ADAPTATION OF THE VISUAL SYSTEM TO THE TEMPORAL STATISTICS OF NATURAL IMAGES

Marco Buiatti

Gatsby Computational Neuroscience Unit

University College London

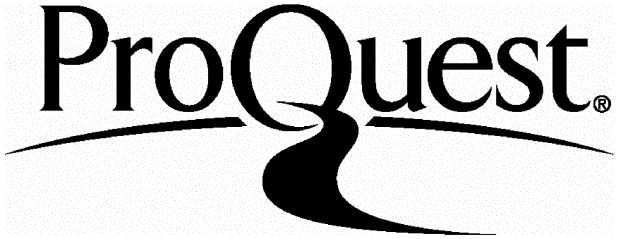
17 Queen Square

London WC1N 3AR

United Kingdom

M. PHIL. DEGREE

February 19, 2002



Abstract

This thesis focuses on ways of dynamically adapting the visual system to the temporal statistics of natural images. Recent experiments suggest that adaptation relies on a coding strategy that has the effect of normalising the neuronal activity with respect to the standard deviation of the stimulus. We investigate the hypothesis that such a nonlinear coding strategy could represent a key mechanism in natural vision both for adaptation and efficient representation of the stimulus. We model the process of variance normalisation in a simple way, and we train it on time series mimicking the typical visual input of a human photoreceptor in a natural environment. Simulations confirm our hypothesis: variance normalisation adapts the wide natural range of light intensities into the limited neuronal one and efficiently removes almost all the redundancy present in the temporal structure of natural images. Moreover, despite the long-range correlations present in the natural input, the integration time corresponding to the optimal normalisation is surprisingly short and compatible with the one observed experimentally.

The importance of this result seems to lie in the simplicity of the model. In order to prove its computational power, it is shown that it efficiently removes the redundancy of time series of very different nature - financial time series - also characterised by long-range correlations. This result suggests that such a method efficiently removes the redundancy of a wide variety of data, no matter how much correlated they are.

Contents

1 Introduction **6**

1.1 Premise 6

1.2 Thesis overview 12

2 Temporal adaptation in the early visual system: Experimental findings and a theoretical framework **14**

2.1 Introduction 14

2.2 Temporal adaptation in the early visual system 15

2.2.1 Adaptation to the mean light intensity 16

2.2.2 Adaptation to contrast 19

2.3 Temporal adaptation: An information theoretic approach . . 28

2.3.1 Biological constraints: noise and saturation 29

2.3.2 A computational strategy: redundancy reduction . . . 31

2.3.3 Redundancy reduction revisited 34

2.4 How does adaptation work in a natural environment? 36

3 Temporal statistics of natural images **38**

3.1 Introduction 38

3.2 Time series of natural images 42

3.2.1 How they were recorded 42

3.2.2 Statistical structure of the time series 43

3.3 Discussion 54

4 Variance normalisation **62**

4.1 Introduction 62

4.2 Adaptation to the mean light level 64

4.2.1 How to calculate the optimal filter 66

4.2.2 The filter's structure 67

4.2.3 Statistical properties of the response 71

4.3 Variance normalisation 75

4.3.1 Variance normalisation in time: the model 76

4.3.2 How to set the integration time of variance normalisation 77

4.3.3 Result 1: Dynamical adaptation of the response 81

4.3.4 Result 2: High-order redundancy reduction 83

4.4 Discussion 90

5 Redundancy reduction in financial time series 95

5.1 Introduction 95

5.2 The statistical properties of financial time series 96

5.3 Redundancy reduction on the FIB 30 index 98

5.3.1 Statistical analysis of the financial time series 100

5.3.2 Redundancy reduction through variance normalisation 101

5.4 Discussion 108

6 Conclusions 109

Chapter 1

Introduction

1.1 Premise

Adaptation is one of the first phenomena to have been investigated in neuroscience. Its importance stems from the basic need of every organism to adapt to the outside world to survive. Since the natural environment is a very complex, highly variable system, adaptation is not an easy task to perform. The idea underlying most of the investigation on adaptation is that the early visual system has developed and evolved to process the natural stimuli it receives. In particular, the idea underlying the work of this thesis is that the coding strategies of the first stages of visual processing are adapted to the specific statistical structure of the natural visual input.

The natural visual input can be seen as an ensemble of spatial patterns that continuously change in time. Its peculiarity lies in two main characteristics:

a) During the day, natural light intensities generally span more than nine orders of magnitude (Victor, 1999);

b) Natural light intensities are highly correlated both in space (Field, 1987; Ruderman and Bialek, 1994; Ruderman, 1994) and in time (Dong and Atick, 1995a; van Hateren, 1997).

In other words, though they span a huge range of values, natural light intensities are highly redundant, and thus partially predictable. On the other hand, single neurons have a limited capacity for receiving and sending information, and only a finite number of them sends messages to the higher areas of the brain. Hence, the visual input has to be coded and transmitted in non trivial ways to overcome two major problems:

1) How can the neurons in the early visual system represent such a wide

range of light intensity within their limited dynamic range of activity?

2) How is the information contained in the visual input encoded to be clearly interpreted and exploited by the higher areas of the brain?

The aim of this thesis is to give a contribution to the understanding of these questions concerning the processing of the temporal structure of the natural visual input. So, we will focus our discussion on temporal processing, and discard the spatial component of the input. We will come back to the limits of this approach later on.

The above questions are not easy to verify experimentally because of the complexity of natural images: the neural response to their wide range of values and their variable statistics is very difficult to characterize, both because it may be difficult to find neurons that have a decently high and continuous response to such widely variable stimuli, and because it is very difficult to correlate the response to the features of such a non-stationary input. In fact, in order to properly characterize the neural response, the overwhelming majority of the experiments have been carried out with simple stimuli (moving bars, drifting gratings etc.) because they stimulate the neuronal

activity in a much clearer way, and the analyses of the relation between the stimulus and the response is much less ambiguous. These experimental results already give a deep insight on the neural strategies really performed during everyday vision. In particular, the knowledge about the processing of the temporal statistics of natural images (the subject of investigation of this thesis) is increasingly detailed. It is well known that the first stages of visual processing adapt to the mean light level by coding the visual input through the 'Weber-Fechner' law (Victor, 1999). It is also known that the later stages of early visual processing (retinal ganglion cells in the vertebrate retina, large monopolar cells in the fly) also adapt to temporal contrast by a non-linear *contrast gain control* (Shapley and Victor, 1978). More recently, another series of experiments suggested that contrast gain control could be a way to normalise the output signal with respect to the variance of the visual input (Brenner et al., 2000; Fairhall et al., 2000, 2001). The discovery of this neural strategy could represent a partial answer to the first question, namely how the visual system has developed to overcome the physical constraints of its neurons. Still, since the stimuli used have a range much narrower than the natural one, and are far less correlated, these results cannot lead to any certain claim about visual processing in a natural environment.

The answer to the second question still centers around the original proposal suggested by Attneave (1954) and Barlow (1961) more than 40 years ago: one of the major principles underlying the coding strategies of the early visual system could be the need to build the most efficient representation of its input, namely a representation that displays the maximum content of information about the input. Since the visual environment is characterised by a complex structure and statistical regularity, and these are reflected in its redundancy, one goal of the early steps in neural processing could be to exploit this redundancy for an informationally optimal representation. Paraphrasing a recent revisitation of this concept (Barlow, 2001), coding should convert hidden redundancy into a manifest, explicit, immediately recognizable form. This can be done by separating the predictable part from the unpredictable part of the input, namely, separating the redundant component from the random one. Several investigations suggest that this could effectively be what early neural processing does. Most of them focused on the reduction of spatial redundancy (Barlow, 1961; Srinivasan et al., 1982; Atick and Redlich, 1992; van Hateren, 1992; Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998), only a few focused on temporal redundancy (Srinivasan et al., 1982; Dong and Atick, 1995b; Dan et al., 1996) and the overwhelming majority limited their analysis to the second-order

statistics.

The work described in this thesis attempts to show that, concerning temporal processing, we already have a partial answer to both questions. The hypothesis we try to verify is that the variance normalisation that seems to be the key mechanism underlying contrast gain control is not only an efficient coding strategy to fit the wide natural range into the limited neuronal one; it is also a very efficient strategy to remove almost all the redundancy (not only the second-order one) present in the temporal structure of natural images, thus reinforcing the idea that Barlow's proposal is realistic. We try to test this hypothesis by modeling the adaptation to the mean as a linear filtering, and the adaptation to contrast as a variance normalisation, as it emerges from the experiments. In order to simulate adaptation in natural vision, we then train the model on a data set of time series mimiking the input of a single photoreceptor in a natural environment. The choice of this data set, accurately recorded by van Hateren (1997), is not casual: van Hateren used it to study how the early visual system of the blowfly processes natural light intensities (van Hateren, 1997). In fact, he showed that the large monopolar cells of the fly continuosly adapt their dynamic range to the widely varying natural one so that the response distribution maintains

its compact, almost Gaussian shape independently from the variability of the input. We will show that a) the output of our model shows a dynamical adaptation very similar to that of large monopolar cells; b) variance normalisation decorrelates the input almost completely, also the redundancy of order higher than the second; c) despite the long-range correlations in the natural input, the optimal integration time is surprisingly short and compatible with the one observed experimentally. In other words, our simulations show why the visual system is able to adapt to the natural input in almost real time.

The last chapter of the thesis shows that, while trying to model a biological process, we ended up building a very simple and efficient way to separate the predictable from the unpredictable part of a signal. In order to verify the computational power of our model, we apply it to data that is deliberately very different from natural images - financial time series. The results are very similar to those obtained with natural time series.

1.2 Thesis overview

Chapter 2 is devoted to an overview of the experimental results on contrast adaptation and the theoretical framework that constitutes the basis of our

work. The general properties of natural images and the specific ones of the natural time series we use to simulate natural vision are discussed in *Chapter 3*. *Chapter 4* describes how the model is built, the training on natural time series and the statistical analysis of the results. Finally, *Chapter 5* shows the efficiency of our model when applied to time series of completely different nature - financial time series. A discussion on the overall results and further developments are drawn in the *Conclusions*.

Chapter 2

Temporal adaptation in the early visual system: Experimental findings and a theoretical framework

2.1 Introduction

This chapter is devoted to describing the experimental findings and the theoretical framework that inspired our model. Hence, it will present:

a) An overview of the biological phenomenon we wish to model. We will describe the main features of temporal adaptation emerging from years of experimental efforts, and we will try to characterise the properties of the coding mechanism that seems to underlie such process.

b) A theoretical framework that suggests the functional principles underlying temporal adaptation, and the mathematical tools to quantify its computational efficiency. We will describe some of the models of early sensory processing that have been based on the same theory, highlighting their merits, their limits and how we try to overcome them.

2.2 Temporal adaptation in the early visual system

The neural code of the early visual system is not a fixed set of rules linking the response to the stimulus with a static transformation independent from the stimulus itself. Instead, it depends significantly on the overall properties of the visual stimulus. In this section, we will briefly overview the current state-of-the-art on the adaptation of the neural code in the early visual sys-

tem to the temporal properties of its input. We will initially focus on the response of the vertebrate retina to time-varying visual stimuli, but we will see that the adapting mechanisms that emerge are not animal specific and are common to the motion sensitive neurons of the fly, suggesting that adaptation is driven by very general principles. Most of the following overview on retinal adaptation is taken from the reviews by Victor (1999) and Meister and Berry (1999).

2.2.1 Adaptation to the mean light intensity

A basic problem that the early visual system has to face is to operate over at least a 10^9-fold range of natural light intensities. The remarkable ability of the retina to accomplish this task is primarily due to the properties of the photoreceptors. In the lower half of the intensity range, signalling is accomplished primarily by rods, while in the upper half it is accomplished primarily by cones. To a first approximation, photoreceptor responses depend in a linear way on their photon catch. The behaviour is very close to linear for dim flashes whose intensity does not fluctuate over more than a decade. But even considering the subdivision in rods and cones, every photoreceptor should provide useful signals over a 10^5-fold range. Over most

of the operating range of the retina, contrast[1] changes of one part in 100 are readily detected. Were this to be accomplished by strictly linear photoreceptors, their outputs would need to be precise to within one part in 10^7. This wide dynamic range is incompatible with strict linearity. Instead, the sensitivity of rods and cones decrease with increasing illumination, in a manner in which the size of the response to a fixed change in contrast remains approximately constant. This relation, known as "Weber-Fechner law" of adaptation, implies that the retina produces approximately the same response for two visual displays that are related by a simple proportional scaling of all intensity values.

Such behaviour is of clear practical utility: because the intensity of the light illuminating the natural world changes over many orders of magnitude every day, so does the absolute intensity reflected by objects in the scene. However, the surface reflectance of these objects remains the same, and thus the relative ratios of intensities received from different parts of the scene are approximately independent of the illuminant. Through the adaptation to

[1]The conventional definition of contrast of a spot of light is Weber's one:

$$Contrast = \frac{L_{spot} - L_{background}}{L_{background}}.$$

However, contrast has been associated to many other similar measures. See (Tadmor and Tolhurst, 2000) for an accurate discussion on this.

the average intensity, photoreceptors encode the invariant features of objects and discard, for the most part, information about the absolute light level.

The same kind of response (approximately linear for small fluctuations around the mean light intensity, and following Weber law when variations are larger) is found in horizontal, bipolar and ganglion cells. (Given their highly non-linear behaviour, the role of amacrine cells in this adaptation process is still unclear, but we are not interested in their function here.). The time course of adaptation to the mean in all these neurons is remarkably short, being of the order of tens of milliseconds. Such a short time scale means that the retina is able to adapt its sensitivity to the mean light level in almost real time. This is an important feature of temporal adaptation, and will be one of the basic parameters of our model.

Along with the sensitivity, other aspects of the retinal response change with the average light level as well. In dim light, the time course of the response slows down considerably: in photoreceptors, a 20-fold decrease in intensity at the low end of the operating range is associated not only with a 5-fold increase in sensitivity, but also with a 2.5-fold lengthening of the latency-to-peak response; analogously, in ganglion cells, a brief flash pro-

duces a burst of spikes with longer latency and longer duration. Thus, retinal cells integrate the visual input over a longer time interval before reporting it to the brain. This averaging may be required to attenuate the effects of neural noise under conditions where the signal is small, but it comes at the cost of impaired time resolution. In the case of ganglion cells, spatial integration is also altered in dim light: the receptive field loses its antagonistic surround region, and subsequently the area in which light excites an ON-type ganglion cell expands somewhat. Again, this may be a strategy to enhance the visual signal by collecting as much light as possible, at the expense of some spatial resolution (Srinivasan et al., 1982). All these effects are also observed in human psychophysics (Shapley and Enroth-Cugell, 1984), suggesting that retinal processing largely accounts for the perceptual effects of light adaptation. However, in this thesis we will exclusively focus on the temporal integration of the visual stimulus, neglecting the spatial component. For a complete review on the subject, see Victor (1999).

2.2.2 Adaptation to contrast

The variability of the visual stimulus cannot be accounted for only by its mean: we expect that the early visual system has developed to adapt to the *whole* statistics of its input. In fact, it has been found that the visual sys-

tem also adapts to the spatial and temporal contrast of the light intensity. Here, we will review the main experimental results on temporal contrast adaptation in some detail, for it will be the crucial mechanism underlying our model.

While adaptation to the mean begins in photoreceptors, there is evidence that contrast adaptation does not occur before bipolar cells (Sakai et al., 1995), and is mainly implemented in the last part of the retinal path. Evidence for temporal contrast adaptation was found by Shapley and Victor (1978) in the *in vivo* response of cat retinal ganglion cells to simple spatial patterns (sine gratings or rectangular spots) modulated in time by a sum of sinusoids, whose amplitude and frequency was systematically varied. Analysing the response in the frequency domain, they showed that it is significantly altered by an increase in contrast: its shape is sharpened and shifted to high temporal frequencies, while its amplitude grows less than proportionally with contrast. The time course of this process was found to be of the order of 100 ms only, suggesting that retinal ganglion cells adapt to contrast almost as fast as they adapt to the mean. Shapley and Victor (1978) pointed out that this mechanism, called *contrast gain control*, cannot be reduced to trivial mechanisms like static saturation and adaptation

to the mean, clearly inconsistent with their data. They suggested instead that the origins of such a dynamic non-linearity could rely in the input of a network of (highly non-linear) amacrine cells. Recent investigations (Kim and Rieke, 2001) show that the anatomical basis of this process are indeed complex: contrast adaptation in ganglion cells includes contributions from mechanisms both acting on the currents reaching the ganglion cell soma and intrinsic to spike generation in the ganglion cell; moreover, inputs from bipolar cells appear to be relevant too. Despite its complex nature, contrast gain control seems to be an adapting mechanism that shows very similar properties in all vertebrate retinas.

The results described until this point all have a dramatic limit: the stimuli used in most experiments were simple spatiotemporal patterns that are very far from the natural ones. Following Barlow's idea (to which we will come back in the next section), the coding strategies of the early visual system have developed to adapt to the statistical properties of the natural environment (Barlow, 1961). So, we expect that the best way to characterise the neuronal response is to correlate it with the statistical properties of a whole distribution of signals as input, rather than with the particular features of individual stimuli. This statistical approach has been widely used

in the last decade, leading to results that are more easily connected with vision in a natural world.

Within this framework, new interesting results on contrast adaptation come from Smirnakis et al. (1997): to assess the dependence of the neural response to the contrast independently from the mean, they measured the *in vitro* response of rabbit and salamander retinal ganglion cells to a white noise stimulus (randomly refreshed every $2msec$) whose mean was held constant, and whose standard deviation was periodically switched between two different values. The interval between two switches was of several tens of seconds. The stimulus was a single spatially uniform field. The neural response was characterised by the first-order (linear) Wiener kernel, computed by correlating the firing rate and the preceding stimulus intensity. Results show that immediately after an increase in contrast, the linear kernel shrinks both in shape and amplitude; while within a few tens of milliseconds the shape seems to reach a stationary state, the amplitude keeps on decreasing within a time scale of seconds. At the same time, the mean spiking rate shows an initial abrupt increase followed by a roughly exponential decrease to a new steady value, considerably higher than the one at lower contrast. This result is two-fold: while it shows a short-term effect that recalls the contrast

gain control found by Shapley and Victor (1978) with very different stimuli, it reveals a truly new effect of long-term adaptation of the mean spiking rate.

The interpretation of the two processes has given rise to an interesting debate. In their paper, Smirnakis and collaborators claim that the slow (10-20 seconds) exponential decay of the mean spiking rate and the slow drop in evoked response amplitude are due to the same mechanism, to which they attribute the function of adapting to the scene statistics. They consider the contrast gain control a nonlinear feature of the light response too fast to be related to changes in the scene statistics.

Shapley gives his interpretation in a shortly following paper (Shapley, 1997): the rapid shrinkage in the time scale of the first order response is analogous to the effect of contrast gain control that Victor and him found 20 years before (Shapley and Victor, 1978), and contrast adaptation relies on this non-linear transformation; on the other hand, there is no clear evidence of the adaptive role of the long-term effect showed by Smirnakis et al. (1997), and the association between the slow decrease of the linear response amplitude and the exponential decay of the rate is dubious.

A breakthrough in the comprehension of the multiple mechanisms and time scales of temporal contrast adaptation has been given by the recent work of Fairhall et al. (2000, 2001): they replicated and extended the experiments of Smirnakis et al. (1997) on the motion sensitive neurons of the fly visual system, correlating the response with the temporal properties of the stimulus by systematically varying its time scale and amplitude. In order to investigate the existence of two very different time scales, they used a velocity stimulus $S(t)$ that is the product of two factors:

$$S(t) = s(t) \cdot \sigma(t), \tag{2.1}$$

where $s(t)$ is a normalised Gaussian white noise refreshed every $\tau_s = 2msec$ and $\sigma(t)$ is the envelope of the standard deviation, which varies on a characteristic time scale $\tau_a >> \tau_s$.

They initially repeated the experiments of Smirnakis and collaborators (this corresponds to choosing a piece-wise constant $\sigma(t)$ in Eq. 2.1) by varying the intervals between high and low contrast. Besides confirming Smirnakis et al. (1997)'s results, they showed that the time scale of adaptation of the rate is not absolute, but is a function of the time scale established in the

experiment: the time constant that characterizes the rate exponential decay is simply proportional to the interval length between two switches. This effect is certainly surprising, because it reveals a long-term mechanism that extends over a time scale that is two orders of magnitude larger than the one involved in fast adaptation. Moreover, it seems to be an important feature of early neural coding because it has been found in different species (salamander, rabbit and fly) and modalities (intensity and velocity detection). Nevertheless, the functional role of this effect is still obscure, and Fairhall and collaborators seem to agree with Shapley (1997) on the implausibility of its adaptive role (we will come back to this issue later on).

They then accurately investigated the relation between the stimulus and the neural response by using three different input pattern (besides the piecewise constant, a random one and a periodic one, so that the response didn't depend on particular features of the stimulus like the abrupt variations of the first one). They computed the response by assuming that the probability of firing a spike depended on the previous $100msec$ of the stimulus history. Since the stimulus space would then be an intractable high-dimensional object, they projected the stimulus on its most relevant direction, the spike-triggered average, and computed the neural response with respect to the

projected stimulus (see Appendix A for details). The results they find for the three different stimuli are shown in Figure 2.1: the response function varies widely with the stimulus statistics (center figures). But the plot of the rate response normalised by the average rate versus the stimulus normalised by its standard deviation shows that responses surprisingly overlap (bottom figures). This means that the system rapidly and continuously adjusts its coding strategy, rescaling the input/output relation in such a way that the response is invariant to changes of the input variance. The data show that this variance normalisation occurs within a time scale of $200ms$, that is almost as rapidly as the response is measurable, suggesting again that temporal contrast adaptation occurs almost as rapidly as adaptation to the mean.

Given these results, a detailed description of temporal contrast adaptation seems to be possible. Such adaptation appears to rely on two separate mechanisms involving two very different time scales and having two very different functions. Again, we cite Fairhall et al. (2000), who proposed that the neural response to a stimulus with well-separated time scales in the form of Eq. 2.1 could take the general form of a rate times timing code, where

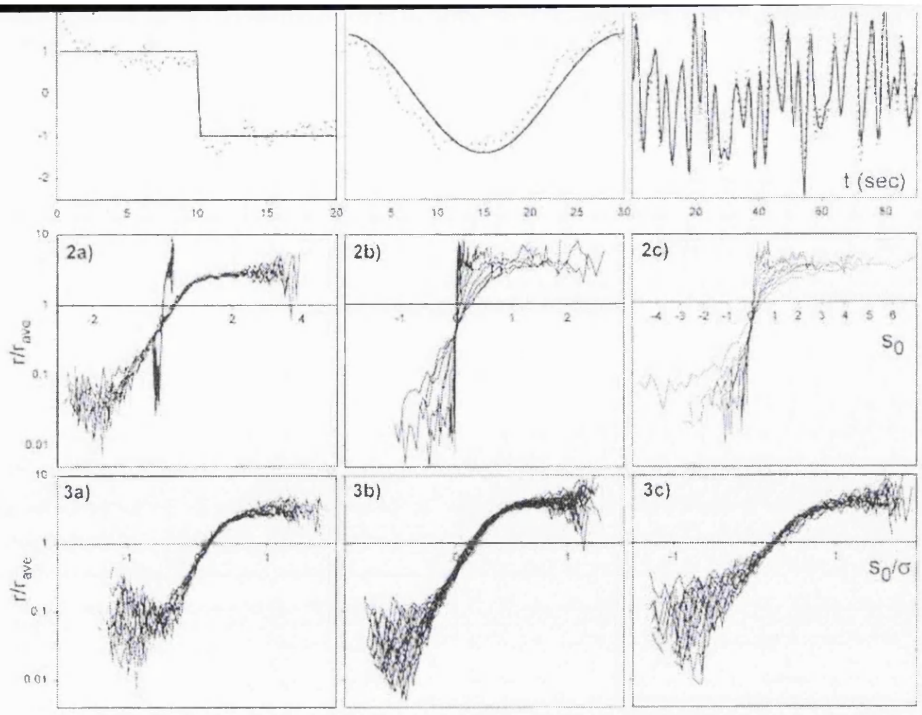


Figure 2.1: Input/output relations for (a) switching, (b) sinusoidal and (c) randomly modulated experiments. The three top figures show the modulation envelope $\sigma(t)$, in log for (b) and (c) (solid), and the measured rate (dotted), normalised by mean and standard deviation. The three center figures show input/output relations calculated in non-overlapping bins throughout the stimulus cycle, with the input s_0 in units of the standard deviation of the whole stimulus. The three bottom figures show the input/output relations with the input rescaled to units of the local standard deviation. From (Fairhall et al., 2000).

the response may be approximately modeled as

$$r(t) = R[\sigma(t)] \cdot g(s(t)), \tag{2.2}$$

where the function R modulates the overall rate and depends on the slow dynamics of the variance envelope, while the precise timing of a given spike in response to fast events in the stimulus is determined by the non-linear input/output relation g, which depends only on the normalised quantity $s(t)$. Adaptation to the input statistics seems to rely on g, that seems to

maximise information transmission about the fast components of the stimulus through the apparent normalisation by the local standard deviation. The same variance normalisation had been previously found with slightly different stimuli by Brenner et al. (2000), and seems to be a fundamental strategy of adaptation. Its time course and properties are compatible with the well-known contrast gain control, suggesting that the two processes are in fact the same mechanism.

The function R modulating the rate varies on much slower time scales, and cannot be taken as an indicator of the extent of the system's adaptation to a new ensemble. Rather, R appears to function as an independent degree of freedom, capable of transmitting information, at a slower rate, about the slow stimulus modulations.

2.3 Temporal adaptation: An information theoretic approach

So far, we have seen that there is evidence for several adaptation mechanisms characterizing temporal processing in the early visual system. At this

point, we can come back to the questions that we posed in the Introduction and ask: Is it possible to individuate any general principle underlying these mechanisms? What drives the coding strategies of the early visual system? By trying to answer this question, we will set the theoretical basis of our work.

2.3.1 Biological constraints: noise and saturation

One of the first, basic problems that the visual system has to face is simply given by its physical constraints: while the natural light intensity spans more than nine orders of magnitude, each neuron has a limited dynamic range of activity; in order to avoid both low signal-to-noise ratio and saturation, an efficient coding strategy should be able to continuously match such limited range with the wide dynamic range of light intensities. This suggests that the coding strategy should adapt to the statistics of its natural input in such a way to fully exploit all the dynamic range of the neuron. Laughlin (1981) was one of the first to show that this is indeed what seems to occur. He presented a very simple version of this problem, in which a single cell must encode contrast variations with a graded voltage response. Assuming that the range of voltages is constrained by fixed limits, the optimal coding

strategy for information transmission would be the one that equally exploits all the finite range of voltages. Laughlin measured the (very skewed) distribution of contrasts as seen through an aperture size of a fly photoreceptor as it moves through the woods, and computed the static non-linear contrast-to-voltage conversion that would reshape such distribution into a flat one. This was compared to the input/output relation of the second order neurons (the large monopolar cells, LMC) in the fly visual system, and the match was very good.

The perhaps more important limit of Laughlin's simple experiment is that the response function is static. Instead, as we have seen in the previous section, the response continuously adjusts to the local range of activity of the time-varying stimulus. In a more recent experiment (briefly indicated in the introduction), van Hateren (1997) shows that the response adaptation in time is so quick that the output range is always fully exploited, no matter what the range of the input is. In order to mimic natural vision, he measured time series representative of what each photoreceptor of a real visual system would receive in a natural environment, and recorded the activity of the fly photoreceptors and LMCs responding to these time series. His results show that, while photoreceptors code the widely varying natural intensities

with a logarithmic transformation, the distribution of the output of LMCs has an approximately Gaussian shape that is almost independent from the distribution of the input. This suggests the existence of a non-linear gain control that fits the response into the limited dynamic range of LMCs, by continuosly changing the parameters of the response. It is compelling to believe that such gain control relies on the same adaptation mechanism (described in the previous section) that performs variance normalisation in the motion-sensitive cells of the fly.

2.3.2 A computational strategy: redundancy reduction

Still, these experiments say little on the mechanisms employed to overcome the constraints given by the limited range. Moreover, once the input is encoded, it has to be decoded by higher areas of the brain. In other words, the representation elaborated by the early visual system should be sent in a form that is easily used by higher neural processing. This simple observation set the basis for the original proposal suggested by Attneave (1954) and Barlow (1961) more than 40 years ago: one of the major principles underlying the coding strategies of the early visual system could be the need to build the most efficient representation of its input, namely a representation

that displays the maximum content of information about the input. Since the natural visual input is highly redundant (we will see this in detail in the next chapter), one goal of the early steps in neural processing could be to exploit this redundancy for an informationally optimal representation. Barlow formulated this proposal specifically thinking about the early visual system: since each retinal ganglion cell has a limited information capacity, the best strategy to overcome such limitation would be to make the signals of the single neurons as statistically independent as possible, so that each one sends a different message and the overall information is maximised. Given that each ganglion cell codes information for the light intensity at a certain spatial location of the visual scene, the optimal coding strategy would be the one that most reduces the spatial redundancy present in natural images.

Thanks to its simple logic and high predictive power (but see the comments at the end of this section), the principle of redundancy reduction has inspired a large number of investigations on the early visual system, finding several experimental confirmations. For example, Srinivasan et al. (1982) suggested that both the center-surround receptive field and the temporal response of X retinal ganglion cells enable the decorrelation of the spatio-temporal input. Atick and Redlich (1992) investigated whether the hypoth-

esis of second order redundancy reduction could predict the contrast sensitivity response observed in psychophysical experiments. They constructed the optimal contrast response assuming that it is linear and that it would whiten the spatial power spectrum of natural images (that is characterised by a $\propto 1/k^2$ decay (Field, 1987)). They showed that, if quantum noise is included, this prediction matches very well with contrast sensitivity curves observed in psychophysical experiments. A similar paradigm was used to predict the spatiotemporal receptive fields of fly LMCs van Hateren (1992), obtaining results consistent with physiological experiments.

Atick and co-workers also extended their theory to temporal processing, by first measuring the spatio-temporal correlation structure of visual scenes (Dong and Atick, 1995a), and then predicting the linear temporal filter that would whiten the temporal spectrum of natural images (Dong and Atick, 1995b). Based on experimental results, they predicted that the lateral geniculate nucleus is concerned with improving efficiency of visual representation through active temporal decorrelation of the retinal signal in much the same way that the retina improves efficiency by spatially decorrelating incoming images. They tested their prediction experimentally (Dan et al., 1996) by recording the response of the cat LGN neurons to time-varying

natural visual stimuli. Their prediction was confirmed: while the response to natural scenes showed no linear correlations, the response to white noise was not completely decorrelated. This suggests that the coding strategy has specifically adapted to decorrelate the natural visual input.

2.3.3 Redundancy reduction revisited

Though redundancy reduction is a very attractive idea, its use has been sometimes misleading and contradictory, to the point that, in a very recent paper, Barlow felt the need to revisit and partially correct the original proposal (Barlow, 2001). Inspired by Barlow's new vision, we draw two important observations:

a) Though the original idea was right in drawing attention to the importance of redundancy in sensory messages because this can often lead to crucially important knowledge of the environment, it was wrong in emphasizing the main technical use for redundancy, which is compressive coding. The reason is that it is knowledge and recognition of the redundancy, not its reduction, that matters. Thus, coding should convert hidden redundancy into a manifest, explicit, immediately recognizable form, rather than reduce it or eliminate it. Following this idea, the spatial and temporal decorre-

lation observed in the early visual system shouldn't be seen as a way to reject useless redundant information; it should rather be seen as part of a strategy of separation of different components of the signal, in this case the high frequency one from the low frequency one, where the latter is likely to be represented by other nerve fibers. Within this view, temporal adaptation could be seen as a way to separate the high frequency, unpredictable component from the low frequency, predictable one. We believe that this hypothesis is a realistic one.

b) Even though all the studies we cited claim to investigate redundancy reduction, almost all of them limit their analyses to second order redundancy. As we will see in the next chapter, natural scenes contain much more information than is captured by the power spectrum or the autocorrelation function. We believe that the coding strategy of the visual system is also influenced by the redundancy of order higher than the second. Specifically, we believe that the variance normalisation mechanism we have described in the previous section is nothing but a strategy to remove the higher order redundancy given by correlations in contrast. This will be the main hypothesis underlying our model of temporal adaptation.

2.4 How does adaptation work in a natural environment?

Let us summarize our knowledge on the adaptation of the retina to the temporal structure of its input. Given the experimental results that we have just described, adaptation seems to be implemented in a two-step process:

1) The retina adapts to the mean light intensity by coding the input with the Weber-Fechner law;

2) The retina further adapts to the contrast by a non-linear transformation that has the effect of normalising the response with respect to the local standard deviation of the stimulus.

Unfortunately, as we pointed out in the introduction, it is very difficult to characterise the neural response to natural stimuli. Thus, we don't know whether such adaptive mechanisms work in the same way in a natural environment with a much wider range and much stronger and long-ranged correlations. On the other hand, the few experiments carried on with natural time-varying stimuli suggest that both adaptation to a wide input range

(van Hateren, 1997) and decorrelation indeed occur in a natural environment (Dan et al., 1996; van Hateren, 1997). Still, in these latter experiments, the complexity of the natural stimuli didn't allow a detailed analysis of the adaptive mechanisms performed.

The work of this thesis aims to prove that the link between the above results is very strong. The main hypothesis is that the two-step adapting strategy outlined above is sufficient not only to adapt the limited dynamic neuronal range to the wide range of the natural visual input, but also to fulfill the computational goal of removing most of the redundancy of natural time series.

In order to understand how complex the temporal structure of natural images is and how natural images differ from the stimuli that were used in the experiments, we will analyse the statistical structure of the natural time series we use (Chapter 3). In Chapter 4, we will try to prove the above hypothesis by modeling the two-step adaptation process in a very simple way, and then training it on time series of natural images.

Chapter 3

Temporal statistics of natural images

3.1 Introduction

The idea that the development and evolution of the early visual system is strongly linked to the statistics of its input led many people to study the statistics of natural images. The standard approach is to analyse the statistical properties of large groups of natural scenes, implicitly assuming the existence of an image *ensemble* from which the single images are drawn. Despite the different environments from which the images were taken and the vast experimental differences in data collection, it has been possible to

identify some statistical features that are surprisingly ubiquitous. The best known one is the robust scaling of the second-order statistics: the power spectrum of natural scenes (averaged over all orientations) takes the form of a power-law in the spatial frequency:

$$S(k) = A/k^{2-\eta}, \tag{3.1}$$

where k is the magnitude of spatial frequency, η is the 'anomalous exponent' (usually small), and A is a constant which determines the overall image contrast (Field, 1987; van Hateren, 1992; Ruderman and Bialek, 1994; Ruderman, 1994). This result provides evidence for a certain symmetry in ensembles of natural images: scale invariance (Ruderman, 1994). Scale invariance implies simply that the image statistics do not change with the angular scale. Pictures of such an ensemble will have the same ensemble statistics regardless of the lens' focal length. More generally, the new ensemble may be *self-affine* to the original one, meaning that the new images must also be multiplied by a suitable constant after rescaling to make the statistics identical to the original ones. If $Q[\phi(\alpha x)]$ is any ensemble statistics of $\phi(x)$ on scale α, then scale invariance implies that

$$Q[\phi(x)] = Q[\alpha^{\nu}\phi(\alpha x)], \tag{3.2}$$

where ν is an universal exponent (i.e. it is independent of both α and Q). Thus in a scale-invariant ensemble we can make the replacement $\phi(x) \rightarrow \alpha^{\nu}\phi(\alpha x)$ for all instances of ϕ in any expectation value.

This is a strong statement. It greatly restricts the form of the image distribution. Such a property also gives us some intuition about natural scenes instead of a mere quantification of their statistics. For instance, it reinforces the notion that objects in the natural world can appear at any angular scale in an image (i.e., they can be any distance away), which is one plausible mechanism for producing scale invariance.

A reasonable explanation of the real cause of scaling has been given by Ruderman (1997): natural images are composed primarily of statistically independent objects which occlude one another. Further, natural environments tend to arrange themselves so that the image regions corresponding to these objects are power-law in size. Combined, these two properties give rise to scaling universally.

These results concern the spatial correlations in the natural images. Much less is known about the temporal part, that is the part in which we

are interested. Dong and Atick (1995a) showed that spatial and temporal correlations are strongly intertwined: the general form of the spatiotemporal power spectrum is

$$S(k,\omega) \propto k^{-m-1} F(\omega/k), \tag{3.3}$$

where k is the spatial frequency, ω is the temporal frequency, $F(\omega/k)$ is a non-trivial function of the ratio ω/k, and m is the exponent of the static power spectrum (if we compare it with Eqn. (3.1), $m = 2 - \eta$).

The link between this result and the temporal structure of the light intensity hitting a single photoreceptor is non-trivial, because it depends on many factors: the velocity of objects, that of the eyes, the acceptance angle of the photoreceptor. However, the intertwining between spatial and temporal components suggests that the temporal structure of the photoreceptor's input should also be influenced by the overall scaling of natural scenes in some way. van Hateren (1997) directly investigates the temporal structure of the light intensity hitting a photoreceptor by recording time series of natural intensities as it is described in the next section. He shows that the power spectrum of the time series is also a power-law function, claiming that it could be derived from Eqn. (3.3). This result confirms the scaling structure of the second-order statistics.: the scaling of natural scenes strongly influ-

ences the temporal structure of the visual input.

In the following of the chapter, we will give a detailed description of the statistical properties of the time series recorded by van Hateren, and we will show, starting from his result, that the scaling of natural scenes strongly influences the second- as well as the higher-order temporal structure of the visual input.

3.2 Time series of natural images

3.2.1 How they were recorded

The natural time series we will use consist of 12 recordings of natural intensities, each 45 minutes long, kindly made available by van Hateren on his web site (http://hlab.phys.rug.nl/tslib/index.html) [1]. The 12 time series were recorded by a photodetector that has a spectral sensitivity similar to the photopic sensitivity of the human eye, an angular resolution of a few arcminutes, i.e., comparable to that of human foveal and parafoveal vision, and a temporal resolution of $1.2kHz$. The total system is linear in intensity over more than four orders of magnitude. The optical system was mounted

[1] For the technical details of the recordings, see (van Hateren, 1997).

on a headband and worn by a freely walking person. Since the device follows the direction of the gaze of the head (and not that of the eyes), the subject wore marked glasses, and was told to keep the markers at a fixed position in the visual field. Recordings were made by different subjects walking in various environments (woods, fields, near lakes, and residential areas) under various weather conditions (sunny, overcast and foggy). Obviously, they must be considered as only a crude approximation of what would result when real eye dynamics are taken into account. Yet, such time series are likely to be close enough to the natural input of photoreceptors to enable a meaningful analysis of light adaptation in a natural environment[2].

3.2.2 Statistical structure of the time series

Photoreceptors are known to have a sensitivity that is approximately proportional to the logarithm of the light intensity (Bownds and Arshavsky,

[2]Reinagel and Zador (1999) studied how voluntary eye movements in humans change the statistics of the foveal photoreceptors' input. They observed that image patches selected for viewing had a higher spatial contrast and a steeper two-point correlation function than patches selected at random. However, it is evident from the figures they show that such an effect is quite weak and doesn't substantially modify the shape of the correlation function. This suggests that the main difference between van Hateren's time series and the real input of the subject's photoreceptors could be that the real input is slightly less correlated than the recorded one, but its correlational structure shouldn't be substantially altered. The difference is likely to be even smaller than the one measured by Reinagel and Zador because van Hateren sampled the natural scenes following the position of the head rather than randomly.

1995). This is the main reason why we study the statistical properties of the logarithm of the light intensity,

$$x(t) = \log(I(t)), \tag{3.4}$$

rather than the intensity itself. Here time t is a discrete variable taking the values $t_k = k\Delta$, where k is a positive integer and $\Delta = 0.8333$ msec is the time interval between two consecutive data recordings. In most of the following analyses, we show the statistical properties of three of the 12 time series. This to show that, while some statistics are shared among the time series, some others aren't, and the eye has to cope with all of them. In order to show the differences, the three time series were among the most different ones, and can be considered representative of the other nine.

Probability distribution

Figure 3.1 shows the log-intensity histograms of three of the 45 minutes long time series. As can be seen, the histograms are significantly different from zero for a broad range of values, up to seven log-units. Their roughly uniform distribution over this logarithmic range means that the original data are very skewed to low values. More importantly, the shape and width of the histograms vary significantly among different time series. This is an evident

sign of non stationarity: the data are far from being drawn independently from a single probability distribution.

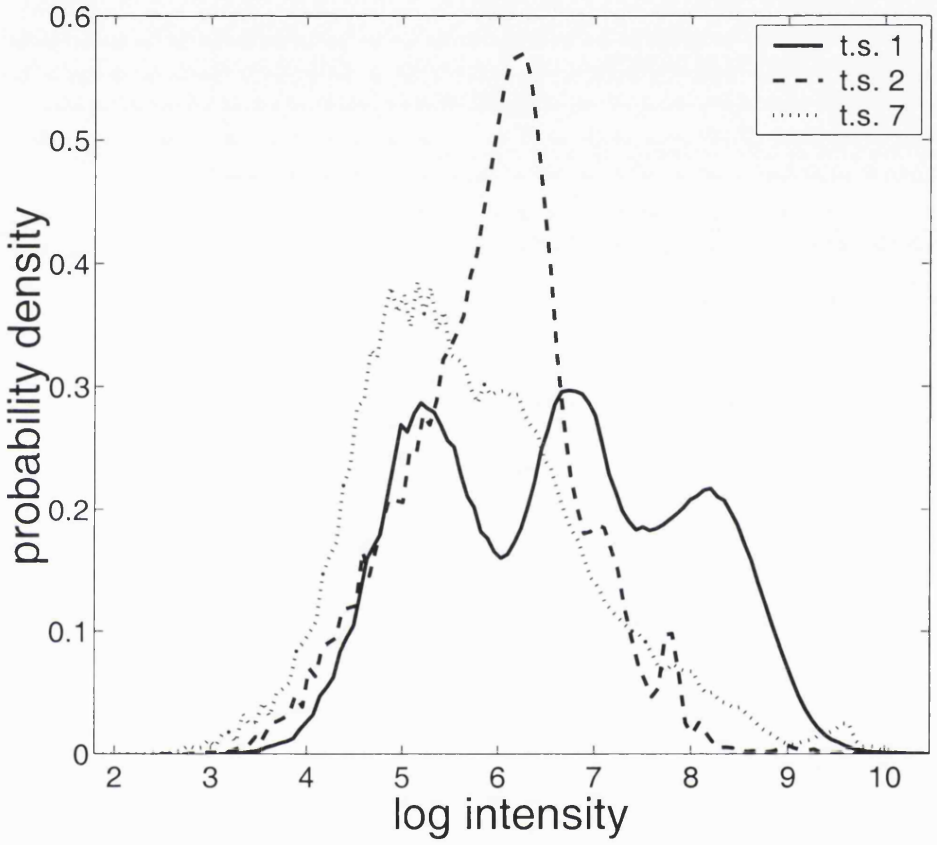


Figure 3.1: Histograms of the log-intensity of the first, second and seventh time series (45 minutes each).

Second-order statistics

We study the second-order statistics beginning from the power spectrum, defined as

$$S(f_n) = \frac{1}{N^2}\langle \hat{x}(f_n) \cdot \hat{x}(f_n)^* \rangle, \tag{3.5}$$

where $\hat{x}(f_n)$ is the Fourier transform of the log-intensity,

$$\hat{x}(f_n) = \sum_{k=1}^{N} x(t_k) e^{-2\pi i k n/N}, \tag{3.6}$$

$t_k = k\Delta$, $f_n = n/(N\Delta)$, $0 \leq n \leq N/2$, and $\Delta = 0.8333$ msec is the sampling interval of the time series; the brackets $\langle \ldots \rangle$ denote averaging over segments of length N overlapping by one half of their length. This procedure significantly reduces the variance of the power spectrum estimate with respect to taking a very long single segment (Press et al., 1987). As in (van Hateren, 1997), power spectrum was evaluated by the periodogram estimate (3.5) without any prior windowing.

Figure 3.2 shows that the power spectrum of three time series has a power-law decay, with a power very close to 1. This result confirms the temporal scaling already shown by van Hateren (1997) for the first time series, and previously suggested by Dong and Atick (1995a). For frequencies

higher than approximately 100 Hz, the power spectrum starts to deviate from a power law: as suggested by van Hateren (1997), this is due to the low-pass filtering effect of the spatial aperture of the light detector, in combination with the upper limit of angular velocities produced by the subject carrying the detector. The homogeneity of the decay among different time series is quite striking.

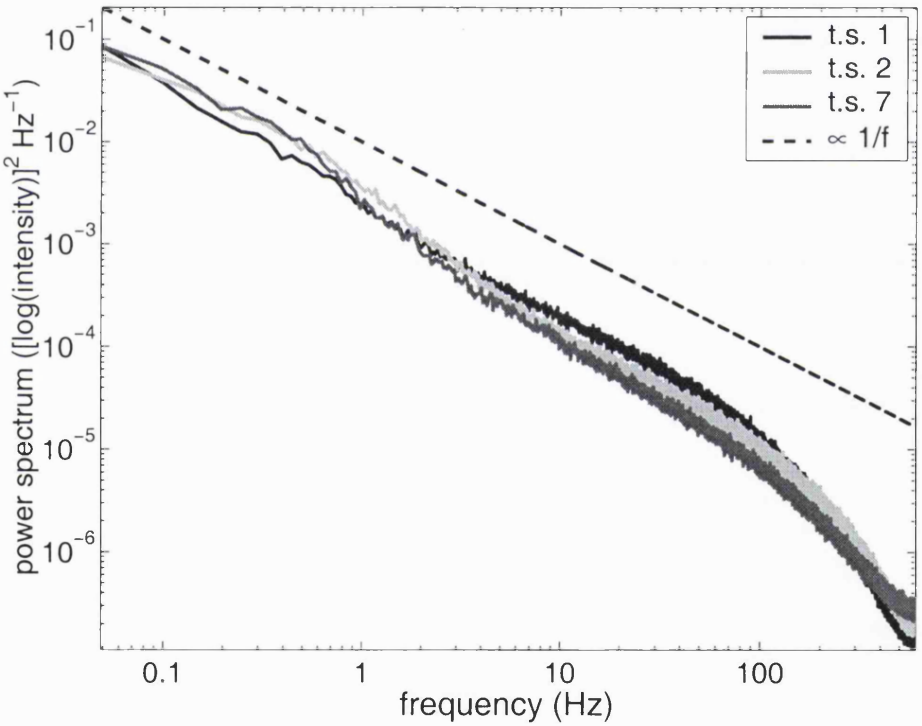


Figure 3.2: Log-log plot of the power spectrum of the log-intensity of the first, second and seventh time series (45 minutes each), averaged over segments of 20.48 seconds (24576 time steps).

Does the square windowing intrinsic in the periodogram estimate introduce any artifacts? In Figure 3.3, we compare the periodogram estimate with the power spectrum estimate obtained by preliminarly windowing the data with a Welch and a Hanning window: while they look indistinguishable for high frequencies, the Welch and Hanning estimates deviate from a power law at very low frequencies. As shown in the same figure, this effect tends to disappear when the estimate is calculated over a longer segment so that sampling at lower frequencies is allowed. This suggests that such effect is presumably due to the presence of a peak of the power spectrum around zero frequency which alters the Welch and Hanning estimates at immediately higher frequencies because their smooth leakage function is not zero one frequency bin away from its center. This does not happen with the periodogram estimate because its leakage function is exactly zero one bin away. This is the reason why we showed the periodogram estimate. The effect remains after subtracting the mean, suggesting that the peak at very low frequencies is due to correlations extending over a very long range.

Since we will mostly deal with time rather than frequency, we show the

autocorrelation function, $C_x(s)$, defined as

$$C_x(s) = \frac{\langle x(t)x(t+s)\rangle - \langle x(t)\rangle^2}{\langle x^2(t)\rangle - \langle x(t)\rangle^2}, \tag{3.7}$$

where $\langle \cdots \rangle$ indicates temporal averaging. The autocorrelation function has a very slow decay, even if in this case the decay is not uniform neither within, nor among the time series (Figure 3.4). In particular, the decay for the time series 2 and 7 seems to be characterised by a transition between two different power laws. Despite this non-uniformity, however, the decay is of a power-law kind up to a very long range for all the time series, indicating the presence of correlations extremely extended in time.

Higher-order statistics

Although they are ignored in most of the literature, higher order statistics also give very important information: Figure 3.5 shows that the quadratic correlation function of the log-intensity, $C_{xx}(s)$, defined as

$$C_{xx}(s) = \frac{\langle x^2(t)x^2(t+s)\rangle - \langle x^2(t)\rangle^2}{\langle x^4(t)\rangle - \langle x^2(t)\rangle^2}, \tag{3.8}$$

decays very slowly, again approximately as a piece-wise power-law, revealing that statistics of order higher than two are also long-range correlated. Though the correlation function in Figure 3.4 and Figure 3.5 look quite similar, we will see in the next chapter that the long-range correlated structure of the quadratic fluctuations around the mean light level is non-trivial since it is only partially reduced when the second-order correlations are removed.

Mutual information

In order to have a more general measure of the two-point correlations, we analyse the mutual information of the signal at a single time with respect to the same signal s time steps before:

$$I(x,s) = \sum p(x(t), x(t-s)) log \frac{p(x(t), x(t-s))}{p(x(t)) \cdot p(x(t-s))}. \tag{3.9}$$

This gives us a measure of the two-point correlation at any delay or, in other words, how predictable is the signal at time t from the signal at time $t-s$.

In order to calculate the mutual information, the data for every time series were binned in 100 bins of the same size, and the distributions $p(x)$ and $p(x(t), x(t-s))$ were evaluated. Even with our large data set (3240000 data

points for each time series), Eqn. (3.9) gives a biased estimate for the mutual information, due to the finite sample size. An idea of this finite-size bias can be given by the mutual information obtained after randomly reshuffling the data, that should be zero for an infinite data set. Evaluating it over the time series, this bias turns out to be of the order of 10^{-3} nats, which is more than two orders of magnitude smaller than the mutual information of the original data; therefore, we can be quite confident that the finite-size effect is very small. Since in the next chapter we will compare the mutual information of different variables for which the bias can be different, we subtract it from the mutual information, obtaining the unbiased mutual information I_u.

Figure 3.6 shows the results. The unbiased mutual information of the original signal, $I_u(x, s)$, is very high, and decays as a power law. Such power-law decay has been found in the mutual information between two pixel values versus their distance in natural images (Ruderman, 1994), and between two letters as a function of distance in English texts (Ebeling and Poschel, 1994): it is - again - common to systems characterised by infinitely extended correlations.

Comparison among the distributions

The amount of data needed to get a reasonably good approximation of the equilibrium distribution of a variable depends on the correlation between consecutive data points. The more (positive) correlation, the longer one needs to measure before the estimated distribution approaches the true one. A measure of how well the distribution of a variable, based on binned data, approaches the true distribution will therefore give an indication of how much correlated this variable is. We measured how the distance between estimated distributions from different time series decreases with the length of the data segment from which they are estimated. The reason why we also used this method is because it gives an estimate of the correlations between consecutive data points at all orders, and not only the pair-wise ones. Such estimate could not be measured with the mutual information because even the three-point mutual information (how the signal at time t depends on the signal at times $t - t_1$ and $t - t_2$ for any time lags t_1 and t_2) would require a data set much larger than ours.

The values of each time series were binned, and the probability of a data-point falling in bin i, $\rho(i)$, estimated as $\rho(i) = n_i/T$, where n_i is number of observed data points falling in bin i, and T is the total measuring time, or

the total number of data points. To correct for the non-uniform shape of the overall distribution, the data were binned in K bins of variable width chosen such that a data point had equal probability to end up in any of the bins. Bin i was given by x being between x_{i-1} and x_i, where $x_0 = -\infty$ and $\int_{x_{i-1}}^{x_i} p(x)dx = 1/K$, where we used the data of all time series to estimate $p(x)$.

As measure for the distance, D, we used the sum of the square of the difference between the estimated distributions.

$$D_{a,b} = \sum_{k=1}^{K} [\rho_a(k) - \rho_b(k)]^2, \tag{3.10}$$

where ρ_a and ρ_b indicate the estimated distribution for the a^{th} and b^{th} time series respectively. The only reason why we choose a quadratic distance is its analytical simplicity, but we expect our results to hold for a different distance measure. This distance was measured for estimated distributions based on T consecutive data points from different time series for different values of T. Figure 3.7 shows $D(x)$, computed by averaging $D_{a,b}$ over all 66 pairs of time series, plotted against T. Also plotted is $D(rnd(x))$, computed as $D(x)$ after randomly reshuffling x. In appendix B, it is shown

that, under the assumption that each data point is drawn independently from the equilibrium distribution, the expected value of $D_{a,b}$, $\overline{D}_{a,b}$, satisfies $\overline{D_{a,b}} = 2(K-1)/(KT)$. Figure 3.7 shows that the agreement between $D(rnd(x))$ and the analytical prediction $\overline{D_{a,b}}$ is very good.

As the figure shows, the distance between the estimates of log-intensity distribution is larger than that for the independently drawn one. Moreover, it falls off more slowly than $1/T$. This reflects the long-range correlation structure in the statistics of x. The fact that for any length T the slope in the log-log plot of distance versus time is larger than -1 is due to the power-law decay of the temporal correlations in x. Correlations with a finite time scale would give rise to a slope of -1 for T larger than that time scale.

3.3 Discussion

The main result of this statistical analysis is that van Hateren's time series have long-range correlations that extend on a very long temporal scale (at least tens of seconds). The power-law decay of the mutual information implies that light intensity at time t is highly predictable from light intensity at previous times. At the same time, as it is evident from the variety of the distributions of the single time series, light intensities span a huge range of

values. These two characteristics - wide range of intensities and long-range correlations - distinguish the natural time series from the standard experimental stimuli.

What is the optimal strategy to code such a complex stimulus into a limited range of response? A suggestion comes again from Ruderman and Bialek (1994), who investigate the non-Gaussian nature of the histogram of local contrast, defined as $\phi(x) = \log[I(x)/I_0]$, where I_0 is chosen for each image so that the average contrast is zero. They show that, while no linear transformation on the images produces Gaussian distributions, the histogram distribution of the local contrast normalised by the local mean and the local standard deviation produces a distribution that is very close to a Gaussian.

In the next chapter, we will model the neuronal adaptation to the temporal structure of natural time series through a transformation similar to the spatial normalisation proposed by Ruderman and Bialek to have a Gaussian distribution. It will be shown that the long-range correlations that emerge in the higher-order statistics do not depend on the second-order statistics only, but contain information that the power spectrum and the autocorrelation

function cannot describe.

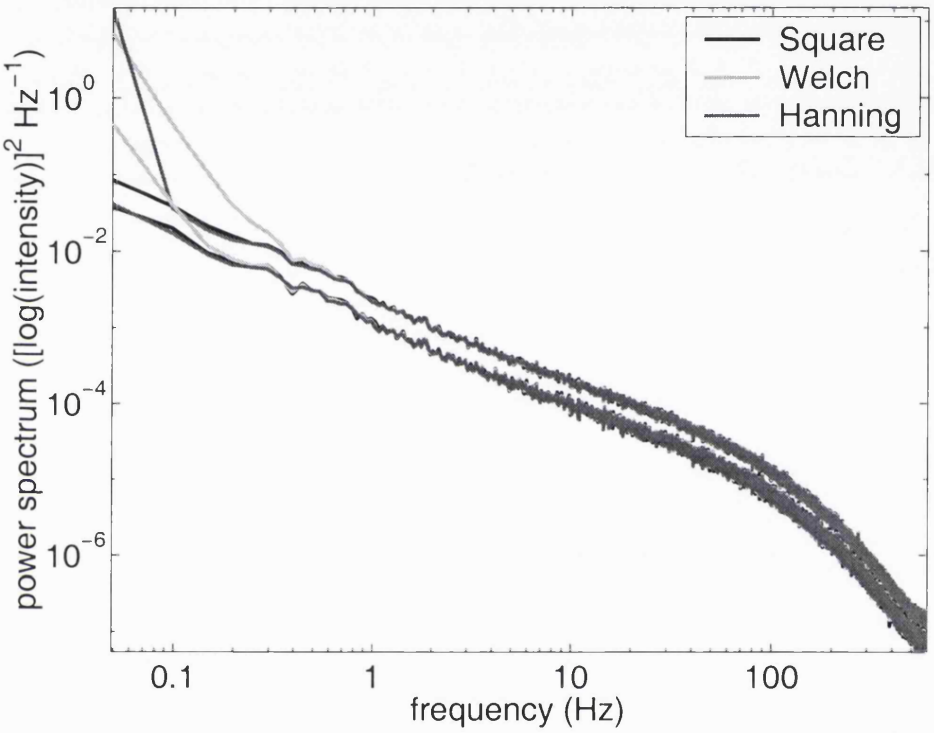


Figure 3.3: Log-log plot of the power spectrum of the log-intensity of the first time series, calculated using the periodogram (square windowing), Welch windowing and Hanning windowing. The three top curves refer to the estimates averaged over segments of 20.48 seconds (24576 time steps), while the three bottom curves show the estimates averaged over segments of 40.96 seconds (49152 time steps), at the same frequency values of the estimates evaluated over the shorter segments.

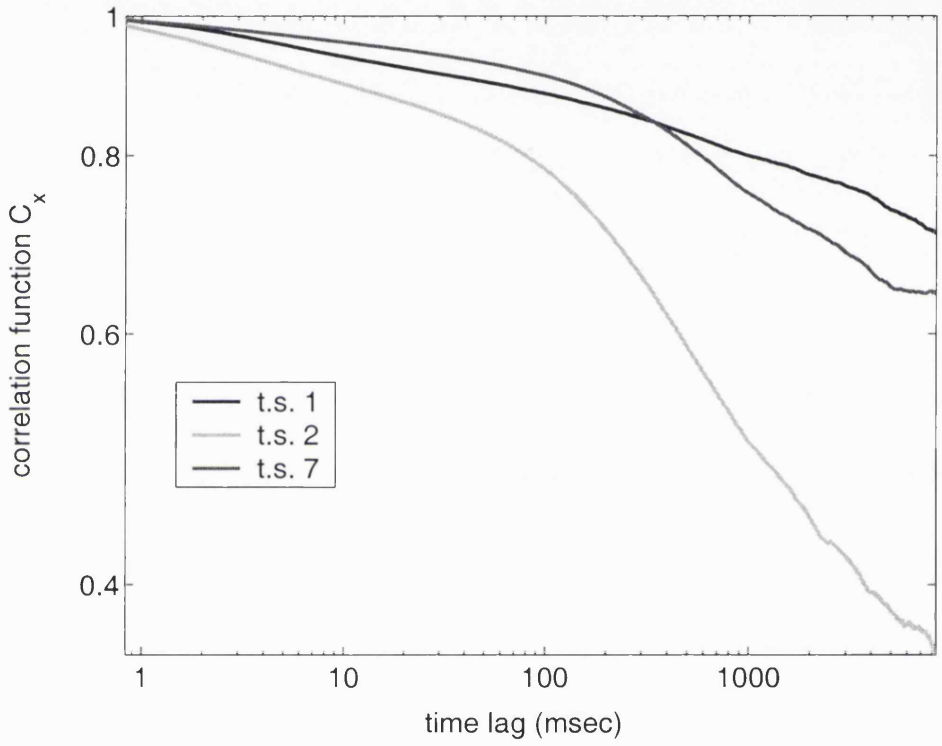


Figure 3.4: Log-log plot of the autocorrelation function of the log-intensity of the first, second and seventh time series (45 minutes each).

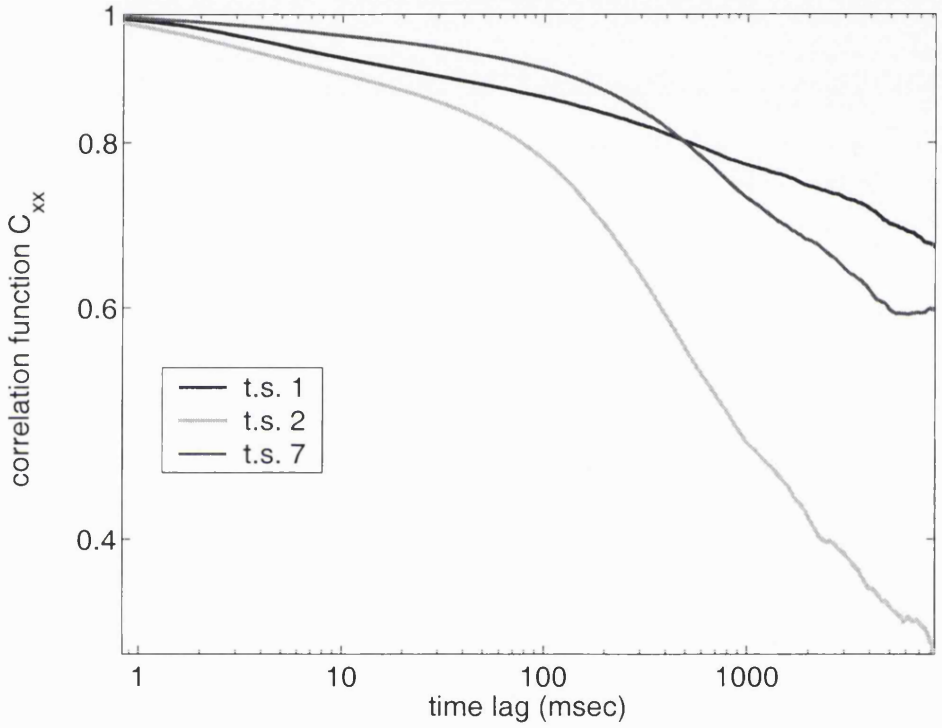


Figure 3.5: Log-log plot of the correlation function $C_{xx}(s)$ of the first, second and seventh time series (45 minutes each) .

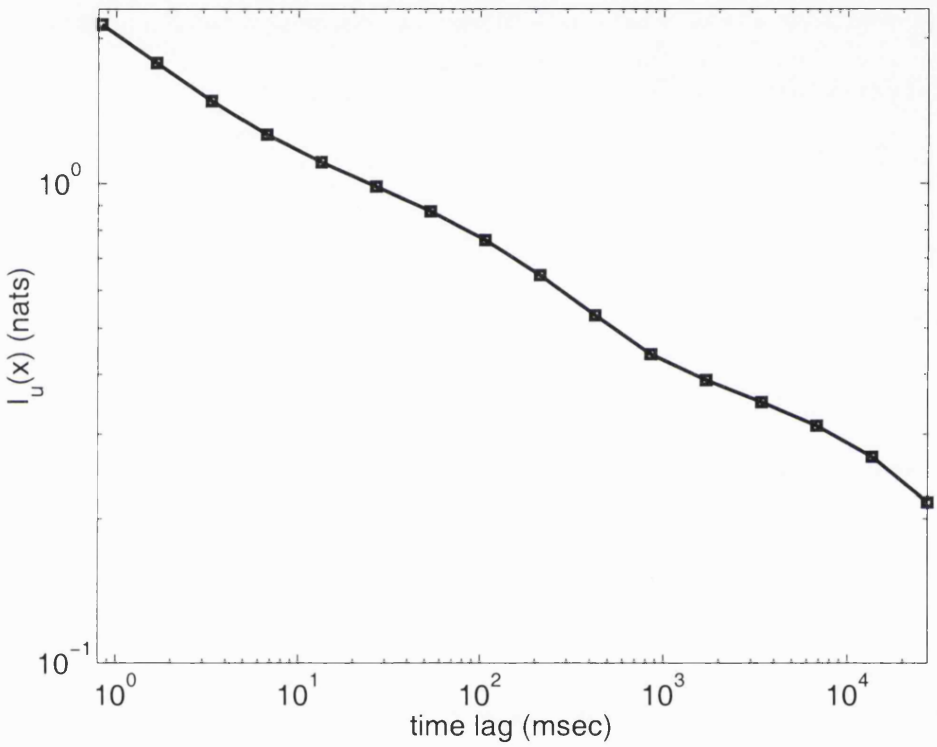


Figure 3.6: Log-log plot of the unbiased mutual information, averaged over all time series.

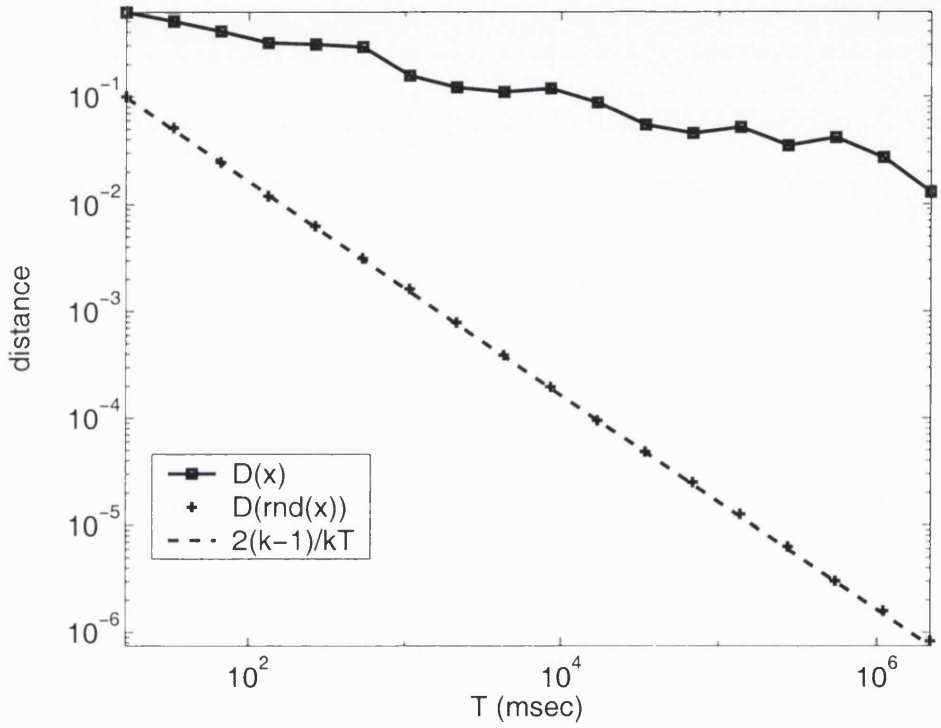


Figure 3.7: Log-log plot of the distance between the distributions of x for increasing length of the segment from which they were estimated, randomly reshuffled x ($rnd(x)$) averaged over all time series, and expected distance for independently drawn data.

Chapter 4

Variance normalisation

4.1 Introduction

In Chapter 2, we saw that adaptation to the temporal statistics of light intensity seems to be implemented in a two-step process:

1) The retina adapts to the mean light intensity by coding the input with the Weber-Fechner law;

2) The retina further adapts to light contrast by a non-linear transformation that has the effect of normalising the response with respect to the local standard deviation of the stimulus.

As we already pointed out, we don't know whether such adaptive mechanisms are sufficient to adapt to the natural environment where the light intensity has a much wider range and much stronger and long-ranged correlations than the standard stimuli used in a laboratory.

In this chapter, we will try to prove that they indeed do. The main hypothesis is that the two-step adapting strategy outlined above is sufficient not only to adapt the limited dynamic neuronal range to the wide range of the natural visual input, but also to fulfill the computational goal of removing most of the redundancy of natural time series.[1]

We will test this hypothesis by building a simple model that accounts for both mean and contrast adaptation, and by studying the statistical properties of its output taking the time series of natural images described in the previous chapter as input. The basic idea underlying the model is that neurons can filter out the predictable part of their input by predicting it from the history of the same input in the past. In this way, they only have to transmit the unpredictable part of the input, that has a much narrower

[1]The work presented in this chapter has been written in an article form to be submitted for publication (Buiatti and van Vreeswijk, 2002).

range of values and is therefore easier to code. This idea of *predictive coding* was first introduced by Srinivasan et al. (1982) as a strategy employed by the early visual system to encode a signal in a way that minimizes the effects of intrinsic noise. We use the same idea to simulate the two steps of adaptation mentioned above: linear predictive coding (the same strategy adopted by Srinivasan et al. (1982)) simulates adaptation to the mean light intensity, while a successive nonlinear coding strategy - variance normalisation, equivalent to the prediction of the variance from the signal's history - simulates adaptation to contrast. Unlike Srinivasan et al. (1982), in this study we ignore the effect of noise; the consequences of its unavoidable presence are discussed at the end of this chapter.

4.2 Adaptation to the mean light level

As we have seen in Chapter 2, the most pronounced adaptive response to a change in the mean light level is characterised by the so-called Weber-Fechner law: the retina produces approximately the same response for two visual displays that are related by a simple proportional scaling of all intensity values (Meister and Berry, 1999). In order to keep our model simple, we assume that this process fully describes the adaptation to the mean. We

model such process by assuming that the early visual system

(a) performs a logarithmic transformation of the light intensity $I(t)$, and

(b) transmits the difference between the current log-intensity, $x(t)$, and its best linear prediction estimated over its past values, $x_p(t)$, defined as

$$x_p(t) = \sum_{k=1}^{\tau_L} a(k)x(t-k), \tag{4.1}$$

where $a(k)$ is the optimal linear filter and τ_L its length. The transmitted signal is just the prediction error, $y(t)$, given by

$$y(t) = x(t) - \sum_{k=1}^{\tau_L} a(k)x(t-k). \tag{4.2}$$

We estimate the optimal filter, $a(k)$, by minimizing the average square prediction error, ϵ, on van Hateren's time series:

$$\epsilon = \langle y^2(t) \rangle, \tag{4.3}$$

where $\langle \cdots \rangle$ indicates temporal averaging. This is the least-square criterion for the linear regression of Eq. (4.2), and can be performed analytically, as

it is shown in the next section.

Given that the filter $a(k)$ is normalised, it is easy to verify that the response satisfies the Weber-Fechner law: If all the light intensity in the visual scene changes by a factor G, the new response $y_G(t)$ will be given by

$$\begin{aligned} y_G(t) &= log(G\cdot I(t)) - \sum_{k=1}^{\tau_L} a(k) log(G\cdot I(t-k)) \\ &= x(t) + log(G) - \sum_{k=1}^{\tau_L} a(k)x(t-k) - log(G)\cdot \sum_{k=1}^{\tau_L} a(k) \\ &= y(t). \end{aligned} \tag{4.4}$$

4.2.1 How to calculate the optimal filter

The analytic solution of the minimization of (4.3) is obtained by solving the set of equations for $a(k)$

$$\langle x(t)x(t-s)\rangle = \sum_{k=1}^{\tau_L} \langle x(t-k)x(t-s)\rangle a(k) \tag{4.5}$$

where $1 <= s <= \tau_L$. This is solved by a matrix inversion. We evaluated it numerically with a program that implements the Levinson-Durbin algorithm (Press et al., 1987). We accurately checked that such matrix inversion doesn't introduce any artifact, namely that the matrix is always full-rank (the condition number, defined as the ratio of largest to smallest singular value of the matrix, is always smaller than 10^4, whose reciprocal is far larger than the computer's floating point precision) and that the residual of the solution of Eq. (4.5) is negligible, being of the order of 10^{-15}. Moreover, other algorithms (LU, Gaussian elimination) give exactly the same result.

4.2.2 The filter's structure

The value of τ_L represents the time interval over which the neuron integrates the past signals to predict the new one. In other words, it represents the time scale over which the prediction is performed. We estimate τ_L by studying how it influences the error in (4.3) and the structure of the filter $a(k)$.

It is evident from Figure 4.1 that, while for low values of τ_L, an increase in τ_L significantly improves the prediction, the error reaches an almost sta-

tionary value for $\tau_L \geq 20$ time steps, corresponding to $\simeq 17$ msec.

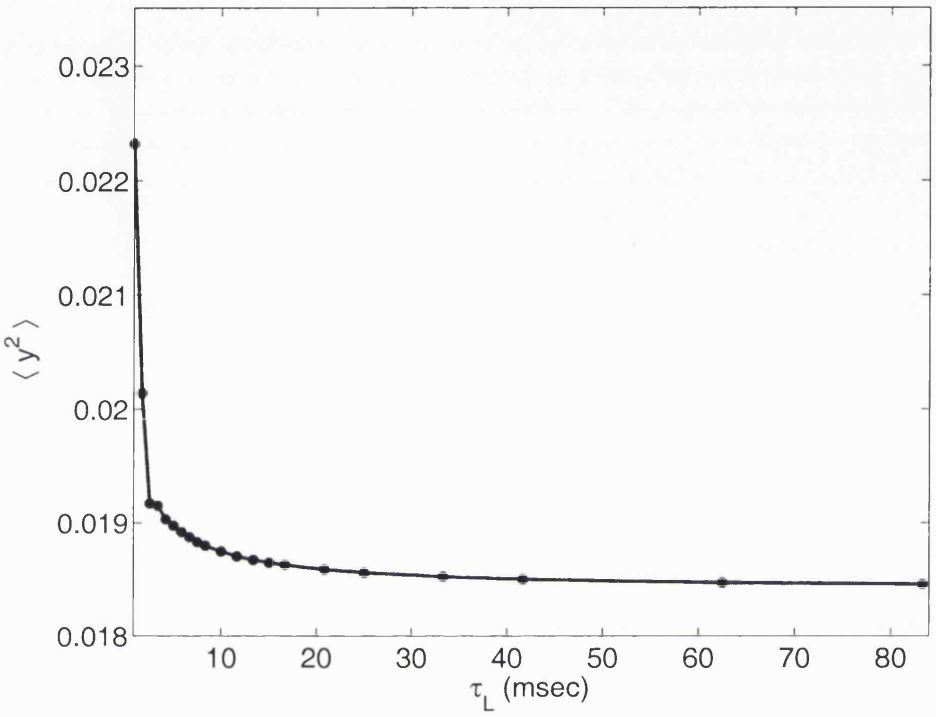


Figure 4.1: Prediction error $\langle y^2(t) \rangle$ versus τ_L. The average is performed on the whole first time series (45 minutes).

Figure 4.2 shows the optimal filter obtained by minimizing (4.3) with different values of τ_L. It is evident that the main structure of the filter $a(k)$ is given by its first components, while the ones with larger k quickly tend to zero. The apparently anomalous high value of the last component is an artifact due to the finite length of the filter. While for $\tau_L = 10$ time steps ($\simeq 8$ msec) this effect still alters significantly the filter, it quickly becomes

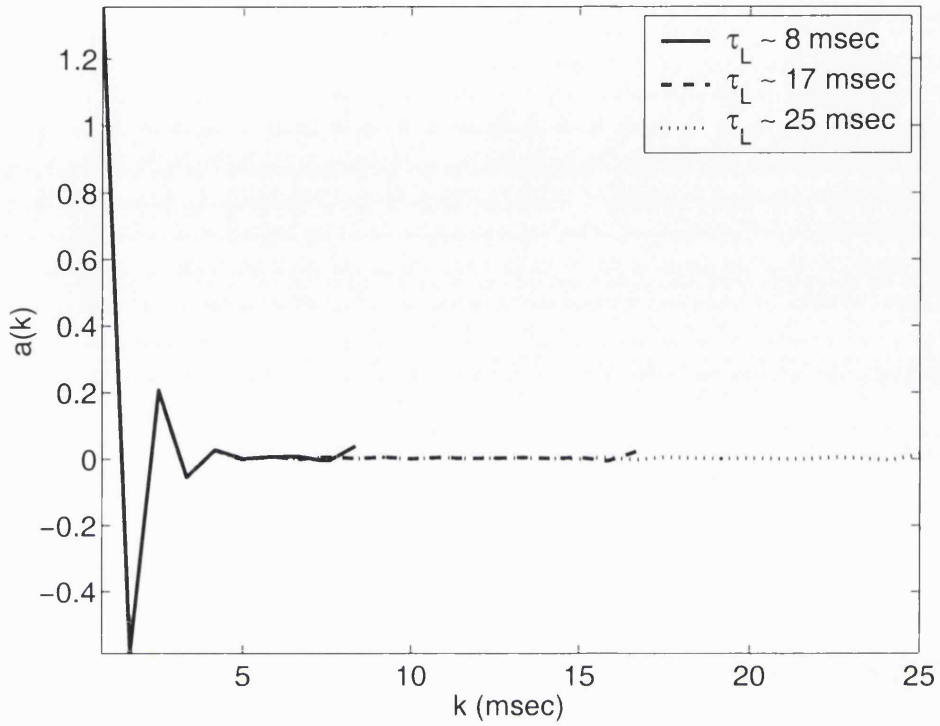


Figure 4.2: Linear filter $a(k)$ vs k for $\tau_L = 10, 20, 30$ time steps, respectively corresponding to roughly $8, 17, 25$ msec. First time series.

irrelevant for higher values of τ_L. In fact, the relevant components of the filters for $\tau_L = 20$ time steps ($\simeq$ 17 msec) and $\tau_L = 30$ time steps ($\simeq$ 25 msec) are almost indistinguishable.

Given these observations, we conclude that the linear prediction (4.1) can be performed almost at its best with an integration time as short as 20 time steps, corresponding to roughly 17 msec. This time scale is consistent

with the very short time scale within which adaptation to the mean occurs in the retina (Victor, 1999). Therefore, we set $\tau_L = 20$ time steps ($\simeq 17$ msec) for all the analyses carried out in the rest of the paper, always checking that our results continue to be valid for larger values of τ_L. Surprisingly, considering the difference in statistics between intensity series of as much as 45 minute, the same filter, $a(k)$, is obtained, except for estimation errors, using arbitrarily chosen stretches of as little as 1 second of data. The filter uncovers structures in the input that are *constant* in time.

It is worth observing that our procedure is the same one adopted by Srinivasan et al. (1982) when they consider the noise-free case. In spite of that, the optimal filter that we found has a longer time scale and a more complex structure. This difference is due to the different statistical properties of the input. They used an exponentially correlated signal, characterised by a finite correlation time: the information between the past and the present signal decays very quickly with time, and the best prediction of the present signal is made on the immediately preceding ones. We used van Hateren's natural time series that, as we showed in the previous chapter, are characterised by a very slow, power-law like decay of the correlations with time lag, i.e. a virtually infinite correlation time: in this case, it seems to make

sense to perform the prediction on a larger stretch of the signal's history.

4.2.3 Statistical properties of the response

Is this adaptation process effectively decorrelating the input? We answer to this question by analysing the statistical properties of the response $y(t)$:

a) Figure 4.3 shows that the error autocorrelation function $C_y(s)$, defined as

$$C_y(s) = \frac{\langle y(t)y(t+s)\rangle - \langle y(t)\rangle^2}{\langle y^2(t)\rangle - \langle y(t)\rangle^2}, \tag{4.6}$$

is around zero for almost all $s \neq 0$. The tiny bump that is visible for $s \simeq 17$ msec is due to the finite order of the linear filter; if we use $\tau_L = 30$ time steps ($\simeq 25$ msec), the bump at $s \simeq 17$ msec is replaced by an even smaller one around $s \simeq 25$ msec. In fact, in the limit $\tau_L = \infty$, linear correlations in the prediction error $y(t)$ completely disappear for any time lag s[2]. Thus, linear

[2]This result can be demonstrated analytically as follows:

$$\begin{aligned}
\langle y(t)y(t-s)\rangle &= \langle (x(t) - \sum_{k=1}^{\infty} a(k)x(t-k)) \cdot (x(t-s) - \sum_{j=1}^{\infty} a(j)x(t-j-s))\rangle \\
&= \langle x(t)x(t-s)\rangle - \sum_{k=1}^{\infty}\langle x(t-k)x(t-s)\rangle a(k) + \\
&- \sum_{j=1}^{\infty}[\langle x(t)x(t-j-s)\rangle - \sum_{k=1}^{\infty}\langle x(t-k)x(t-j-s)\rangle] \cdot a(j),
\end{aligned}$$

filtering is able to remove all the second order correlations of the input (also shown in Figure 4.3).

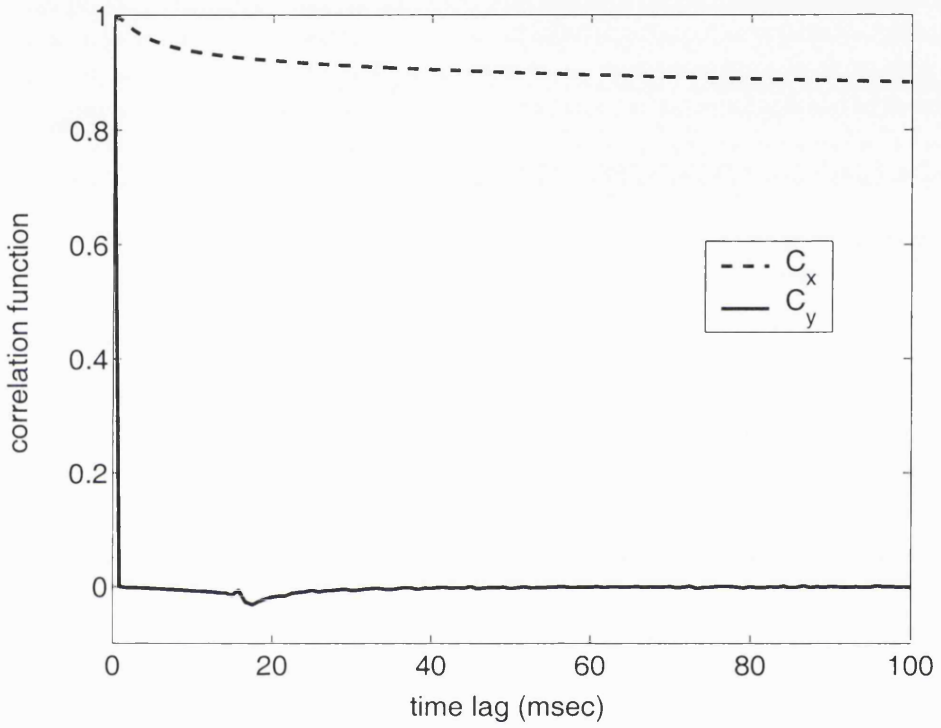


Figure 4.3: Linear correlation functions C_x of the log-intensity and C_y of the response from linear filtering, versus time lag (in msec). First time series.

$$\propto \quad \delta(s) \tag{4.7}$$

where the last equality comes from the fact that, since $\tau_L = \infty$, equation (4.5) holds for any $s > 0$, and from the assumption of invariance of the autocorrelation function to translations.

b) Figure 4.4 shows the quadratic correlation function $C_{yy}(s)$, given by

$$C_{yy}(s) = \frac{\langle y^2(t)y^2(t+s)\rangle - \langle y^2(t)\rangle^2}{\langle y^4(t)\rangle - \langle y^2(t)\rangle^2} \tag{4.8}$$

that indicates the correlations between the fluctuations around the mean. In this case, after an initial abrupt decrease, the quadratic correlation function decays very slowly for increasing time lag s: linear filtering is not sufficient to remove the higher-order redundancy of the input.

In order to understand how fourth order correlations influence the distribution of the output $y(t)$, we examine the probability distribution of the variable

$$\tilde{y}_T(t) = \frac{y(t) - \langle y(t)\rangle_T}{\sqrt{\langle (y(t) - \langle y(t)\rangle_T)^2\rangle_T}} \tag{4.9}$$

where $\langle \ldots \rangle_T$ indicates the average over a window from time $t - T$ to time $t+T$. The probability distribution $p(\tilde{y}_T(t))$ is obtained by sliding the window along the whole time series. The variable $\tilde{y}_T(t)$ indicates the statistical behaviour of the variable $y(t)$ within the time scale T. Figure 4.5 shows that for $T = 10$ time steps ($\simeq$ 8 msec), the distribution of $\tilde{y}_T(t)$ is very close to a Gaussian; this means that up to this very short time scale, linear filtering is sufficient to substantially decorrelate the signal, and code it in

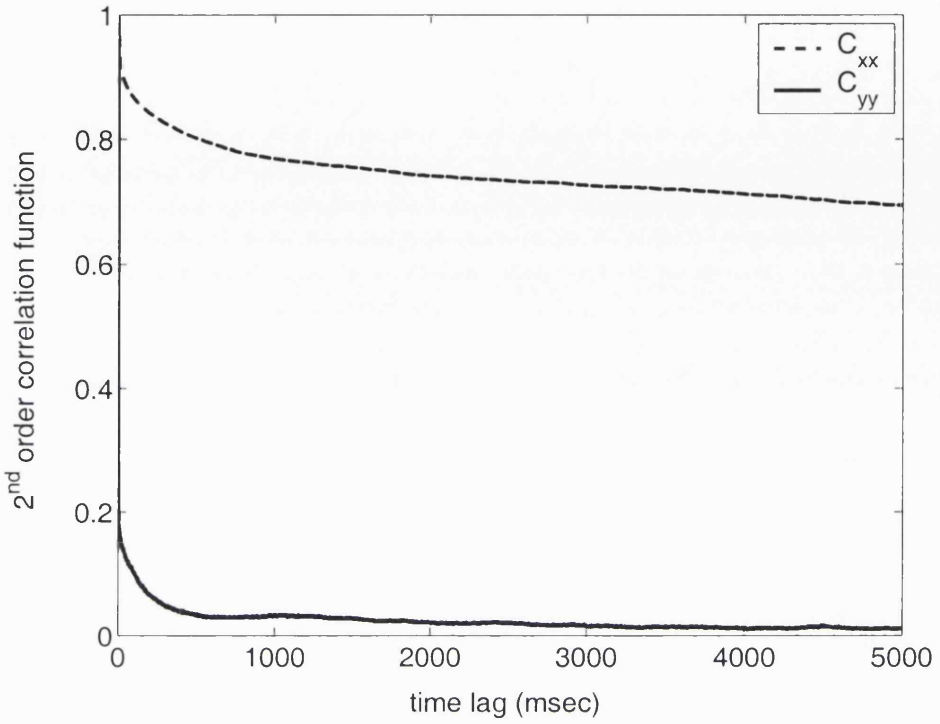


Figure 4.4: Quadratic correlation functions C_{xx} of the log-intensity and C_{yy} of the response from linear filtering, versus time lag (in msec). First time series.

a fairly compact distribution. But the same figure shows that already for $T = 100$ time steps ($\simeq 83$ msec), the distribution's peak becomes narrower, and the tails lengthen. This effect is caused by the strong fluctuations in the variance of the input that, as we showed above, linear filtering cannot completely decorrelate. A further increase in T causes the distribution to get closer and closer to $p(y)$, characterised by a very high central peak and very long tails. Such a sparse distribution is indeed not optimal to fully

exploit a limited range of sensitivity.

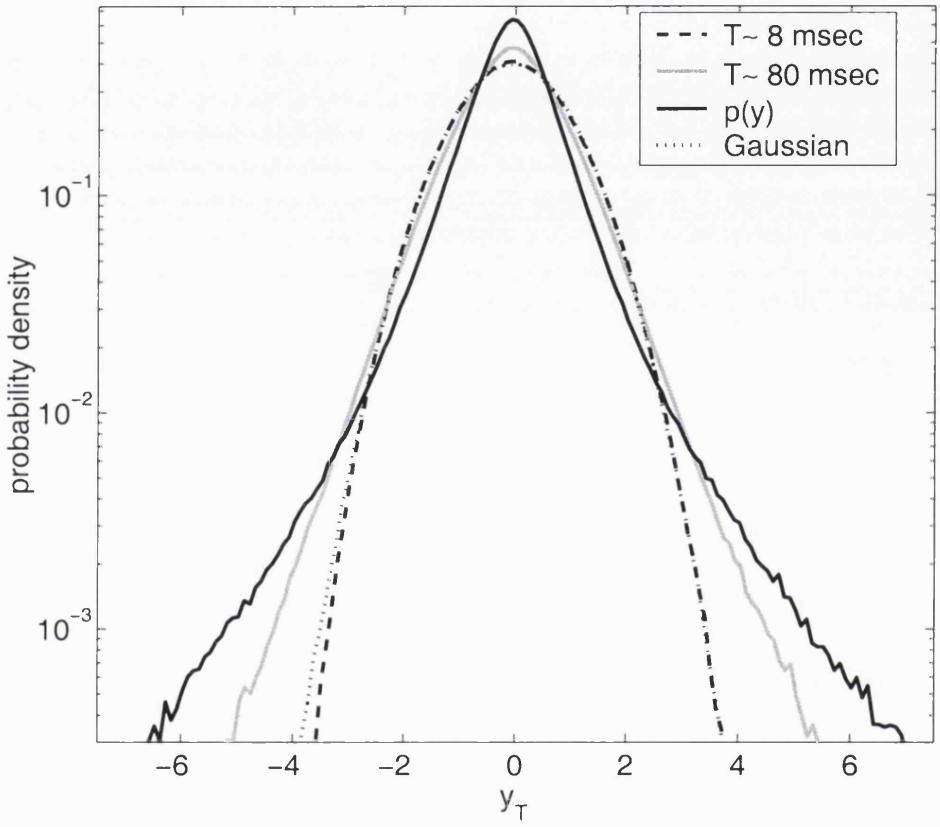


Figure 4.5: Histograms of the prediction error y and of the scaled variable $\tilde{y}_T$ for varying T, compared to a Gaussian. Abscissa in units of standard deviation. First time series.

4.3 Variance normalisation

As we have seen in Chapter 2, the early visual system seems to adapt to changes in temporal contrast in such a way that the form of the response

function depends only on the stimulus normalised by its local standard deviation, rather than on the stimulus itself (Brenner et al., 2000; Fairhall et al., 2000, 2001). Our aim is to model this mechanism in a very simple way, and to study whether it is sufficient to code the natural input in an efficient and compact way and to remove its higher order redundancy.

In this section, we will first describe our model. We will then set the value of its main parameter (the integration time of the normalisation) by maximising information transmission. We will finally show that it efficiently codes the widely variable range of the input into a compact output distribution in almost real time, and that it removes most of the redundancy of the natural time series.

4.3.1 Variance normalisation in time: the model

The model is directly inspired by the experimental results on contrast adaptation: the response, $r(t)$, is given by the filtered signal, $y(t)$, normalised by an estimate of the local standard deviation, $C(t)$:

$$r(t) = \frac{y(t)}{C(t)}. \tag{4.10}$$

The estimate $C(t)$ is evaluated by performing a weighted sum of the square of the filtered signal $y(t')$ at times $t' < t$:

$$C^2(t) = A \sum_{n=1}^{\infty} exp(-\frac{n}{\tau_N}) \cdot y^2(t-n). \tag{4.11}$$

The time constant τ_N corresponds to the integration time of the normalisation, and represents the time scale of variance adaptation. The choice of decreasing exponential weights comes from the idea that normalisation relies more on the recent values than on the later ones. However, this choice is not crucial, since a uniform normalising window (a finite set of equal weights) yields to very similar results. The constant A is chosen such that

$$\langle r^2 \rangle = 1. \tag{4.12}$$

This constraint corresponds to a biologically plausible constraint on the total power of the signal.

4.3.2 How to set the integration time of variance normalisation

In the language of information transmission, the requirement of matching the wide range of light intensities with the limited range of neuronal activity

corresponds to maximising the information transmission between x and r. In order to have maximum information transmission, we should maximize the mutual information between the stimulus and the response. Since our model is noise free, maximizing mutual information is equivalent to maximizing the entropy of the overall response. Since this calculation is practically impossible, we limit ourselves to setting the value of τ_N that maximizes the entropy of the single-valued probability distribution,

$$S_r(\tau_N) = -\sum_r p(r, \tau_N) \log(p(r, \tau_N)). \tag{4.13}$$

The evaluation of (4.13) on the natural time series shows that the value of τ_N for which the entropy is maximum is very small, and even though it varies from one time series to another, it mostly lies in the range between 10 and 20 time steps ($\simeq$ 8 to $\simeq$ 17 msec). Thus, we can set τ_N to the value that maximizes the entropy averaged over all time series. As shown in Figure 4.6, the averaged entropy is maximum for $\tau_N = 16$ time steps, corresponding to $\simeq$ 13 msec, and further on we will use this value in our simulations. We also checked that the entropy value for $\tau_N = 16$ time steps computed on every single time series is always very close to the maximum value.

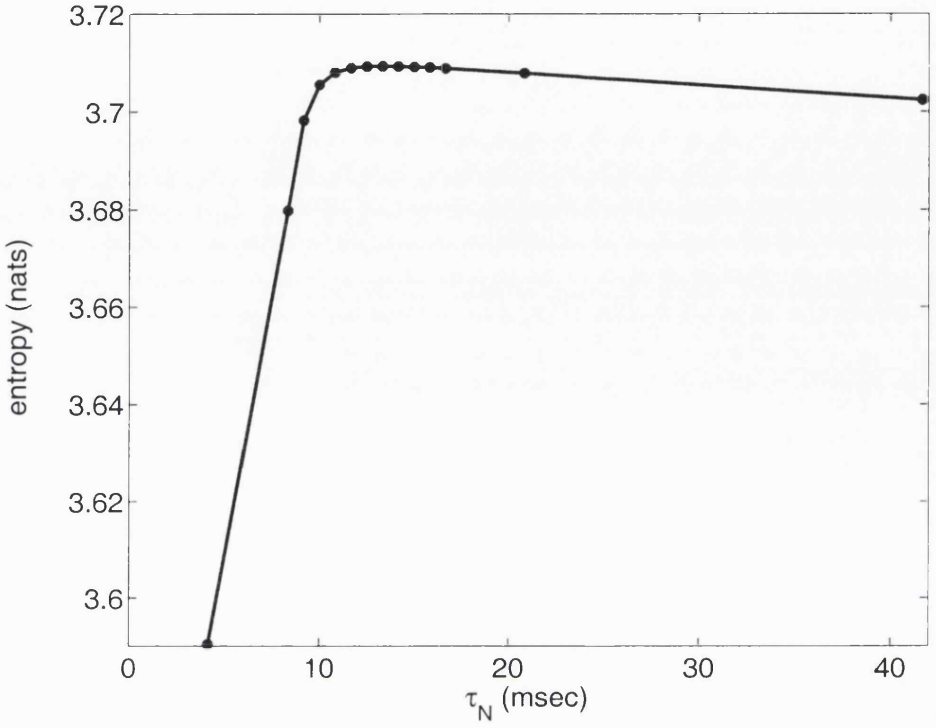


Figure 4.6: Entropy $S_r(\tau_N)$ vs τ_N (average over all time series).

Given the constraint (4.12), the distribution that maximizes the entropy is the Gaussian distribution. Is variance normalisation going in the same direction? In order to answer this question, we measure how the kurtosis $\kappa_r(t)$, defined as

$$\kappa_r(t) = \frac{\langle r^4(t) \rangle}{\langle r^2(t) \rangle^2} - 3, \tag{4.14}$$

varies with τ_N. In this case, results are more heterogeneous from one time series to another. However, for most of the time series, the value of τ_N for which the entropy is maximized is approximately the same for which the

kurtosis is closer to that of a Gaussian distribution, i.e. zero (this is shown in Figure 4.7 for the first time series). When we depart from that value, either decreasing or increasing, the kurtosis rapidly increases. Correspondingly, the tails of the distribution get longer and longer and the shape is more and more peaked, eventually becoming for high τ_N very similar, up to a multiplicative constant, to the distribution of the linearly filtered signal y.

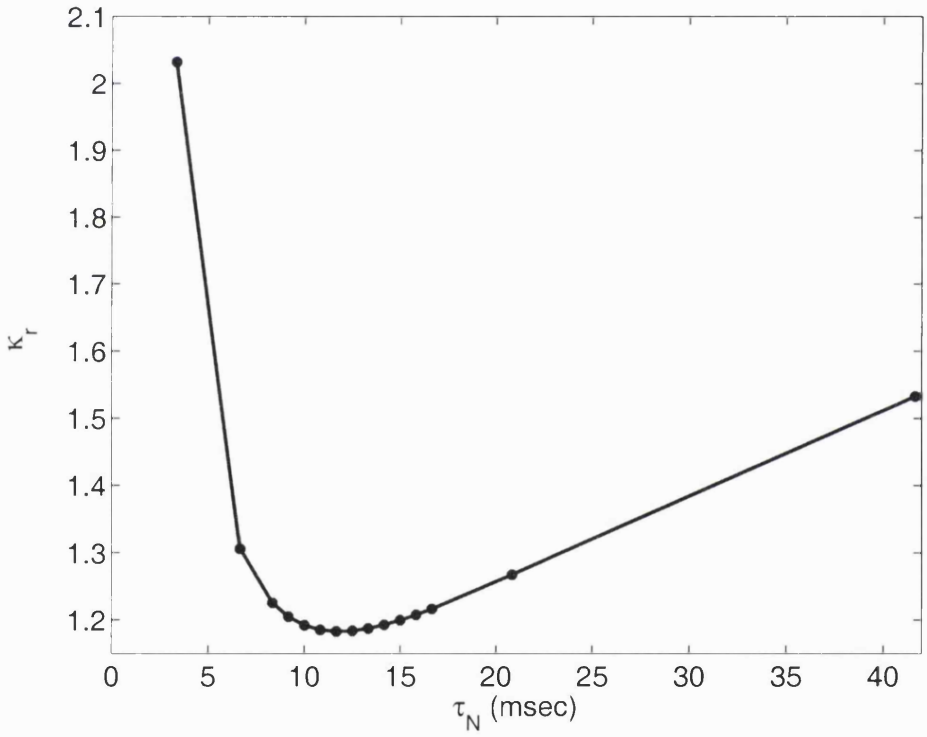


Figure 4.7: Kurtosis κ_r vs τ_N of the first time series. The minimum corresponds to $\tau_N \simeq 12$ msec, while $S_r(\tau_N)$ of the same time series reaches its maximum for $\tau_N \simeq 11$ msec.

4.3.3 Result 1: Dynamical adaptation of the response

Is variance normalisation sufficient to adapt the response to the widely varying natural stimulus in real time, as it is suggested from van Hateren (1997) experiment? Figure 4.8 shows the histograms of the natural input x and of the responses y and r for three segments of 1 minute each. As expected, the histograms of the natural input vary considerably in width and shape. The histograms of the response y all have a similar shape, characterised by a high peak and long tails; moreover, the width is highly variable. As discussed before, this is due to the inefficiency of the linear filtering to adapt to the wide variations of the local variance. On the contrary, the distributions of the variance normalised response almost overlap: variance normalisation efficiently adapts to the varying stimulus by exploiting at best the limited dynamical range of the response in a very short time. So, our model successfully reproduces the same kind of dynamical adaptation that occurs in the LMC's of the fly when stimulated with the same natural input, as shown by van Hateren (1997) in Figure 4.

In order to show the efficiency of the dynamical adaptation of r across different time scales, we show in Figure 4.9 the distributions of y and r computed on one second stretch and on a whole time series. While the distribu-

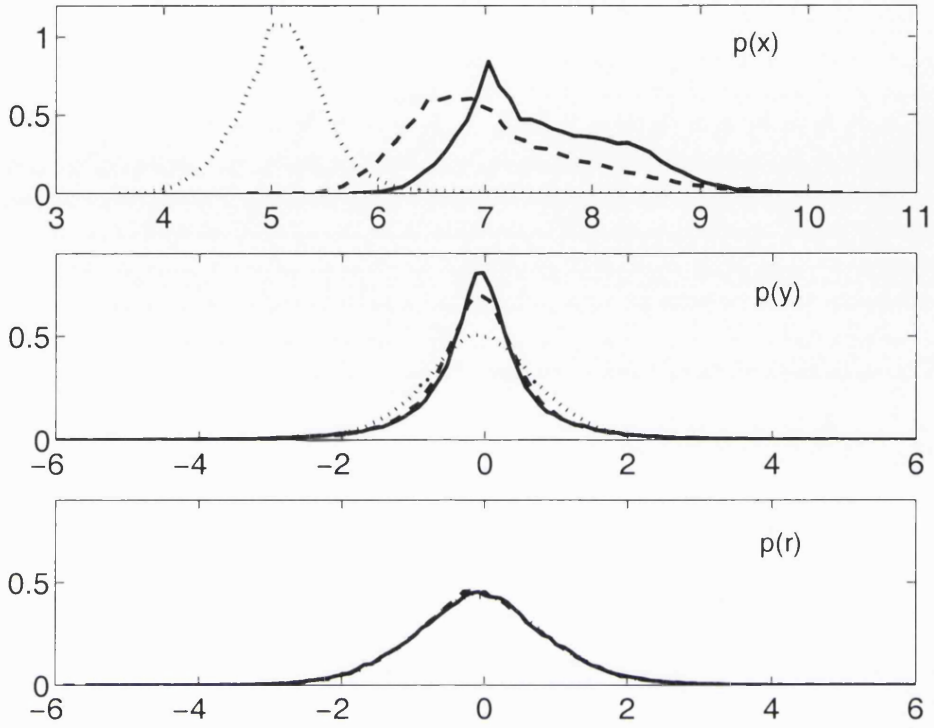


Figure 4.8: Probability density of x, y and r for three segments of 1 minute extracted from the first time series. Segments correspond to minute 15 (solid lines), 33 (dashed lines) and 43 (dotted lines) of the first time series. The abscissa is in log-intensity units for x, and in units of standard deviation for y and r.

tions for one second are quite similar and quite compact, the distributions of the whole time series are strikingly different: while $p(y)$ is long-tailed and highly-peaked, $p(r)$ is very similar to a Gaussian in the central part, and slowly departs from a Gaussian only in the extreme part of the tails (note that the plots are in semi-logarithmic scale). Dynamical adaptation occurs up to very long time scales.

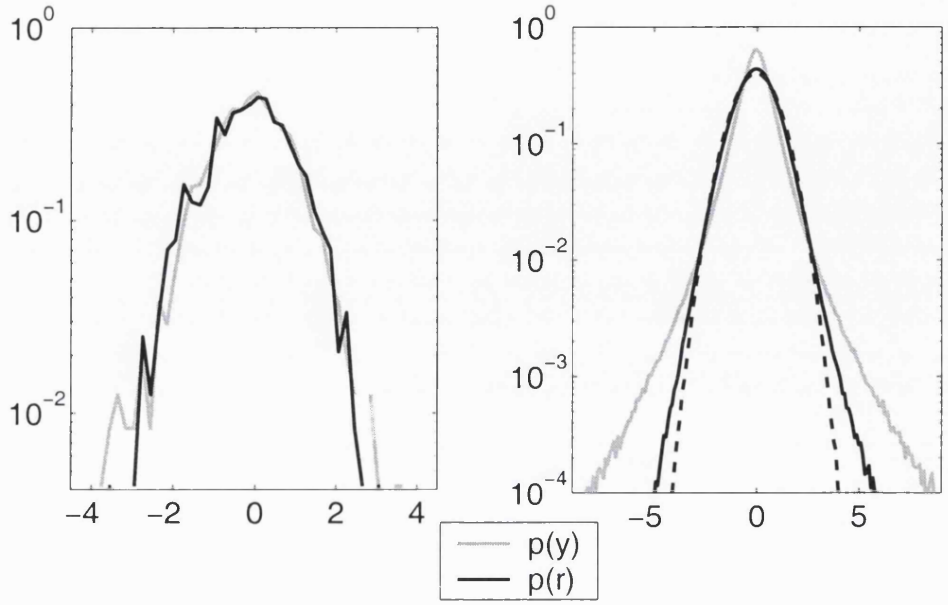


Figure 4.9: Semi-logarithmic plot of the probability densities of y and r for $T = 1$ second (left) and $T = 45$ minutes (right). The abscissa is in units of standard deviation. The dashed curve in the right figure is a Gaussian distribution. First time series.

4.3.4 Result 2: High-order redundancy reduction

The second important result concerns the efficiency of variance normalisation in removing the high-order redundancy still present in the linearly filtered response y. We begin our analysis by showing the quadratic correlation function, defined as in the previous chapter:

$$C_{rr}(s) = \frac{\langle r^2(t)r^2(t+s)\rangle - \langle r^2(t)\rangle^2}{\langle r^4(t)\rangle - \langle r^2(t)\rangle^2}. \tag{4.15}$$

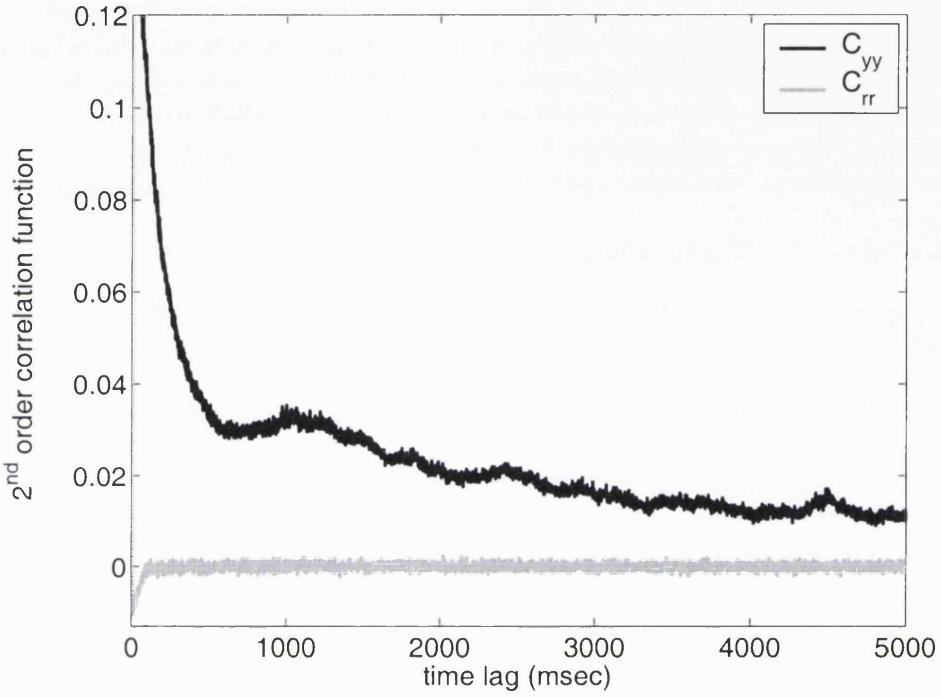


Figure 4.10: Quadratic correlation functions C_{yy} of the response from linear filtering, and $C_{rr}(s)$ of the response from variance normalisation versus time lag (in msec). First time series. Note that both correlation functions are normalised to one at zero time lag.

It is evident from Figure 4.10 that, while the linearly filtered signal y still has slow-decaying 4^{th} order correlations, the fluctuations of r are almost uncorrelated. This is a very important result: despite the very long correlation time of the fluctuations around the mean in the input, variance normalisation completely decorrelates them, and does it by integrating over

a very short time scale.

Mutual information

In order to have a more general measure of the pair-wise redundancy, we analyse the mutual information of the signal at a single time with respect to the same signal s time steps before, using the same definition of the previous chapter:

$$I(x,s) = \sum p(x(t), x(t-s)) \log \frac{p(x(t), x(t-s))}{p(x(t)) \cdot p(x(t-s))}, \tag{4.16}$$

and similarly for y and r. As in the previous chapter, the distributions are evaluated on every time series (3240000 data points) by binning the data set in 100 bins of the same size, and we correct for the finite sample effect by subtracting from the calculated mutual information the mutual information obtained after randomly reshuffling the data. It is worth observing that the finite-size bias for the variables y and r is even smaller (between 10^{-3} and $10^{-4} nats$) than the one for x.

Figure 4.11 shows the resulting unbiased mutual information I_u for the three variables. As shown before, the unbiased mutual information of the

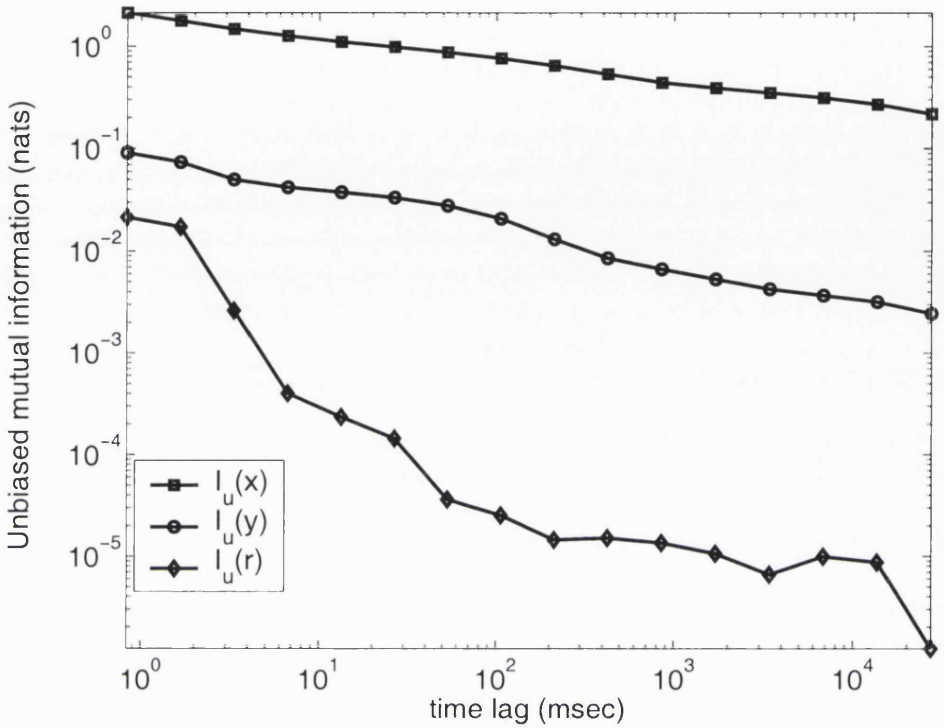


Figure 4.11: Log-log plot of the unbiased mutual information vs time lag, averaged over all time series.

original signal, $I_u(x, s)$, is very high, and decays as a power law. The unbiased mutual information of the linearly filtered signal, $I_u(y, s)$, is much lower (i.e. the mutual information is much closer to that of the randomly reshuffled data), but it still decays very slowly, scaling again approximately as a power law: long-range correlations are still there. However, the unbiased mutual information of the normalised signal, $I_u(r, s)$, collapses to values very close to zero in a very short time interval: variance normalisation decorrelates

the pair-wise redundancy at any order, getting rid of virtually all long-range correlations.

Comparison among the distributions

Finally, we measured the distance between the distributions estimated from different time series for the variables x, y, as well as r, to have an estimate of their redundancy at any order (as explained in detail in the previous chapter).

We measured the distance, D, between the distribution using the same form introduced in Chapter 3:

$$D_{a,b} = \sum_{k=1}^{K} \left[\rho_a(k) - \rho_b(k)\right]^2 , \tag{4.17}$$

where ρ_a and ρ_b indicate the estimated distribution for the a^{th} and b^{th} time series respectively. The variables x, y and r were binned using the same method described in Chapter 3. Figure 4.12 shows $D_{a,b}$, averaged over all 66 pairs of time series, plotted against T, for all three variables. Also plotted is the expected value of $D_{a,b}$, $\overline{D_{a,b}}$, for two time series, under the assumption that each data point is drawn independently from the equilibrium distribution. In appendix B, it is shown that, for independently drawn data, $\overline{D}$

satisfies $\overline{D} = 2(K-1)/(KT)$. As for x, the figure shows that the agreement between $D(rnd(r))$ and the analytical prediction $\overline{D_{a,b}}$ is very good.

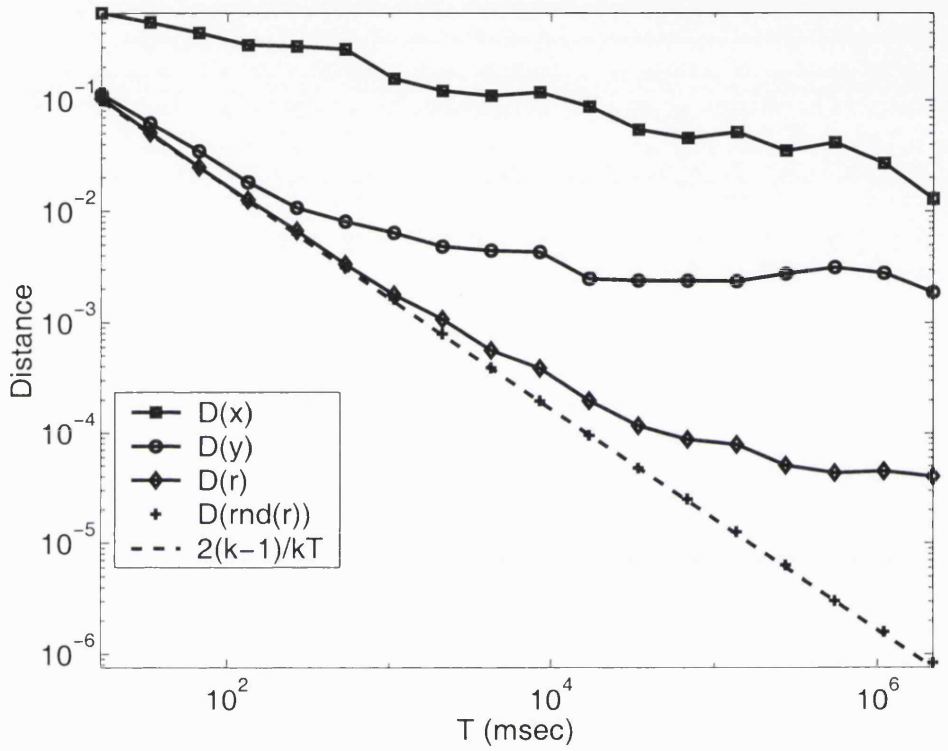


Figure 4.12: Log-log plot of the distance between distributions for data x, y, r, randomly reshuffled r ($rnd(r)$) for increasing length of the segment from which they were estimated, and expected distance for independently drawn data.

As we have seen in the previous chapter, the long-range correlation structure in the statistics of x causes the distance between the estimates of log-intensity distribution to fall off more slowly than $1/T$.

The distance between the estimates for the y distribution are smaller, and for T less than 100 msec it falls of roughly as $1/T$; for large T however, it falls of more slowly. This is a signature for the fact that the linear filtering is able to approximately decorrelate the input only for short stretches. Since for longer segments the long-range, higher order correlations affect the statistics significantly, the distance between estimates of the distribution of y based on longer series are much larger than expected from the assumption of independence.

The distance of the estimates of the distribution of r follows the theoretical distance much more closely. For up to 1 second there is no significant difference between these two. Only for longer time series do the residual correlations contribute sufficiently to increase the distance between the estimates markedly from that between estimates based on independently drawn samples. Still, even for the largest T shown here, the distance between estimates of r are more than an order of magnitude smaller than those of the estimates of y, showing that the variance normalisation has eliminated almost all residual correlation in y.

4.4 Discussion

The main result of the work presented in this chapter is that our initial hypothesis is substantially confirmed. While linear filtering decorrelates second-order correlations, it cannot adapt to fluctuations in variance. Therefore, its distribution results to be long-tailed and highly-peaked. This effect, as it has been shown by Baddeley (1996), can be simply caused by the variability of the local intensity variance. On the other hand, the simple nonlinear transformation given by variance normalisation yields surprising results:

1) The variance normalised response dynamically adapts to the widely varying natural input by coding it into a compact, almost Gaussian distribution in almost real time, reproducing the adaptation of the large monopolar cells of the fly stimulated with the same signal (van Hateren, 1997);

2) Variance normalisation effectively removes not only the redundancy given by the correlations of the fluctuations around the mean, but also almost all the temporal redundancy of natural images. This is shown by the vanishing mutual information, that gives an estimate of the overall pair-wise correlations, and by the distance measure among distributions, that gives a

general estimate of the overall redundancy at any order.

The other important result concerns the estimate of the time scales of the two adaptations given by the optimisation of the two steps. Despite the long-range correlations in the natural input, both optimal integration times are very short, being in the range between 10 and 20 msec. Despite the simplicity of the model, both estimates appear to be very consistent with the time scales of adaptation measured in the experiments. In particular, the optimal time scale of variance normalisation is compatible with the fast one identified in (Fairhall et al., 2000, 2001).

In order to verify that this is the only time scale involved in this process, we performed a second variance normalisation on the signal r. Simulations show that a second normalisation has substantially no effect on the statistics of r, no matter what the integration time is. This result reinforces the idea that the slow adaptation of the rate to changes in input statistics identified in (Smirnakis et al., 1997) and (Fairhall et al., 2000, 2001) has little to do with adaptation to the rapid, unpredictable variations in the input; rather, it is more likely to act as an independent process, transmitting information about the slow modulations of the long-range correlated, predictable part of

the signal (see (Fairhall et al., 2001) for a detailed discussion of this aspect).

Finally, there are a few issues that are worth to be discussed, and that potentially lead to further interesting investigations:

1) We showed that variance normalisation produces an almost Gaussian output. We claimed that this is consistent with an optimal coding strategy because the Gaussian distribution is optimal if we assume a constraint on the overall power of the neural activity. This assumpion seems to be plausible, and the experimental results of van Hateren (1997) on the Gaussian response of large monopolar cells seem to confirm it. However, other constraints can be taken into account, yielding to very different optimal response distributions. For example, Levy and Baxter (1996) argue that coding schemes with the largest representational capacity are not, in general, optimal when energy expenditures are taken into account, while a plausible constraint could be set on the average firing rate, giving rise to exponential distributions. Baddeley et al. (1997) show that such coding strategies can be found in the primary visual cortex of anaesthetised cat and in inferior temporal areas of awake monkey. It seems to be hard to identify a universal constraint that models the biological ones at best. The idea is that the most relevant con-

straint depends on the area and the function of the neurons we are trying to model. In this sense, our model seems to fit fairly well what happens in the early stages of visual processing, but should be modified if higher areas of the brain are taken into account.

2) Given its simplicity, the model does not account for some basic features of neuronal signals. As pointed out in the beginning of this chapter, noise is a fundamental factor to be accounted for. Srinivasan et al. (1982)) showed that noise significantly alters the coding mechanism by lengthening the neural temporal response. We expect that noise would lengthen and smoothen the linear filter of our model, worsening the efficiency of the adaptation mechanism that we proposed. Further investigation is indeed required in this direction.

3) The mechanism that we proposed to simulate the variance normalisation shown by Fairhall et al. (2001) is to let the amplitude of the linear filter vary in time with the local variance. However, Smirnakis et al. (1997) showed that the temporal response changes both in amplitude and time scale when the variance changes. Another future investigation could then be model the adaptation to the local statistics by also letting the integration time of the

response vary in time. The combination of the two adapting mechanisms (varying amplitude and varying integration time) could compensate the effect of noise to give results very similar to what we obtained with our model.

4) Variance normalisation has already been suggested as an adaptation strategy implemented in the primary cortex to normalise the signals about the light intensity coming from neighbouring spatial locations signalled by neighbouring neurons (Carandini et al., 1997; Simoncelli and Schwartz, 1999; Schwartz and Simoncelli, 2001). The final task is to build a model that integrates temporal adaptation with spatial coding. We believe that this work is an important starting point to investigate in this direction.

Chapter 5

Redundancy reduction in financial time series

5.1 Introduction

The previous chapter showed that variance normalisation is a very good way to model temporal adaptation in the early visual system. Besides the biological result, it also showed that we ended up building a very simple and efficient method to separate the predictable part from the unpredictable part of a signal. In order to verify the computational power of our model, we apply it to data that is deliberately very different from natural images - financial time series. We will first present a brief introduction to the

statistical properties of the financial data. We will then apply our method to the analysis of a financial index of the Italian stock exchange. We will show that, similarly to what has been obtained with natural time series, our method efficiently separates the random component from the correlated component of the financial data, suggesting that it can be considered a useful tool for the statistical analysis of any time series.

5.2 The statistical properties of financial time series

Since the 1950s, the analysis and modeling of financial markets have become an important research area of economics and financial mathematics. The researches pursued have been very successful, and nowadays a robust theoretical framework characterizes these disciplines. More recently, a group of physicists became interested in the analysis and modeling of financial markets by using tools and paradigms of their own discipline. Here we summarise the main results of their investigations.

Despite the huge variety of the financial data, there are some general statistical properties that are surprisingly ubiquitous. In any financial mar-

ket, the autocorrelation function of returns[1] is a monotonic decreasing function with a very short correlation time. High-frequency data analyses have shown that correlation times can be as short as a few minutes in highly traded stocks or indices (Mantegna and Stanley, 1996; Liu et al., 1999). A fast decaying autocorrelation function is also observed in the empirical analysis of data recorded transaction by transaction. By using as time index the number of transactions emanating from a selected origin, a time memory as short as a few transactions has been detected in the dynamics of most traded stocks of the Budapest emerging financial market (Palagyi and Mantegna, 1999).

The short-range memory between returns is directly related to the necessity of absence of continuous arbitrage opportunities in efficient financial markets. In other words, if correlations were present between returns (and then between price changes), this would allow one to devise trading strategies that would provide a net gain continuously and without risk. The continuous search for and the exploitation of arbitrage opportunities from traders focused on this kind of activity drastically reduce the redundancy in the time series of price changes. Another mechanism that lowers the

[1]Returns are defined as the logarithm of price changes. See also Eq. (5.1).

redundancy of stock price time series is related to the presence of the so-called "noise traders". With their action, noise traders add into the time series of stock price information which is unrelated to the economic information, decreasing the degree of redundancy of the price changes time series.

It is worth pointing out that not all the economic information present in stock price time series disappears due to to these mechanisms. Indeed the redundancy that needs to be eliminated concerns only price changes and not any of its nonlinear functions. The absence of time correlations between returns does not mean that returns are identically distributed over time. In fact, different authors have observed that nonlinear functions of return such as the absolute value or the square are correlated over a time scale much longer than a trading day. Moreover, the functional form of this correlation seems to be power-law up to at least 20 trading days approximately (Liu et al., 1997, 1999; Raberto et al., 1999).

5.3 Redundancy reduction on the FIB 30 index

The financial data that are the object of our analysis consist of time series of a financial index of the Italian stock exchange, the FIB 30. Technically, it indicates the value of the *future* on the index MIB 30, an index linked to

the 30 most important titles traded in the Milan stock exchange. The time series contain the value of the index at every minute for 197 consecutive working days during the year 1999. Since a working day consists of 496 minutes, the overall length of the time series is of 97712 values.

As it is common in the literature (Liu et al., 1997; Mantegna et al., 1999), rather than studying the index $z(t)$, we analyse the statistical properties of the logarithmic increments $g(t)$ of the index (usually called returns),

$$g(t) = \ln z(t + \Delta t) - \ln z(t), \tag{5.1}$$

where Δt is the time lag and time t is a discrete variable taking the values $t_k = k\Delta$, where k is a positive integer and $\Delta = 1$ minute is the time interval between two consecutive data recordings. $g(t)$ can be seen as the relative price change $\Delta g/g$ in the limit $\Delta t \to 0$. Here we set $\Delta t = 1$ minute; we accurately checked that we obtain similar results for other choices of Δt.

Over the day, the market activity shows a strong "U-shape" dependence with high activity in the morning and in the afternoon, and much lower activity over noon. Since we are interested in the long-range redundancy

rather than in this specific intra-day pattern of the market activity, we analyse the normalised function

$$G(t) = g(t)/A(t), \tag{5.2}$$

where $A(t)$ is the mean value of $|g(t)|$ at the same time of the day averaged over all days of the data set.

5.3.1 Statistical analysis of the financial time series

Figure 5.1 shows the autocorrelation function $C_G(s)$, defined by substituting $x(t)$ with $G(t)$ in Eq. (3.7). As expected from the previous results described in the introduction, it is very close to a δ function: returns are linearly decorrelated already within two minutes of transactions.

Figures 5.2 and 5.3 show the square correlation function $C_{GG}(s)$, defined by substituting $x^2(t)$ with $G^2(t)$ in Eq. (3.8), and the autocorrelation function of the absolute value of $G(t)$, $C_{|G|}(s)$, defined by substituting $x(t)$ with $|G(t)|$ in Eq. (3.7). As expected, they are both long-range correlated over at least two days of transactions. In fact, Figures 5.2 and 5.3 show that they can be fitted surprisingly well by power-law functions with a unique exponent. This is not always the case (see for example (Liu et al., 1997)).

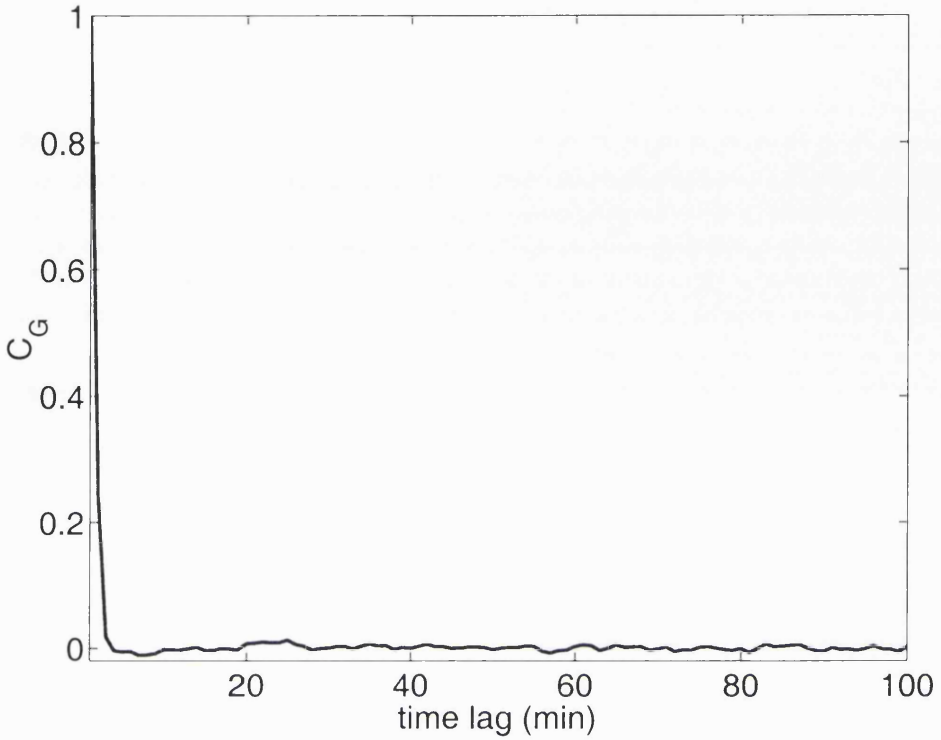


Figure 5.1: Autocorrelation function $C_G(s)$, versus time lag s in minutes.

5.3.2 Redundancy reduction through variance normalisation

Linear filtering and variance normalisation

We compute the variable indicating the residual from the predicted mean, $y(t)$, by substituting $x(t)$ with $G(t)$ in Eq. (4.1), where the linear filter $a(k)$ is estimated by minimising the average square prediction error over the entire FIB 30 time series. Following the same procedure used to set the

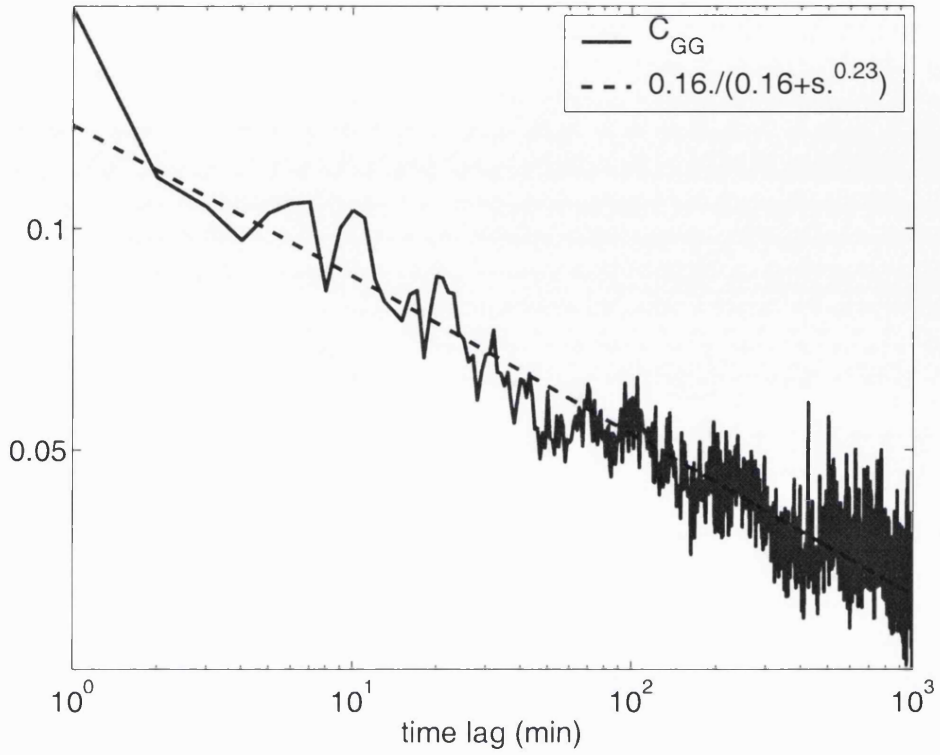


Figure 5.2: Quadratic correlation function $C_{GG}(s)$ versus time lag s in minutes.

integration time of the linear filter in the previous chapter, we estimate that $\tau_L = 20$ minutes allows a linear prediction very close to the optimal one, and we set $\tau_L = 20$ minutes in all the simulations presented in this chapter.

Analogously, we compute the normalised variable $r(t)$ as in Eq. (4.10). Following the method described in the previous chapter, we set τ_N to the value that maximizes the entropy $S_r(\tau_N)$ defined as in Eq. (4.13). As it is

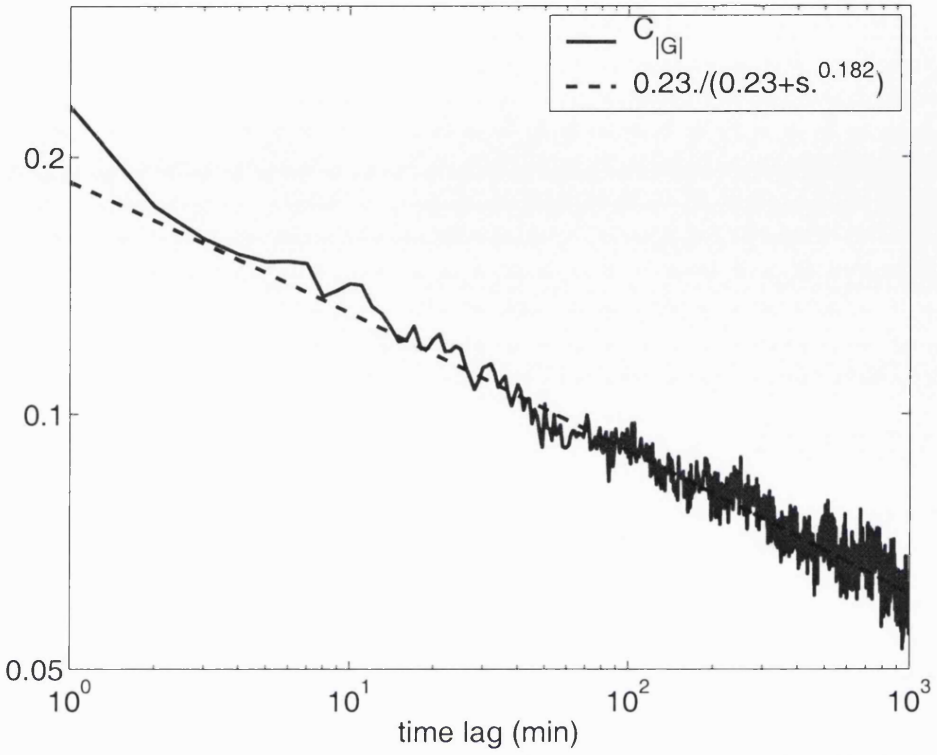


Figure 5.3: Correlation function of the absolute value of $G(t)$, $C_{|G|}(s)$ versus time lag s in minutes.

evident from Figure 5.4, the entropy maximum is very close to the kurtosis minimum: in fact, the value $\tau_N = 39$ minutes that maximizes the entropy is exactly the same value that minimizes the kurtosis $\kappa_r(\tau_N)$. Thus, we compute $r(t)$ by setting $\tau_N = 39$ minutes.

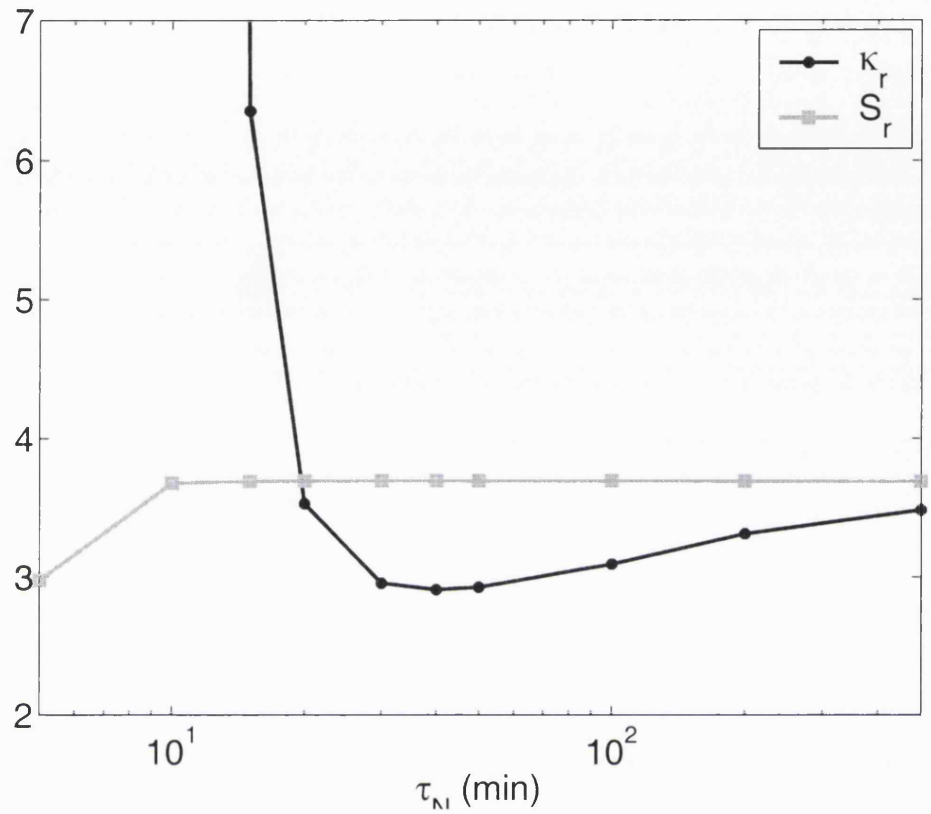


Figure 5.4: Kurtosis $\kappa_r(\tau_N)$ and entropy $S_r(\tau_N)$ versus integration time of variance normalisation τ_N.

Results

Since $G(t)$ is already linearly decorrelated, we expect that linear filtering doesn't substantially alter the original time series. We expect that redundancy will be removed only by variance normalisation.

Figure 5.5 shows that both linear filtering and variance normalisation

remove the already tiny linear correlations present in the variable $G(t)$.

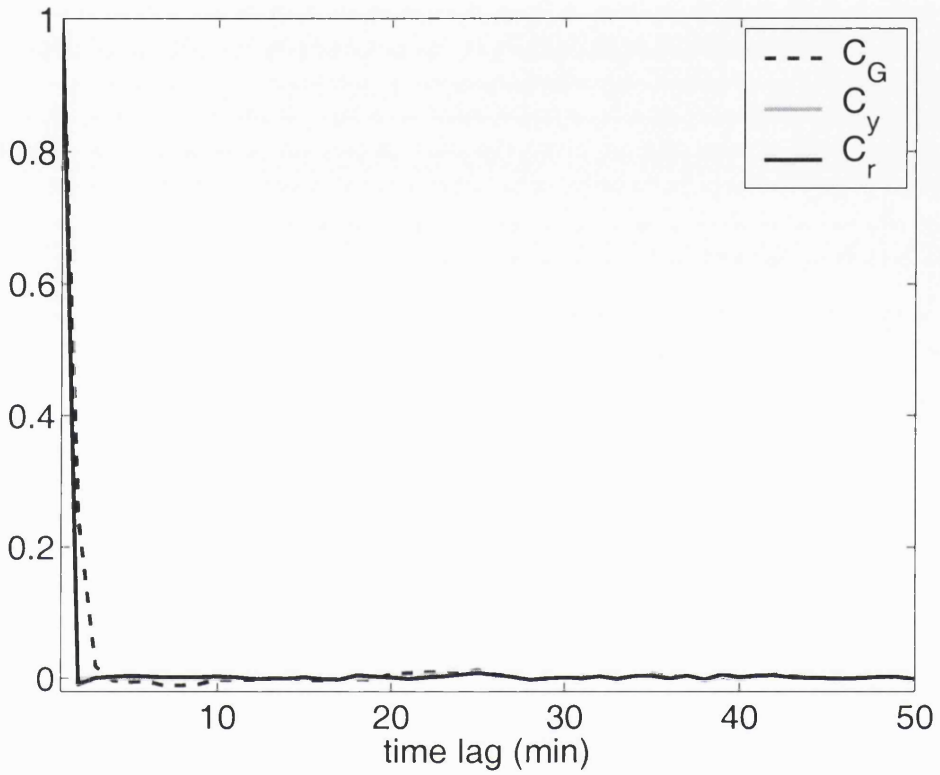


Figure 5.5: Autocorrelation function $C_G(s)$, $C_y(s)$ and $C_r(s)$ versus time lag s in minutes.

Figure 5.6 clearly shows that the redundancy present in the data has a very small second-order component: the linearly filtered variable $y(t)$ has almost the same amount of quadratic correlations as the original data $G(t)$. On the other hand, quadratic correlations in $r(t)$ fade away after a few minutes: variance normalisation efficiently removes them.

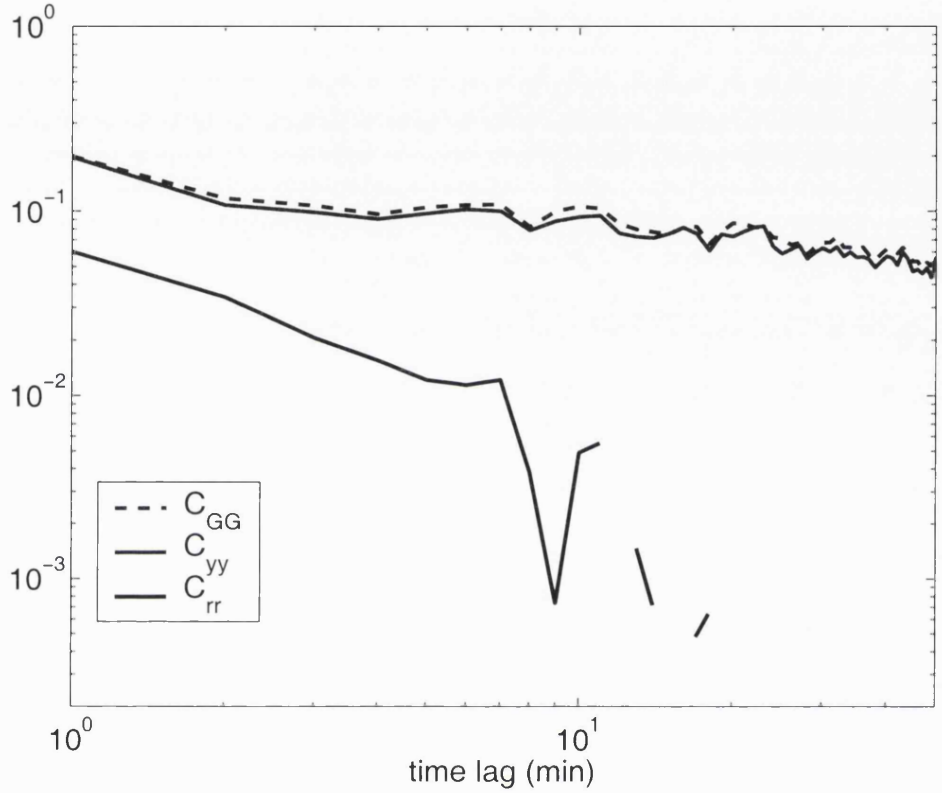


Figure 5.6: Quadratic correlation function $C_{GG}(s)$, $C_{yy}(s)$ and $C_{rr}(s)$ versus time lag s in minutes.

Figure 5.7 shows the unbiased mutual information as defined in Eq. (3.9) and corrected for the finite sampling size as in Chapter 3. Similarly to what happens for van Hateren's time series, $I_u(G)$ decays very slowly, approximately as a power law. $I_u(y)$ is just a bit smaller than $I_u(G)$, and follows the same approximate power-law decay: linear filtering hardly removes any redundancy. Instead, $I_u(r)$ decays very quickly to negligible values, so low

that they are indistinguishable from those computed from the reshuffled data (in fact, some values are not represented in the figure because they are negative): variance normalisation removes almost all the pair-wise redundancy.

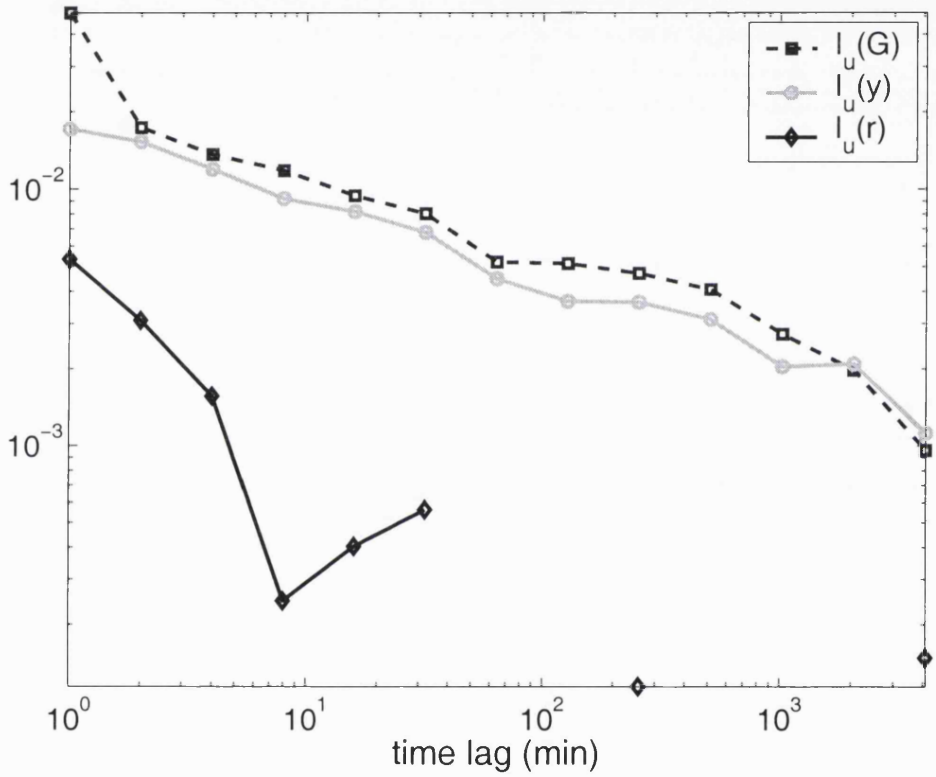


Figure 5.7: Unbiased mutual information $I_u(G)$, $I_u(y)$ and $I_u(r)$ versus time lag s in minutes.

5.4 Discussion

In order to prove the computational power of our normalisation method, we deliberately chose a data set whose origin is very different from that of time series of light intensities. The main feature shared by the two data sets is that they are highly correlated. The extended correlations present in the financial time series were a good test for our simple redundancy reduction scheme. Our method proved that even long-range power-law correlations persisting over very long periods of time can be removed by a normalisation based on a "memory" of only approximately 40 minutes of data. In other words, despite the long-range correlations, the magnitude of the price change can be significantly predicted by using as little as 40 minutes of its past history. Though further investigation is needed to generalise this variance normalisation method, we hope that this very simple result can be of some use in the analysis of time series of any nature.

Chapter 6

Conclusions

The main aim of the work presented in this thesis was to investigate the key mechanisms underlying adaptation of the early visual system to the temporal statistics of its input. The hypothesis that we tried to prove is that the coding strategy based on variance normalisation that seems to underlie the later stages of adaptation (Fairhall et al., 2000, 2001) is sufficient to overcome two major problems that the visual system has to face:

1) Fitting the wide range of natural intensities into the limited dynamic range of neuronal activity almost in real time;

2) Separating the unpredictable part from the redundant part of the

natural input by removing almost all its correlations.

By modeling variance normalisation with a very simple coding scheme, we were able to prove that both results are achieved, even when the input is widely varying and highly redundant as natural time series are, moreover reproducing some experimental findings on the visual system of the fly (van Hateren, 1997). Moreover, we proved that, despite the long-range correlations of the input, the "memory" needed for an optimal normalisation is very short. This is consistent with the very short adaptation time scales found in the literature (Shapley and Victor, 1978; Fairhall et al., 2000, 2001).

As pointed out in the discussion of Chapter 4, these results stimulate further investigations in many directions. In order to represent more accurately the neuronal activity, the model should account for the generation of spikes, noise should be included, and a time-varying integration time of the filter could be taken into account. Moreover, analogous results on spatial variance normalisation in the primary cortex (Simoncelli and Schwartz, 1999; Schwartz and Simoncelli, 2001) suggest that such a strategy is likely to be widely implemented in the brain; a next step could be to investigate the possible combination of spatial and temporal adaptation.

Nevertheless, the importance of this result seems to lie in the simplicity of the model: variance normalisation alone can remove the redundancy of highly correlated data. In other words, the redundant part of highly correlated time series can be efficiently represented by estimating the mean and the variance over a very short history. This result is confirmed by training the model on time series of very different origin - financial time series - suggesting that time series of any nature can be analysed in the same way.

Acknowledgements

I am very grateful to Carl van Vreeswijk who gave me the chance, the tools and the confidence to undertake this work. Most of the work presented in this thesis is the result of the close and friendly collaboration with him.

I want to warmly thank Geoff Hinton for having offered me an accurate supervision of this thesis. His comments and critics have been invaluable for writing this work.

I am also very grateful to Manuela Piazza because her clever comments and witty questions have helped me to reinforce the main idea underlying this work, and present it in a clear and unambiguous way.

Finally, I wish to thank the two referees for the accurate and helpful

work of comment and revision from which I believe the thesis has benefited a lot.

Appendix A: The input/output relation

This appendix describes the method as illustrated by Fairhall et al. (2001) to compute the input/output relation between the stimulus and the neuronal response. The idea is to identify features of the stimulus that modulate the probability of occurrence of individual spikes, $P(spike|stimulus)$; they do not consider patterns of spikes, although the same methods can be easily generalised. The space of stimulus histories of length $\sim$ 100 msec, discretised at 2 msec (as the stimulus was in the experiment of Fairhall and collaborators), leading up to a spike has a dimensionality $\sim$ 50, too large to allow adequate sampling of $P(spike|stimulus)$ from the data, so the dimensionality of the stimulus description must be reduced.

The simplest way to do so is to find a subset of directions in stimulus space determined to be relevant for the system, and to project the stimulus onto that set of directions. These directions correspond to linear filters. Such a set of directions can be obtained from the moments of the spike-conditional stimulus: the first such moment is the spike-triggered average, or reverse correlation function (Rieke et al., 1997). It has been shown (Brenner et al., 2000) that for H1, under these conditions, there are two relevant dimensions: a smoothed version of the velocity, and also its derivative. The rescaling observed in steady state experiments was seen to occur independently in both dimensions, so without loss of generality they use as filter the single dimension given by the spike-triggered average. The stimulus projected onto this filter will be denoted by s_0.

The filtered stimulus is passed through a nonlinear decision process akin to a threshold. The input/output relation $P(spike|s_0)$ (Brenner et al., 2000) is calculated by using Bayes' rule:

$$\frac{P(spike|s_0)}{P(spike)} = \frac{P(s_0|spike)}{P(s_0)}. \qquad \text{(A-1)}$$

The spike rate $r(s_0)$ is proportional to the probability of spiking, $r(s_0) \propto$

$P(spike|s_0)$, leading to the relation

$$\frac{r(s_0)}{\overline{r}} = \frac{P(s_0|spike)}{P(s_0)}, \tag{A-2}$$

where $\overline{r}$ is the mean spike rate. $P(s_0)$ is the prior distribution of the projected stimulus, which we know. The distribution $P(s_0|spike)$ is estimated from the projected stimulus evaluated at the spike times, and the ratio of the two is the nonlinear input/output relation.

Appendix B: Distance between two estimates with independently drawn samples.

Here we derive the square distance between two estimates of a distribution where the events are independently drawn.

The events take values i, with $i \in [1, K]$, with a probability q_i for the value i. The first estimated distribution is constructed from N samples, with n_i samples having value i. For the second distribution M samples are used, with m_i samples taking the value i. The estimated probabilities, $p_{1,i}$

and $p_{2,i}$, are given by $p_{1,i} = n_i/N$ and $p_{2,i} = m_i/M$ respectively.

We now calculate the average square distance, $\overline{D}_{1,2}(N, M)$, between two estimated distributions based on N and M samples respectively. $\overline{D}_{1,2}(N, M)$ is given by

$$\overline{D}_{1,2}(N, M) = \langle \sum_{i=1}^{K} (p_i^1 - p_i^2)^2 \rangle_{N,M}. \tag{B-1}$$

Here the brackets denote averaging over all possible outcomes $\{n_i\}$ and $\{m_i\}$ with probability q_i that a sample takes value i, and the constraints $\sum_i n_i = N$ and $\sum_i m_i = M$. Defining $U(N)$ and $V(N, M)$ as

$$U(N) = \langle \sum_{i=1}^{K} \left(\frac{n_i}{N}\right)^2 \rangle_N \tag{B-2}$$

and

$$V(N, M) = \langle \sum_{i=1}^{K} \frac{n_i m_i}{NM} \rangle_{N,M}, \tag{B-3}$$

$D_{1,2}(N, M)$ can be rewritten as

$$D_{1,2}(N, M) = T_2(N) + T_2(M) - 2T_1(N, M). \tag{B-4}$$

$U(N)$ and $V(N < M)$ are given by

$$U(N) = N! \prod_{i=1}^{K} \sum_{n_i=0}^{\infty} \frac{q_i^{n_i}}{n_i!} \delta_{\Sigma n, N} \sum_{k=1}^{K} \left(\frac{n_k}{N}\right)^2 \tag{B-5}$$

and

$$V(N, M) = N!M! \prod_{i=1}^{K} \sum_{n_i=0}^{\infty} \frac{q_i^{n_i}}{n_i!} \sum_{m_i=0}^{\infty} \frac{q_i^{mi}}{m_i!} \delta_{\Sigma n, N} \delta_{\Sigma m, M} \sum_{k=1}^{K} \frac{n_k m_K}{NM}. \tag{B-6}$$

Here δ is the Kronecker delta and we have used Σn and Σm to denote $\sum_i n_i$ and $\sum_i m_i$ respectively.

Because of the constrains U and V are not easily calculated, but one can obtain them from the characteristic functions. U can be evaluated using the characteristic functions $\tilde{U}$, defined by

$$\tilde{U}(x) = \sum_{N} \frac{N^2 x^N}{N!} U(N). \tag{B-7}$$

For every x this characteristic function satisfies

$$\tilde{U}(x) = \sum_{k=1}^{K} \sum_{n_k=0}^{\infty} \frac{(xq_k)^{n_k}}{n_k!} n_k^2 \prod_{j \neq k} \sum_{n_j=0}^{\infty} \frac{(xq_j)^{n_j}}{n_j!}$$

$$
\begin{aligned}
&= \sum_{k=1}^{K} \exp[(1-q_k)x] \sum_{n_k=0}^{\infty} \frac{(xq_k)^{n_k}}{n_k!} n_k^2 \\
&= (x + Q^{(2)}x^2)e^x, \qquad \text{(B-8)}
\end{aligned}
$$

where we have used $\sum_k q_k = 1$ and $Q^{(2)}$ is given by $Q^{(2)} = \sum_k q_k^2$. Combining this with the definition of $\tilde{U}$ and collecting terms with the same powers in x one obtains that

$$U(N) = Q^{(2)} + N^{-1}(1 - Q^{(2)}). \qquad \text{(B-9)}$$

Similarly, using $\tilde{V}(x,y) \equiv \sum_{N,M} \frac{NMx^N y^M}{N!M!} V(N,M)$, one shows that

$$V(N,M) = Q^{(2)}. \qquad \text{(B-10)}$$

Combining these results we obtain that, average over sampling of the distribution, the distance between two estimated probability distribution, based on N and M samples respectively, is given by

$$D_{1,2} = \left(\frac{1}{N} + \frac{1}{M}\right)\left(1 - \sum_{k=1}^{K} q_k^2\right). \qquad \text{(B-11)}$$

Bibliography

Atick, J. and Redlich, A. (1992). What does the retina know about natural scenes? *Neural Computation*, 4:196–210.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61:183–193.

Baddeley, R. (1996). Searching for filters with 'interesting' output distributions: an uninteresting direction to explore? *Network: Computation in Neural Systems*, 7:409–421.

Baddeley, R., Abbott, L., Booth, M. C., Sengpiel, F., Freeman, T., Wakeman, E. A., and Rolls, E. T. (1997). Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society of London B*, 264:1775–1783.

Barlow, H. (1961). Possible principles underlying the transformation of

sensory messages. In Rosenblith, W., editor, *Sensory Communication*, pages 217–234. MIT press, Cambridge, MA.

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

Bownds, M. and Arshavsky, V. (1995). What are the mechanisms of photoreceptor adaptation? *Behavioral and Brain Sciences*, 18:415–424.

Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, 26:695–702.

Buiatti, M. and van Vreeswijk, C. (2002). Variance normalisation: a coding mechanism for redundancy reduction in natural vision. In preparation.

Carandini, M., Heeger, D., and Movshon, J. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17:8621–8644.

Dan, Y., Atick, J. J., and Reid, R. C. (1996). Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *Journal of Neuroscience*, 16:3351–3362.

Dong, D. and Atick, J. (1995a). Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6:345–358.

Dong, D. and Atick, J. (1995b). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6:159–178.

Ebeling, W. and Poschel, T. (1994). Entropy and long-range correlations in literary english. *Europhysics Letters*, 26:241–246.

Fairhall, A. L., Lewen, G. D., Bialek, W., and de Ruyter van Steveninck, R. (2000). Multiple timescales of adaptation in a neural code. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, MA.

Fairhall, A. L., Lewen, G. D., Bialek, W., and de Ruyter van Steveninck, R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, 412:787–792.

Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4:2379–2394.

Kim, K. and Rieke, F. (2001). Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. *Journal of Neuroscience*, 21:287–299.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C*, 36:910–912.

Levy, W. and Baxter, R. (1996). Energy efficient neural codes. *Neural Computation*, 8:531–543.

Liu, Y., Cizeau, P., Meyer, M., Peng, C.-K., and Stanley, H. (1997). Correlations in economic time series. *Physica A*, 245:437.

Liu, Y., Gopikrishnan, P., Cizeau, P., Meyer, M., Peng, C.-K., and Stanley, H. (1999). Statistical properties of the volatility of price fluctuations. *Physical Review E*, 60:1390.

Mantegna, R. and Stanley, H. (1996). Turbulence and financial markets. *Nature*, 383:587–588.

Mantegna, R. N., Palagyi, Z., and Stanley, H. E. (1999). Applications of statistical mechanics to finance. *Physica A*, 274:216–221.

Meister, M. and Berry, M. (1999). The neural code of the retina. *Neuron*, 22:435–450.

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–610.

Palagyi, Z. and Mantegna, R. (1999). Empirical investigation of stock price dynamics in an emerging market. *Physica A*, 269:132–139.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1987). *Numerical Recipes.* Cambridge University Press, Cambridge, U.K.

Raberto, M., Scalas, E., Cuniberti, G., and Riani, M. (1999). Volatility in the italian stock market: an empirical study. *Physica A*, 269:148–155.

Reinagel, P. and Zador, A. (1999). Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, 10:1–10.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: exploring the neural code.* MIT Press, Cambridge, MA.

Ruderman, D. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548.

Ruderman, D. (1997). Origins of scaling in natural images. *Vision Research*, 37:3385–3398.

Ruderman, D. and Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73:814–817.

Sakai, H., Wang, J., and Naka, K. (1995). Contrast gain control in the lower vertebrate retinas. *Journal of General Physiology*, 105:815–835.

Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4:819–825.

Shapley, R. (1997). Adapting to the changing scene. *Current Biology*, 7:R421–R423.

Shapley, R. and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. *Progr. Ret. Res.*, 3:263–346.

Shapley, R. and Victor, J. (1978). The effect of contrast on the transfer properties of cat retinal ganglion cells. *Journal of Physiology*, 285:275–298.

Simoncelli, E. P. and Schwartz, O. (1999). Image statistics and cortical normalization models. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems 11*, pages 153–159. MIT Press, Cambridge, MA.

Smirnakis, S. M., Berry, M., Warland, D. K., Bialek, W., and Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial structure. *Nature*, 386:69–73.

Srinivasan, M., Laughlin, S., and Dubs, A. (1982). Predictive coding: a

fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B*, 216:427–459.

Tadmor, Y. and Tolhurst, D. (2000). Calculating the contrasts that retinal ganglion cells and lgn neurones encounter in natural scenes. *Vision Research*, 40:3145–3157.

van Hateren, J. (1992). Theoretical predictions of spatiotemporal receptive fields of fly lmcs, and experimental validation. *Journal of Comparative Physiology A*, 171:157–170.

van Hateren, J. (1997). Processing of natural time-series of intensities by the visual system of the blowfly. *Vision Research*, 37:3407–3416.

van Hateren, J. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, 265:359–366.

Victor, J. (1999). Temporal aspects of neural coding. *Network: Computation in Neural Systems*, 10:R1–R66.