the responses of multiple neurons, again coupled under the conditional independence assumption.

We then calculated the MI between the spikes evoked in groups of cells and the stimulus clusters immediately preceding the spikes. IC spikes from single cells provided on average 9.2 times more the information than MGB cells, and 7.6 times more than AI cells about the identity of the preceding acoustic segment (Fig. 4.2). This suggests that in contrast with IC cells, AI and MGB neurons poorly discriminate between stimuli grouped on the basis of spectro-temporal energy structure. The underlying reason is the high sensitivity of AI neurons to small perturbations in the stimuli, as illustrated in Fig. 1B. This is in contradistinction to the way they code the identity of the stimuli, abbot which they convey almost half the information, compared to the same IC cells (Fig. 2.8 on chapter 2). Therefore, AI neurons distinguish between complex stimuli much better than would be expected based on the information they confer about the acoustic structure.

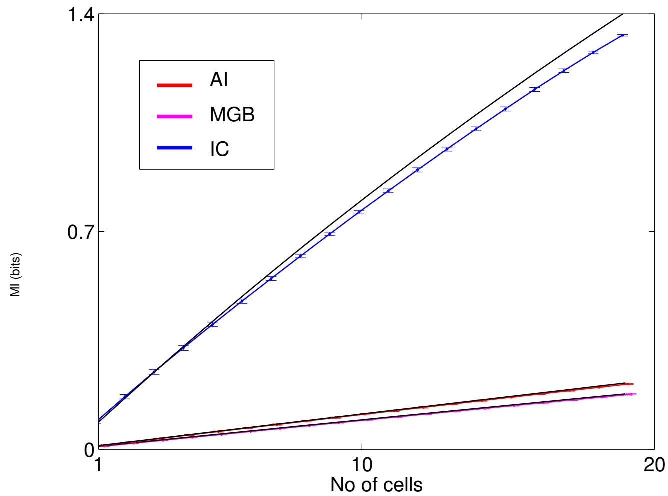

Figure 4.15: Information conveyed by single spikes about acoustic segments as a function of the number of neurons: IC-blue, MGB-magenta (partially covered by AI), A1-red. The black curves denote the expected information obtained from independent neurons. Error bars designate the standard error of the mean MI for several subsets of the same size. For each set size, analysis was repeated for 20 subsets and for 5 different randomization seeds.

To estimate redundancy, MI from a group of cells was again compared with the MI expected from independent cells, as explained in section 3.2.6. The normalized redundancy index (Fig. 4.2) reveals a significantly larger deviation from independence for IC cells than for MGB and cortical cells. These results suggest that in the coding of short-term spectro-temporal structure, as in the coding of stimulus identity, neural representations change along the ascending auditory system in a way that reduces redundancy among neurons.

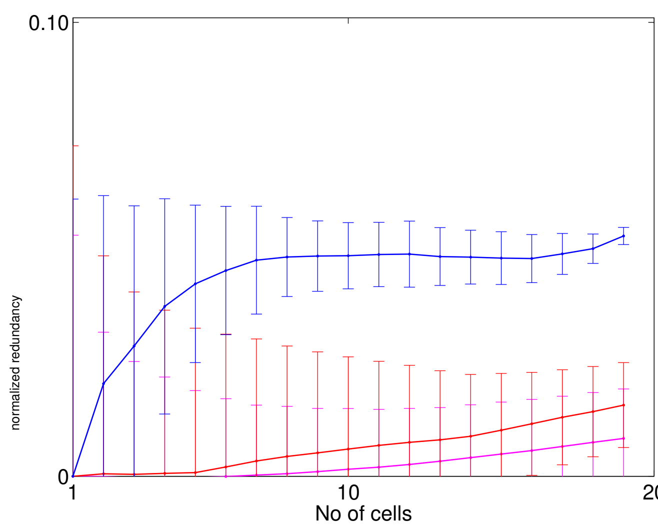

Figure 4.16: Normalized redundancy: the difference between the obtained information and the baseline of expected information from independent cells, normalized by the baseline.

4.3 Summary

This chapter has investigated the interactions among small groups of auditory neurons, and their relation to coding of acoustic stimuli. The unique setup of our dataset, namely, responses to the same set of stimuli in a number of auditory processing stations, allowed us to compare these interactions in different brain regions. We developed rigorous quantitative measures of information redundancy among sets of neuronal responses and applied them to the electrophysiological data.

Our main results are threefold. First, we showed that small groups of IC cells tend to be more redundant in the information they convey about the stimulus identity than AI and MGB cells. In other words, cells of higher regions in the processing hierarchy code stimuli in a more informationally independent manner. Secondly, we showed that this redundancy is significantly correlated with the BFs the IC cells but not with the BFs of AI or MGB cells. This means that frequency characterization poorly captures the type of processing neurons in MGB and AI perform. Finally, AI and MGB cells convey an order of magnitude less information about the spectro-temporal structure of the stimuli as compared with IC neurons. This suggest that neurons in MGB and AI succeed in coding the identity of the stimuli but without coding well the precise acoustical structures in it.

The low redundancy in AI and MGB, and the lack of correlation of this redundancy with the BFs of the cells, has strong implications, which go far beyond the assertion that BF responses are predictive for complex sounds. The reason is that current accepted methods of characterizing AI neurons, primarily the spectro temporal receptive field (STRF), imply redundancy between neurons that share spectro temporal characteristics. Although we find such redundant neurons in IC we do not find them in MGB or AI, demonstrating that STRF characterization in AI misses crucial aspects of neuronal coding even for stimuli as simple as those used in our work.

Chapter 5

Extracting relevant structures

A key problem in understanding auditory coding is to identify the acoustic features that neurons at various levels of the system code. If we can map the relevant stimulus features and trace how they change along the processing hierarchy, we can understand the processing properties of the system.

A principled approach for extracting relevant features was proposed by Tishby and co-authors (Tishby, Pereira, & Bialek, 1999), with the *Information Bottleneck* (IB) framework. This powerful approach aims at identifying structures of a variable X that have functional importance, by compressing X in a way that preserves information about another variable Y. In the context of the current problem it can be used to compress acoustic stimuli while preserving information about the distribution of neural responses, and use it to identify the stimulus aspects to which system is sensitive.

Unfortunately, the IB approach is insufficient for characterizing the processing that takes place in a brain region like the cortex. To understand the reason why, consider for example a case where one measures cortical activity in response to acoustic stimuli, and maps the acoustic features to which the cortical neurons respond. Such an analysis does not characterize cortical processing but rather the processing performed by the whole chain of processing stations that ends in the cortex. In fact, many of the features that are revealed this way do not reflect cortical processing, but processing that takes place at lower levels. For example, in the visual system, high level cells are often sensitive to the location of an object in the visual field (e.g. (Levy, Hasson, Avidan, Hendler, & Malach, 2001)), but such spatial sensitivity is computed by the narrow spatial sensitivity of the receptors in the retina. Similarly, in the auditory system, cortical neurons are frequency sensitive, but the frequency content of a stimulus is already determined at the receptor level in the cochlea. The conclusion is that when we aim to identify the processing that occurs at the cortex, we should search for features relevant to cortical neurons and not lower level neurons.

This problem is in fact common to all unsupervised learning tasks, and is caused by an inherent problem with the definition of unsupervised learning and relevance.

Namely, data contain several overlapping and competing structures, and the "right" structures depend on the task at hand. We will show here how this problem can be alleviated by using additional "non-interesting" data, such as neural activity in early processing stations. Using this type of irrelevant side data is the subject of the current chapter.

This chapter is organized as follows: Section 5.1 describes the information bottleneck approach. Section 5.2 explains how IB can be extended to use side information that allows to ignore irrelevant structures. Finally section 5.4 describes applications of IBSI in various domains.

5.1 Information bottleneck

5.1.1 Formulation

The information bottleneck method was first presented in (Tishby et al., 1999), and later extended in a series of papers (Slonim & Tishby, 2000; Friedman, Mosenzon, Slonim, & Tishby, 2001; Slonim, Friedman, & Tishby, 2001, 2002). This framework has proven powerful for numerous applications, such as clustering the objects of sentences with respect to the verbs (Pereira, Tishby, & Lee, 1993), documents with respect to their terms (Baker & McCallum, 1998; Hoffman, 1999; Slonim & Tishby, 2000), genes with respect to tissues (Friedman et al., 2001; Sinkkonen & Kaski, 2001), and visual stimuli with respect to spike patterns (2001-49, 2002) and vice versa (Dimitrov & Miller, 2001; Gedeon, Parker, & Dimitrov, 2002). We provide here a short review of this general approach. The reader is referred to (Slonim et al., 2002) for a fuller description.

In IB we are given a pair of variables X and Y, and their joint distribution $p(x, y)$. The goal is to compress X into another variable T, while preserving information about the variable Y. This compression is achieved via soft clustering; that is, a stochastic mapping of x's into t's, which we denote $p(t|x)$. This stochastic mapping determines the joint distribution of T, Y via the Markov relation $p(y|t) = \sum_x p(y|x)p(x|t)$. IB operates to learn a mapping that minimizes the information $I(X;T)$, while at the same time maximizing the information $I(T;Y)$ This is formally cast as a weighted tradeoff optimization problem

$$\min I(X;T) - \beta I(T;Y) \tag{5.1}$$

where β is the tradeoff parameter between compression and information preservation. It can also be shown (Gilad-Bachrach, Navot, & Tishby, 2003) that this problem is equivalent to a constrained optimization problem of minimizing $I(X;T)$ under a lower bound on $I(T;Y)$. In the constrained optimization problem β takes the role of a Lagrange multiplier.

The full formal optimization problem is written as

$$\mathcal{L} = I(X;T) - \beta I(T;Y) - \sum_x \lambda_x (\sum_t p(t|x) - 1) \tag{5.2}$$

where λ_x are Lagrange multipliers and the last term is used to enforce the normalization constraint $1 = \sum_t p(t|x)\ \forall x$.

The target function 5.2 can be differentiated w.r.t. $p(t|x)$ to find an optimal mapping for any predefined value of the parameter β. Fortunately this yields a set of self-consistent equations that the optimum obeys

$$\begin{aligned} p(t|x) &= \frac{p(t)}{Z} e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \\ p(t) &= \sum_x p(t|x)p(x) \\ p(y|t) &= \frac{1}{p(t)} \sum_x p(y|x)p(t|x)p(x) \end{aligned} \tag{5.3}$$

where

$$Z = \sum_t p(t) e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \tag{5.4}$$

is a normalization factor, known in statistical mechanics as the partition function.

5.1.2 IB algorithms

A series of algorithms was developed for solving the IB variational problem of Eq. 5.2. The self consistent equations 5.3 were used in (Tishby et al., 1999) to devise an iterative algorithm, in the spirit of the Blahut algorithm used in rate distortion theory (Blahut, 1972; Csiszar, 1974). This iterative algorithm named iIB by (Slonim et al., 2002) operates to optimize three free variables: $p(t|x)$, $p(t)$ and $p(y|t)$. It is based on the fact that the IB target function is convex w.r.t. each of the three variables independently but is not jointly convex. The iIB algorithm repeatedly fixes two of the free variables to their current values and optimizes over the third one. The pseudo code of iIB appears in figure 5.1.

Iterative IB

Input:

Joint distribution $p(x,y)$.
Trade-off parameter β .
Cardinality parameter K and a convergence parameter ϵ .

Output:

A (typically "soft") partition T of X into K clusters.

Initialization:

Randomly initialize $p(t \mid x)$ and find the corresponding $p(t),\ p(y|t)$ through Eq. 5.3.

Main loop

repeat

$p_{i+1}(t|x) \leftarrow \frac{p_i(t)}{Z_{i+1}(x,\beta)} e^{-\beta D_{KL}[p(y|x)||p(y|t)]}$, $\forall\ t \in T,\ \forall\ x \in X$

$p_{i+1}(t) \leftarrow \sum_x p(x) p_{i+1}(t|x)$, $\forall\ t \in T$

$p_{i+1}(y|t) = \frac{1}{p_{i+1}(t)} \sum_x p_{i+1}(t|x) p(x,y)$, $\forall\ t \in T,\ \forall\ y \in Y$

until (stopping criterion)

Figure 5.1: Pseudo-code of the iterative IB (iIB) algorithm. The stopping criterion used in (Slonim, 2002) is to stop if $\forall\ x \in X,\ JS_{\frac{1}{2},\frac{1}{2}}[p_{i+1}(t|x), p_i(t|x)] \leq \epsilon$, where JS is the Jensen-Shannon divergence defined in A.2.2. In principle, this procedure only guarantees convergence to a local extremum, hence it should be repeated several times with different initializations, and the solution which minimizes the target function $L = I(T;X) - \beta I(T;Y)$ should be chosen.

In many applications, hard clustering solutions can be more easily interpreted. This is particularly true when dealing with categorical data where averaging elements is meaningless. A series of hard clustering algorithms that operate to optimize the IB functional were suggested based on widely known clustering methods such as hierarchical clustering and K-means.

Among these hard clustering methods, the one that was found to be specifically useful is *sequential-IB* (sIB) (Slonim et al., 2001). In sIB, one starts with a random assignment of elements to clusters, and then iteratively improves the clustering by almost-greedy steps. At each such step, an element is taken randomly out of its cluster and is assigned to a cluster such that the overall score will be maximally improved. A pseudo code of this algorithm appears in Figure 5.2.

Sequential IB

```
Input:
    Joint distribution p(x, y).
    Trade-off parameter β.
    Cardinality value K.

Output:
    A partition T of X into K clusters.

Initialization:
    T ← random partition of X into K clusters

Main Loop:
    while not done
        done ← TRUE
        for every x ∈ X :
            Remove x from current cluster t(x)
            t^new(x) ← argmin_{t∈T} ΔL({x}, t)
            if t^new(x) ≠ t(x)
                done ← FALSE .
            Merge x into t^new(x)
        end for
    end while
```

Figure 5.2: Pseudo-code of the sequential-IB (sIB) algorithm. In principle, this procedure only guarantees convergence to a local extremum, hence should be repeated for different initializations, and the solution which maximizes the target function $L = I(X;T) - \beta I(T;Y)$ should be chosen.

5.2 Relevant and irrelevant structures

5.2.1 The problem

A fundamental goal of machine learning is to find regular structures in given empirical data, and use them to construct predictive or comprehensible models. This general goal, unfortunately, is very ill defined, as many data sets contain alternative, often conflicting, underlying structures. For example, documents may be classified either by subject or by writing style; spoken words can be labeled by their meaning or by the identity of the speaker; proteins can be classified by their structure or function - all are valid alternatives. Which of these alternative structures is "relevant" is often implicit in the problem formulation.

The problem of identifying "the" relevant structures is commonly addressed in supervised learning tasks, by providing a "relevant" label to the data, and selecting features that are discriminative with respect to this label. As described in the previous section, an information theoretic generalization of this supervised approach has been proposed in (Pereira et al., 1993; Tishby et al., 1999) through the "information bottleneck method" (IB).

An important condition for this approach to work is that the auxiliary variable indeed corresponds to the task. In many situations, however, such a "pure" variable is not available. The auxiliary variable may in fact contain alternative and even conflicting structures. We show here that this general and common problem can be alleviated by providing "negative information"; that is, information about "non-important", or irrelevant, aspects of the data that can interfere with the desired structure during learning.

As an illustration, consider a simple nonlinear regression problem. Two variables x and y are related through a functional form $y = f(x) + \xi$, where $f(x)$ is in some known function class and ξ is noise with some distribution that *depends on* x. When given a sample of (x, y) pairs with the goal of extracting the relevant dependence $y = f(x)$, the noise ξ - which may contain information on x and thus interfere with extracting y - is an irrelevant variable. Knowing the joint distribution of (x, ξ) can of course improve the regression result.

The problem of identifying stimulus features that are relevant for neural activity, presented at the beginning of this chapter is a more real-life example. Another real world example can be found in the analysis of gene expression data. Such data, as generated by DNA-chips technology, can be considered as an empirical joint distribution of gene expression levels and different tissues, where the tissues are taken from different biological conditions and pathologies. The search for expressed genes that testify to the existence of a pathology may be obscured by genetic correlations that also exist in other conditions. Here again a sample of irrelevant expression data, taken for instance from a healthy population, can enable clustering analysis to focus on the pathological features alone, and ignore spurious structures.

These examples, and numerous others, are all instantiations of a common problem: in order to better extract the relevant structures information about the irrelevant components of the data should be used. Naturally, various solutions have been suggested to this basic problem in many different contexts (e.g. spectral subtraction, weighted regression analysis). The section below presents a general unified information theoretic framework for such problems, extending the original information bottleneck variational problem to deal with discriminative tasks of this nature.

5.2.2 Information theoretic formulation

To formalize the problem of extracting relevant structures consider first three categorical variables X, Y^+ and Y^- whose co-occurrence distributions are known. Our goal is to uncover structures in $P(X, Y^+)$, that do not exist in $P(X, Y^-)$. The distribution $P(X; Y^+)$ may contain several conflicting underlying structures, some of which may also exist in $P(X, Y^-)$. These variables stand for example for a set of stimuli

X, a set of neural responses Y^+ from a brain region whose code we wish to explore, and an additional set of neural responses Y^-. Other examples can be a set of terms and two sets of documents or a set of genes and two sets of tissues with different biological conditions. In all these examples Y^+ and Y^- are *conditionally independent* given X. We thus make the assumption that the joint distribution factorizes as $p(x, y^+, y^-) = p(x)p(y^+|x)p(y^-|x)$.

The relationship between the variables can be expressed by a Venn diagram (Figure 5.3), where the area of each circle corresponds to the entropy of a variable (see e.g. (Cover & Thomas, 1991) p.20 and (Csiszar & J.Korner, 1997) p.50 for discussion of this type of diagrams) and the intersection of two circles corresponds to their mutual information.

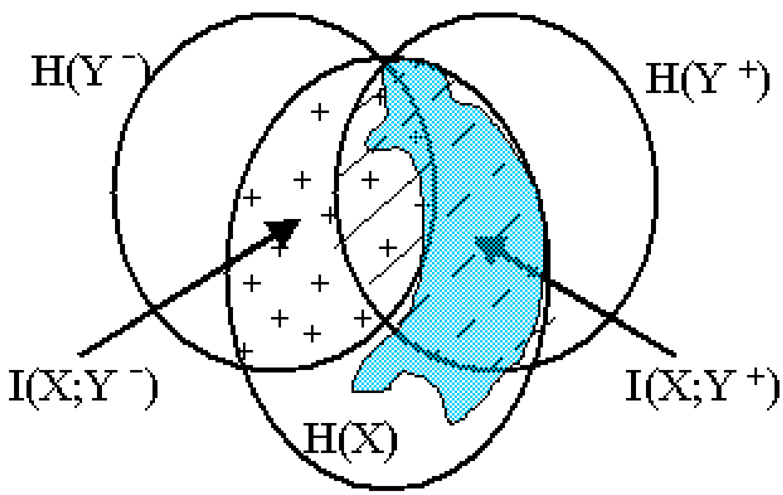

Figure 5.3: A Venn diagram illustrating the relations between the entropy and mutual information of the variables X,Y^+,Y^-. The area of each circle corresponds to the entropy of a variable, while the intersection of two circles corresponds to their mutual information. As Y^+ and Y^- are independent given X, their mutual information vanishes when x is known, thus all their overlap is included in the circle of X.

To identify the relevant structures in the joint distribution $p(x, y^+)$, we aim to extract a compact representation of the variable X with minimal loss of mutual information about the relevant variable Y^+, and *at the same time* with maximal loss of information about the irrelevance variable Y^-. The goal of *information bottleneck with side information* (IBSI) is therefore to find a stochastic map of X to a new variable T, $p(t|x)$, in a way that maximizes its mutual information about Y^+ and minimizes the mutual information about Y^-. In general one can only achieve this goal perfectly in the asymptotic case and the finite case leads to a sub optimal compression, an example of which is depicted in the blue region in figure 5.3. These constraints can be cast into a single variational functional that we aim to minimize

$$\mathcal{L} = I(X;T) - \beta \left[I(T;Y^+) - \gamma I(T;Y^-) \right] \ . \qquad (5.5)$$

This functional consists of three terms that quantify *compression* ($I(T;X)$), information preservations ($I(T;Y^+)$), and information removal ($I(T;Y^-)$). The Lagrange multiplier β determines the tradeoff between compression and information extraction while the parameter γ determines the tradeoff between preservation of information about the relevant Y^+ variable and removal of information about the irrelevant one Y^-. In some applications, such as in communication, the value of γ may be determined by the relative cost of transmitting the information about Y^- by other means (see (Wyner, 1975)). In others, as shown for example in (Globerson, Chechik, & Tishby, 2003), real systems exhibit phase transitions phenomena at designated values of γ, which allow to identify the transition points in the parameter space and reveal the qualitatively different regimes of the system's states space

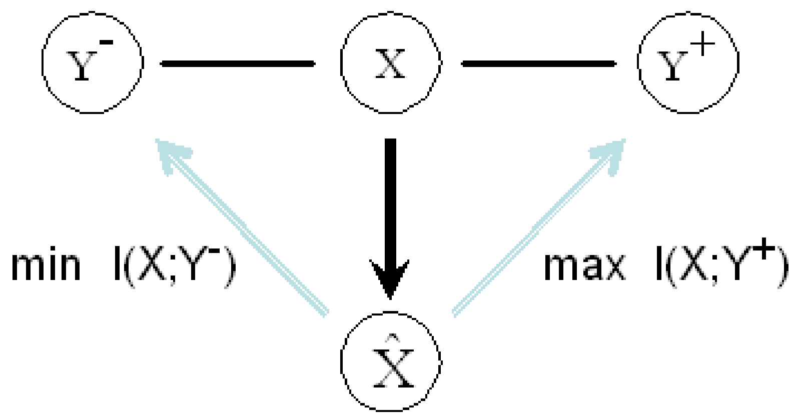

Figure 5.4: A graphic representation of discriminative IB. Given the three variables X,Y^+ and Y^-, we seek a compact stochastic representation T of X which preserves information about Y^+ but removes information about Y^-. In this graph Y^+ and Y^- are indeed conditionally independent given X.

The structure of the data in the original information bottleneck formulation is characterized by the information curve; namely, the values of the IB functional as a function of different β values. In IBSI the data are similarly characterized by the functional values but this time as a function of two free parameters, β and γ, creating a two dimensional manifold. Such a manifold is demonstrated in Figure 5.5.

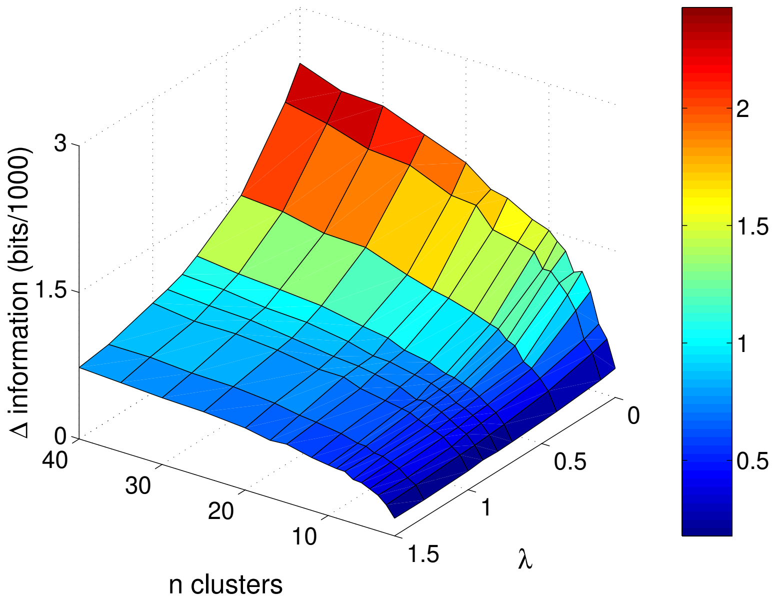

Figure 5.5: The information plane. The value of the IBSI functional as a function of the two Lagrange coefficient values β and γ. The information plane plotted in this figure was calculated using face images as described in section 5.4.4.

The IB variational problem of Eq. 5.1 is a special case of our current variational problem with $\gamma = 0$, namely, no side or irrelevant information is available. In this case only the distributions $p(t|x)$, $p(t)$ and $p(y^+|t)$ are determined.

5.2.3 Solution characterization

The complete Lagrangian of this constrained optimization problem is given by

$$\begin{aligned}\mathcal{L}[p(t|x)] &= I(X;T) - \beta\left[I(T;Y^+) - \gamma I(T;Y^-)\right] \\ &\quad - \sum_x \lambda(x) \sum_t p(t|x)\end{aligned} \tag{5.6}$$

where $\lambda(x)$, are the normalization Lagrange multipliers that enforce the constraint $\sum_t p(t|x) = 1$. Here, the minimization is performed with respect to the stochastic mapping $p(t|x)$, taking into account its probabilistic relations to $p(z)$, $p(y^+|z)$ and $p(y^-)$.

Theorem 5.2.1 : *The extrema of $\mathcal{L}$ obey the following self consistent equations*

$$\begin{aligned}p(t|x) &= \frac{p(t)}{Z} e^{-\beta\left(D_{KL}[p(y^+|x)||p(y^+|t)] - \gamma D_{KL}[p(y^-|x)||p(y^-|t)]\right)} \\ p(t) &= \sum_x p(t|x)p(x) \\ p(y^+|t) &= \frac{1}{p(t)} \sum_x p(y^+|x)p(t|x)p(x) \\ p(y^-|t) &= \frac{1}{p(t)} \sum_x p(y^-|x)p(t|x)p(x)\end{aligned} \tag{5.7}$$

where

$$Z = \sum_t p(t) e^{-\beta\left(D_{KL}\left[p(y^+|x)||p(y^+|t)\right] - \gamma D_{KL}\left[p(y^-|x)||p(y^-|t)\right]\right)} \tag{5.8}$$

is a normalization factor.

Proof: Following the Markov relation $p(y|x,t) = p(y|x)$, we write

$$\begin{aligned} p(y,t) &= \sum_x p(y,t|x)p(x) \\ &= \sum_x p(y|t,x)p(t|x)p(x) = \sum_x p(y|x)p(t|x)p(x) \end{aligned} \tag{5.9}$$

where $p(x) = \sum_{y^+}\sum_{y^-} p(y^+, y^-, x)$, and obtain for the second term of Eq. 5.6

$$\begin{aligned} \frac{\delta I(T;Y^+)}{\delta p(t|x)} &= \frac{\delta}{\delta p(t|x)} \sum_{t'}\sum_{y^+}\sum_{x'} p(y^+|x')p(t'|x')p(x') \times \\ &\quad \times \log\left(\frac{p(y^+|t')}{p(y^+)}\right) \\ &= p(x)\sum_{y^+} p(y^+|x) \log\left(\frac{p(y^+|t)}{p(y^+|x)}\frac{p(y^+|x)}{p(y^+)}\right) \\ &= -p(x)D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\quad +p(x)D_{KL}[p(y^+|x)||p(y^+)] \end{aligned} \tag{5.10}$$

a similar differentiation for the remaining terms yields

$$\begin{aligned} \frac{\delta}{\delta p(t|x)}\mathcal{L} &= p(x)\log\left(\frac{p(t|x)}{p(t)}\right) \\ &\quad -p(x)\beta D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\quad -p(x)\beta\gamma D_{KL}[p(y^-|x)||p(y^-|t)] \\ &\quad +p(x)\lambda(x,y^+,y^-) \end{aligned} \tag{5.11}$$

where

$$\begin{aligned} \lambda(x,y^+,y^-) &= \frac{\lambda(x)}{p(x)} + \\ &\quad \beta\left(D_{KL}[p(y^+|x)||p(y^+)] - \gamma D_{KL}[p(y^-|x)||p(y^-)]\right) \end{aligned} \tag{5.12}$$

holds all terms independent of t. Equating the derivative to zero then yields the first equation of proposition 5.2.1. The remaining equations hold due to the Markov relations $p(y^+|t,x) = p(y^+|x)$, $p(y^-|t,x) = p(y^-|x)$. $\square$

The formal solutions of the above variational problem have an exponential form which is a natural generalization of the solution of the original IB problem (Eq. 5.2).

5.2.4 IBSI and discriminative models

The side information formalism naturally raises the question of its relation to discriminative models.

Two types of discriminative models come to mind in this context. First, one may consider a learning task whose goal is to discriminate between samples from Y^+ and Y^-. This task, widely studied in the context of binary classification is inherently different from the task we are interested in here. In the second type of tasks, we are interested in discriminative training of mixture models for $p(X, Y^+)$ and $p(X, Y^-)$. This is the subject of the section here.

Consider a mixture model in which we generate an observation y for a given x, by the following procedure: At first, every x is assigned to a cluster t according to a prior distribution $\pi(t)$. This cluster determines a distribution $p(y|t(x))$, parametrized by $\theta(y|t(x))$. Then, samples of (x, y) pairs are generated as follows: First, a value of x is chosen according to a prior distribution $p_x(x)$. Then, multiple samples of y are drawn from the distribution $p(y|t(x))$. We assume that y takes discrete values.

Let us denote the empirical count generated by this process by $n^+(x_i, y_j)$. We also assume that an additional count was generated by a similar process, and denote it by $n^-(x_i, y_j)$. The cardinality $|X|$ is the same for the two counts, but the remaining parameters are not necessarily equivalent.

In *generative training*, one aims at finding the parameters $\theta(y|t)$ (clusters' centroids) and $t(x)$ (cluster assignments), such that the likelihood of the count given the parameters is maximized. The likelihood of the n^+ count is formally written as

$$\begin{aligned} & L(n^+ \mid \pi, \theta^+, {p_x}^+, t(x)) \\ = & \prod_{i=1}^{|X|} \pi(t(x_i)) \prod_{l=1}^{n} {p_x}^+(x_l)\theta(y|t(x_l)) \\ = & \prod_{i=1}^{|X|} \pi(t(x_i)) \prod_{j=1}^{|X|} \prod_{j=1}^{|Y|} \left[{p_x}^+(x_i)\theta(y_j|t(x_i))\right]^{n(x_i,y_j)} \end{aligned} \tag{5.13}$$

and similarly for the n^- count. Searching for the parameters that maximize this likelihood can be achieved using standard maximum likelihood algorithms for mixture models, such as Expectation Maximization (EM) (Dempster, Laird, & Rubin, 1977). These are applied *separately* on the counts n^+ and n^- to estimate the parameters that maximize each of the likelihood terms.

In *discriminative training*, one aims at finding the parameters that maximize the log likelihood ratio

$$\max_{\pi,\theta^+,\theta^-,{p_x}^+,{p_x}^-,t(x)} \log\left(\frac{L(n^+|\pi, \theta^+, {p_x}^+, t(x))}{L(n^-|\pi, \theta^-, {p_x}^-, t(x))}\right) \tag{5.14}$$

where maximization allows to fit the distribution parameters p_x and θ separately for each of the counts, **but the prior $\pi(t)$ and the assignment $t(x)$ is common to both counts**.

This log likelihood ratio equals

$$
\begin{aligned}
R &= \log\left(\frac{L(n^+|\pi,\theta^+,{p_x}^+,t(x))}{L(n^-|\pi,\theta^-,{p_x}^-,t(x))}\right) = \qquad (5.15)\\
&= \sum_{i=1}^{|X|}\pi(t(x_i)) + \sum_{i=1}^{|X|}\sum_{j=1}^{|Y^+|} n^+(x_i,y_j^+)\log\left({p_x}^+(x_i)\theta^+(y_j^+|t(x_i))\right)\\
&\quad -\sum_{i=1}^{|X|}\pi(t(x_i)) - \sum_{i=1}^{|X|}\sum_{j=1}^{|Y^-|} n^-(x_i,y_j^-)\log\left({p_x}^-(x_i)\theta^-(y_j^-|t(x_i))\right)
\end{aligned}
$$

In the limit of large counts $n \equiv \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|} n(x_i,y_j) \to \infty$ we have $\frac{1}{n}n(x_i,y_j) \to p(x_i,y_j)$, and the likelihood ratio converges to

$$
\begin{aligned}
R &\to n\sum_{i=1}^{|X|}\sum_{j=1}^{|Y^+|} p^+(x_i,y_j)\log\left({p_x}^+(x_i)\theta^+(y_j|t(x_i))\right) \qquad (5.16)\\
&\quad -n\sum_{i=1}^{|X|}\sum_{j=1}^{|Y^-|} p^-(x_i,y_j)\log\left({p_x}^-(x_i)\theta^-(y_j|t(x_i))\right)\\
&= nE\left[\log\left(\frac{{p_x}^+(x_i)\theta^+(y_j|t(x_i))}{{p_x}^-(x_i)\theta^-(y_j|t(x_i))}\right)\right]
\end{aligned}
$$

where expectation is over the joint distribution $p(x,y+,y-)$. We conclude that this discriminative training of mixture models aims at maximizing the expected log likelihood ratio, which has two terms: the priors $\frac{{p_x}^+(x_i)}{{p_x}^-(x_i)}$ and the conditionals $\frac{\theta^+(y_j|t(x_i))}{\theta^-(y_j|t(x_i))}$.

For the purpose of comparison, we now rewrite the information preservation and removal terms of IBSI (i.e. taking $\beta \to \infty$)

$$
\begin{aligned}
I(T;Y^+) &- \gamma I(T;Y^-) = \qquad (5.17)\\
&= \sum_t\sum_{y^+}\sum_{y^-} p(t,y^+,y^-)\log\left(\frac{p(y^+|t)}{p(y^+)}\right)\\
&\quad -\gamma\sum_t\sum_{y^+}\sum_{y^-} p(t,y^+,y^-)\log\left(\frac{p(y^-|t)}{p(y^-)}\right)\\
&= \left\langle \log\left(\frac{p(y^+|t)}{p(y^-|t)^\gamma}\frac{p(y^-)^\gamma}{p(y^+)}\right)\right\rangle_{p(t,y^+,y^-)}
\end{aligned}
$$

This means that when $\gamma = 1$, and the marginal distributions are equal ${p_x}^+(x) = {p_x}^-(x)$, the two expressions become similar. Moreover, when the cardinality of Y^+ and Y^- is the same and the marginal distributions of Y^+ and Y^- are uniform, the last equation becomes identical to the discriminative training of the mixture model. In the general case however such an equivalence does not hold.

5.2.5 Multivariate extensions

The above setup can be extended to the case of multiple variables on which multi-information should be preserved about $\{y_1^+, ..., y_{N^+}^+\}$ and variables on which multi-information should be removed about $\{y_1^-, ..., y_{N^-}^-\}$, as discussed in (Friedman et al., 2001). This yields

$$\log \frac{p(t|x)}{p(t)} \propto -\log(Z) - \sum_i \gamma_i^+ D_{KL}\Big[p(y_i^+|x)||p(y_i^+|t)\Big] + \sum_i \gamma_i^- D_{KL}\Big[p(y_i^-|x)||p(y_i^-|t)\Big] \quad (5.18)$$

which can be solved together with the other self-consistent conditions, similarly to Eq. 5.8.

Iterative IBSI

Input:
Joint distributions $P(x, y^+)$, $P(x, y^-)$
Trade-off parameters β,γ
Number of clusters K
Stopping parameter ϵ

Output:
A soft partition $P(T|X)$ of X into T using K clusters.

Main:
Initialize $p(t|x)$ and set i=0.
Calculate $p_i(t)$,$p_i(y^+|t)$ and $p_i(y^- \mid t)$ through Eq. 5.7.
repeat

$i \leftarrow i + 1$

$p_i(y^+|t) \leftarrow \frac{1}{p_{i-1}(t)} \sum_x p_{i-1}(t|x)p(x, y^+)$, $\forall t \in T, \ \forall y^+ \in Y^+$

$p_i(y^-|t) \leftarrow \frac{1}{p_{i-1}(t)} \sum_x p_{i-1}(t|x)p(x, y^-)$, $\forall t \in T, \ \forall y^- \in Y^-$.

$p_i(t|x) \leftarrow \frac{p_{i-1}(t)}{Z^i} \times$

$\exp\left(-\beta(D_{KL}[p(y^+|x)||p_i(y^+|t)] - \gamma D_{KL}[p(y^-|x)||p_i(y^-|t)])\right)$

$\forall t \in T, \ \forall x \in X$

$p_i(t) \leftarrow \sum_x p(x)p_{i+1}(t|x)$, $\forall t \in T$

until $\forall\, x \in X, \ JS[p_i(t|x), p_{i-1}(t|x)] \leq \epsilon$.

Figure 5.6: Pseudo-code of the iterative IBSI algorithm. JS denotes the Jensen-Shannon divergence defined in A.2.2, and Z the partition (normalization) function. This procedure only guarantees convergence to a local extremum, hence should be repeated for different initializations and a solution that maximizes $\mathcal{L} = I(T; X) - \beta I(T; Y^+) + \beta\gamma I(T; Y^-)$ should be chosen.

5.3 IBSI algorithms

The common formulation of IB and IBSI allows us to adopt a series of algorithms that were originally developed for IB. A detailed description of these original algorithms can be found in (Slonim et al., 2002), together with a comparison of their performance on text categorization data. Our main interest in this section is the novel difficulties that the IBSI formulation poses. We therefore focus on two algorithms. The first is an iterative Blahut-Arimoto style algorithm that iterates between the self consistent equations 5.7 derived above. The second is a heuristic sequential K-means algorithms, that was shown by (Slonim et al., 2002) to achieve good empiric results for the IB case ($\gamma = 0$).

5.3.1 Iterating the fix point equations

The optimization problem defined in Eq. 5.5 requires finding the optimum over four parameters $p(t|x)$, $p(t)$, $p(y^+|t)$ and $p(y^-|t)$. Fortunately, even though the Lagrangian 5.6 is not jointly convex with respect to these parameters, it is convex w.r.t. $p(t|x)$ and $p(t)$ separately under some conditions. This convexity allows us to use the fixed points equations of 5.7 in an iterative manner to approach the optimum. This is achieved by iteratively fixing all the equations but one, and optimizing over the non freezed parameter. This calculation is performed for all values of $x \in X$, $t \in T$, $y^+ \in Y^+$ and $y^- \in Y^-$, and is repeated until convergence.

This algorithm, *iterative-IBSI*, whose pseudo code is given in figure 5.6, is similar in form but inherently different from the iterative-IB (iIB) algorithm described in (Slonim et al., 2001). This is due to the additional term in the IBSI target function, that makes the problem concave rather than convex with respect to $p(y^-|t)$. Instead of iteratively optimizing over the four free parameters $\{p(t|x), p(t), p(y^+|t), p(y^-|t)\}$, we use the Markovian relation for $p(y|t)$ explicitly, and optimize over $p(t|x)$ and $p(t)$. While convergence is always achieved with the original IB algorithm, convergence for IBSI is only guaranteed for a limited range of γ values. The following theorem establishes that such a range of parameters indeed exists.

Theorem 5.3.1 : Convergence of iterative IBSI
There exists a range of γ values for which iterative-IBSI converges to a stationary fixed point of the IBSI functional.

The proof follows the following line: We define an auxiliary (Lyapunov) function $\mathcal{F}$, and show that it is bounded from below and convex w.r.t. the relevant parameters for some non-empty region of γ. We then show that $\mathcal{F}$ decreases at each step of the algorithm unless it has reached a fixed point.

We begin the proof by defining the free energy functional for the IBSI

Definition 5.3.2 : IBSI Free energy
The free energy for a given $p(x,t)$ and $p(t)$ is defined as

$$\mathcal{F} \equiv -\langle \log(Z) \rangle = -\sum_{x,t} p(x)p(t \mid x) \log Z(x,\beta,\gamma) \ , \tag{5.19}$$

where $Z(x,\beta,\gamma)$ is defined by

$$\begin{aligned} -\log Z(x,\beta,\gamma) &= \log \frac{p(t|x)}{p(t)} + \beta D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\quad -\beta\gamma D_{KL}[p(y^-|x)||p(y^-|t)] \ . \end{aligned} \tag{5.20}$$

At the extremum points, when all four equations of 5.7 hold, Z is the normalization (partition) function of $p(t|x)$, as follows from equation 5.7. Substituting Z into 5.19 we have

$$\begin{aligned} \mathcal{F} &= \sum_{x,t} p(x,t) \log \frac{p(t|x)}{p(t)} + \beta \sum_{x,t} p(x,t) D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\quad -\beta\gamma \sum_{x,t} p(x,t) D_{KL}[p(y^-|x)||p(y^-|t)] \end{aligned} \tag{5.21}$$

Lemma 5.3.3 : *$\mathcal{F}$ equals the IBSI functional $\mathcal{L}$ up to an additive constant.*

Proof: We rewrite the average of $D_{KL}[p(y|x)||p(y|t)]$ using the Markov relation $p(y|t) = \sum_x p(y|x)p(x|t)$ to obtain

$$\begin{aligned} \sum_{x,t} p(x,t) D_{KL}[p(y|x) \ || \ p(y|t)] &= \sum_{x,t} p(x,t) \sum_y p(y|x) \log \left(\frac{p(y|x)}{p(y|t)} \right) \\ &= \sum_x p(x) \sum_y p(y|x) \log(p(y|x)) \\ &\quad - \sum_t p(t) \sum_y \log(p(y|t)) \sum_x p(y|x)p(x|t) \\ &= -H(Y|X) - \sum_t p(t) \sum_y \log(p(y|t)) p(y|t) \\ &= -H(Y|X) + H(Y|T) \\ &= -H(Y|X) - I(Y;T) + H(Y) \\ &= I(X;Y) - I(Y;T) \quad . \end{aligned} \tag{5.22}$$

Therefore

$$\begin{aligned} \mathcal{F} &= I(T;X) + \beta D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\qquad -\beta\gamma D_{KL}[p(y^-|x)||p(y^-|t)] \\ &= I(T;X) + \beta(I(X;Y^+) - I(T;Y^+)) \\ &\qquad -\beta\gamma(I(X;Y^-) - I(T;Y^-)) \\ &= \mathcal{L} + \beta I(X;Y^+) - \beta\gamma I(X;Y^-) \quad , \end{aligned} \tag{5.23}$$

Thats is, $\mathcal{F}$ equals $\mathcal{L}$ up to an additive constant that depends on the empirical distributions $p(X, Y^+)$ and $p(X, Y^-)$. $\square$

Lemma 5.3.4 : *$\mathcal{F}$ is bounded from below.*

Proof: Following the lemma 5.3.3, we denote $c \equiv \beta I(X; Y^+) - \beta\gamma I(X; Y^-)$ and write

$$\begin{aligned} \mathcal{F} &= c + \mathcal{L} \\ &= c + I(T; X) - \beta I(T; Y^+) + \beta\gamma I(T; Y^-) \\ &\geq c + 0 - \beta I(T; Y^+) + 0 \\ &\geq c - \beta H(Y^+) \end{aligned} \tag{5.24}$$

and the last term is a constant. $\square$

Two observations should be made at this point. First, $\mathcal{F}$ is not convex w.r.t. $p(y^-|t)$, but concave. Secondly, $\mathcal{F}$ is not jointly convex w.r.t. all its parameters. With all this, we now characterize a weaker form of convexity that guarantees convergence of the algorithm.

Lemma 5.3.5 : *Under the Markov relations $p(y^+|t) = \sum_x p(y|x)p(x|t)$, $p(y^-|t) = \sum_x p(y^-|x)p(x|t)$, there exists a range of γ values for which $\mathcal{F}$ is convex with respect to $p(t)$ and to $p(t|x)$ independently.*

Proof: The convexity of $\mathcal{F}$ w.r.t. $p(t)$ stems directly from the concavity of the log function, and holds for all values of γ.

From the definition of the $\mathcal{F}$ (equation 5.21), it is a weighted linear combination of D_{KL} terms, where all the terms with negative coefficients are weighted by γ. We can therefore use the convexity of D_{KL} to infer the convexity of this linear combination.

Since $D_{KL}[p||q]$ is a convex function of q (see e.g. (Cover & Thomas, 1991), Chapter 2), $D_{KL}[p(y^+|x)||p(y^+|t)]$ is convex with respect to $p(y^+|t)$ for all y and t. Thus, under the Markov relation $p(y^+|t) = \sum_x p(y^+|x)p(x|t)$, $D_{KL}[p(y^+|x)||p(y^+|t)]$ is a convex function of $p(x|t)$, and its second derivative w.r.t. $p(x|t)$ is strictly negative. So is its derivative w.r.t $p(t|x)$ Similarly, $-D_{KL}[p(y^-|x)||p(y^-|t)]$ is concave w.r.t. $p(x|t)$, and its second derivative is positive. Since all the terms with positive second derivatives are weighted by γ, there exist a critical value $\gamma^{max} > 0$ for which the second derivative of the linear combination is strictly negative, thus the functional is convex w.r.t. $p(x|t)$. $\square$

We are now ready to prove theorem 5.3.1.

Proof: Convergence of iterative-IBSI

At each step of the algorithm, we first use the Markov relation to set $p(y^+|t)$ and $p(y^-|t)$, and then set $p(x|t)$ and $p(t)$ such that the first derivative of $\mathcal{L}$ vanishes. Since $\mathcal{F}$ equals $\mathcal{L}$ up to a constant, it also zeroes the first derivative of $\mathcal{F}$. Moreover, since

$\mathcal{F}$ is convex w.r.t. $p(t|x)$ and $p(t)$, it decreases the value of $\mathcal{F}$. In addition, since $\mathcal{F}$ is bounded from below, the algorithm must converge, and reach the stopping criterion. Finally, consider the case where iterative-IBSI reached a fixed point of $\mathcal{F}$; that is, subsequent steps of the algorithm do not change $\mathcal{F}$ any longer. This only happens when the fixed points equations are satisfied, i.e. the algorithm has reached a local minima of $\mathcal{L}$. $\square$

Note however that this only proves the convergence of the algorithm to a stationary point w.r.t the target function $\mathcal{L}$ and not w.r.t. $p(t|x)$. If several local minima exist with the same value it does not rule out the possibility that the distribution cycles between them on consecutive steps of the algorithm.

5.3.2 Hard clustering algorithms

As in the case of IB, various heuristics can be applied, such as deterministic annealing - in which increasing the parameter β is used to obtain finer clusters; greedy agglomerative hard clustering (Slonim & Tishby, 1999); or a sequential K-means like algorithm (sIB) (Slonim et al., 2002). The latter provides a good compromise between top-down annealing and agglomerative greedy approaches and achieves excellent performance, and its pseudo-code is given in Figure 5.7.

Sequential IBSI

Input:
Joint distributions $p(X, Y^+)$, $p(X, Y^-)$
Trade-off parameters β,γ
Number of clusters K

Output:
A hard partition $P(T|X)$ of X into T using K clusters.

Main:
Randomly initialize $p(t|x)$ and set i=0.
while not *Done*
  $Done \leftarrow TRUE$.
  for every $x \in X$:
    Remove x from current cluster, $t(x)$.
    $t^{new}(x) = argmin_{t \in T} \Delta L_{max}(\{x\}, t)$
    if $t^{new}(x) \neq t(x)$,
      $Done \leftarrow FALSE$.
    Merge x into $t^{new}(x)$
  end for
end while

Figure 5.7: Pseudo-code of sequential IBSI algorithm. The hard clustering is represented in the mapping $p(t|x)$, whose values are either 0 or $1/|X|$. JS denotes the Jensen-Shannon divergence defined in A.2.2, and Z the partition (normalization) function. This procedure should in principle be repeated for different initializations and then the solution that optimizes the target function $\mathcal{L} = I(T; X) - \beta I(T; Y^+) + \beta\gamma I(T; Y^-)$. should be chosen

5.4 Applications

We describe several applications of IBSI. First we illustrate its operation using a simple synthetic example. Then we apply it to two "real world" problems: hierarchical text categorization in the domain of information retrieval (section 5.4.3) and feature extraction for face recognition (section 5.4.4). Finally its application to auditory neural coding is described in section 5.4.5.

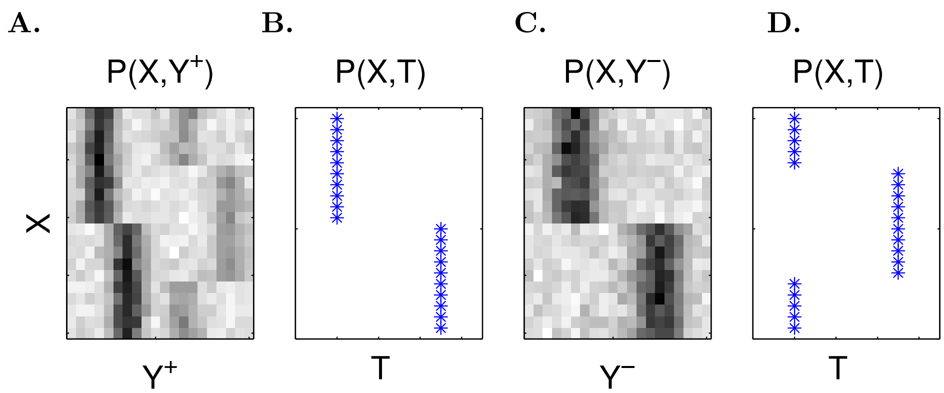

Figure 5.8: Demonstration of IBSI operation. **A.** A joint distribution $P(X, Y^+)$ that contains two distinct and conflicting structures. **B.** Clustering X into two clusters using the information bottleneck method separates upper and lower values of X, according to the stronger structure. **C.** A joint distribution $P(X, Y^-)$ that contains a single structure, similar in nature to the stronger structure $P(X, Y^+)$. **D.** Clustering X into two clusters using IBSI successfully extracts the weaker structure in $P(X, Y^+)$.

5.4.1 A synthetic illustrative example

To demonstrate the ability of our approach to uncover weak but interesting hidden structures in data, we designed a co-occurrences matrix contains two competing sub-structures (see figure 5.8A). For demonstration purposes, the matrix was created such that the stronger structure can be observed on the left and the weaker structure on the right. Compressing X into two clusters while preserving information on Y^+ using IB ($\gamma = 0$), yields the clustering of figure 5.8C, in which the first half of the x's are all clustered together. This clustering follows from the strong structure on the left of 5.8A.

We now create a second co-occurrence matrix, to be used for identifying the relevant structure, in which each half of X yield similar distributions $P(y^-|x)$. Applying our discriminative clustering algorithm now successfully ignores the strong but irrelevant structure in $P(Y^+, X)$ and retrieves the weak structure. Importantly, this is done in an unsupervised manner, without explicitly pointing to the irrelevant structure.

5.4.2 Model complexity identification

A fundamental problem in modeling data is to identify model complexity levels that well describe given data. This question, already raised in the 12^{th} century by William of Occam, has large effect on models' accuracy. In chapter 1, we described its consequences for density and mutual information estimation, and explained its consequences in terms of accuracy and reliability of estimation. This issue is exemplified by the well known problem of overfitting models to data, and has prompted numerous attempts to elucidate the relation between training error and generalization error (see e.g. chap 7 in (Hastie et al., 2001)). The most widely known approaches are *Minimal description length* (MDL (Rissanen, 1978) and its related *Bayesian information criterion*, BIC, and *Akaike information criterion*, AIC) and VC theory (Vapnik, 1982, 1995). These approaches focus on the supervised learning scenario, where the basic tradeoff between a good description of the data and a good generalization can be naturally quantified. In the context of unsupervised learning, I claim that there is not necessarily a single level of complexity that is the "correct" one for a system. For example, many natural data sets have several scales and multi resolutions, as illustrated in figures 5.9.

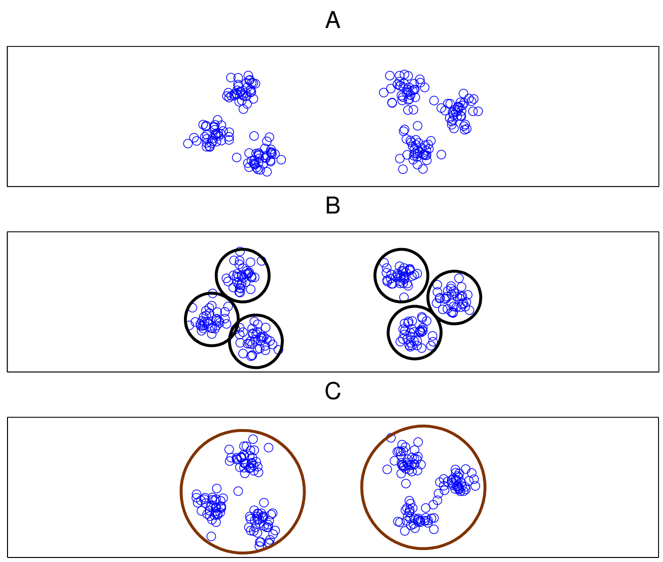

Figure 5.9: Multi resolution in clustering data. Raw data on the upper panel reveals interesting structures both at high resolution (black circles, panel B) and low resolution (brown circles, panel C).

In the context of modeling data with side information, the question of identifying the best resolution of the data, becomes a question of identifying the resolution that provides a good characterization of $P(X, Y^+)$, but not $P(X, Y^-)$. To test whether IBSI can discover such salient resolutions, a synthetic example was created, in which Y^+ contained two relevant resolutions (for 2 and 4 clusters), while Y^- only contained structure at the higher resolution of 4 clusters. The results of running IBSI on this data are demonstrated in figure 5.10

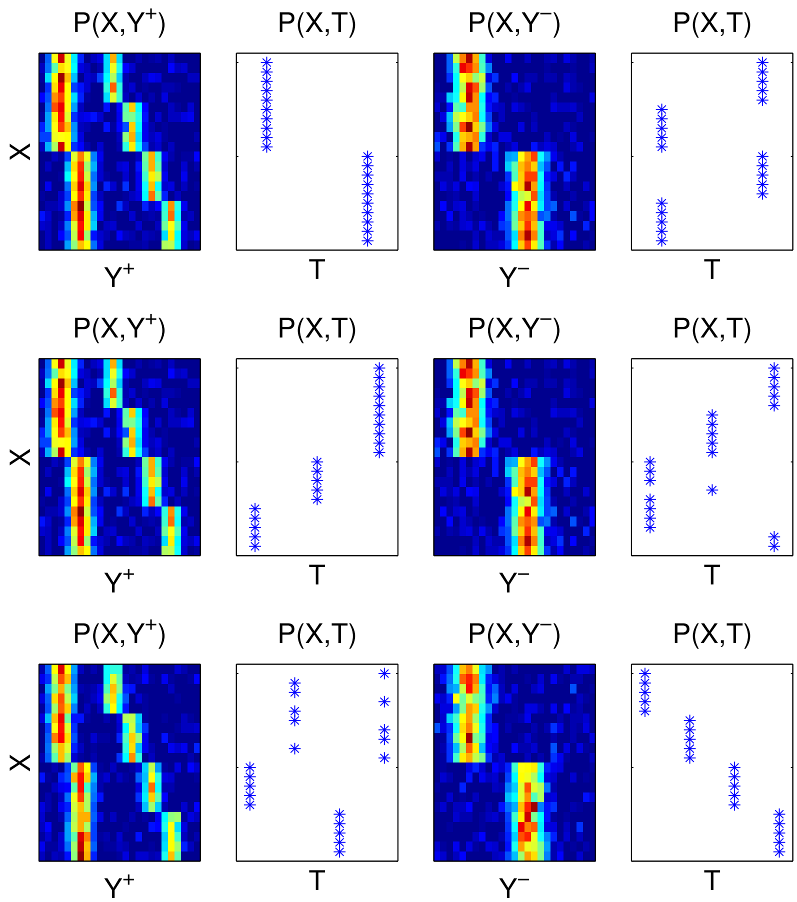

Figure 5.10: Demonstration of IBSI operation, for 2,3 and 4 clusters. Panels as in the previous figure.

Indeed the saliency of each resolution can be measured by the value of the IBSI functional as a function of number of clusters. Figure 5.11 shows that it can identify the relevant resolution in the current data.

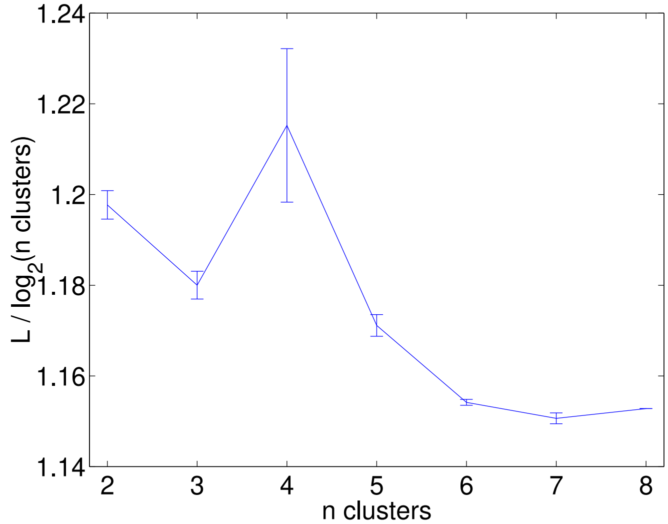

Figure 5.11: Saliency of IBSI operation, for 2 to 8 clusters. Y-axis measures the value of the function $L = I(T;Y^+) - I(T;Y^-)$, normalized by $H(T)$ which is an upper bound on the information. The resolution obtained with 4 clusters is found to best characterize the data.

The last two examples were designed for demonstration purposes, thus the irrelevant structures is strongly manifested in $P(X;Y^-)$. The next examples shows that our approach is also useful for real data, in which structures are much more covert.

5.4.3 Hierarchical text categorization

Text categorization is a fundamental task in information retrieval. Typically, one has to group a large set of texts into sets of homogeneous subjects. Recently, Slonim and colleagues showed that the IB method achieves categorization that predicts manually predefined categories with great accuracy, and largely outperforms competing methods (Slonim et al., 2002). Clearly, this unsupervised task becomes more difficult when the texts have similar subjects, because alternative categories are extracted instead of the "correct" one.

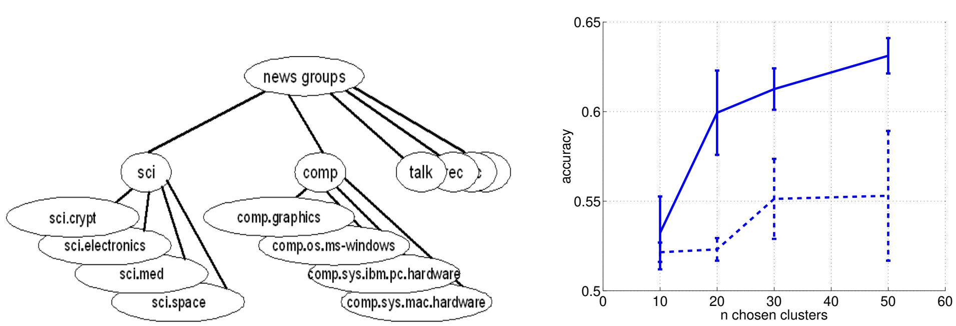

Figure 5.12: **A.** An illustration of the 20 newsgroups hierarchical data we used. **B.** Categorization accuracy vs. number of word clusters k. $N = 100$.

This problem can be alleviated by using side information in the form of additional documents from other categories. This is specifically useful in hierarchical document categorization, in which known categories are refined by grouping documents into subcategories. (Dumais & Chen, 2000; Vinokourov & Girolani, 2002). IBSI can be applied to this problem by operating on the terms-documents co-occurrence matrix while using the other top-level groups to focus on the relevant structures. To this end, IBSI is used to identify clusters of terms that will be used later to cluster a group of documents into its subgroups,.

While IBSI is targeted at learning structures in an unsupervised manner, we have chosen to apply it to a labelled dataset of documents to asses its results compared with manual classification. Labels are not used by our algorithms during learning and serve only to quantify performance. We used the *20 Newsgroups database* collected by (Lang, 1995) preprocessed as described in (Slonim et al., 2002). This database consists of 20 equal sized groups of documents, hierarchically organized into groups according to their content (figure 5.12). We aimed at clustering documents that belong to two newsgroups from the super-group of computer documents and have very similar subjects *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*. As side information we used all documents from the super-group of science (*sci.crypt*, *sci.electronics*, *sci.med*, *sci.space*).

To demonstrate the power of IBSI we used double clustering to separate documents into two groups. The goal of the first clustering phase is to use IBSI to identify clusters of terms that extract the relevant structures of the data. The goal of the second clustering phase is simply to provide a quantitative measure for the quality of the features extracted in the first phase. We therefore performed the following procedure. First, the 2000 most frequent words in these documents were clustered into N clusters using IBSI. Then, word clusters were sorted by a single-cluster score $D_{KL}[p(y^+|t)||p(y^+)] - \gamma D_{KL}[p(y^-|t)||p(y^-)]$, and the k clusters with the highest score were chosen. These word-clusters were then used for clustering documents. The performance of this process is evaluated by measuring the overlap of the resulting clusters with the manually classified groups. Figure 5.12 plots document-clustering accuracy for $N = 100$, as a function of k. IBSI ($\gamma = 1$) is compared with the IB method (i.e. $\gamma = 0$). Using IBSI successfully improves mean clustering accuracy from about 55 to 63 percents.

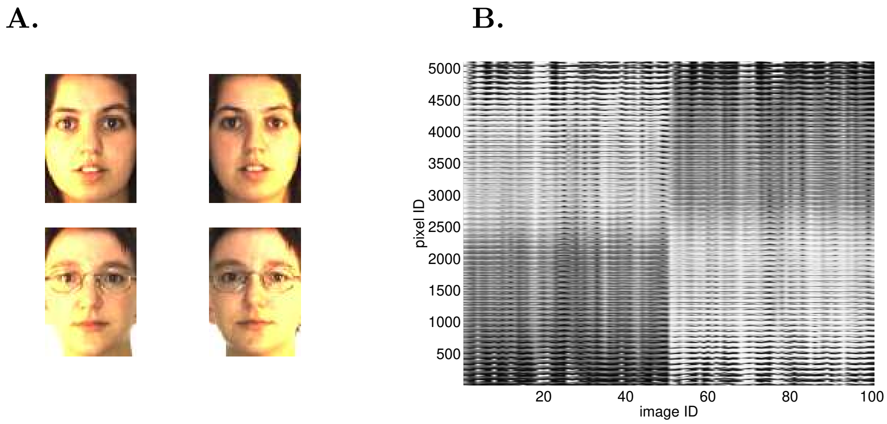

Figure 5.13: **A.** Samples of four face images, each woman face with two different light sources. **B.** Matrix of joint distribution $P(X, Y^+)$ used for IBSI. Each column of the matrix corresponds to a different image in the bank. The 50 left columns correspond to images with a left source light.

5.4.4 Face images

To further demonstrate the applicability of IBSI to diverse types of real-world data, we applied it to the problem of extracting features of face images. We used the AR database of faces (24, 1998), and focused on images that contained a strong light source either from the right or the left. These illumination conditions imply strong statistical structures in the set of images. The four examples of images we used are shown in figure 5.13A. Figure 5.13B shows the matrix of all the face images $P(X; Y^+)$, where each column vector of the matrix corresponds to a different picture in the database. The effects of light source are apparent: The columns on the left half of the matrix (corresponding to images with a left light source) appear similar. Moreover, the rows (corresponding to pixels) in the upper part of the matrix are similar (and so are the rows in its lower part). This suggests that clustering pixels according to their distribution across images is expected to reflect the effects of light source.

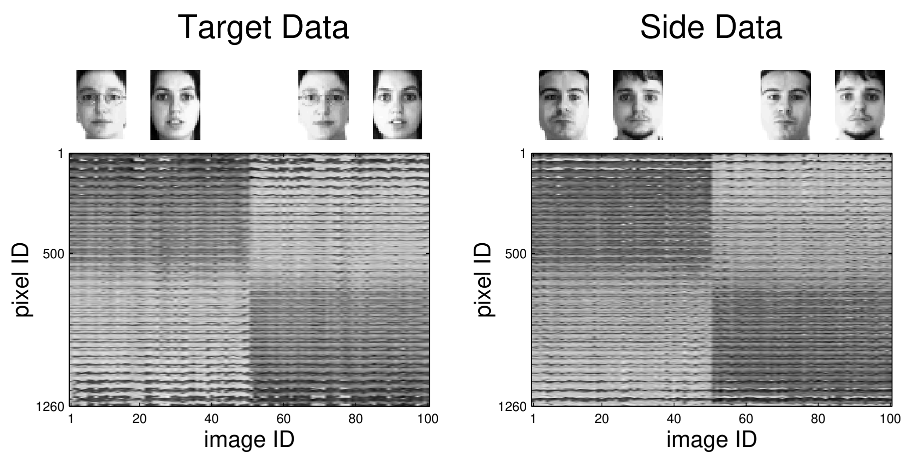

Figure 5.14: Face images of women (main data, left panel) and men (side data, right panel). Each image was reshaped to a column vector and all vectors were concatenated to a matrix with n_{images} columns and n_{pixels} rows.

We applied IBSI to this problem using an additional set of men's faces as side information data $P(X;Y^-)$. This choice of side data actually makes it harder for our method to extract relevant features because of the numerous structures common to men's and women's faces, and images of other illuminated objects would have been more appropriate to the task. Figure 5.15 depicts the results obtained when clustering pixels into 4 clusters, for varying levels of the parameter γ. On each panel clusters appear at different colors, where the colors code the level of differential information $I(T;Y^+) - \gamma I(T;Y^-)$ for each cluster (see the color-bar on the right).

When side data are ignored ($\gamma = 0$), clustering pixels extract a structure that reflects illumination conditions. However, for positive γ values, the effects of light direction is diminished and more salient features are enhanced such as the eyes and the area around the mouth.

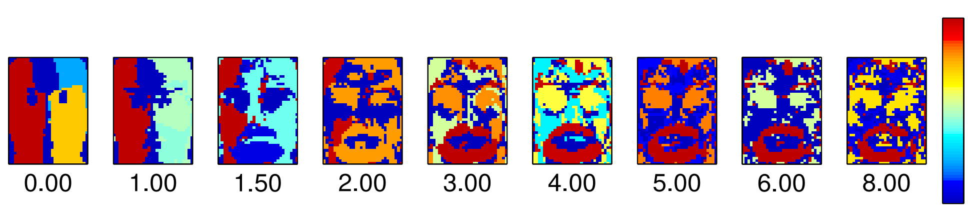

Figure 5.15: Clusters obtained with IBSI for various values of γ. In each panel, four clusters are depicted in different colors, where the colors code the level of differential information $I(T;Y^+) - \gamma I(T;Y^-)$ for each cluster. Similar results are obtained with different number of clusters

5.4.5 Auditory coding

IBSI was used to characterize the processing performed by cortical neurons while filtering out the processing that takes place in lower processing stages.

We calculated the empirical joint distributions of 15 stimuli and neural responses (spike counts) for all cortical neurons. For each cortical neuron we also calculated the same joint distributions for an ANF model neuron that has the same BF as the cortical neuron. We then applied IBSI using positive information terms for all cortical cells and negative ones for the ANF model cells, using an equal $\gamma = 1$ parameter for all irrelevant data and $\beta^{-1} = 0$. This yields a target function in the form of Eq. 5.18

$$\begin{aligned} \mathcal{L} &= -\left[I(T;Y_1^{AI}) + \ldots + I(T;Y_{45}^{AI})\right] \\ &\quad +\gamma\left[I(T;Y_1^{ANF}) + \ldots + I(T;Y_{45}^{ANF})\right] \end{aligned} \tag{5.25}$$

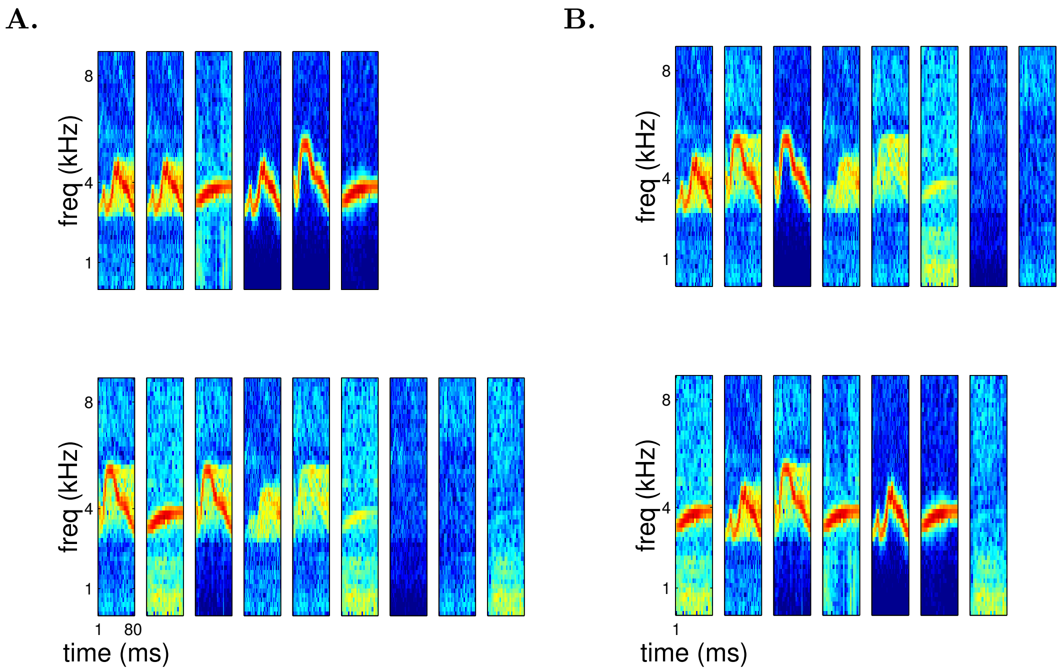

Figure 5.16: Clusters of stimuli obtained with IBSI with 45 AI cells as relevant variables and 45 ANF cells as irrelevant ones. **A.** $\gamma = 0$ **B.** $\gamma = 1$.

An example of the results for 2 clusters is given in figure 5.16. Adding the irrelevant data affects stimuli clusters in the following way. While with $\gamma = 0$ the stimuli that have low energy (noise and background) are all clustered together (see the second cluster in A), this is no longer true when ANF cells are used as irrelevant variables Fig. 5.16B. Unfortunately with this data, it is difficult to determine which are the common relevant features in the stimuli that are clustered together. The major limitations of the current data are two folds: First, the total number of stimuli is small, which makes it more difficult to perform a stability analysis or estimate clusters significance. More importantly, it is highly complex and of high dimensionality. Therefore, even if we know that several stimuli are considered similar by some neurons, it does not provide enough hints about the features to which these neurons respond. In this sense, the IBSI framework is expected to be more beneficial in settings where a centroid of a cluster can be defined in a meaningful way, for example when stimuli belong to some parametric family. Therefore we purposely refrain from drawing conclusions about auditory coding from this experiment, and use it only to demonstrate the potential of IBSI to neural coding investigation.

5.5 Extending the use of side information

This chapter has described how non interesting data can be used to unlearn the structure of the noise in an unsupervised task that uses the information bottleneck framework. While the method described here relies on clustering as the method of dimensionality reduction, the use of side information in the form of irrelevance variables is not limited to clustering but can be extended to various dimensionality reduction methods. This can be formalized by again considering the target function

$$\mathcal{L}(T) = I(X;T) - \beta \left(I(T;Y^{+}) - \gamma I(T;Y^{+}) \right) \tag{5.26}$$

which is to be maximized over all the (possibly stochastic) mappings $X \rightarrow T$.

We have extended the IB framework to handle continuous variables, and described its complete formal solution for the case of Gaussian variables (Chechik, Globerson, Tishby, & Weiss, 2003). For Gaussian-IBSI the problem reduces to one of finding generalized eigenvector of the covariance matrices $\Sigma_{X|Y^{+}}$ and $\Sigma_{X|Y^{-}}$ (Chechik & Globerson, 2003). This result shows the connection between Gaussian-IB and Gaussian IBSI to the well studied problems of *canonical correlations* (Thompson, 1984; Borga, 2001) and the lesser studied *generalized canonical correlation*.

Using side information in the form of irrelevant data, can be further generalized to use information measures other than $I(X;Y)$. Consider the following formalism suggested by (Globerson, 2003). Let $\vec{f}(x_1, \ldots, x_n)$ be a reduced representation of the data $X = (x_1, ..., x_n)$, and let $Q(f(X), Y)$ be some quality function that measures how well the dimensionality reduction f preserves the structures in the joint relation X, Y. The goal of relevant feature extraction is then to maximize

$$Q(f(X), Y^{+}) - \gamma\, Q(f(X), Y^{+}) \ . \tag{5.27}$$

When f is a clustering of x, $f(X) \equiv T$ and $Q(f(X), Y)$ is the mutual information $I(T;Y)$ this formalism reduces to IBSI with infinite β. However this formalism allows us to extend the side information idea to other settings by choosing other quality functions and features. One such extension was presented in (Globerson et al., 2003), extending an approach presented in (Globerson & Tishby, 2003) and named *sufficient dimensionality reduction*. In this approach continuous statistics are extracted which can be thought of as a set of weighting functions. The quality of these statistics is quantified using the notion of *information in measurement*, which replaces the Shannon mutual information used in IBSI.

5.6 Summary

We presented an information theoretic approach for extracting relevant structures from data, by utilizing additional data known to share irrelevant structures with the relevant data. Naturally, the choice of side data may considerably influence the solutions obtained with IBSI, simply because using different irrelevant variables is equivalent to asking different questions about the data analyzed. In practice, side data can be naturally defined in numerous applications, in particular in exploratory analysis of scientific experiments. For example, it is most suitable to investigating changes in the neural code following learning, or coding differences between brain regions both in terms of the difference between the stimuli that are coded and the code words that are used. It is expected that the use of irrelevance data as side information will be extended to additional dimensionality reduction methods and neural coding applications.

Chapter 6

Discussion

The primary goal of this thesis was to identify computational principles that govern information processing and representation in the auditory system. For this purpose, we set to develop formal and quantitative methods that could identify such principles in a set of electrophysiological recordings from three brain regions.

We started the investigation by discussing methods to extract information from spike trains and reviewed in detail both the theoretical motivation and the practical considerations of reducing the dimensionality of spike trains and representing them by simpler statistics. We tested six different dimensionality reduction methods and compared the level of information they achieve. In all of these methods, IC neurons conveyed about twice more information about the identity of the presented stimulus than AI and MGB neurons.

We found that the maximal information can almost always be extracted by considering the distribution of temporal patterns of spikes. Surprisingly, the first spike latency carries almost the same level of information. In contrast, spike counts convey only half of the maximal information level. These results reveal a surprising observation regarding the nature of the neural code in the different stations. It shows that even though spikes in the cortex are not precisely locked in time to the stimulus (as in the IC), the occurrence of a single spike, the first spike since the onset, conveys the maximal information that could be achieved with any other methods tested.

We then advanced to investigate how small groups of neurons in different brain areas interact to represent the auditory stimuli. For this purpose we developed measures of informational redundancy in groups of cells, and described their properties. These measures can be reliably estimated in practice from empirical data using stimulus conditioned independence approximation. Since redundancy is biased by the baseline single-unit information level, we study this effect and show how it can be reduced with a proper normalization. Finally, we discussed redundancy biases due to a ceiling effect on the maximal information and the way to correct for these biases.

Applying these methods to our data we obtained three main results. First, we showed that small groups of IC cells are more redundant in the information they convey about stimulus identity than AI and MGB cells. In other words, cells of higher regions in the processing hierarchy tend to code features of the stimuli in a more independent manner. These findings put forward redundancy reduction as a possible generic organization principle of sensory systems. This principle was suggested 40 years ago by Barlow from theoretical considerations, and an empirical evidence for such a process is presented here for the first time.

Secondly, we showed that redundancy is significantly correlated with the best frequency (BF) of IC cells but not with the those of AI or MGB cells. This means that frequency characterization does not capture well the type of processing that AI and MGB cells perform. Finally, we found that AI and MGB cells convey an order of magnitude less information about the spectro-temporal structure of the stimuli as compared to IC neurons. This suggests that AI cells succeed to code well the identity of the stimuli without coding the precise acoustical structures in it.

The low redundancy in AI and MGB, and the lack of correlation of this redundancy with the best frequency of the cells has strong implications, which go far beyond the statement that BF responses are not predictive for complex sounds. The reason is that currently accepted methods of characterizing AI neurons, primarily the spectro temporal receptive field (STRF), imply redundancy between neurons that share spectro temporal characteristics. Although we find such redundant neurons in IC we do not find them in MGB or AI, showing that STRF characterization in AI misses crucial aspects of neuronal coding even for simple stimuli as used in our work.

Coding in an independent manner relates to the issue of *specialization* of neuronal responses. The standard meaning of specialization in the literature is that neurons respond only to a restricted set of stimuli. In this sense, for example, face neurons in the infero temporal cortex are specialized. In the context of primary auditory cortex, as discussed in (Nelken, 2002), most of the evidence today does not support the view that neurons in AI are more specialized than neurons in lower auditory stations. For example,(Middlebrooks, Clock, Xu, & Green, 1994) showed that neurons in AI do not have specialized receptive fields for space,(Kowalski, Depireux, & Shamma, 1996b, 1996a; Depireux, Simon, Klein, & Shamma, 2001) used analysis of STRF in the ferret auditory cortex and demonstrated a rather simple structure in most of them (quadrant separability). The type of results quantified here suggests a more delicate type of specialization, since instead of mapping the parametric set of stimuli to which neurons respond, we measure the ability of neurons to discriminate between stimuli that are acoustically highly similar (see figure 1.9 in section 1.3.5). This type of informational specialization is manifested in the form of low redundancy. The fact that we do not find cortical neurons that are as redundant as IC neurons (even though they respond to stimuli and convey considerable amount of information about them) suggests that the informational redundancy measure is the more interesting measure of specialization.

The term *redundancy reduction* was originally coined by Barlow(Barlow, 1959b, 1959a, 1961). As explained in Chapter 1, he suggested that a principal goal of sensory processing is to achieve an efficient code by compressing sensory inputs to obtain parsimonious representations. Barlow later suggested (Barlow, 2001) that the actual goal of the system is rather *redundancy exploitation*: During this process the statistical structures in the inputs are extracted and coded. This revised idea leads to inherently different consequences, and predicts that higher areas will contain many units, specialized to respond to complex structures of the stimuli, and largely independent. The results presented in our study are in agreement with this revised view. We find that neurons in higher processing stations are less informative but more independent than those at the lower levels, presumably because they convey information about more complex structures in the inputs.

Reducing redundancy during information processing while mapping stimuli to a higher dimensional feature space may provide better discrimination among complex stimuli, as in independent component analysis (ICA) (Bell & Sejnowski, 1995) and support vector machines(Vapnik, 1995). Redundancy reduction may therefore be a generic organizational principle of sensory systems that allows for easier readout of stimulus aspects that are behaviorally relevant.

The last part of this dissertation addresses the problem of identifying the features that are relevant for neural responses. A major difficulty in this task is that the characterization of the processing performed in a single brain region requires filtering out the processing that takes place earlier in the processing hierarchy. This is an instance of a generic problem in unsupervised learning, of identifying relevant structures in data that contain many competing structures. We presented a formal definition of the problem in the framework of distributional clustering, as well as its analytical and algorithmic solutions. We showed its applicability in a variety of data domains as texts clustering and feature extraction for face recognition. Our results lay the groundwork for developing additional dimensionality reduction methods of data that use irrelevance data, and have already been extended to various problems and learning techniques such as the study of neural coding in evolving agents (Avraham, Chechik, & Ruppin, 2003), data mining of the web (Gondek & Hofmann, 2003) linear projections of Gaussian variables and their relation to canonical correlation analysis (Chechik et al., 2003; Chechik & Globerson, 2003), spectral based feature selection (Shashua & Wolf, 2003) and continuous embedding of categorical variables (Globerson et al., 2003), as applied for example to feature extraction for face recognition.

Appendix A

Information Theory

A.1 Entropy

Shannon (Shanon, 1948) developed the concept of entropy to measure the uncertainty of a discrete random variable. Suppose X is a discrete random variable that obtains values from a finite set $x_1, ..., x_n$, with probabilities $p_1, ..., p_n$. We look for a measure of how much choice is involved in the selection of the event or how certain we are of the outcome. Shannon argued that such a measure $H(p_1, ..., p_n)$ should obey the following properties

1. H should be continuous in p_i.
2. If all p_i are equal then H should be monotonically increasing in n.
3. If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

Shannon showed that the only H that satisfies these three assumptions is of the form

$$H = -k \sum_{i=1}^{n} p_i \log p_i \tag{A.1}$$

and termed it the entropy of X, since it coincides with the notion of entropy defined in certain formulations of statistical mechanics. k is a constant that determines the units of measure, and can be absorbed in the base of the log. The current thesis adheres to the computer science literature and uses the log in base 2. To summarize, we define entropy as

Definition A.1.1 : Entropy
The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{A.2}$$

We will also sometimes use the notation $H[\mathbf{p}]$ to denote the entropy of a random variable that has a probability distribution $\mathbf{p}$. Given several random variables we then define

Definition A.1.2 : Joint Entropy
The *joint entropy* $H(X,Y)$ of a pair of discrete random variables X and Y with a joint distribution $p(x,y)$ is defined by

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) \tag{A.3}$$

Definition A.1.3 : Conditional entropy
Let X and Y be discrete random variables with joint distribution $p(x,y)$ and conditional distributions $p(x|y)$, then the entropy conditioned on a single symbol is defined by

$$H(X|Y=y) = -\sum_x p(x|y) \log p(x|y) \quad . \tag{A.4}$$

The *conditional entropy* is defined by

$$\begin{aligned} H(X|Y) &= \sum_y p(y) H(X|Y=y) \\ &= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x,y) \log p(x|y) \quad . \end{aligned} \tag{A.5}$$

Several properties of the entropy worth mentioning.

Theorem A.1.4 : Properties of H(X)
The entropies $H(X)$ of a discrete random variable X that can obtain the values $x_1, ..., x_n$, and the joint entropy $H(X,Y)$, obey the following properties

1. *Non-negativity* $H(X) \geq 0$
2. *Upper bound* $H(X) \leq \log(n)$
3. *Chain rule:* $H(X,Y) = H(X) + H(Y|X)$
4. *Conditioning reduces entropy* $H(X|Y) \leq H(X)$
5. $H(p)$ *is concave in* p

A.2 Relative entropy and Mutual information

The entropy of a variable is a measure of the uncertainty in its distribution. The relative entropy is a measure of the statistical distance between two distributions

Definition A.2.1 : Relative Entropy
The *relative entropy* or the *Kullback Leibler divergence* between to probability functions $p(x)$ and $q(x)$, is defined by

$$D_{KL}[p||q] = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{A.6}$$

The KL divergence appears in statistics as the expected value of the log likelihood ratio. It therefore determines the ability to discriminate between two states of the world, yielding sample distributions $p(x)$ and $q(x)$.

We also use sometimes a variant of D_{KL}

Definition A.2.2 : Jensen-Shannon divergence
The *Jensen-Shannon divergence* between to probability functions $p_1(x)$ and $p_2(x)$, is defined by

$$JS_\pi[p||q] = \pi_1 D_{KL}[p_1||p] + \pi_2 D_{KL}[p_2||p] \tag{A.7}$$

with $\{\pi_1, \pi_2\}$ being prior probabilities $\pi_i > 0$, $\sum_i \pi_i = 1$, and p is the weighted average $p = \pi_1 p_1 + \pi_2 p_2$.

Theorem A.2.3 : Properties of D_{KL}
Let $p(x)$ and $q(x)$ be two probability distributions, Then

1. $D_{KL}[p||q] \geq 0$ *with equality iff* $p(x) = q(x) \forall x$.
2. $D_{KL}[p||q]$ *is convex w.r.t the pair* (p, q).

Definition A.2.4 : Mutual Information
The *mutual information* $I(X;Y)$ of two random variables X and Y is the KL divergence between their joint distribution and the product of their marginals

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad . \tag{A.8}$$

By this definition the mutual information provides some measure of the dependence between the variables. From the non negativity of the D_{KL} we obtain

Theorem A.2.5 : Non negativity of I(X;Y)
Let X and Y be two discrete random variables, then

$$I(X;Y) \geq 0 \tag{A.9}$$

and equality iff X and Y are independent.

Theorem A.2.6 : Properties of the mutual information
Let X and Y be two discrete random variables, then their mutual information $I(X;Y)$ obeys

1. *Symmetry* $I(X;Y) = I(Y;X)$.
2. $I(X;Y) = H(X) - H(X|Y)$.
3. $I(X;X) = H(X)$.
4. *Chain rule:* $I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, ..., X_{i-1})$.
5. *Data processing inequality: if* (X, Y, Z) *form a Markov chain, then* $I(X;Y) \geq I(X;Z)$. *As a consequence,* $I(X;Y) \geq I(X; f(Y))$ *for any function* f *of* Y.

A.3 Extensions

While the above notions were defined for discrete variables, entropy and mutual information can be extended to continuous variables (Shanon, 1948; Cover & Thomas, 1991). This issue is beyond of the scope of the current manuscript. Also, the notion of information can be extended to more than two variables using the view that information measure the KL distance from independence

Definition A.3.1 : Multi Information
The *multi information* $I(X_1; \ldots; X_n)$ of n random variables is the KL divergence between their joint distribution and the product of their marginals

$$I(X_1; \ldots; X_n) = \sum_{x_1,...,x_n} p(x_1, ..., x_n) \log \frac{p(x_1, ..., x_n)}{\prod_i p(x_i)} \quad . \tag{A.10}$$

By this definition the multi information provides some measure of the dependence between all the variables. From the non negativity of the D_{KL} we obtain that the multi information is non negative. The properties of the multi information measure are further discussed in (Studenty & Vejnarova, 1998).

Appendix B

Table of symbols

AI	Auditory cortex
BF	Best Frequency
$D_{KL}[p\|\|q]$	The Kullback Liebler divergence (Definition A.2.1)
$H(X)$	The entropy of a discrete variable X (Definition A.1.1)
$I(X;Y)$	The Mutual information of two variables X and Y (A.2.4)
$I[p]$	The mutual information of variables with a joint distribution p
IC	inferior colliculus
$JS[p\|\|q]$	The Jensen-Shannon divergence (A.2.2)
MGB	Medial Geniculate body of the thalamus
MI	Mutual information
n	Sample size
N	Number of variables
p	Probability distribution. $p(X,Y)$ is the joint distribution of X and Y
$\hat{p}$	Probability distribution that is estimated from empirical data
R	Neural responses (a random variable)
S	Stimulus (a random variable)
STRF	Spectro-Temporal Receptive Field
$T(R)$	A statistic of the responses

References

Abeles, M., Bergmann, H., Margalit, H., & Vaadia, E. (1993). Spatiotemporal firing patterns in the frontal cortex of behaving monkies. *The Journal of Neurophysiology*, *70*, 1629-1638.

Aertsen, A., & Johannesma, P. (1981). The spectro-temporal receptive field: A functional characteristics of auditory neurons. *Biological Cybernetics*, *42*, 133-143.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, *3*, 213-251.

Atick, J., & Redlich, A. (1990). Towards a theory of early visual processing. *Neural Computation*, *2*, 308-320.

Attneave, F. (1954). Some information aspects of visual perception. *Psychological Review*, *61*, 183-193.

Avraham, H., Chechik, G., & Ruppin, E. (2003). Are there representations in embodied evolved agents? taking measures. In *Proceedings of the 7th European conference on artificial life.* UK.

Baker, L., & McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21th conference on research and development in information retrieval (SIGIR-98).* New York: ACM press.

Barlow, H. (1959a). The coding of sensory messages, chapter XIII. In Thorpe & Zangwill (Eds.), *Current problems in animal behavior* (p. 330-360). Cambridge University Press.

Barlow, H. (1959b). Sensory mechanisms, the reduction of redundancy, and intelligence. In *Mechanisation of thought processes* (p. 535-539). London: Her Majesty's stationary office.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication.* Cambridge, MA: MIT Press.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*, 241-253.

Bar-Yosef, O., Rotman, Y., & Nelken, I. (2001). Reponses of neurons in cat primary auditory cortex to bird chirps: Effects of temporal and spectral context. *Journal of Neuroscience*.

Becker, S. (1996). Mutual information maximization: Models of cortical self organization. *Network: Computation in Neural Systems*, *7*, 7-31.

Bell, A., & Sejnowski, T. (1995). An information maximization approach to blind seperation and blind deconvolution. *Neural Computation*, *7*, 1129-1159.

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

Blahut, R. (1972). Computation of channel capacity and rate-distortion function. *IEEE Transactions on Information Theory*, *18*(4), 460-473.

Borga, M. (2001, January). *Canonical correlation: A tutorial.* On line tutorial http://people.imt.liu.se/magnus/cca.

Brenner, N., Strong, S., Koberle, R., Steveninck, R. de Ruyter van, & Bialek, W. (2000). Synergy in a neural code. *Neural Computation*, *13*(7), 1531-1552.

Calvin, W. (1995). Cortical columns, modules, and Hebbian cell assemblies. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (p. 269-272). Cambridge, MA: MIT Press.

Casseday, J., Fremouw, T., & Covey, E. (2002). The inferior colliculus. In D. Oertel, R. R. Fay, & A. N. Popper (Eds.), *Integrative functions in the mammalian auditory pathway* (Vol. 15, p. 238-318). New York: Springer-Verlag.

Chechik, G. (2003). Spike-timing dependent plasticity and relevant mutual information maximization. *Neural Computation*, *15*(7), 1481-1510,.

Chechik, G., & Globerson, A. (2003). *Information bottleneck and linear projections of Gaussian processes.* (Tech. Rep. No. 4). Leibniz Center for Research, School of Engineering and Computer Science, The Hebrew University of Jerusalem.

Chechik, G., Globerson, A., Tishby, N., & Weiss, Y. (2003). *Information bottleneck for Gaussian variables.* (Submitted)

Chhikara, R., & Folks, J. (1989). *The inverse Gaussian distribution: Theory, methodology and applications.* New York: Marcel Dekker.

Cover, T., & Thomas, J. (1991). *The elements of information theory.* New York: Plenum Press.

Csiszar, I. (1974). On the computation of rate-distortion functions. *IEEE Transactions on Information Theory*, *20*(1), 122-124.

Csiszar, I., & J.Korner. (1997). *Information theory: Coding theorems for discrete memoryless systems* (2nd ed.). New York: Academic Press.

Dan, Y., Alonso, J., Usrey, W., & Reid, R. (1998). Coding of visual information by precisely correlated spikes in the LGN. *Nature Neuroscience*, *1*, 501-507.

Dayan, P., & Abbot, L. (2002). *Theoretical neuroscience.* Cambridge, MA: MIT Press.

DeCharms, R., Blake, D., & Merzenick, M. (1998). Optimizing sound features for cortical neurons. *Science*, *280*(5368), 1439-1444.

Degroot, M. (1989). *Probability and statistics* (2nd ed.). Reading MA: Addison Wesley.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1-38.

Depireux, D., Simon, J., Klein, D., & Shamma, S. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *The Journal of Neurophysiology*, *85*(3), 1220-1234.

Devroye, L., & Lugosi, G. (2001). *Combinatorial methods in density estimation.* New York: Springer-Verlag.

Dimitrov, A., & Miller, J. (2001). Neural coding and decoding: Communication channels and quantization. *Network: Computation in Neural Systems*, *12*(4), 441-472.

Dumais, S., & Chen, H. (2000). Hierarchical classification of web content. In N. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 23th conference on research and development in information retrieval (SIGIR-00)* (p. 256-263). New York: ACM press.

Eggermont, J., Johannesma, P., & Aertsen, A. (1983). Reverse correlation methods in auditory research. *Quarterly Reviews of Biophysics*, *16*, 341-414.

Ehret, G., & Merzenich, M. (1988). Complex sound analysis (frequency resolution filtering and spectral integration) by single units of the inferior colliculus of the cat. *Brain Research*, *13*(2), 139-163.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). 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.

Fishbach, A., Nelken, I., & Yeshurun, Y. (2001). Auditory edge detection: A neural model for physiological and psychoacoustical responses to amplitude transients. *The Journal of Neurophysiology*, *85*, 2303-2323.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, *222*, 309-368.

Friedman, N., Mosenzon, O., Slonim, N., & Tishby, N. (2001). Multivariate information bottleneck. In J. Breese & D. Koller (Eds.), *Uncertainty in artificial intelligence: Proceedings of the seventeenth conference (UAI-2001)* (p. 152-161). San Francisco, CA: Morgan Kaufmann.

Gat, I., & Tishby, N. (1999). Synergy and redundancy among brain cells of behaving monkeys. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems* (Vol. 11). Cambridge, MA: MIT Press.

Gawne, T., & Richmond, B. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, *13*(7), 2758-2771.

Gedeon, T., Parker, A., & Dimitrov, A. (2002). Information distortion and neural coding. *Canadian Applied mathematics Quarterly.*

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2003). An information theoretic tradeoff between complexity and accuracy. In *Proceedings of the conference on learning theory.* Washington.

Globerson, A. (2003). *Sufficient dimensionality reduction with irrelevant statistics.* (A talk at the computer science department, the Hebrew University)

Globerson, A., Chechik, G., & Tishby, N. (2003). Sufficient dimensionality reduction with irrelevant statistics. In *Uncertainty in artificial intelligence: Proceedings of the nineteenth conference (UAI-2003).* San Francisco, CA: Morgan Kaufmann.

Globerson, A., & Tishby, N. (2003). Sufficient dimensionality reduction. *Journal of Machine Learning Research*, *3*, 1307-1331.

Gochin, P., Colombo, M., Dorfman, G. A., Gerstein, G., & Gross, C. (1994). Neural ensemble coding in inferior temporal cortex. *The Journal of Neurophysiology*, *71*, 2325-2337.

Gondek, D., & Hofmann, T. (2003). Conditional information bottleneck clustering. In *3rd ieee international conference on data mining, workshop on clustering large data sets.*

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning.* Springer.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.

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

Heil, P., & Irvine, D. (1996). On determinants of first-spike latency in auditory cortex. *NeuroReport*, *7*, 3073-3076.

Hoffman, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 21th conference on research and development in information retrieval (SIGIR-98)* (p. 50-57). New York: ACM press.

Hopfield, J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature*, *376*(6535), 33-36.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106-154.

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

Jordan, M. (Ed.). (1998). *Learning in graphical models.* Cambridge, MA: MIT press.

Kingman, J. (1993). *Poisson processes.* Oxford University Press.

Kowalski, N., Depireux, D., & Shamma, S. (1996a). Analysis of dynamic spectra in ferret primary auditory cortex II. Prediction of unit responses to arbitrary spectra. *The Journal of Neurophysiology*, *76*, 3524-3534.

Kowalski, N., Depireux, D., & Shamma, S. (1996b). Analysis of dynamic spectra in ferret primary auditory cortex I. Characteristics of single unit responses to moving ripple spectra. *The Journal of Neurophysiology*, *76*, 3503-3523.

Krishna, B., & Semple, M. (2001). *A computational model for first-spike latency and variability in the auditory nerve.* (The Association for Research in Otolaryngology MidWinter Meeting 2001)

Lang, K. (1995). NewsWeeder: learning to filter netnews. In *Proceedings of the 12th international conference on machine learning* (p. 331-339). San Mateo, CA: Morgan Kaufmann.

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience*, *4*, 533-539.

Linsker, R. (1988). Self organization in a perceptual network. *Computer*, *21*, 105-117.

Linsker, R. (1989). An application of the principle of maximum information preservation to linear systems. In D. Touretzki (Ed.), *Advances in neural information processing systems* (Vol. 1, p. 186-194). San Mateo, CA: Morgan Kaufman.

Linsker, R. (1992). Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, *4*, 691-702.

Linsker, R. (1997). A local learning rule that enables information maximization for arbitrary input distribution. *Neural Computation*, *9*, 1661-1665.

Martinez, A., & Benavente, R. (1998). *The AR face database* (Tech. Rep.). CVC Technical Report.

Meister, M. (1996). Multi neuronal coding in retinal signaling. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 609-614.

Meister, M., Lagnado, L., & Baylor, D. (1995). Concerted signaling by retinal ganglion cells. *Science*, *270*, 1207-1210.

Middlebrooks, J., Clock, A., Xu, L., & Green, D. (1994). A panoramic code for sound location by cortical neurons. *Science*, *264*(5160), 842-844.

Middlebrooks, J., Xu, L., Furukawa, S., & Mickey, B. (2002). Location signaling by cortical neurons. In D. Oertel, R. Fay, & A. N. Popper (Eds.), *Integrative functions in the mammalian auditory pathway* (Vol. 15, p. 319-357). New York: Springer-Verlag.

Miller, G. (1955). Note on the bias in information estimates. In H. Quastler (Ed.), *Information theory in psychology, problems and methods II-B* (p. 95-100). The Free Press.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Nadal, J., Brunel, N., & Parga, N. (1998). Nonlinear feedforward networks with stochastic outputs: Infomax implies redundancy reduction. *Network: Computation in neural systems*, *9*, 207-217.

Nadal, J., & Parga, N. (1994). Nonlinear neuron in the low noise limit: A factorial code maximizes information transfer. *Network: Computation in neural systems*, *5*, 565-581.

Nelken, I. (2002). Feature detection by the auditory cortex. In D. Oertel, R. R. Fay, & A. N. Popper (Eds.), *Integrative functions in the mammalian auditory pathway* (Vol. 15). Springer-Verlag.

Nelken, I., Chechik, G., King, A., & Schnupp, J. (2003). *The neural code in primary auditory cortex.* (In preparation)

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14.* Cambridge, MA: MIT Press.

Nirenberg, S., Carcieri, S., Jacobs, A., & Latham, P. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, *411*, 698-701.

Palombi, P., & Caspary, D. (1996). GABA inputs control discharge rate primarily within frequency receptive fields of inferior colliculus neurons. *The Journal of Neurophysiology*, *75*(6), 2211-2219.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, *15*, 1191-1253.

Panzeri, S., & Schultz, S. (2001). A unified approach to the study of temporal correlational and rate coding. *Neural Computation*, *13*(6), 1311-1349.

Panzeri, S., Schultz, S. R., Treves, A., & Rolls, E. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *266*, 1001-1012.

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

Pearl, J. (1988). *Probabilistic inference in intelligent systems.* San Mateo, CA: Morgan Kaufman.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *31st annual meeting of the association for computational linguistics* (p. 183-190). Columbus, Ohio: Morgan Kaufman.

Popper, A., & Fay, R. (1992). *The mammalian auditory pathway: Neurophysiology.* New York: Springer Verlag.

Powell, T., & Mountcastle, V. (1959). Some aspects of the functional organization of the cortex of the postcentral gyrus of the monkey: A correlation of findings obtained in a single unit analysis with cytoarchitecture. *Bulletin of the Johns Hopkins Hospital*, *105*, 133-162.

Reich, D., Mechler, F., & Victor, J. (2001). Independent and redundant information in nearby cortical neurons. *Science*, *294*, 2566-2568.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Rolls, E. T., Treves, A., & Tovee, M. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in primate termpoal visual cortex. *Experimental Brain Research*, *114*, 149-162.

Samengo, I. (2001). Independent neurons representing a finite set of stimuli: Dependence of the mutual information on the number of units sampled. *Network: Computation in Neural Systems*, *12*, 21-31.

Samengo, I., & Treves, A. (2000). Representational capacity of a set of independent neurons. *Physical Reviews E*, *63*, 1-14.

Schneidman, E., Slonim, N., Tishby, N., Steveninck, R. deRuyter van, & Bialek, W. (2002). *Analyzing neural codes using the information bottleneck method* (Tech. Rep.). Leibniz Center for Research, School of Engineering and Computer Science, The Hebrew University of Jerusalem.

Schnupp, J., Mrsic-Flogel, T., & King, A. (2001). Linear processing of spatial cues in primary auditory cortex. *Nature*, *414*, 200-204.

Schreiner, C., & Langner, G. (1997). Laminar fine structure of frequency organization in the auditory midbrain. *Nature*, *388*, 383-386.

Seshardri, V. (1993). *The inverse Gaussian distribution.* Oxford.

Shafer, G., & Pearl, J. (1990). *Reading in uncertain reasoning.* San Mateo, CA: Morgan Kaufman.

Shanon, C. (1948). A mathematical theory of communication. *The Bell systems technical journal*, *27*, 379-423,623-656.

Shashua, A., & Wolf, L. (2003). *Sparse spectral-based feature selection with side information* (Tech. Rep. No. 57). Leibniz Center for Research, School of Engineering and Computer Science, The Hebrew University of Jerusalem.

Singer, W., & Gray, C. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Reviews in Neuroscience*, *18*, 555-586.

Sinkkonen, J., & Kaski, S. (2001). Clustering based on conditional distribution in an auxiliary space. *Neural Computation*, *14*, 217-239.

Slonim, N., Friedman, N., & Tishby, N. (2001). Agglomerative multivariate information bottleneck. In *Advances in neural information processing systems.* Cambridge, MA: MIT Press.

Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR-02).* New York: ACM Press.

Slonim, N., & Tishby, N. (1999). Agglomerative information bottleneck. In S. Solla, D. Cohn, & T. Lynn (Eds.), *Advances in neural information processing systems* (Vol. 12). Cambridge, MA: MIT Press.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In N. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 23th conference on research and development in information retrieval (SIGIR-00)* (p. 208-215). New York: ACM press.

Spangler, K., & Warr, W. (1991). The descending auditory systems. In R. Altschuler, R. Bobbin, B. Clopton, & D. Hofmann (Eds.), *Neurobiology of hearing: The central auditory system* (p. 27-45). New York: Raven.

Steveninck, R. deRuyter van, lewen, G., Strong, S., Koberele, R., & Bialek, W. (1997). Reproduciblity and variability in neural spike trains. *Science*, *275*, 1805-1808.

Studenty, M., & Vejnarova, J. (1998). The multi-information function as a tool for measuring stochastic dependence. In M. Jordan (Ed.), *Learning in graphical models* (p. 261-297). Cambridge, MA: MIT Press.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation.* (Vol. 47). Thousands Oak, CA: Sage publications.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520-522.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of 37th allerton conference on communication, control and computation.* Allerton House, Illinois.

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

Ukrainec, A., & Haykin, S. (1996). A modular neural network for enhancement of cross-polar radar targets. *Neural Networks*, *9*, 143-168.

Uttley, A. (1970). *Information transmission in the nervous system.* London: Academic press.

Vaadia, E., Haalman, I., Abeles, M., Bergman, H., Prut, Y., Slovin, H., & Aertsen, A. (1995). Dynamics of neural interactions in monkey cortex in relation to behavioral events. *Nature*, *373*, 515-518.

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

Vapnik, V. (1982). *Estimation of dependences based on empirical data.* Berlin: Springer-Verlag.

Vapnik, V. (1995). *The nature of statistical learning theory.* Berlin: Springer-Verlag.

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

Vinokourov, A., & Girolani, M. (2002). A probabilistic framework for the hierarchic organization and classification of document collections. *Journal of intelligent information systems*, *18*(23), 153-172. (Special Issue on automated text categorization)

Vreeswijk, C. van. (2001). Information trasmission with renewal neurons. In J. Bower (Ed.), *Computational neuroscience: Trends in research.* Elsevier Press.

Warland, D., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *The Journal of Neurophysiology*, *78*, 2336-2350.

Wyner, A. (1975). On source codding with side information at the decoder. *IEEE Transactions on Information Theory*, *21*, 294-300.

Yeung, R. (1997). A framework for linear information inequalities. *IEEE Transactions on Information Theory*, *43*, 1924-1934.