Computational Models of Primary Visual Cortex and the Structure of Natural Images

vorgelegt von
Diplom-Informatiker (Dipl.-Inf.)

Hauke Bartsch

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

Promotionsausschuss:

Vorsitzender:	**Prof. Dr.-Ing. R. Orglmeister**
Berichter:	**Prof. Dr. rer. nat. K. Obermayer**
Berichter:	**Prof. Dr. sci. nat. F. Wysotzki**

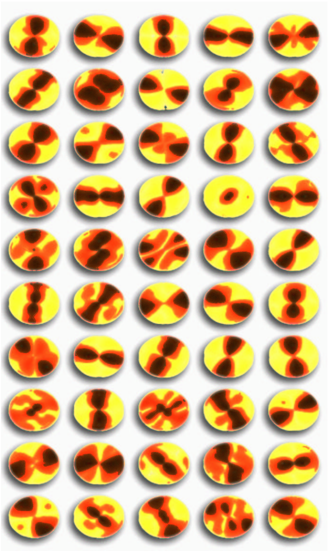
Tag der wissenschaftlichen Aussprache: **2003-12-01**

Berlin **2004 / D 83**

Hauke Bartsch

January 28, 2003

Computational Models of Primary Visual Cortex and the Structure of Natural Images



PhD Thesis

Department of Electrical Engineering and Computer Science
Technical University of Berlin, Germany

Abstract

Understanding the function of the brain appears to be mainly an introspective task. Nevertheless, large parts of our brain are used as a direct interface to our environment. They constantly acquire information and control our actions. The dynamic interweaving of brain and environment opens a dual route to the understanding of cortical processing. Based on that insight the present work focuses on the following two issues:

a) analyzing the functional role of cortical networks, and

b) analyzing the special requirements imposed onto the brain by the statistics of its input, namely the statistics of natural images.

After an introduction into both, natural images and the mammalian visual system, we first analyze the connection scheme found in the primary visual cortex at different levels of abstraction. We start by deriving a system of coupled differential equations describing a column of excitatory and inhibitory neurons. Phenomena like contrast invariant orientation tuning and contrast saturation are investigated which have been found important features of cortical neurons. The initial model is extended to explain also response properties related to contextual effects. There stimuli presented outside the classical receptive field of the neuron can modulate its response. Principle difficulties in having cross-orientation modulations by iso-orientation specific patchy connections are shown. In explaining cross-orientation modulations we analyze the effect of the distinctive spatial layout of the cortex. We found that two opposite effects contribute to the observed contextual modulation; (i) local inhibition that is induced by a local change in input (leads to suppression), and (ii) dis-inhibition.

The second part deals with the input of the visual system, namely pictures of scenes encountered in the surrounding world. We formulate the hypothesis that higher order features in spatial pattern can be described in terms of intrinsic invariance and symmetry and introduce a mathematical formulation of smooth local symmetries. Applications for object classification, image alignment, and landmark detection illustrate the principle advantage of our structure analysis over methods of shape analysis.

Two new algorithms are introduced to efficiently learn higher order features. The first one introduces a centralized Gaussian mixture model to extract

second-order features estimating the density of the data. The obtained code is shown to outperform other known linear codes by being well distributed and by showing a high population sparseness, both are preferable properties in coding of natural images.

Originating from geometrical considerations of manifolds in high dimensional spaces we introduce a non-linear transformation and by this a family of feature spaces that are shown to be useful to detect correlation of a specific order in the data. Moreover it is shown that these correlations can be learned in the feature space by linear methods. This general property of the transformation is interesting for a large class of algorithms in the field of explorative data analysis. In the context of independent component analysis this transformation defines a feature space in which the assumption of independent sources can be fulfilled for a set of over-complete basis functions.

Ziel dieser Arbeit ist es, zum Verständnis der Strukturen des menschlichen Gehirns beizutragen. Grosse Teile unseres Gehirns funktionieren als eine direkte Schnittstelle zu unserer Umwelt. Zwischen Umwelt und Gehirn werden kontinuierlich Informationen verarbeitet und Handlungen initiiert. Das dynamische Wechselspiel zwischen Umwelt und Gehirn macht es notwendig, auch bei der Analyse der Verschaltungsstrukturen im Gehirn ihre jeweiligen Eingaben, hier meist sensorische Signale, zu untersuchen. Auf der Grundlage einer Dualität von Gehirnstrukturen und sensorischen Signalen beschäftigt sich diese Arbeit mit den folgenden beiden Themen:

a) der Analyse der Funktion kortikaler Schaltkreise und

b) den speziellen Anforderungen, die bei der Verarbeitung von Bildern an das Gehirn, insbesondere deren statistischen Eigenschaften gestellt werden.

Nach einer kurzen Einführung in die Statistik natürlicher Bilder und die Anatomie und Funktion des visuellen Systems der Säugetiere untersuchen wir die Verschaltungsstrukturen, die im ersten visuellen Areal gefunden werden. Insbesondere die Phänomene der kontrastinvarianten Antwort auf orientierte Gitter als optische Stimuli und das Sättigungsverhalten der Zellen bei hohen Kontrasten in der Eingabe werden analytisch und durch Computersimulationen unterstützt untersucht. Das verwendete Differentialgleichungsmodell für gekoppelte Zellpopulationen wird schrittweise erweitert, um auch kontextabhängige Effekte untersuchen zu können. Hierbei hängt die Antwort einer Zelle von den Reizen in ihrer weiteren Umgebung ab. Wir zeigen unter anderem, dass zwei unterschiedliche Effekte zu den Kontextmodulationen beitragen: zum einen lokale Hemmung (*Inhibition*), die durch eine Änderung in der Struktur der Eingabe bestimmt werden, und des weiteren durch den Effekt der *Dis-inhibition*.

Die Analyse von Bildern unserer Umgebung ist das Hauptziel im zweiten Teil der Arbeit. Sie sind der Input in das visuelle System. Wir formulieren die Hypothese, dass wichtige Eigenschaften von Bildern durch ihre inhärenten Invarianzen und Symmetrien definiert werden. Um diese Hypothese zu testen, führen wir ein mathematisches Mass für lokale Symmetrien in räumlichen Mustern ein. Anwendungen auf den Gebieten der Objektidentifikation, der Objektausrichtung und der Landmarkenfindung unterstreichen die Vorzüge der Strukturanalyse gegenüber einer reinen Formanalyse.

Zwei neue Algorithmen werden vorgestellt um Eigenschaften höherer Ordnung in Bildern zu lernen. Der erste basiert auf dem Modell eines zentralisierten *Gauss'schen-Mixture Modells* und extrahiert Merkmale dadurch, dass er ein Modell der Verteilungsfunktion der Daten lernt. Die gelernten Merkmale sind denen anderer Modelle erster Ordnung hinsichtlich der Populationsantworteigenschaften überlegen.

Ausgehend von geometrischen Überlegungen zu Mannigfaltigkeiten in hochdimensionalen Räumen führen wir eine Transformation in einen nicht-linearen Merkmalsraum ein, in dem Korrelationen beliebiger Ordnung mit linearen Methoden gelernt werden können. Im Kontext der *Independent Component Analysis* als einem Beispiel für einen Algorithmus der explorativen Datenanalyse kann die Transformation dazu benutzt werden, um über-komplette Basisfunktionen zu lernen.

Contents

1. Introduction

> *"Vision is our primary sensory channel for interaction with the outside world. It allows us to recognize familiar faces and creatures, and objects; it allows us to orient ourselves in space and to navigate from place to place. It is a pathway for esthetic enjoyment and for information transmission. The visual system is one of the many miracles of nature."*
>
> *(Shapley and Enroth-Cugell, 1984)*

The brain is responsible for our ability to do complicated things like singing, playing chess or writing a thesis. If it is able to perform complex things we would expect that the brain itself is complex. But how can we measure its complexity? An equally valid question is: How can we measure the complexity of our environment? Both questions are related because large parts of the brain are involved in the analysis of sensory information. In analyzing either the device (our brain) or the sensory information we hope to learn something about the successful interplay between both.

At the end of the 19-th century Ramón y Cajal was the first who established that the building blocks of the brain, the nerve cells, act as independent units. We understand that a large part of the complexity of the brain is due to the connections of these units. But complexity is rather more than number of connections. A fully connected neural network has a large number of connections, but it takes only a single line in Matlab[1] to define a simple network with a large number of connections

$$\mathbf{y} = \tanh\left(\mathrm{ones}(10^{14})\mathbf{x} + \Theta\right).$$

It would be huge, in fact its number of connections is about the same (10^{15}) as the expected number of connections in our brain, but intuitively we would not suspect it of being complex. A similar reasoning can be applied to our sensory information. Writing a program that enumerates all possible combinations of

[1] Matlab (©The MathWorks, Inc.) is a tool for doing numerical computations with vectors and matrices.

(discreet) stimuli is easy but it will tell us little about the complexity of our surrounding world. What is the basis of our intuition about what is complex and what not?

There are measures for the complexity of objects. One is the measure of *Kolmogorov complexity* (Solomonoff, 1997; Kolmogorov, 1965; Chaitin, 1966). Considering that Shannon's information theory (Shannon and Weaver, 1948) is concerned with the average information of a random source Kolmogorov complexity of an object is a form of absolute information of the individual object. It can be defined as the size (number of binary digits, or bits) of the shortest program that without additional data, computes the object and terminates. In this sense the network specified above is not complex because of its short description length. Unfortunately this measure is of rather theoretical use because there is no way to produce *the* shortest program (or even to recognize that a program is the shortest possible). However it is useful in the context of comparing different programs and appears widely in disguise of Ocam's Razor or the minimum description length principle.

Kolmogorov complexity

Other concepts that are related to complexity are structure and redundance. Redundancies are repeating parts of an object. If we have detected redundancies we have also found the essential parts, the structure in the data. An example is the sequence 01010101 and its representation as $01 * 4$. The redundancy defines the structure of the sequence and using this structure we can represent the sequence in a compact way[2].

Back to our starting point. To learn something about how we perform complex tasks we need to find the essential structures in either the brain or the environment. We know that the brain is highly ordered; into areas, layers, functional columns, and distinct neuron classes. Defining computer models of the part of the human brain that is concerned with vision is the topic of the first part of this thesis. The second part deals with the dual problem of finding structure in the respective sensory channel the visual stimuli.

By analyzing both the brain and its input we hope to deepen our understanding for the function of the brain, how it is organized and how it can handle wast amounts of information so astonishing efficient.

[2] Yes, we have to add the length of the algorithm that performs the $*$ operation, and the length of the blueprint that was used to build the computer that performs the algorithm, and the description length of the basic physical laws and constants that define the universe in which the machine is build that performs the algorithm which produces the sequence. That is meant by Kolmogorov complexity being not practical.

1.1. Scope and Goals

Building a model about what (we think) is relevant to information processing is mostly based on a lot of assumptions about what (we think) is irrelevant. So any results obtained by models have to be viewed in the light of the underlying assumptions. In this sense doing biological modeling is helpful about proving or disproving of concepts (about the question how biology can work) but not about how it works. All statements directed towards the function of the brain can be enclosed with the phrase: 'If I would be brain thats how I would make it.'

To name only some of the facts which are not in the scope of this thesis: we will neglect any temporal receptive field structure of neurons as found by Ringach, Sapiro and Shapley (1997) and DeAngelis, Ohzawa and Freeman (1995), the fact that vision is an active process and any detailed modeling of neurons at the level of channels, synapses, or spikes. Instead, after an introduction into basic findings of the statistics of natural images and the general layout of the primate visual system we will head directly for (i) the interaction of populations of neurons (Chapter 3 on page 40) and (ii) the functional role of lateral connections in analyzing natural stimuli (Chapter 4 on page 78).

Especially in the later chapters we will assume that the reader is familiar with some standard algorithms of machine learning and data analysis. Introduction into these algorithms will be very brief and the reader is directed to standard books about brain theory and neural networks as for example the always very helpful Arbib (1998).

1.2. Plan of the Manuscript

This manuscript is concerned with the functional architecture of the primary visual cortex (visual area V1, striate cortex) which serves as an excellent model for the human sensory system.

In Figure 1.1 the general arrangement of chapters is laid out. The first part of the thesis reviews the relevant anatomical and physiological findings together with various functional forms proposed in literature.

In the second part of the manuscript computational models are developed which address the interplay of orientation selective neuron populations (*i*) locally to one *column* of visual cortex, (*ii*) between different columns (*hypercolumn model*) and (*iii*) between different hypercolumns (*lattice model*).

The third part analyses in a more formal framework how the statistics of the visual input influences the shape of the response characteristics of model neurons. In analyzing natural images we make predictions about the receptive

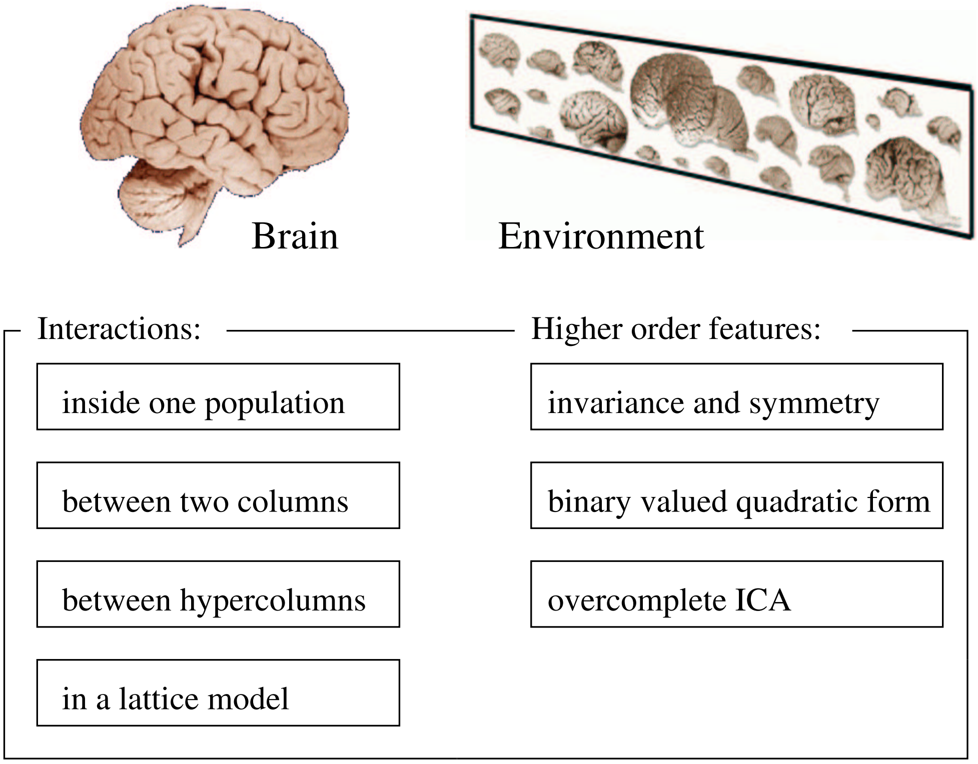


Figure 1.1.: Overview of chapters

fields of cortical neurons. One more technical grounds an algorithm is given which can solve the problem of overcomplete source separation.

1.3. The Input of the Visual System

There are reasons to believe that the brain is tuned to the special requirements of its input signals. A benefit of this could be to overcome the limitations of the receptors, e.g., of the finite sampling of the signal by the retinal light receptors (Ruderman and Bialek, 1992). Another reason could be that in a constantly changing environment a system that can adapt is more effective in terms of information transmission and representation (Barlow, 1961). Natural selection would work in favor of systems with superior performance thus the evolution to highly efficient and error tolerant systems.

But adaptation of the brain also happens on shorter than evolutionary time scales. During ontogeny in sensory areas of many species a refinement of the neuronal connections is observed. The principles behind these processes are still under debate. Candidate theories to explain the changes of the cortex during ontogenesis are based on either intra-cortical mechanisms or on constrains imposed on the system by the specific form of the input. A main result of the refinement during ontogeny is the topographic layout of many sensory areas in the adult cortex. Nearby neurons tend to code for nearby stimuli[3]. The formation of topological maps was intensely studied and serves as a archetypical system for the study of developmental processes.

topography

Map Formation by Intra-cortical Constrains

A intra-cortical property that has been used to explain topographic mappings is the overall wiring length (Allman and Kaas, 1974; Koulakov and Chklovskii, 2001). Short overall wiring lengths are favorable for the speed of processing and the metabolism of the animal. Arguments against this purely intra-cortical explanation of the cortical layout are: (i) The traveling speed of action potentials can be varied independently on the wiring length, e.g. by changing the diameter of the wire. (ii) there are examples for animals with a 'salt-and-petter' organization of the respective sensory area (visual system of the rat) thus wiring length minimization may not be a crucial requirement[4]. Our understanding about the metabolic constrains of the system is limited. One assumes that in the order of 20% of our energy is used up by the central nervous

[3] The visual cortex displays a spatially topographic layout, the auditory cortex a frequency topographic layout.

[4] Rats rely mostly on their auditory and olfactory senses and the rat auditory cortex is, for example, organized (frequency) topographic.

system. Keeping the energy consumption of the brain low may have been a worthwhile strategy favored by natural selection in the human evolution.

Map Formation Influenced by the Stimulus

The topographic layout of neurons processing sensory information can also be explained by peculiarities of the stimulus. Models that explore local correlations in the data also explain the emergence of topographic maps. Takeuchi and Amari (1979) carried out a one-dimensional analysis of a continuous version of a neural activity model. They showed that when the width of the input stimuli is smaller than the extent of the lateral interactions in cortex an ordered map results. This ansatz to explain the intra-cortical structures was motivated first by the finding of prominent local correlations in images thus by the observation that most images contain redundant information.

Experiments indicated that the initial orientation and ocular-dominance maps are largely independent on visual experience (Crair, Gillespie and Stryker, 1998). This indicates that the general layout of the cortical maps is genetically predefined. After an initial period the stimulus can strongly influence the maps (Singer, 1981) and the cortex remains versatile in adult animals.

Because map formation is a very general tool used at large by the cortex it is very likely that its basics are layed down as early as observed perhaps even enforced by genetical factors (Kaschube, Wolf, Geisel and Löwel, 2002). Other features of the neurons may depend stronger on the specific type of input. Roe, Pallas, Hahm and Sur (1990) performed a drastic experiment along these lines. It is plain that neurons in the auditory cortex have to detect different features than the neurons in the visual cortex which results in different observed lateral connection schemes in both areas. Long-range connections in the auditory cortex, for example, show not the patch like terminals of the long-range connection in primary visual cortex. In their work Roe et al. (1990) demonstrate that re-routing the information of the retina to the auditory system results in an orderly map of visual space in the auditory cortex. Lateral connections change accordingly and that visual information can be processed by the auditory cortex in much the same way as in the visual cortex. From these experiments one can derive that the sensory inputs can direct the formation of cortical circuitry to a large extent.

Structure and Redundancy of the Stimulus

An important relation can be drawn between the redundancies and the structure of the input. Whereas redundancy describes the part of the code that does not transport (additional) information structure describes the essential parts of the input. If a part of the input has a structure it will be easily compressible and therefore it contains redundant information. Thus, in exploring redundancies in the code one searches for structure. The idea of structure detection is connected with the measure of Kolmogorov complexity (see the Introduction of this thesis) and the principle of *minimum description length*.

Kolmogorov complexity

minimum description length

1.3.1. A Statistical Description of Images

In order to explain the structure of the brain that is used for the processing of visual information we like to review the literature on the statistics of *natural images*, that is to say of pictures of scenes encountered in the surrounding world.

natural images

One way to analyse the structure of natural images is to describe them in terms of statistics. A statistical description of images assumes that the images presented by the environment are instantiations of random vectors. In the mammalian eye regular spaced light sensitive detectors receive the reflected light of objects and report the light intensity. Using rough numbers for the resolution and sensitivity of the optical systems (see Section 2.1 on page 17) we can derive a number of distinguishable visual inputs which will be in the order of

$$100^{(10^6)} = 10^{2,000,000}, \tag{1.1}$$

10^6 neurons each with a sensitivity range of one to one hundred. Although this is a large number[5] we can be certain that the visual input that arrives from our environment is only a fraction of that huge state space of possible images. First of all, there is simply not enough time for an individual to sample the state space. The fact that we nevertheless can handle vision quite well implicates structure in the input. So, there are reasons to believe that the images we perceive are highly coherent, that there are properties by which we can distinguish natural images from just possible images.

In our above calculation we counted, for example, all possible random patterns as equally possible images. But images are seldom random (Field, 1987), because nearby pixels tend to be highly correlated. Also the probability to encounter any known image by a random process is obviously very low. In Appendix A on page 174 we compare the entropy of a set of natural images with the entropy of possible images. Indeed we found, that the distribution of natural images is far away from that of a uniform distribution (the assumption for possible images).

We will see that the information content of natural images is considerably reduced by some basic properties. The remaining state space appears to be

[5] The number of atoms in the universe is in the order of 10^{72}.

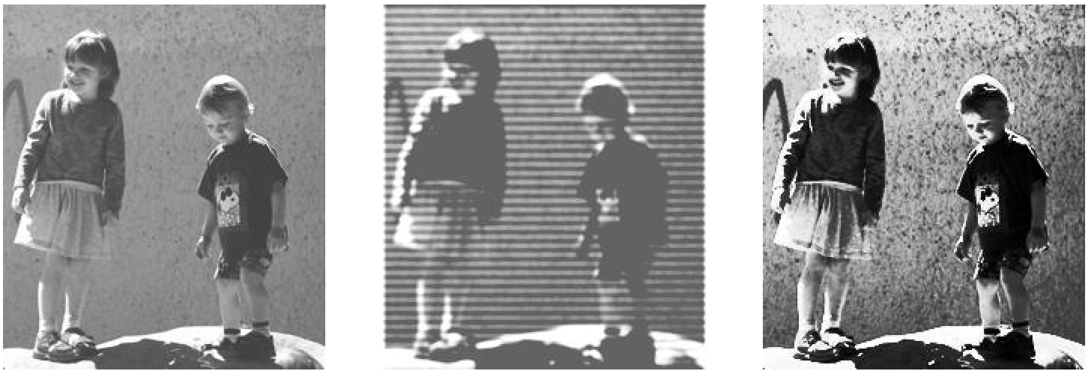


Figure 1.2.: Drastic changes on the level of single pixel does not impair our perception

complicated and to handle that complexity we will introduce in this thesis higher order statistics.

1.3.2. First Order Statistics

first order statistics

To describe the first order statistics, e.g., the gray level histogram on images, is found of not being of much use. One can strongly change the histogram of an image, for example, by changes in illumination without much affecting its perception (see Section 1.2). Ruderman (1994) found linear tails when plotting $\log$(number of occurrences of gray level) versus $\log$(gray level).

log statistics

The log statistic is often used if one suspects the presents of power laws (Gisiger, 2001). Lets consider the following function $y = ax^{\alpha}$, where a and α are real and constant and x is a variable. By taking the log of both sides, one obtains $\log y = \log a + \alpha \log x$, which is when plotted on a log-log scale a straight line of slope α. The line intersects the ordinate axis at $\log a$. Important in this context is the scale-invariant property of power laws. Replacing the variable x by $z = \beta x$ we obtain for the log measure $(a\beta^{\alpha})z^{\alpha}$ which is again a power law with exponent α. Only the constant of proportionality has changed from a to $a\beta^{\alpha}$.

Another reason to use log statistics in image analysis is that one is usually interested in studying the light intensity arriving on the lens. By using the log of the gray-values one obtains linear relationship between gray level and intensity.

Also the histogram of an image should be invariant under multiplication of the gray level by a constant. One way to achieve this invariance is to study $\log(I/I_0)$ instead of I, where I_0 is the mean gray level of the image. The low information content of single pixels indicates that information which is contained

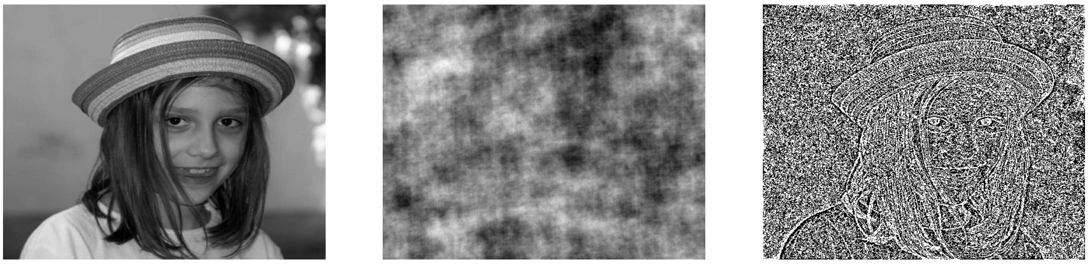


Figure 1.3.: Surrogate data demonstrates that information is contained in the higher order statistics. Randomizing the phase (*middle*) destroys higher order information but keeps mean and variance. Randomizing the amplitude (*right*) destroys mean and variance

in the statistics of single pixel is highly redundant. Even if the amplitude component of an image is completely destroyed recognition of objects is possible (see the example in Figure 1.3). This highlights that important information is conveyed in the higher order statistics of natural images. What is detected by lower order statistics is mostly redundant information and systems engaged in analyzing images should incorporate some kind of adaptation to light levels.

The importance of higher order statistics is reflected in the success of local edge based image coding (compression) algorithms by the means of wavelets over compact coding schemes derived from non-local features like Fourier-spectra or PCA.

1.3.3. Second Order Statistics

Second order statistics deal in contrast to the first order statistics with the statistics of combined events. One assumes the image I to be a continuous function from $\mathbb{R}^2$ into $\mathbb{R}$.

co-occurrence

Co-occurrences describe the complete second order statistics. From the assumption of translational invariance it follows that we can look onto the statistics with respect to an arbitrary pixel. A function of this kind can be defined as

$$\mathrm{coo}(i, j, x) \quad = \quad P(I_0 = i \;\&\; I_x = j) \tag{1.2}$$

for x being a position in the image and i, j gray levels. By this measure the relative frequency of the occurrence of specific pairs of gray values can be measured.

covariance
correlation

Covariances or correlations can also be computed with respect to a center pixel. But in contrast to the measure of co-occurrence they express the tendency of two features (pixel) to vary together (in their gray values)

$$\mathrm{cov}(X,Y) = E\left((X-E(X))(Y-E(Y))\right) \quad (1.3)$$

$$\rho(X,Y) = \mathrm{cov}(X,Y)/\left[E\left((X-E(X))^2\right)E\left((Y-E(Y))^2\right)\right]^{1/2} \quad (1.4)$$

here $E(.)$ is the expectation over many images. The measure of correlation is widely used to analyse effects of distance and orientation on the co-variability of the image intensities. Results indicate that images are mostly smooth, in other words they have finite spatial correlations with occasional rapid changes in contrast (edges). Field (1987) analysed the spatial frequency of a number of natural images. He found that when averaging over all orientations, the power at a given frequency in the images was proportional to 1/frequency (according to Billock (2000) $1/f^{\beta}$ with a β of $0.9 \ldots 1.2$). This indicates that nearby positions in the images are highly correlated (because the Fourier transform of an image can be converted by the Wiener-Khintchin theorem (Connor, 1982) to the auto-correlation function). It also shows that natural images are highly non-Gaussian.

$1/f$

Also Ruderman and Bialek (1994) found that distributions of local quantities such as contrast are scale invariant and have nearly exponential tails which reflects that there is no typical scale at which objects are seen.

Arguments for Higher Order Features

Turiel and Parga (2000) decomposed pixels in the image, into sets, the *fractal components* of the image, so that each set contains only points characterized by a fixed strength of the singularity of the contrast gradient in its neighborhood. They found that under changes in scale each fractal component exhibits its own transformation law and scaling exponent, e.g., how sharp or soft a change in contrast is at a given point can be quantified in terms of the value of the scaling exponent at that site. This indicates that there is not a well-defined scale for the components of an image but at the same time the scene is not scale invariant globally.

fractal components

Baddeley (1997) analysed in detail the correlation structure of different sets of natural images. Open and urban landscapes were used to estimate the degree of correlation between image intensity measurement pairs as a function of both distance and orientation. He found that psychophysical findings on distance estimation (Cormack and Cormack, 1974) can be explained by the slower decaying rates of horizontal correlations compared to more vertical[6] correlations.

There are different ways to explain the approximately power law fashion of the observed correlation structure. One possibility suggested by Field (1987) is

[6] More specific, the direction of smallest decay was image-set specific and in the range of $20° - 45°$.

that images are self-similar or fractal (see also (Billock, 2000)). This implicates that there is no special scale for objects. Baddeley (1997) considers another model. There, idealized randomly sized and shaped "objects" are viewed at a number of random distances. Within each object the image pixels are perfectly correlated, but across the boundary of an object the characteristics are completely uncorrelated. If all object sizes are equally probable the model explains the observed correlation of natural images. Alvarez, Gousseau and Morel (1999) also showed that the size of objects in natural images exhibit a scale invariant property. Objects were defined as connected components where the contrast does not exceed a certain threshold.

The simple explanations for the observed power laws raise the question about how informative the average correlation is, and what other methods can be used to extract informations hidden in the images. One goal of this thesis is to introduce a *less* averaged version of the correlation model. Its essence is that more than one correlation matrix is learned simultaneously and the whole model seeks to explain the underlying causes of the input, its constituting structure.

Second order moments (correlations) mostly reveal the locally smooth nature of natural images, thus its redundant information. Contrarily edges represent higher order features (see Figure 1.3) and contain important information about the visual scene. This can be made plain by computing the relative frequencies of sets of image patches that appear in a scene. According to Shannon and Weaver's (1948) definition of information most information is contained in image patch configurations that appear with the lowest frequencies[7]. In Figure 1.4 on the facing page, center the likelihood of image patches is computed as the negative log frequency of the summed absolute response of a set of ICA-filters for the image shown left. The filter-bank is used in order to reduce the dimensionality of the problem, in other words, the possible number of images over which we have to integrate for computing the likelihood. It turns out, that predominantly edges are among the least probable image features, thus they convey the largest amount of information in the images. This also corresponds to the findings of Geman and Koloydenko (1998), that edges are "the most probable non-background" micro-image configurations.

1.3.4. Decomposition into Basis Sets

Other studies deal with a decomposition of images into linear combinations of basis images. An image X is considered as a random vector of size $1 \times MN$

[7] Information is connected to surprise.

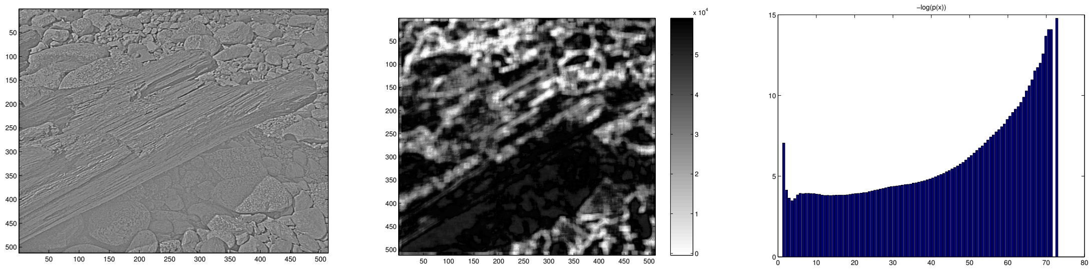


Figure 1.4.: *Center:* Likelihood of patches x from the image *left* coded in gray value. *Right:* the corresponding histogram of occurrences of image features obtained from 10 images. The probability of a single patch was computed by counting the number of summed absolute responses to a filter bank of edge detector units (obtained by FastICA on the same image stack). Edges are the least probable image features thus contain the highest amount of information

and is written as:

$$X = \sum_{i=1}^{n} a_i \mathbf{A}_i = \mathbf{a}A \tag{1.5}$$

where a_i is a random variable, $\mathbf{A}_i$ is a fixed image also called a *basis image* and A is a matrix containing $\mathbf{A}_i$'s as columns. Note, that the number of basis images can be larger or smaller than the dimension of X (that is A needs not to be a square matrix). *basis image*

One uses a condition on the a_i's and learns the $\mathbf{A}_i$'s which describe the images. In principal component analysis one assumes the a_i's as being pairwise uncorrelated. The $\mathbf{A}_i$'s are obtained by looking for eigenvalues of the covariance matrix of the images or image patches. For natural images the $\mathbf{A}_i$ are non-local and resemble a Fourier basis (Olshausen and Field, 1996).

The assumption that the responses of the set of filters should be pairwise statistically independent lead to spatially restricted Gabor-like receptive fields which encode lines at certain positions in visual space (Bell and Sejnowski, 1996). Also from the assumption of sparseness in the neuronal response localized edge detectors can be obtained (Olshausen and Field, 1996). Sparseness is related to the idea that in neuronal assemblies most neurons should be silent most of the time to save metabolic energy. Because of the similarity of the obtained filter to receptive fields of simple cells in primary visual cortex one concludes that the primary visual cortex performs in order to reduce the redundancy in its input (Barlow, 1961).

A non-linear extension of the model above is the introduction of polynomial basis functions. Here $\mathbf{A}_i$ is modeled as a *basis function* $\phi_i(X)$: *basis function*

$$\mathcal{F}(X) \;=\; \sum_{i=1}^{n} a_i \phi_i(X) \tag{1.6}$$

If the basis functions are, for example, cross products of k or fewer coordinates of the input vector X then $\mathcal{F}(X)$ is a polynomial of degree k. In the context of *sigma-pi units* neuronal nets these polynomial classifiers[8] have been 're-named' *sigma-pi units* (Σ - sum, $\prod$ - product) or *high-order nets*. Starting with chapter 4 on page 78 we will analyse models of this kind in order to learn non-linear basis functions for natural images.

Summarizing: Natural images can be well characterized by their local contrast, and efficient algorithms should capture the invariances and redundancies found in the images. In the next chapter we will see that most of the listed properties of images comply with anatomical structures or physiological findings in the mammalian visual system. Further aspects of the statistics of natural scenes are reviewed in Atick and Redlich (1992), Field (1994) and Ruderman (1994).

[8] An elementary two-class discrimination is performed by comparing the output to a threshold.

2. The Mammalian Visual System

The visual system of humans is organized as a serially connected system called the *visual pathway* (see Figure 2 on the following page). The retina of each eye consists of a plate having three layers of cells, one of which contains the over 125 million light-sensitive receptor cells, or rods and cones. The two retinas send their output to two peanut-size nests of cells deep within the brain, the *lateral geniculate bodies*. These structures in turn send their fibers to the striate or primary visual cortex (V1). From there, after being passed from layer to layer through several sets of synaptically connected cells, the information is sent to several neighboring higher visual areas (V2-V*) (see Figure 2.1).

visual pathway

In terms of information processing we will see that the retina computes a compact code which is transmitted through the bottleneck of the optic fibers into the relay station LGN. In the LGN also massive back-projections from the next stage of cortical processing are present and may influence the signal from the retina. Little is known about the functional role of this back-projection.

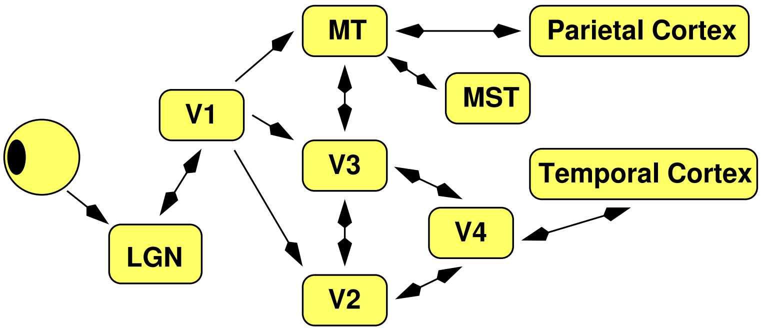
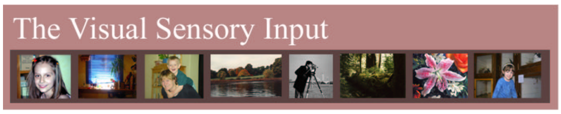


Figure 2.1.: General flow of information between the first areas of visual information processing

The next section is based on Hubel's (1995) excellent book.

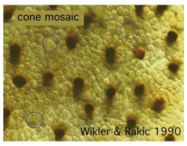
The Visual Pathway

The Visual Sensory Input

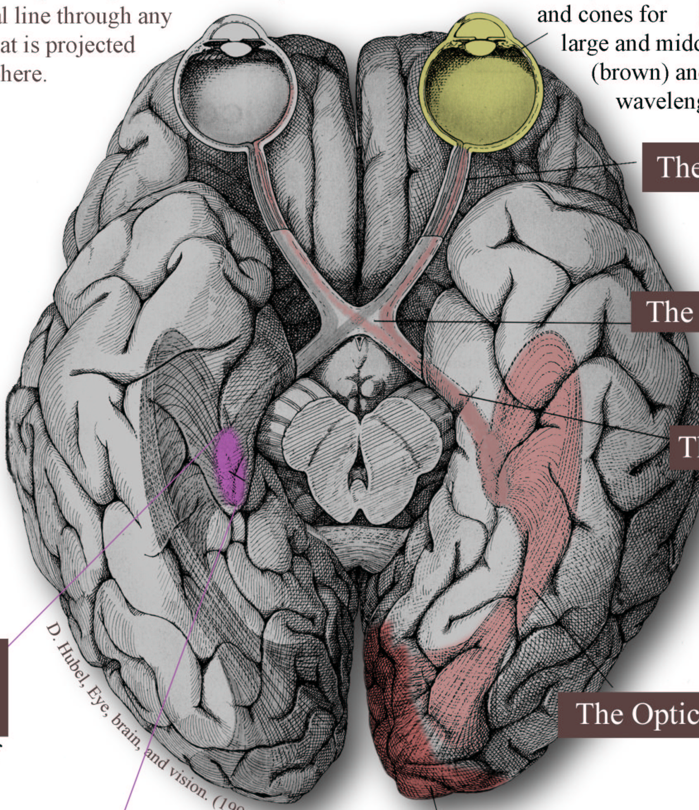


Our eyes are directed in such a way that their areas of sight overlap. The visual pathway ensures that all points to the left of a vertical line through any point we are looking at is projected onto the right hemisphere.

The Eye



A light sensitive receptor grid. Arrangement of rods (yellow), and cones for large and middle wavelength (brown) and cones for short wavelength (circle).



D. Hubel, Eye, brain, and vision. (1995)

The Optic Nerve

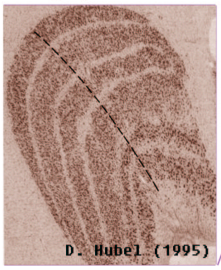
The Optic Chiasm

The Optic Tract

The Optic Radiation

The Lateral Geniculate Nucleus

Segregating layers of different receptive field organization and eye preferrence.



D. Hubel (1995)

The Primary Visual Cortex

The primary visual cortex is the part of the neocortex that receives visual input from the retina. Because it is very similar to the rest of the neocortex in its anatomical structure, it is widely believed that understanding the structure and function of the primary visual cortex will provide fundamental insights into how the neocortex operates.

Hauke Bartsch, Nov. 2002

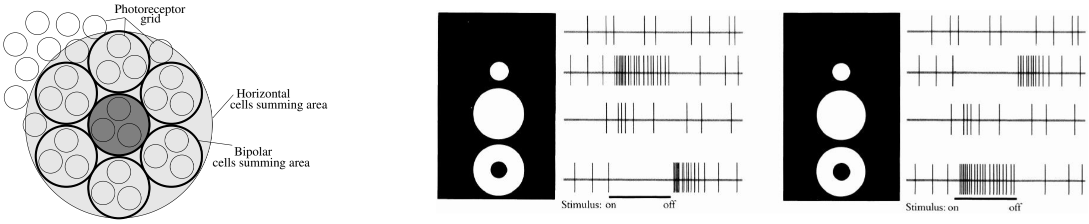


Figure 2.2.: *Left:* Spatial arrangement of receptor cells and summing area of retinal ganglion cells (RGC) and horizontal cells in the retina. *Center:* Response of on-center off-surround RGC to different stimuli. *Right:* Responses of an off-center RGC (center and right figures are adapted from Hubel (1995))

2.1. The Retina

The neural signal which leaves the retina consists of trains of impulses carried by the axons of retinal ganglion cells. One can define the response of the retina to visual stimulation as the change of rate in the firing impulses. Since the work of Hartline (1940), Barlow (1953), and Kuffler (1953) one knows that each retinal ganglion cell generate responses to stimulation over a limited area of the retina, and this area is defined as the receptive field of that ganglion cell. Kuffler (1953) found that the receptive fields of cat retinal ganglion cells consists of two concentric zones which he called the center and surround. The center and surround were mutually antagonistic[1]. In on-center cells in which the center caused excitatory responses to increments of light, the surround would cause inhibitory responses to increments. In off-center cells in which the central region was inhibitory during an increment, the surround would be excitatory during and increment. The on- and off- center cells and their center-surround organization are illustrated in Figure 2.2.

Because of their design retinal ganglion cells are very good at spatial comparisons– judging which of two neighboring regions is brighter or darker. Our efficiency in doing this allows us to distinguish differences in the order of 2%. Most remarkable, this is mainly invariant to the level of illumination. The luminance is an objective measure of the amount of light emanating from a luminous source or reflecting object, weighted by the observer's spectral sensitivity function. Illumination can be expressed in terms of effective quanta of light per unit time per unit area of the surface on which the light is falling. Whereas the apparent brightness, which is our subjective sensation of how light or dark an object is, does not change, its illumination can change dramatically. For

luminance

[1] In terms of information processing a center surround receptive field performs a decorrelation procedure onto the input (Atick and Redlich, 1992).

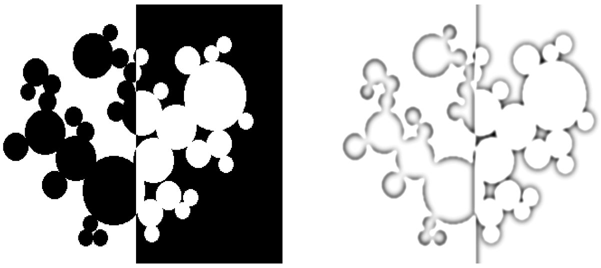


Figure 2.3.: Ganglion cell model output (*right*) for the image which is shown *left*. The simple model consists of a filtering of the input image by a difference of Gaussian function which models the center-surround structure of on-center off-surround retinal ganglion cells

example, when we read a sheet of paper in a room or in daylight we always perceive white paper and black letters, but black letters outdoor send twice as much light to our eyes as the white paper indoors. For us, the important thing is the amount of light *relative* to the amount reflected by surrounding objects.

Most neurons have a limited response range of a factor of one hundred from noise to ceiling. But they encode three to five log units of stimulus level. It follows that in order to achieve the illumination insensitivity the retina has to adapt to its input. The mechanism is to increase the contrast sensitivity and the contrast gain as the illumination increases, finally leveling off to asymptotic values in bright light. A summary of the visual adaptation and retinal gain control mechanisms can be found in Shapley and Enroth-Cugell (1984). Due to the above described mechanisms the contrast is the important quantity. This is nicely illustrated by the Cornsweet illusion (Figure 2.4 on the facing page). The contrast ramp in the middle of the picture defines our perceived brightness in the left and right halfs of the figure. The fact that we receive signals only where contrast is changing is again depicted in Figure 2.3 where the output of a ganglion cell model for a stimulus is shown.

Contrast

Contrast is a physical property of the visual stimulus (for example, a grating); it is the magnitude of luminance variation in the stimulus relative the average luminance. Contrast can be defined by two related formulas. The Rayleight contrast C_R is the mean-to-peak amplitude of the grating divided by the mean; the Weber contrast C_W or Weber fraction is the peak-to-peak amplitude divided by the luminance at the trough of the luminance profile.

$$C_W = (L_{\text{object}} - L_{\text{background}})/L_{\text{background}}, \quad C_R = (L_{\max} - L_{\min})/(2L_{\text{mean}}) \quad (2.1)$$

For low contrasts as used in most experiments both measurements are approximately the same.

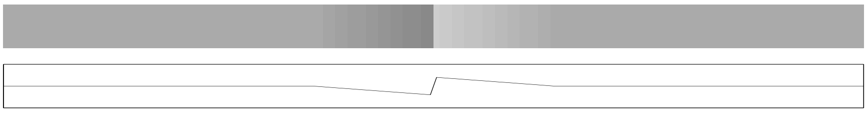


Figure 2.4.: Brightness depends on border contrast. This is an illustration of the Craik-O'Brien-Cornsweet illusion. The entire right half of the field is apparently brighter than the left half, yet the luminance of the two half fields are equal away from the border between them. Its luminance profile is drawn underneath the image

Adult humans are sensitive to a broad range of spatial scales ranging from very course scales (< 0.1 cycles per degree) to frequencies as high as 60 cycles per degree (Billock, 2000; Owsley, Sekuler and Siemens, 1983).

One should keep in mind that the retina is highly developed and performs numerous computations which we do not cover here. One example is lateral inhibition which performs in order to 'sharpen the edges' in the retinal image or separation of the visual information into the different channels, e.g., motion, form, and color. Knowledge about the statistics of the data can also be used to increase the spatial resolution of the retina (Ruderman and Bialek, 1992). In terms of efficient coding this can be understood as forming a compact code at the early stage of visual information processing. Only this pre-processing allows the retina to send its information coming from 125 million receptor cells through roughly one million fibers into the lateral geniculate nucleus.

2.2. The Visual Pathway and the LGN

Our eyes are directed in such a way that their areas of sight overlap. The visual pathway ensures that all points to the right of a vertical line through any point we are looking at is projected onto the left hemisphere.

The optic fibers coming from the retina cross and distribute at the optic chiasm before reaching the left and right lateral geniculate nucleus (LGN) (see the Figure on page 16). Fibers from the left half of the left retina go to the geniculate on the same side, whereas fibers from the left half of the right retina cross at the optic chiasm and go to the opposite geniculate. Similar, the output of the two right half-retinas ends up in the right hemisphere. This peculiarity of the brain that each hemisphere is dealing with the opposite site of the environment is found not only in vision but also in motor control or auditory cortex.

lateral geniculate nucleus

The LGN is a layered structure which receives topographically organized input from both retinae and projects to the cerebral cortex (see the Figure on page 16). It consists of several layers of neurons separated by intervening layers of axons and dendrites. The 1.5 million cells comprising the layers are of different types. The so-called magno-(=large) cellular and parvo-(=small) cellular cells are believed to be the anatomical segregation of the pathways conveying form and movement signals. The layers are further distinguished into having different input (from ipsi-lateral and contra-lateral eyes), and different receptive field organization (ON- and OFF-types of cells) which is inherited from the retinal organization.

Contrast-response curves from the LGN and cortical potentials are quite different from those for the retina in that way that amplitudes increase approximately linearly with log contrast over a 2-log-unit range (1 to 100%) (Ohzawa, Sclar and Freeman (1985), cat data). But apart from the contrast normalization we will assume the LGN as behaving as a simple relay station of information coming from the retina to the primary visual cortex. This picture may be wrong because the LGN receives much more back-projecting fibers from the cortex than input fibers from the retina. It is not know yet what kind of information processing happens in this region.

2.3. The Primary Visual Cortex

The primary visual cortex corresponds to Brodmann's area 17 at the posterior tip of the brain. It is also known as *striate cortex* because of the highly distinctive layering structure that shows up in a Nissl stain (which marks cell bodies only). Because of its location on the the upper and lower lips of the calcarine ("spur-shaped") sulcus, the striate cortex is also known as the *calcarine cortex*.
V1 Yet another name is *visual area one* or *V1*.

Neurons arranged vertically to the surface of the cortex (neuronal columns) often have common properties (Hubel and Wiesel, 1962). Neuron in one column have, for example, overlapping receptive fields which correspond to the same region of the retina of one eye.

A carful estimation of the size of a cortical column of the monkey brain was given by Peters and Sethares (1996) (based on clusters of apical dendrites). The modules are spaced at an average center-to-center distance of $56\mu m$ and contain in the order of 200 neurons.

The neuronal columns for one eye are grouped together and form (dependent on the species) stripes or patches which are known as *ocular dominance* columns. Within the ocular dominance columns sub-columns of neurons which are sensitive to particular orientations in space – known as *orientation columns*, can be found. To complicate matters further, color-sensitive columns known as blobs pierce the centers of the ocular dominance columns resulting in inter-blob neurons which are orientation-sensitive, and blob neurons, which are not. Moreover, blobs are specialized on the basis of wavelength (”color”) sensitivity. Neurons are also selective to the spatial frequency of the stimuli (Das and Gilbert, 1999). Across the surface of the cortex the orientation preference, response latency and temporal frequency vary systematically. With respect to other parameters, as for example the spatial phase of the stimulus no systematical changes are observed (DeAngelis, Ghose, Ohzawa and Freeman, 1999).

In the later chapters we will focus on additional features primary neurons may respond to in order to capture the structure of natural images. What is common between all found features maps is the restricted focus of the neurons in retinal space (termed *receptive field*). The maps are formed in a topographic manner retaining spatial relationships. In summary the cortex appears to be a substrate of interweaved, more or less topographically organized feature maps.

2.3.1. Classical Receptive Field Measurements

By stimulating the visual field with random dot pattern one finds a position in retinal space for which a neuron responds. A small stimulus is centered at this position and during an increase of the stimulus diameter the response of the neuron is measured. The point where the response of the measured cell saturates or starts to decline (see end–stopping on page 31) defines an area which is termed the *classical receptive field* of that neuron.

classical receptive field

One should keep in mind, firstly, that the concept of the classical receptive field is not as well defined as it seems. There are other measurement approaches by which the receptive field may be defined differently. For example: use initially large stimulus sizes and shrink the diameter up to the point, where the response of the cell starts to decline. Secondly, the classical receptive field is stimulus dependent, neurons were reported which strongly respond to center–surround stimulus configurations where neither stimulus component alone was effective (Sillito, Grieve, Jones, Cudelro and Davis, 1995). In Section 2.3.4 on page 29 we list some more observed context effects. One has to be careful in comparing the different experiments in terms of the type of animal used and also about the specific measurement protocol because it was demonstrated that receptive field properties can change through time (Gilbert and Wiesel, 1990; DeAngelis et al., 1995).

Oriented Stimuli

The majority of neurons in the primary visual cortex is known to respond best to oriented bars or gratings at certain positions in visual space, they are orien-

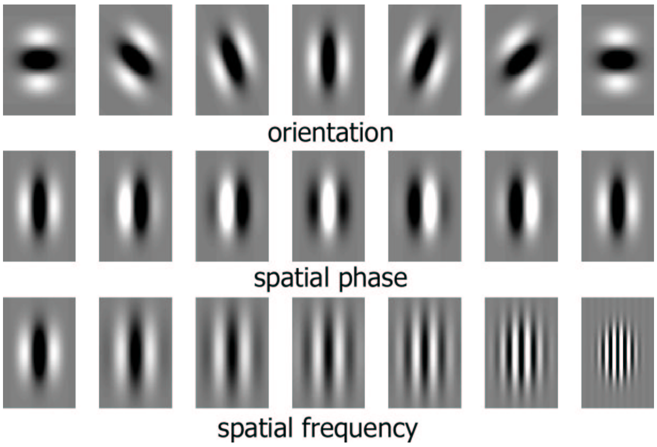


Figure 2.5.: Cortical simple cells respond to and can therefore be defined by their preference to stimulus orientation, spatial phase and spatial frequency

tation selective[2] (Hubel and Wiesel, 1962). Depending upon how they respond to grating stimuli they are classified as either simple or complex cells. For simple cells the cells' response depend on the stimulus in an approximately linear fashion. Troyer, Krukowski and Miller (2002) observed that the input to a simple cell obtained for a drifting grating stimulus can more exactly be described as the sum of two terms, a linear term and a non-linear term. The linear term represents the temporal modulation of the input, and the non-linear term represents the mean input which grows with stimulus contrast. The sensitivity of simple cells depends also upon the spatial phase (position) of the grating. Valois, Albrecht and Thorell (1982) and de Valois and de Valois (1988) found that at each eccentricity the human visual system is sensitive to a spatial frequency range of three to five octaves. In a paper based on these findings Lee (1996) derived a family of self-similar 2D Gabor wavelets that are suitable to model and analyse the linear characteristics of simple cell receptive fields.

simple cells

complex cells

Complex cells' responses are insensitive to the spatial phase of the stimulus. The distinct orientation which elicits maximum response is called the preferred orientation of that neuron. Orientation preference is often measured as the half–width at half height of the orientation–tuning curve (see Figure 2.6 on the facing page).

[2] The feature of orientation selectivity depends on the species and on the neuronal layer. In cat the neurons in layer 4 are orientation selective, in monkey they are not.

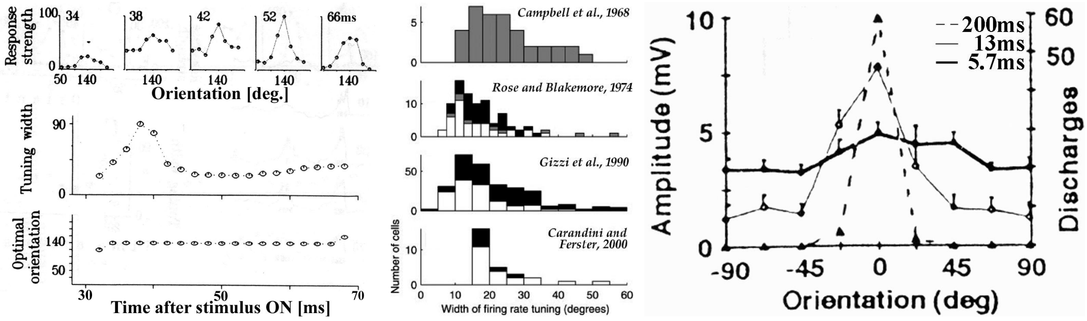


Figure 2.6.: *Left:* Preferred orientation remains constant but selectivity sharpens (figure adapted from Volgushev et al. (1995), cat data). *Center:* Orientation tuning width of simple and complex cells in cat primary visual cortex (figure adapted from Carandini and Ferster (2000)). *Right:* (adapted from Pei et al. (1994))

Response of Cortical Neurons to Contrast

The response of cortical neurons to different levels of contrast is analog to that which has been proposed for retinal light adaptation (see Section 2.1 on page 17). Contrast adaptation allows cortical neurons to maintain a high differential sensitivity to changes in contrast of a stimulus despite the limitations of a restricted response range (Sanchez-Vives, Nowak and McCormick, 2000). However, it is worth noting that some cells show no adaptation behavior. Thus, information about the overall absolute contrast may be transmitted to the cortex. It is not clear if this information is really used for recognition (see Figure 1.3 on page 10). The contrast–response curves can be described largely by a thresholded linear function that saturates well below the maximum firing rate of the neurons or even declines for high contrasts (called super-saturation) (Albrecht and Hamilton, 1982). In Section 3.2.1 and 3.2.2 we will show, how contrast-saturation can be understood as a network effect.

Signaling of Multiple Orientations

Numerous models have been proposed for the generation or the sharpening of orientation selectivity inside V1 (Somers, Nelson and Sur, 1995; Carandini and Rigach, 1997; Ben-Yishai, Hansel and Sompolinsky, 1997; Bartsch, Stetter and Obermayer, 1997; Stetter, Adorján, Bartsch and Obermayer, 1998). They are based on (i) excitation and inhibition arranged in a Mexican-hat like fashion enforcing long–range or global inhibition and more localized excitation. This results in lateral inhibition and is used for a mechanism to sharpen orientation–tuning curves. The model assume (ii) a high intracortical coupling strength to achieve the observed sharp orientation tuning from an initially broad tuned thalamic input. This enforces high competition, a winner–take–all strategy of lateral inhibition and therefore signaling the presence of multiple orientation is difficult (Carandini and Rigach, 1997; Bartsch et al., 1997). Carandini and Rigach (1997) has pointed out that one could test the assumptions made by measuring the response of neurons in V1 to stimuli inside the classical receptive field composed of different orientations. There is little data known about this. DeAngelis, Robson, Ohzawa and Freeman (1992) measured the response of V1 neurons for stimuli that consists of two superimposed gratings at the size of the classical receptive field. They found a reduced activation compared to a single grating at optimal orientation. This is an argument against a simple linear summation.

tilt-illusion

Attraction and repulsion between orientations are observed psychophysically as tilt–illusions and are also reported physiologically in V1 (Gilbert and Wiesel, 1990). It is known that the ability of animals and humans to carry out perceptual tasks, such as discrimination of two similar stimuli, improves with practice for that specific direction only, not for substantial different orientations or special frequencies (Walk, 1978). This suggests that learning is due to changes at early stages of the sensory pathway, where stimuli characterized by very different parameters are represented by different neurons (Mato and Sompolinsky, 1996).

The problem of generating both sharp contrast–invariant orientation tuning and a reliable signaling of multiple orientations can probably be solved by the idea of Adorján, Schwabe, Piepenbrock and Obermayer (2000) to incorporate a rapid change in cortical connection strength during a fixation period ($200ms$) implemented by fast synaptic depression (see Schwabe, Adorján and Obermayer (2000) for a similar ansatz that uses spike–frequency adaptation). Initially, high competition is used to extract first the salient features (by high recurrent coupling) and in a second less competitive phase the precise signaling of multiple orientations becomes possible. It is argued that this could be optimal in terms of information transfer. At the beginning of a fixation period only a limited number of spikes are obtained so the signal–to–noise ratio is to high to reliably detect multiple orientations. Most information can be extracted therefore at the beginning by attempting to extract only one orientation with the robust high competition regime. In later phases more spikes are collected, the signal–to–noise ratio gets better so it is possible to detect multiple orientations.

It has to be shown to what degree dynamic coding is really implemented because there are other ways to cope with high signal–to–noise ratios like coding by a population response. Tsodyks and Sejnowski (1995) demonstrated that instead of integrating over time it is efficient to integrate over the response of many similar neurons which can reflect changes in the stimulus conditions nearly instantaneously.

Another approach to understand the conditions by which models can encode multiple orientations was proposed by Zemel and Pillow (2000). They optimized the recurrent weights so that the intra-cortical computation results in a stable activity profile that resembles a convolution of cell responses to different orientations present in the stimulus. The obtained connectivity profile closely resembles a Mexican-hat like function as most models assume. It was hypothesized that the fine structure of the profile causes the enhanced ability of the model to encode multiple orientations. Unfortunately their model also works nearer to the linear phase (reduced intra-cortical connectivity strength) as the model by Carandini and Rigach (1997) which has the drawback of less pronounced sharpening and more contrast dependent orientation tuning (cf. Section 3.2.1 on page 45).

2.3.2. Complex Stimuli in the CRF

Searching for optimal or nearly-optimal stimuli is traditionally performed manually using a limited set of stimuli. One has to assume that the relevant parameters of the stimulus are known in advance. The selection of the relevant parameters by the experimenter and the search procedure itself introduces an element of subjectivity into such experiments. Because in higher sensory areas even the relevant parameters are unknown a more effective approach is needed.

The assumed relevant stimulus parameters of simple cells in primary visual cortex are orientation and spatial frequency. This motivates many functional models of the visual cortex to use 2D Gabor wavelets to model the receptive field of linear visual cortical neurons (simple cells). Gabor wavelets provide the best trade-off between time resolution and frequency resolution (Gabor, 1946). The validity of this ansatz was verified by careful mappings of the receptive field of the simple cells by Jones and Palmer (1987).

Measuring Receptive Fields by Reverse-Correlation

In their experiment simple cell responses were measured with a micro electrode. The receptive field of a certain cell was measured location for location by projecting a dot-like stimulus on a homogeneous screen the corresponding eye looks to. The method is called *reverse correlation* and measures the spike triggered average response in the presence of a white noise stimulus (deBoer and Kuyper, 1968). The estimate is the best *linear* model explaining the firing rate given the stimulus (see the excellent book of Dayan and Abbott (2001) on this topic). The reverse-correlation can be used to obtain the most

reverse correlation

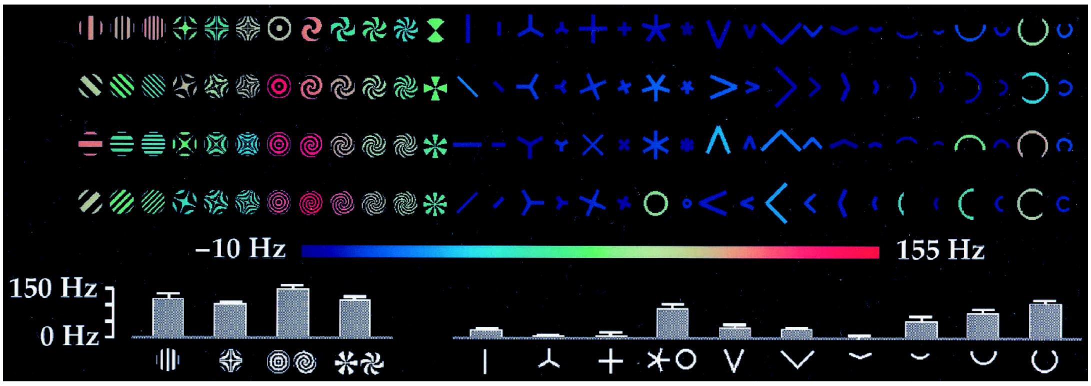


Figure 2.7.: Response of a V2 cell to grating and contour stimuli. Color-coded mean response of an individual cells to 128 stimuli. Stimulus orientation is normalized. The *bar plots* at the *bottom* of the panel show the mean responses + SEM of the given cell to the most effective stimuli. Image from (Hegdé and Essen, 2000)

effective stimulus in that it relates the optimal kernel for firing rate estimation to the stimulus. Given a stimulus with constant energy the most effective stimulus is one that is proportional to the optimal linear kernel. Because of the ansatz reverse-correlation can be computed off-line but is limited to linear or nearly-linear neurons.

Measuring Receptive Fields by Gradient Ascent

gradient ascent method

There is an interesting alternative that should also work for non-linear neurons. The gradient ascent method can be applied, in principle, for neurons in higher cortical areas (Földiàk, 2001). Here, starting with a blank stimulus, white noise is added to the (rapidly changing) stimulus and the change in responds is measured on-line. The stimulus is moved in the direction of larger responses (duration of the experiment is 5- to 10 minutes). Simple cells show the expected bright and dark elongated regions. For complex cells repeated optimization runs result in similar power spectra but the stimuli cannot be aligned pixel-by-pixel indicating the non-linearity of complex cells. The method produces, after initial symmetry breaking, locally optimal solutions. Interestingly these local solutions could not be fitted well by simple 2D Gabor functions. This indicates problems with the idea that complex cells are built by convergent input from simple cells.

Measuring Receptive Fields by Specific Stimuli

Other studies also indicate that relevant features of neurons early in the visual pathway cannot be described by their preference to oriented bars or gratings

Table 2.1.: Table summarizing the most prominent cell types found in primary visual cortex. Neurons are assumed as being inhibitory if they are fast spiking, smooth stellate, and with beaded dendrites. Data are taken from McGuire et al., 1991, Lund and Wu, 1997, Anderson et al., 1993, and Azouz et al., 1997

cell type/layer	+/−	connection type	extend in mm	literature m-macaque, c-cat
pyramidal	+	long-range steppy	$400/3000$	McGuire 1991 (m)
spiny stellate	+	steppy	$150/440$	Lund 1997 (m)
chandelier/2	–	local	$150/200$	Anderson 1993 (m)
small basket/2	–	local	$200/800$	Azouz 1997 (c)
large basket/3	–	local with extens.	$370/1200$	Azouz 1997 (c)
clutch/4	–	local	$270/460$	Azouz 1997 (c)
smooth stellate/3	–	local with extens.		(c)

alone. We explicitly talk here about the neurons in V1 and V2 and we are aware that neurons in later stages of the visual processing are known to be highly specific in their responses to faces or objects.

As found by Hegdè and Van Essen (Hegdé and Essen, 2000; Hegdé and Essen, 1999) most neurons in areas V2 and V1 show stronger responses to complex stimuli than to the optimal grating stimuli presented in their receptive field (see Figure 2.7). As complex stimuli Lie-figures where used. For a more detailed description of the generation of Lie-figures see Section 4.1.3 on page 85. Results in cat V1 found by Shevelev (1999) indicate that around 40% of all neurons studied ($114/289$) gave larger response to a flashed cross, corner or y-like figure centered in the receptive field to an optimal single bar. Various forms of selectivity or invariant sensitivity of neurons to the shape and orientation could also be observed by Versavel, Orban and Lagae (1990) and Dobbins, Zucker and Cynader (1987).

2.3.3. Anatomy of Lateral Connections in V1

A large part of the problems listed in the last section appear because of our little knowledge about the precise couplings between neurons in the visual path. This is in part because of the methodological difficulties involved in tracing neurons over long distances (as for example from the LGN to the visual cortex). It is easier to analyse the anatomical connection in one cortex

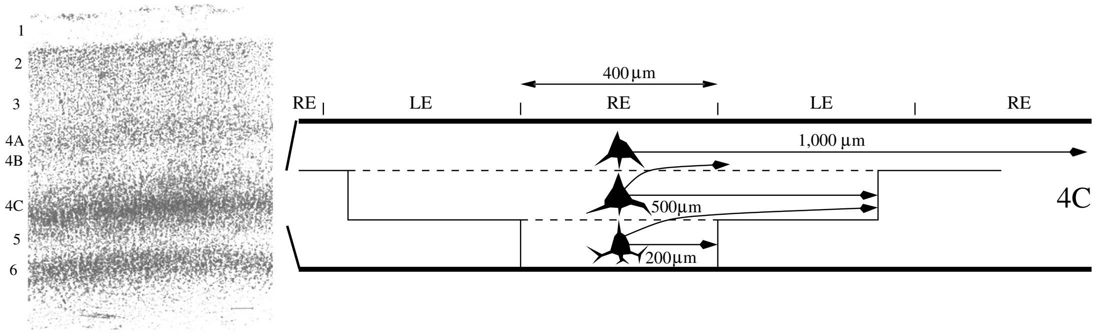


Figure 2.8.: *Left:* Cell body density of the layers of V1 (Nissl staining). *Right:* Lateral spread of axonal projections within 4C. Spiny stellate cells in mid 4C have a mean step size of $250 - 650\mu m$. The step sizes are comparable to the wider projections in upper 4B

areal. Especially the primary visual cortex was studied intensely. One find various types of neurons defined by their response characteristics and on their morphology (spiny = excitatory, smooth = inhibitory). The main groups are excitatory pyramidal cells that together with the more locally connecting spiny stellate cells (also excitatory) constitute 80% of the overall number of cells in the cortex. Inhibitory neurons are more diverse and range from large basket cells to more local chandelier cells.

Whereas the spiny stellates and the basket cells are found to be isotropic, connecting to all neurons in their vicinity the picture gets more interesting in the case of the pyramidal neurons. Lateral connections in layers $1 - 3$ primary visual cortex (macaque monkey) form patch-like terminal zones, and link together neurons sharing common physiological properties. The patches are 200-$300\mu m$ in diameter separated by gaps of similar with and run the full depth of layers 1-3 (Rockland and Lund, 1983). The overall region in which connections are formed is found to be elongated with an average aspect ratio of $1.8 : 1$ and a long axis measuring up to $3.7mm$ (Yoshioka, Blasdel, Levitt and Lund, 1996). Correlation of these zones with optically imaged maps it has been shown that these connections predominantly, but not exclusively, link together points of similar orientation preference, ocular dominance and CO rich or poor compartments (Yoshioka, Blasdel, Levitt and Lund, 1992; Yoshioka et al., 1996; Malach, Amirr, Harel and Grinvald, 1993).

Interestingly, the functional properties in deeper layers of the cortex differ from the ones in the upper layers. Yousef, Bonhoeffer, Kim, Eysel, Tót and Kisvárday (1999) quantitatively analyzed the degree of orientation selectivity of long-range intrinsic connections with respect to the different cortical layers. Using a combination of

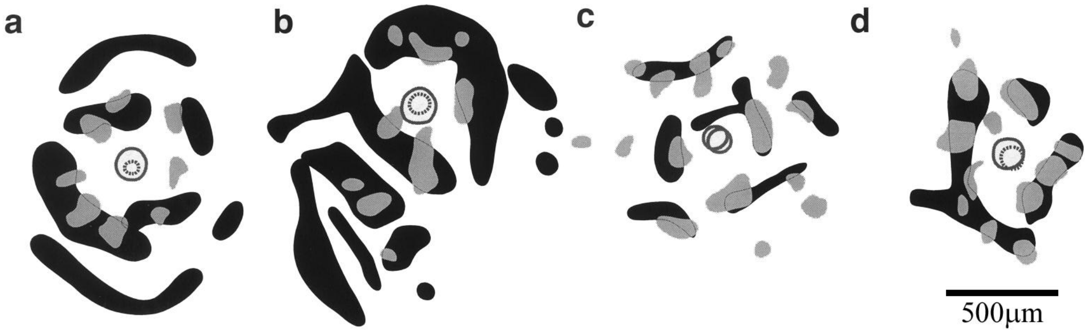


Figure 2.9.: Spatial layout of cells (*a-d*) in macaque primary visual cortex V1. Spatial relationship between bar-shaped terminal fields in layers $4B$-upper $4C\alpha$ (*dark gray stripes*) and patch–like terminal zones in overlying layers $2-3$ (*light gray patches*). Data courtesy by Alessandra Angelucci

optical imaging and injections of both latex micro-spheres and biocytin they analyzed connections in supra-granular, granular, and infra-granular layers of *cat* area 18. Layer 4 lateral networks are found to be in general much shorter (about 50%) than layer 3 networks and display a less clear patchy pattern. Moreover, long range ($> 500\mu m$) connections in layer 4 were distributed almost equally across orientations (iso, 35%; oblique, 34%; cross, 31%), suggesting that the long-range layer 4 circuitry has a different functional role from that of the iso-orientation biased layer $2/3$ circuitry.

Asi, Lund, Blasdel, Angelucci and Levitt (in press) found in macaque monkey that the lateral connections in the deeper layers $4B$-upper and $4C\alpha$ predominantly form bar-shaped terminal fields. These terminal zones have a mean width and length of about 230 and $1050\mu m$, respectively, and are separated by $250\mu m$-wide gaps. The overall labeled field was found to be anisotropic (on average $2.7 \times 1.8mm$). By optical imaging of intrinsic signals it was demonstrated that the labeled regions cover equal areas for either eye, and show a bias in orientation preference. Additional columnar tracer injections involving layers $1-4C\alpha$ reveal an alignment of upper layer terminal patches with lower layer terminal stripes, suggesting a coherent columnar framework, despite laminar differences.

2.3.4. Non-classical Receptive Field Measurements

In the last couple of years some efforts have been made to understand the development and the distinct role of orientation selective receptive fields in

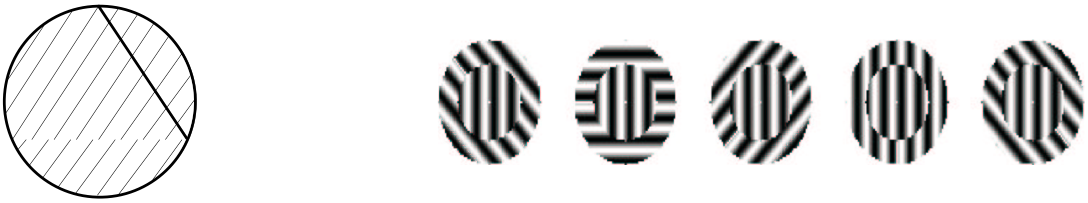


Figure 2.10.: *Left:* Illustration of the context dependency of recognition. There is a triangle inscribed in the circle, one edge is seen and there, one edge is not there but seen and one edge is there but cannot be seen. *Right:* Different configurations of center- surround stimuli. The orientation of the surround stimulus can alter the center response

processing visual information. It was found that the response characteristics of these neurons are dependent on the specific form and size of the stimulus (see Figure 2.5). Surprisingly even stimuli outside the receptive field (see Figure 2.10) of the neurons can alter their response characteristics of neurons. In cat, for example, the receptive field of a neuron near the area centralis is of $\approx \; 2^\circ$ of visual angle. The firing responses however can be modulated by the concomitant stimulation of a surround region up to 10° of relative eccentricity (DeAngelis, Freeman and Ohzawa, 1994; Bringuir, Chavane, Glaeser and Frégnac, 1999).

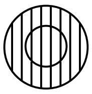
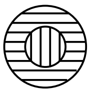
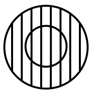
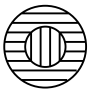
A number of studies in monkeys (Kapadia, Ito, Gilbert and Westheimer, 1995; Sillito et al., 1995; Levitt and Lund, 1997) and cats (Blakemore and Tobin, 1972; Gilbert and Wiesel, 1990; Polat, Mizobe, Pettet, Kasamatsu and Norcia, 1998) have analysed this phenomenon.

contextual effects

Whereas a surround stimulus alone is unable to evoke a response, it can considerably modulate the response of a cell to a stimulus within its classical receptive field. In other words, the response of the cell to a local feature depends on the visual context into which this feature is embedded[3]. Therefore, this class of phenomena is often referred to as contextual effects. Examples for the stimuli used can be seen in Figure 2.10, right. The corresponding figure left shows an *illusory* figure connected with the phenomenon of the different perception of lines depending on local context. Hidden in the figure is a equilateral triangle. Depending on the context the lines of the triangle are clearly visible (heavy line), visible but not present (illusory edge), or not visible but

[3] Sillito et al. (1995) reported neurons which respond for stimulus configuration. According to this a non–optimal oriented center stimulus can elicit a response when it is presented together with a stimulus in the surround.

Table 2.2.: Table summarizing the findings concerning facilitation and suppression depending on the orientation in the non-classical receptive field region

low center contrast	 	• iso-facilitation (Toth et al., 1996, cat) • collinear facilitation (Polat et al., 1993/94/98, cat) • in some cases at all directions suppression (Levitt et al. 1997, macaque) • cross-suppression at all contrasts (Polat et al. 1996, cat)
high center contrast	 	• iso-suppression (Toth et al. 1996, cat, Levitt et al. 1997, macaque) • in other directions/motion suppression often disappears (Levitt et al. 1997, macaque) • in some cases orthogonal surround where most suppressive (Levitt et al. 1997, macaque) • orthogonal flanks reduce the response (Polat et al. 1998, cat) • 80% contrast $2/5$-th showed facilitation (Polat et al. 1998, cat)

present (line hidden in many parallel lines).

The response to stimuli presented within the receptive field can be facilitated or suppressed by surround modulation. For high contrast surround stimuli that matches the preferred orientation the response of the center neuron is reduced. This effect is called *iso–orientation suppression*. For orthogonal orientation, the response is slightly (Levitt and Lund, 1997) or strongly (Sillito et al., 1995) facilitated. This effect is called *cross-orientation facilitation*.

end-stopping

A related effect is called *end–stopping*. For growing stimulus size, the response of a neuron reaches its maximum when the stimulus size fits the dimensions of the classical receptive field. If the stimulus size further increases in length or width, it begins to cover part of the non-classical surround of the cell and causes a suppressive effect.

If the center contrast is low, contextual effects can change their characteristics. Some studies in cats (Polat et al., 1998) and monkeys (Kapadia et al.,

1995) report that iso-orientation suppression turns into facilitation for low center contrast, which is consistent with a fill-in-paradigm. Other studies (Levitt and Lund, 1997) report cross-orientation facilitation to turn into suppression.

Contextual effects depend also on the geometrical properties of the stimulus in the non-classical surround. If the center stimulus is accompanied by two small flanking bars co-aligned with the central oriented stimulus, iso-orientation facilitation has been observed (Polat et al., 1998; Kapadia, Sigman and Gilbert, 1999a).

2.3.5. Texture Segmentation and Line Completion

Reducing the response in presence of an iso–oriented annular surround stimulus might be a possible mechanism for texture-based segmentation, where contour is defined by an abrupt change in the orientation of an elongated texture. Figure 3.12a on page 60 shows an example for a visual scene, in which the boundary between two extended gratings with different orientations pops out. One possible mechanism for the amplification would be an increased response of orientation-selective neurons with receptive fields near the border. Those neurons would see different orientations within and outside their classical receptive fields, and would have an increased response. In contrast, neurons far away from the border would see the same orientation within and outside their classical receptive field and their responses would be decreased by iso-orientation suppression.

But in a different stimulus paradigm, namely if the non-classical receptive field is stimulated by small flanking grating patches or bars instead of full annuli, iso-oriented surround stimuli can also facilitate the response of a neuron (Sengpiel, Sen and Blakemore, 1997; Polat et al., 1998; Kapadia et al., 1999a). This observation, which apparently contradicts the previous findings, could serve as the physiological basis of line-completion, which is schematically illustrated in Figure 3.12b on page 60. Line segments which are aligned are perceptually linked together to parts of a continuous contour, and we perceive an interrupted circle.

A possible anatomical substrate mediating this interactions is orientation specific long-range connections formed by excitatory pyramidal neurons (Rockland and Lund, 1983). Models incorporating these patchy connections show that non-classical receptive field effects can be mediated by these fibers (Pawelzik, Ernst, Wolf and Geisel, 1996; Mundel, Dimitrov and Cowan, 1996; Todorov, Siapas and Somers, 1996; Bartsch et al., 1997; Stetter, Bartsch and Obermayer, 2000b; Bartsch, Stetter and Obermayer, 2001).

A more technical ansatz that models contextual effects was presented by Li (1998) and Li (1999) to explain visual segmentation and contour integration. Visual segmentation can be defined as locating the boundary between different image textures. Contour integration means grouping of local contours into boundaries that may represent underlying objects. Experiments in V1 show that activity levels near simple texture boundaries are robustly higher $10-15ms$ after the initial cell responses (Gallant, Essen and Nothdurft, 1995).

The model uses hypercolumns that are arranged on a discrete grid. Each orientation column is modeled as a coupled pair of an excitatory and an inhibitory neuron. Stimuli were presented to the excitatory neurons only as orientations which are translated into a Gaussian like tuning curve for the hypercolumn. Lateral connections in the model depend on the position of the neurons, on their orientation, and on the type of the pre-synaptic neuron (exc. or inh. neuron). Iso–orientation excitation was implemented explicitly by excitatory connections only to neurons with receptive field positions aligned to the preferred orientation (within a certain angular distance). Also cross–orientation suppression was built in by the coupling of excitatory neurons to inhibitory neurons with receptive field positions that are orthogonal to their preferred orientation (within a certain angular distance).

The connection scheme coincides with observed orientation specific excitatory long–range connections. Long–range cross–orientation connections are not found anatomically, but the large, in comparison to excitatory pyramidal neurons, connection radii of inhibitory basket cells could mediate effective cross-orientation inhibition. In this respect and by using an idealized grid of hypercolumns the model (Li, 1998; Li, 1999) is a phenomenological model. It can not fully explain the emergence of contextual effects by the observed biological circuitry, but it lightens the role of cross–orientation suppression and iso–orientation facilitation in contour integration and texture segmentation tasks performed at early stages of visual information processing.

To understand the role of long–range connections for the formation of contextual effects in V1 we summarize now some approaches. Section 3.3 on page 69 will show that signaling cascades formed by short–range connections only are able to mediate contextual effects.

2.3.6. Origin of Orientation Selectivity

Orientation selective neurons in primary visual cortex show a half-width at half-height tuning of $23^\circ \pm 8^\circ$ ((Carandini and Ferster, 2000), cat data, see Figure 2.6, center). The origin of orientation selectivity is still under debate. The relay cells in the LGN are known to have only moderate elongated receptive fields (mean aspect ratio of 1.26 (cat) (Soodak, Shapley and Kaplan, 1987),

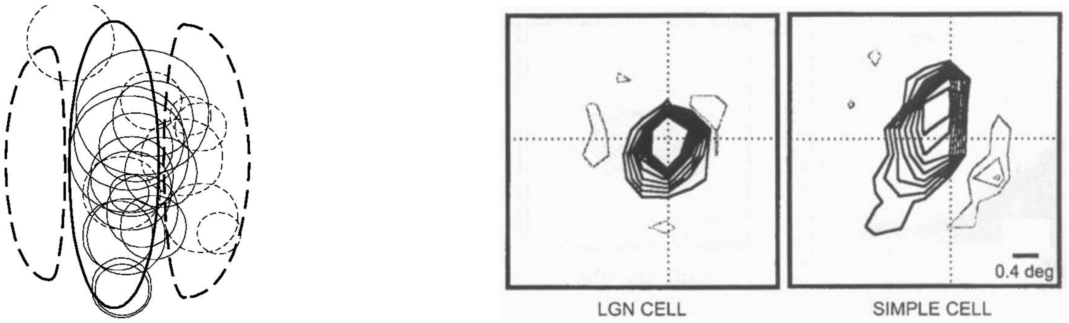


Figure 2.11.: *Left:* Wiring from thalamus to cortex in cat is very precise. The thick lines correspond to the on- and off region of a simple cell in primary visual cortex. The circles represent receptive field centers of geniculate relay cells (type depicted by dashed or solid lines) that fired highly correlated to the simple cell indicating a mono-synaptic connection between them (Reid and Alonso (1995)). *Right:* Mono-synaptically coupled LGN and simple cell with overlapping receptive fields (dotted line). Figure adapted from Alonso et al. (2001)

1.62 (ferret) (Tavazoie and Reid, 2000). So orientation selectivity in V1 can be built by either convergent thalamic input (feed-forward) or by intra–cortical connections (see Figure 2.11).

Simple cells in V1 resemble many features of those of their relay cells, like the sizes of simple cell subfields (Reid and Alonso, 1995), or that responses fall naturally into the same categories as relay cells, including X and Y, or lagged and non–lagged, which supports a dominant input from LGN. Alonso et al. (2001) reported furthermore a surprising precision of connectivity that goes beyond simple retinotopy to include many other response properties, such as receptive-field sign, timing, subregion strength, and size (see Figure 2.11, left). Geniculate cells provide synaptic input to simple cells only when their receptive fields spatially overlap a simple receptive field and match the sign (on or off) of the overlapping subregion (Reid and Alonso, 1995; Alonso et al., 2001). Also there is evidence that cortical inactivation does not change orientation specificity significantly (Chung and Ferster, 1998). For a more complete review on experimental support for the feed–forward model see also Ferster and Miller (2000) and Reid and Alonso (1996).

Objections against a purely feed-forward origin of orientation selectivity comes from Carandini and Ferster (2000). By measuring the orientation tuning of the membrane potential of cat primary visual cortical neurons they found that a substantial sharpening takes place by the spike threshold. Membrane potential itself is only broadly tuned to orientation ($38 \pm 15°$).

Feed-forward models also fail to explain important feature of orientation selectivity, for example, its invariance against changes in stimulus contrast (Sclar and Freeman, 1982; Skottun, Freeman, Sclar, Ohzawa and Freeman, 1987). Furthermore, they do not assign any function to intracortical circuitry, which is known to be much stronger (85%, in number of synaptic connection) than the feed-forward circuitry (15%). In Section 2.3.7 on page 38 we will discuss two models which address a purely intracortical origin of orientation selectivity.

So there are good reasons to believe in both, a feed–forward and an intracortical part of orientation selectivity. It was suggested that the recurrent excitatory circuits of the cortex may amplify an initial feed-forward thalamic signal, sub serving dynamic modifications of the functional properties of cortical neurons (Stratford, Tarczy-Hornoch, Martin, Bannister and Jack, 1996; Carandini and Ferster, 2000). Based on this assumption in Section 3.2 on page 41 we present the evolution of a model that incorporates weakly orientation tuned thalamic input and strong cortical interactions.

2.3.7. Iceberg-Model

One important early model for orientation tuning, first formulated by Hubel and Wiesel (1977), explains orientation selectivity as directly derived from thalamic input in a feed–forward fashion. It has some serious flaws in that is for instance cannot explain the contrast invariant orientation tuning but we include it here because it is one of the most intuitive models. The receptive fields of simple cells consist of elongated subfields with alternating ON- and OFF-response. The Iceberg-model proposes that orientation selectivity is generated by filtering oriented input with the receptive field profile (calculating its overlap with the profile) and feeding the result through a rectifying nonlinearity. The model assumes that orientation selectivity is purely generated by feed-forward processing of input (mediated by the feed-forward fibers from the retina over the LGN to a cortical neuron) and local processing within the neuron. In particular, no function is assigned to the intracortical circuitry as well as feedback-circuitry to the LGN. These connections are neglected in the Iceberg model.

Orientation selectivity is generated in the model framework as follows: Lets assume that the stimulus is an oriented structure such as a sine wave grating. If the orientation and the phase of the sine wave it aligned with orientation and phase of the receptive field profile the total synaptic input will be maximal. If the structures are orthogonal the total synaptic input is zero. This reflects a linear model for the generation of orientation selectivity (see Sec-

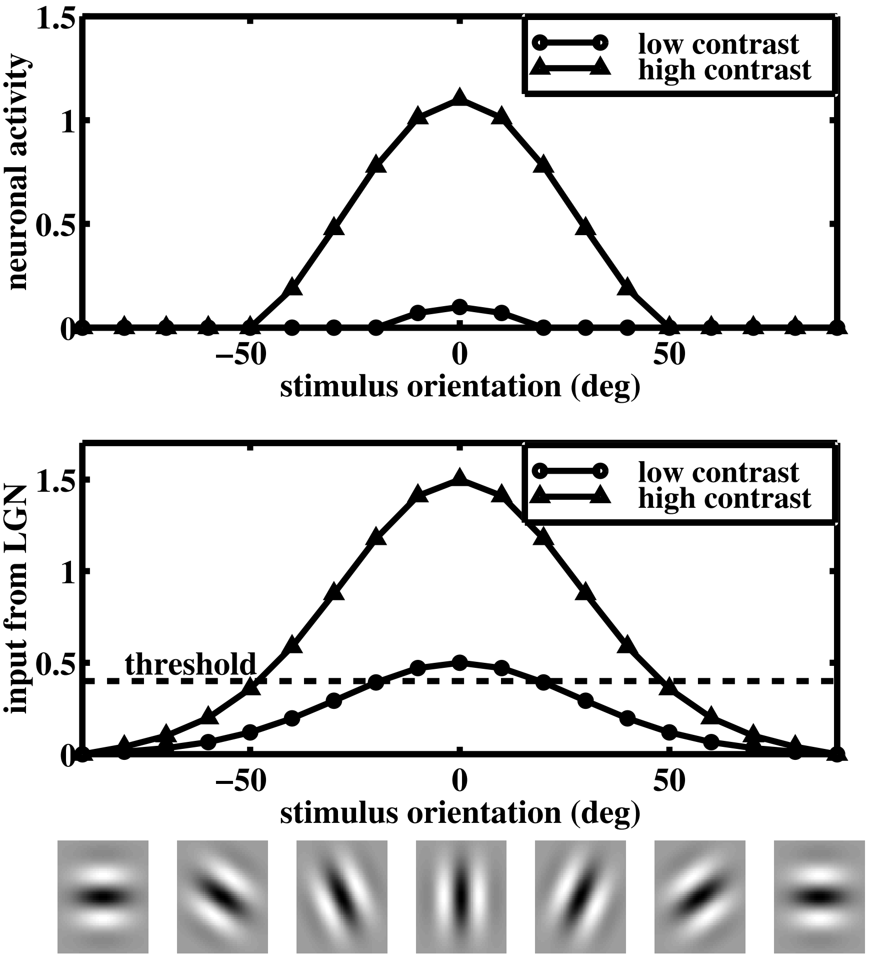


Figure 2.12.: The Iceberg-model of orientation selectivity. The output of the cortical neuron is a thresholded version of its input. The output is orientation selective (receptive field below), but its tuning width increases with stimulus contrast

tion 2.3.2 on page 25). The row at the bottom of Figure 2.12 on the preceding page shows different oriented Gabor-filters. In the shown example, the overlap between the stimulus and the filter is maximal for the middle stimulus and decreases with the difference between orientations. The resulting synaptic input for two different contrast levels of the stimulus grating is plotted in the diagram against the stimulus orientation.

The response in the top diagram of Figure 2.12 on the facing page is orientation selective, but nevertheless the Iceberg model has some serious drawbacks which motivate the consideration of more sophisticated models for cortical function:

- In biology, the sharpness of orientation tuning is independent of stimulus contrast (Sclar and Freeman, 1982). This independence might be important as it ensures that stimulus orientation and stimulus contrast are coded independently of each other in the cortex, which facilitates the independent readout of both quantities. The iceberg-model predicts a strongly contrast-dependent orientation tuning width.

- In biology the orientation tuning curves are sharper than predicted by the iceberg model, and the output of neurons is often more sharply tuned than their input (Volgushev, Pernberg and Eysel, 2000). If rectifying nonlinearities of LGN cells are taken into account in the model instead of the linear filter model assumed here, its simulated tuning curves become even broader.

- The iceberg model does not assign any function to intracortical circuitry, which in fact is even much stronger than the feed-forward circuitry. This leaves the question of which operations are performed by lateral intracortical connections.

In light of these observation, we face the necessity to build a more sophisticated model of cortical function which involves dense recurrent intracortical circuitry (Somers et al., 1995; Ben-Yishai, Bar-Or and Sompolinsky, 1995; Ben-Yishai et al., 1997; Bartsch et al., 1997; Hansel and Sompolinsky, 1998). A sufficiently large cortical circuit may contain millions of densely interconnected neurons, which cannot be modeled at a single neuron level because of a prohibitive computational expense. We face the task to formulate a model framework which allows a simplified yet valuable theoretical description of a large neuron population at a meso-scopic level. One important type of mesoscopic cortical description has been provided by mean-field models of cortical function which will be introduced in Section 3.2 on page 41.

Models for Intra-cortical Origin of Orientation Selectivity

In the hypercolumn model by Adorján, Levitt, Lund and Obermayer (1999), all cortical cells receive thalamic input from a circular symmetric region, hence the thalamic input to individual cortical cells does not provide an orientation bias (but see Tavazoie and Reid (2000)). Cells with receptive fields that are aligned in visual space are assumed to belong to one orientation column. They are activated together whenever there is a line at this position and orientation. In the model all these neurons are strongly connected thus forming dense connectivity inside a model column. By this mechanism the total geniculate input for a neuron depends on the orientation of the stimulus. This results in an orientation preference that corresponds to the orientation of the alignment axis of the receptive fields in visual space. As a testable prediction a complete loss of orientation selectivity is predicted by this model for blocked recurrent excitation.

Another recent approach to explain the emergence of orientation selectivity by a purely intracortical mechanism is (Ernst, Pawelzik, Tsodyks and Sahar-Pikielny, 2000). In their geometrical model intracortical connection strength is high enough to greatly amplify small random fluctuations which produces activity blobs following the maxima in the input. Furthermore, Ernst et al. introduces spatial inhomogeneities modeled as fluctuations of neuron positions. So some neurons at random directions are coupled more strongly than others and will therefore be more easily activated by a specific oriented stimulus. So the position of the activity blobs in the model is determined by both the maximum afferent input and the lateral coupling.

A further model using non-isotropic lateral connections for the intro–cortical generation of orientation preference was presented by Shouval, Goldberg, Jones, Beckerman and Cooper (2000). Already at birth some orientation selectivity is present and further development is then guided by visual experience, findings for this are: (i) Normal development of orientation selectivity can be prevented by rearing animals in the dark (Frégnac and Imbert, 1984; Chapman and Stryker, 1993). Most experiments have found that (ii) in a visual environment with a restricted set of orientations more cells become selective to the orientations prevalent (Sengpiel, Stawinski and Bonhoeffer, 1999). Nevertheless, Gödecke and Bonhoeffer (1996) have shown that two eyes without common visual experience develop similar orientation maps. To explain these contradictory results on the plasticity of orientation selectivity Shouval et al. (2000) assume that an–isotropic lateral connectivity already present at birth forms a scaffold that sets the orientation map, produces broadly tuned cells, and biases the development of orientation selectivity. By this an orientation map can be laid down independently of visual experience, and orientation selectivity then develops through experience–dependent modifications of the feedforward synaptic connections.

The current opinion is that an initially broadly tuned input arrives in cortex via convergent thalamo-cortical connections and is sharpened by intracortical connections. This results in a still broadly to orientation tuned membrane potential which is further sharpened by the spike generation mechanism (Azouz, Gray, Nowak and McCormick, 1997).

2.4. Discussion

Numerous models have been proposed trying to give answers to the possible origin of orientation selectivity. The success of different models using different assumptions about the origin of orientation selectivity is surprising. At least this indicates that the basic assumptions of all of these models must be somehow the same.

We suppose that a basic assumption made by many models is the use of a Mexican–hat like intracortical coupling function and thus induced lateral competition. Lateral competition moves a model from a regime where the neuronal activation is a linear resemblance of the input into a regime where a sparse representation of the input in the neuronal activation is enforced. The models also use a similar level of abstraction, for example, they describe the interaction of populations of neurons. The effect of the non-linear mechanisms of spike generation are left out.

In the proposed models the observed sharp orientation tuning obtained from an initially broadly or even untuned input can only be observed when one assumes relatively high intracortical couplings, e.g., for high competition. Also the contrast–invariance of orientation tuning only works for a high level of competition. The drawback of the regime of high level of competition is that it cannot represent multiple orientations, this works only in the linear, less competitive phase.

Incorporating more knowledge about the cortical circuitry, for example, by using a more realistic neuron model should help to distinguish between the models.

3. From Columns to Hypercolumns and Lattices

Modeling in the framework of mean–field descriptions have proved to be a powerful tool to analyse large networks. Here we show how to initially set up a mean–field model and link its parameters to an underlying model of populations of spiking neurons. The model is changed according to the subject under investigation.

By modeling more explicit different populations of neurons in a single column we arrive at a model that combines sharpening of orientation tuning curves, contrast invariant orientation tuning and saturating contrast response functions.

To explain contextual effects we extend this model first to a full orientation hypercolumn and afterwards to a system of two coupled orientation hypercolumns. One receives input from the center stimulus and the second hypercolumn from the surround. By this arrangement we show principle difficulties in having cross–orientation modulations by iso–orientation specific patchy connections.

We derive a model for analyzing the influence of local cortical connections on the activities of neurons in V1. This time a set of orientation columns is arranged according to a measured orientation map, and orientation columns are connected by local excitatory and slightly more distributed inhibitory fibers.

We demonstrate that (i) sharp and contrast–invariant orientation tuning curves are combined with contrast saturation, (ii) the strength of cortical amplification can be localized in orientation space and (iii) anisotropic contextual suppression by iso–oriented flanking stimuli arises as an emergent property and can be mediated by local connections.

3.1. Introduction

This chapter is grouped into three parts. First, we discuss some effects found in conjunction with the idea of the classical receptive field of neurons in V1. Models for a purely feed–forward or a purely intra–cortical origin of orientation selectivity are discussed. Step by step we introduce a model that combines weekly tuned feed–forward input and recurrent intra–cortical connectivity. This leads to a hypercolumn model with three different neuron populations that models the responses of neuron populations in orientation space. By this we explain orientation sharpening, contrast saturation, and contrast–invariant orientation tuning.

Part two deals with effects observed for complex stimuli in the classical receptive field like end–stopping or the cross–stimulus suppression. We point out the disability of models with high lateral competition (hight recurrent excitation) to represent different orientations simultaneously and refer to possible solutions.

In the third part of this chapter we will discuss contextual effects elicit by stimuli outside the classical receptive field. In a model consisting of two coupled hypercolumns we predict wiring patterns that can explain iso–orientation facilitation and cross–orientation suppression by long–range connections. Based on the predictions of the hypercolumn model a geometrical model is proposed that demonstrates that contextual effects can be mediated by short range connections only.

3.2. A Mean-Field Model of Neuronal Population Activity

Modeling of cortical signal processing can be considerably simplified by taking into account the columnar structure of the cortex. Each cortical column contains hundreds to thousands of neurons, which receive approximately the same afferent and intracortical input and show similar response properties. But even if some response properties (such as direction selectivity in macaque V1) change over depth, there is still a topographic mapping of functional aspects: In most of the cases many nearby neurons show similar selectivities.

Principle and Basic Assumptions

Based on these observations, Ben-Yishai et al. (1995) and Bartsch et al. (1997) have formulated a mean-field description of cortical processing. Firstly, if two neurons k and l within a cortical column receive similar total synaptic inputs, $h_k \approx h_l$, it seems a good approximation to assume that each neuron within a given population α for example, all excitatory ($\alpha =$"e") or all inhibitory neurons ($\alpha =$"i") within a column, receive the same input, $h_{k,\alpha} \equiv h_\alpha$ for all neurons k of the population. In other words, instead of feeding its actual input to every neuron, all neurons within one population are assumed to receive the same mean input. Due to that reason, this kind of models is referred to as *mean-field models*[1]. Secondly, within each population many

[1] This nomenclature has been borrowed from solid state physics, where mean-field models describe the behavior of atomic magnetic moments under the mean magnetic fields of their neighbors instead of the actual fluctuating magnetic field.

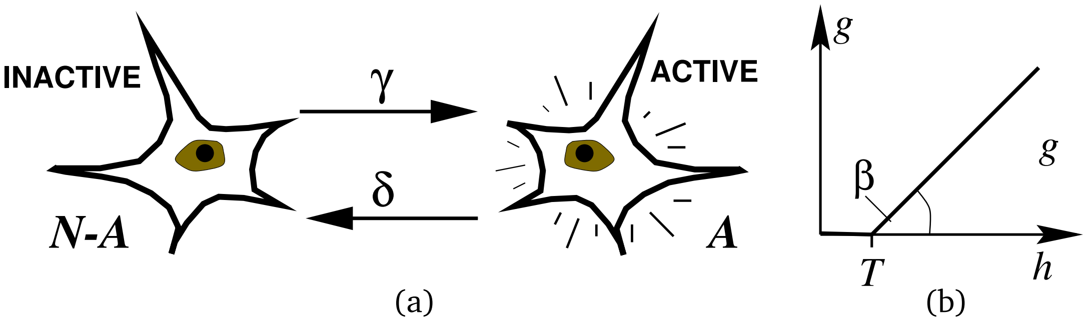


Figure 3.1.: *(a)* Principle of binary stochastic neurons: The neuron randomly toggles forth and back between an active (right) and inactive (left) state. If the neuron is driven by excitatory input, the probability for activation increases and inactivation probability decreases resulting in a higher mean activity. *(b)* Semi-linear activation function $g(h)$.

neurons encode similar input properties. It seems reasonable, that only the overall activity of a whole neuron population is important. The fluctuations within the activities of individual neurons, which can be viewed as non-linear stochastic units, can be neglected at this level. This corresponds to the assumption that important stimulus properties are encoded in population activities (for aspects of population coding see (Paradiso, 1988; Zohary, 1992)). Note that omitting random fluctuation does not imply that we exclude non-random collective phenomena, i.e., the mean population activity can still be strongly time-dependent. For a model that explicitly includes fluctuations we refer to (Tsodyks and Sejnowski, 1995)). Though neurons of one cell type within a cortical column are natural candidates for a neuron population, the framework can be applied in a more general way: Arbitrary sets of neurons with similar inputs and responses can be combined to a population.

Dynamics of a Neuron Population

We are now ready to write down a simplified description of neuronal population dynamics. Lets assume that individual neurons act as binary stochastic units which can flip forth and back between an active (spiking) and inactive (non-spiking) state. The probability per unit time for an inactive neuron to be activated is denoted by the activation rate γ and its inactivation rate by δ (3.1a on this page). Below, we will relate γ and δ to the mean synaptic input of the population: A high excitatory input will lead to a high activation rate and a low inactivation rate which causes a higher fraction of neurons to be active. Low input or inhibitory input in contrast, will reduce the activation rate and the neurons settle down to the inactive state. If we consider a population of N binary stochastic neurons with identical activation and inactivation rates, the number of active neurons A changes within the small time interval Δt according to

$$\Delta A = \gamma \Delta t(N - A) - \delta \Delta t A \; = \; (\gamma N - (\gamma + \delta)A) \; \Delta t. \tag{3.1}$$

In the limit $\Delta t \to 0$, this relation transforms to a rate equation for the fraction of neurons $m = A/N$ that are active at time t:

$$\frac{d}{dt} m = \gamma - (\gamma + \delta)m. \tag{3.2}$$

Now we assume for simplicity that the inactivation rate behaves inversely to the activation rate. This is reasonable because neurons which are strongly driven by input are less likely to stop firing spontaneously. If $\gamma_{\max}$ denotes the maximum possible activation rate, we arrive at $\delta = \gamma_{\max} - \gamma$ and

$$\frac{d}{dt} m = \gamma - \gamma_{\max} m. \tag{3.3}$$

One possible interpretation of the maximum activation is related to the refractory period of biological neurons. If a neuron wants to undergo two subsequent activations, it must at least fire a spike and wait for the refractory period τ until it can be activated to fire the next spike. This means that we can identify the maximum activation rate with $\gamma_{\max} = 1/\tau$. The rate equation for the population activity becomes

$$\tau \frac{d}{dt} m = -m + \tau\gamma =: -m + g, \tag{3.4}$$

where $0 \leq g = \tau\gamma \leq 1$ denotes the activation probability (relative to τ) for the neuron population.

For realistic regimes of operation, Equation 3.4 can be interpreted as the dynamics of a pool of spiking neurons. This view can be motivated as follows: The (absolute) refractory period τ is in the range of 1-2 ms, which means that electrically driven neurons can reach spike rates of approximately $500 - 1000$ Hz. In contrast, the spike rates observed for visually stimulated cortical neurons range around 50 Hz: realistic activation rates are much smaller than the maximal rate, $\gamma \ll \gamma_{\max}$, ($g \ll 1$), and consequently the inactivation rate is very fast: $\delta \approx \gamma_{\max} = 1/\tau$. Each time a neuron becomes activated and fires a spike, it immediately becomes inactivated again with a rate close to its refractory period. In other words, a binary stochastic neuron which is activated fires an individual spike and automatically inactivates again: In the limit of low mean activation, our mean-field model describes a population of spiking (instead of binary) neurons.

Now we have to specify, how the activation probability changes with the synaptic input. If the input of a neuron population falls below a threshold T, all neurons of the population are inactive, otherwise, neurons should be activated the more frequently, the more excitatory input they receive. The simplest function, which preserves this important rectifying nonlinearity present in biological neuronal systems is a semi-linear function, and we can formulate the dependence of the activation probability g on the mean synaptic input h of the population as

$$g(h) = \max(\beta(h - T), 0). \tag{3.5}$$

$g(h)$ is referred to as the activation function of the population (Figure 3.1b). Note that at this point we have made use of the mean-field assumption. In Equation 3.5, h has to be the mean synaptic input of the population. If we had used the actual synaptic input, the activation function would have had different values for each neuron in the population and the ensemble average could not be as easily written down as in Equation 3.4 on the page before.

3.2.1. Modeling Orientation Selectivity with Two Cell Types

In contrast to the considerations of the iceberg model, in (Bartsch, Stetter, Weber and Obermayer, 2000b; Stetter, Bartsch and Obermayer, 2000a) we explored, how the afferent input (transformed visual signals) is processed by a recurrent cortical network.

The primary visual cortex processes visual input locally: Each local visual feature is processed and represented by a patch of cortex about $1{-}2mm$ in diameter, which is called a hypercolumn. The neurons within a hypercolumn are densely connected by lateral intracortical fibers and form a strongly coupled recurrent network. By formulating a mean-field model of a cortical hypercolumn (Ben-Yishai et al., 1995; Ben-Yishai et al., 1997; Bartsch et al., 1997; Bartsch, Stetter and Obermayer, 2000a; Stetter et al., 2000a), one can efficiently describe the dynamics of hundreds of thousands of neurons within a hypercolumn at a population level.

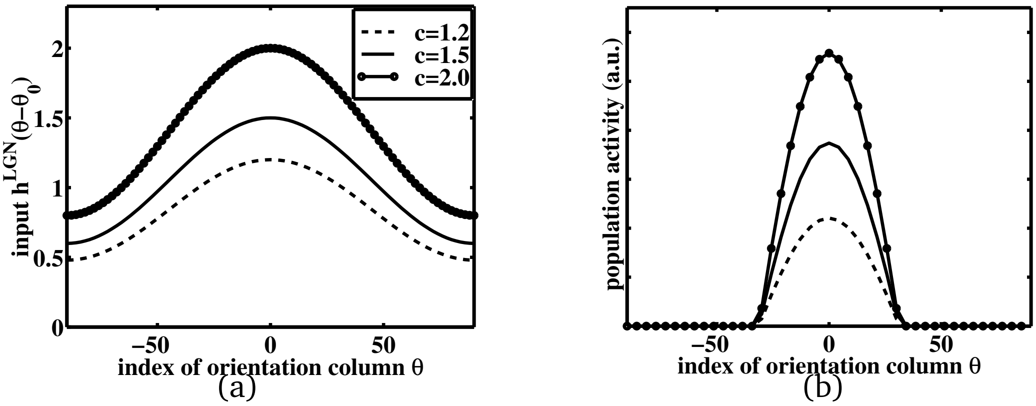


Figure 3.2.: *(a)* Simulation of orientation representation in a hypercolumn with two types of neurons. *(a)* Weakly tuned input $h^{\mathrm{LGN}}(\theta)$ to the hypercolumn for $\theta_0 = 0$ degree stimulation with three log contrast levels. *(b)* Mean activities of the excitatory neurons of the orientation columns in the stationary state for stimulation as shown in *(a)*. The activity pattern is sharply tuned and its width is independent of contrast

Orientation Tuning and Contrast Saturation in a Hypercolumn with Two Cell Types

Figure 3.2 shows how a hypercolumn of recurrently connected cortical orientation columns represents an oriented stimulus. The presence of an oriented contour or grating within the aggregate field of the hypercolumn evokes a broadly tuned input, which is shown in Figure 3.2a, on the current page for three log contrast levels. The resulting activity pattern for three contrast levels are shown in Figure 3.2b, on this page. The activity pattern is more sharply tuned in orientation space than the input, and its tuning width is independent of contrast as observed in biological systems.

Summarizing the results from Stetter et al. (2000a), there is a phase boundary for the strength of the lateral excitation. Above this phase boundary, in the marginal phase, stimulus orientation is represented by a sharp orientation tuning curve, independently of the strength of afferent orientation bias and independently of stimulus contrast. In the marginal phase it can be shown that the contrast response curve cannot saturate. The excitatory activation increases at least proportional with log-contrast, and saturation cannot occur.

In the linear phase the contrast response curve saturates as the result of the activation of inhibitory neurons with a high activation threshold. The model predicts that their contrast threshold coincides with the stimulus contrast at which excitatory neurons begin to saturate.

To explain the phase boundaries of the linear and marginal regimes, we determined them using a numerical simulation. Connection pattern are assumed with local excitation and balanced by flat inhibition, which may be closest to wiring patterns in area 17. Because the linear phase is characterized by a finite slope of the contrast-response curve close to the activation threshold, $c \simeq T_e$, we calculated the contrast gain (the slope of the contrast response curve) at $c = T_e$. Its divergence marks the boundary of the linear regime. The marginal phase is characterized by the generation of a narrow activation blob from initially untuned input. We stimulated a hypercolumn with weakly tuned input ($\varepsilon = 0.01$) and calculated the resulting orientation tuning width $\theta_{c,e}$. A decrease of the orientation tuning width below 90 degree marks the boundary of the marginal phase.

Figure 3.3 on the next page shows the behavior of the contrast gain at the threshold (solid/circles) and the orientation tuning width (crosses) as a function of the connection strength $S \equiv E_0 = E_2 = I_0$. The dotted vertical lines mark the boundary for the linear and for the marginal phase (Stetter et al., 2000a). Analytical and numerical values for the phase boundaries agree well with each other and the simulations demonstrate that the linear and marginal phases do not overlap. For the boundary of the linear phase, the close correspondence of analytical and numerical values is due to the fact, that in the simulation the input was weakly tuned and evoked a flat activation pattern. All orientation columns have similar average activities and can therefore be approximated by a single orientation column. For the boundary of the marginal phase, the small deviation between analytical and numerical values is an artifact of the finite step size in the connection strength used for the simulation. The contrast response curve do not change if the simulations are repeated using a more strongly orientation-biased input ($\varepsilon = 0.3$), whereas orientation tuning curves become continuously sharper with increasing connection strength. In either case, the separation of orientation tuning is contrast dependent outside the marginal phase.

We can summarize as follows: It can be shown, that a hypercolumn model with two neuron populations is not sufficient to account for the representation of stimulus contrast and stimulus orientation as observed in the primary visual cortex of many mammals. In the next section we will introduce a mean-field model, which accounts for the large variety of different inhibitory neuron types observed in the cortex of cats and monkeys (Bartsch, Stetter and Obermayer, 1999a; Bartsch, Stetter and Obermayer, 1999b; Bartsch et al., 2000a).

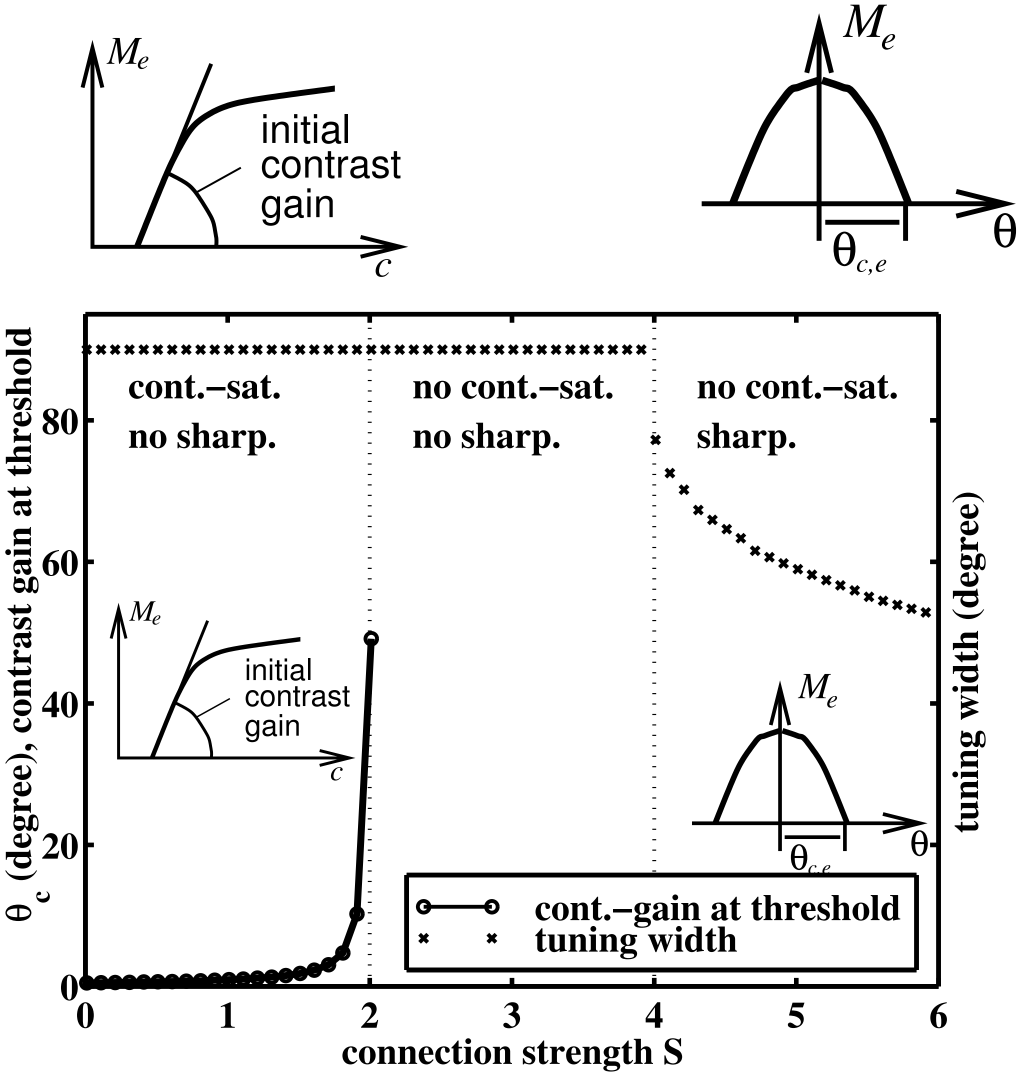


Figure 3.3.: The behavior of the contrast gain at activation threshold (*solid line*) and the orientation tuning width (*crosses*) as a function of the connection strength S for a hypercolumn with cosine shaped excitatory connection scheme and flat inhibition. Vertical dotted lines mark the analytical results. Insets illustrate criteria used for calculation of the curves. Both analytical and numerical results predict that there is no overlapping regime of co-occurrence of contrast saturation and contrast-invariant orientation tuning

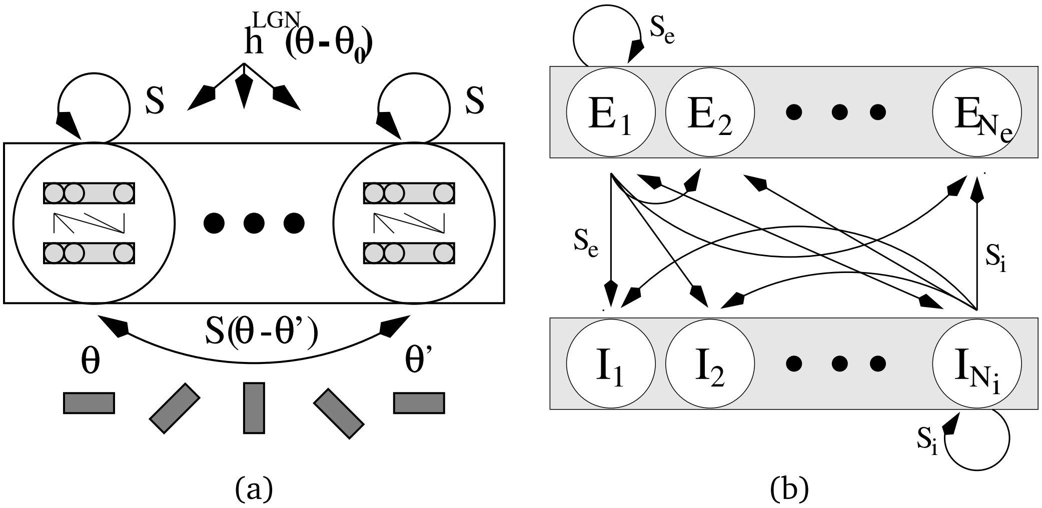


Figure 3.4.: *(a)* Setup of a mean-field hypercolumn with many different cell types. Recurrent couplings depend on the source and target orientations only. *(b)* Structure of a single orientation column. It consists of populations of N_e excitatory and N_i inhibitory cell types with, in general different properties

3.2.2. Hypercolumns with Multiple Populations

Now we want to explore the possible influence of the diversity of neurons in the cortex on its functional characteristics. For this we extend the model of an orientation column with a more complex structure: We take into account the fact that cortical tissue contains many different cell types and combine each of these cell types to a separate population. In general, an orientation column, indexed by its preferred orientation θ, now contains N_e populations with different excitatory neuron types and N_i different populations of inhibitory neurons (Figure 3.4b, on this page). The n-th excitatory population is indexed by (e, n) and the n-th inhibitory population by (i, n). We henceforth refer to the sub-populations as model neurons or simply "neurons".

The strength of recurrent intracortical couplings is assumed to depend only on the source and target orientation columns but not on the particular target neuro n. The mean connection strength from neuron (α, n) within column θ' to neuron β, m in column θ ($\alpha, \beta = e, i$) is given by

$$S^{m,n}_{\beta,\alpha}(\theta, \theta') \equiv S_\alpha(\theta - \theta'). \tag{3.6}$$

The generalization compared to the previous section consists of the fact that different neuron subpopulations can have different mean cellular properties and wiring patterns. To start with a simple case, we keep all properties of the neurons up to their activation functions identical for the present considerations and assume that the neurons differ only in their mean activation thresholds. The activity of neuron (α, n), $\alpha = e, i$ in response to synaptic input h is given by a semi-linear activation function

$$g_{\alpha,n}(h) = \max(\beta_\alpha(h - T_{\alpha,n}), 0), \tag{3.7}$$

where β_α denotes its slope and $T_{\alpha,n}$ its activation threshold. The activities of neurons (e, n) and (i, n) in column θ, $m_{e,n}(\theta, t)$ and $m_{i,n}(\theta, t)$, evolve according to

$$\begin{aligned} \frac{d}{dt} m_{\alpha,n}(\theta, t) &= -m_{\alpha,n}(\theta, t) + g_{\alpha,n}\left(h^{\text{lat}}(\theta, t) + h^{\text{LGN}}(\theta, t)\right) && (3.8) \\ h^{\text{lat}}(\theta, t) &= \sum_{\beta=e,i} \sum_{n} \int\limits_{-\pi/2}^{\pi/2} d\theta' S_\alpha(\theta - \theta') m_{\beta,n}(\theta', t) && (3.9) \\ h^{\text{LGN}}(\theta - \theta_0) &= c(1 - \varepsilon + \varepsilon \cos(2(\theta - \theta_0))). && (3.10) \end{aligned}$$

Note that h^{LGN} and h^{lat} are identical for all subpopulations.

Analytical Treatment of Contrast Saturation

We wish to understand how the contrast-response curve of the orientation column – or a representative subpopulation therein – depends on the distribution of activation thresholds. Again it seems reasonable to analyze an isolated but intrinsically coupled orientation column with N_e excitatory and N_i inhibitory neurons (Figure 3.4b, on the preceding page). In the stationary state, the total synaptic input, H, which is the same for all neurons in the orientation column, is given by

$$H = h^{\text{LGN}} + S_e \sum_{n=1}^{N_e} M_{e,n} - S_i \sum_{n=1}^{N_i} M_{i,n}, \tag{3.11}$$

where, according to Equation 3.7, $M_{\alpha,n} = g_{\alpha,n}(H) \equiv M_{\alpha,n}(T_{\alpha,n}, H)$ are the steady state activations of the model neurons and $S_\alpha \equiv S_\alpha((\theta - \theta') = 0)$ abbreviate the identical intra-column connection strengths between the neurons. Now we assume that the activation thresholds T_e and T_i are distributed over the orientation column according to pdfs $p_e(T_e)$ and $p_i(T_i)$, respectively. In the limit of infinitely many neurons, we can replace the sums in Equation 3.11 by the ensemble averages over the threshold distributions and obtain

$$H = h^{\text{LGN}} + S_e \int_{-\infty}^{\infty} M_e(T_e, H) p_e(T_e)\, dT_e - S_i \int_{-\infty}^{\infty} M_i(T_i, H) p_i(T_i)\, dT_i. \tag{3.12}$$

Because of the definition of the semi-linear transfer function Equation 3.7 on the preceding page, we know that neurons with $T_\alpha \geq H$ are silent and therefore do not contribute to the sums or integrals in Equations 3.11 on the page before and 3.12 on the preceding page. Conversely, for $T_\alpha < H$, the activation function can be replaced by its linear part, $M_\alpha(T_\alpha, H) = \beta_\alpha(H - T_\alpha)$. Therefore we can replace the upper limits of the integrals in Equation 3.12 on the page before by H:

$$H = h^{\mathrm{LGN}} + \beta_e S_e \int\limits_{-\infty}^{H} (H - T_e) p_e(T_e)\, dT_e - \beta_i S_i \int\limits_{-\infty}^{H} (H - T_i) p_i(T_i)\, dT_i. \quad (3.13)$$

Equation 3.13 represents a self-consistent relation between the total synaptic input H and the afferent input h^{LGN}. By solving this equation we can write down an analytical solution for the stationary activation

$$M_e(T_e, H) \equiv M_e(T_e, H(h^{\mathrm{LGN}})) = M_e(T_e, h^{\mathrm{LGN}}) \quad (3.14)$$

as a function of the *external* instead of the total synaptic input, which is the contrast-response function of the neurons. Carrying out the integrals in Equation 3.13 yields

$$\begin{aligned} H &= h^{\mathrm{LGN}} + S_e \beta_e G_e(H) - S_i \beta_i G_i(H) & (3.15) \\ G_\alpha(H) &= \int_{-\infty}^{H} dH' \int_{-\infty}^{H'} dT p_\alpha(T), \quad \frac{d^2}{dT^2} G_\alpha(T) = p_\alpha(T), \quad \alpha = e, i & (3.16) \end{aligned}$$

By defining the function

$$F(H) = H - \beta_e S_e G_e(H) + \beta_i S_i G_i(H) \quad (3.17)$$

Equation 3.15 reduces to $F(H) = h^{\mathrm{LGN}}$ and we can express the steady state activations M_α by

$$M_\alpha(T_\alpha, h^{\mathrm{LGN}}) = \beta_\alpha (H - T_\alpha) = \beta_\alpha \left(F^{-1}\left(h^{\mathrm{LGN}}\right) - T_\alpha \right) \quad (3.18)$$

Equation 3.18 provides an analytical relationship between geniculate input and the response of the recurrent cortical circuit. Note that it only holds for one isolated orientation column and if F is invertible. The latter condition corresponds to the boundary condition for the linear phase.

Figure 3.5 on the facing page illustrates the meaning of Equation 3.18 for the special case of only one excitatory and one inhibitory neuron type and only

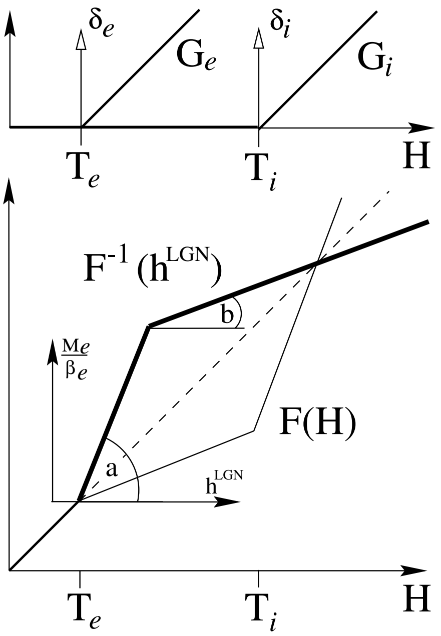


Figure 3.5.: Analytical solution Equation 3.18 on the facing page for one isolated orientation column and δ-peaked threshold distributions. *Top*: The distributions and the resulting second integrals $G_\alpha(H)$. *Bottom*: The function $F(H)$ Equation 3.17 on the preceding page (thin line) and its inverse (thick line) as resulting from the scenario in the top part. The thick line relative to the small coordinate system schematically illustrates the behavior of the contrast response function. Figure from (Bartsch et al., 2000a)

two threshold values T_e and T_i. In this case the two threshold distributions reduce to Kronecker delta functions around the two thresholds, $p_e(T) = \delta(T - T_e)$ and $p_i(T) = \delta(T - T_i)$ and their second integrals become semi-linear functions. $G_\alpha(H) = \max(H - T_\alpha, 0)$ (Figure 3.5 top, on the preceding page). The function $F(H)$ (Equation 3.17 on page 50) becomes

$$F(H) = \begin{cases} H & : \quad H \leq T_e \\ H - \beta_e S_e(H - T_e) & : \quad T_e < H \leq T_i \\ H - \beta_e S_e(H - T_e) + \beta_i S_i(H - T_i) & : \quad H > T_i \end{cases} \tag{3.19}$$

In this scenario, the resulting contrast response function Equation 3.18 on page 50 shows a typical saturating behavior. The gradients of $F^{-1}(h^{\mathrm{LGN}})$ are $a = (1 - \beta_e S_e)^{-1}$, where $a\beta_e$ is the initial contrast gain of the contrast- response function, and $b = (1 - \beta_e S_e + \beta_i S_i)^{-1}$ for higher contrast levels.

Numerical Simulations of Contrast Responses

For the following simulations we assumed threshold distributions, which for excitatory neurons are Gaussian, $p_e(T_e) = \mathcal{N}(\mu_e, \delta_e)$, and for inhibitory neurons are either Gaussian, $p_i(T_i) = \mathcal{N}(\mu_i, \delta_i)$ or bimodal according to two superimposed Gaussian functions $p_i = 0.5(\mathcal{N}(\mu_{i,1}, \delta_{i,1}) + \mathcal{N}(\mu_{i,2}, \delta_{i,2}))$. Inhibitory mean activation thresholds are set to be higher ($\mu_i = 2$) than excitatory mean activation thresholds ($\mu_e = 1$). Also, simulations will use $\beta_e = 0.5$ and $\beta_i = 1$, but the special choice of parameters does not strongly influence the results.

Figure 3.6 on the next page compares the numerical solution of the differential equation 3.8 on page 49 (solid line) with the analytical expression in Equation 3.18 on page 50 (circles) for two unimodal and fairly narrow threshold distributions (histograms) in the linear phase. It demonstrates that the analytical solution approximates the solution of the differential equation very well. The dashed and dash-dotted lines plot G_e and G_i for the distributions used. The behavior of this system can be understood as follows: First, only excitatory neurons are active and, because we operate in the linear phase, act as linear amplifiers. For higher contrast levels, more and more inhibitors become active and reduce the contrast gain. Different from the case of only two thresholds, the contrast-response curve gradually changes its gain over contrast. A gradual contrast saturation can be qualitatively understood as follows: With increasing afferent input h^{LGN}, more and more inhibitory neuron subpopulations are recruited (become active): The increase in number is proportional to $p_i(F^{-1}(h^{\mathrm{LGN}}))$. The more neurons are recruited, the stronger the decrease in contrast gain. In other words, we expect a relationship between the second derivative of the contrast-response function at h^{LGN} and the density of neurons with activation thresholds $T_i = F^{-1}(h^{\mathrm{LGN}})$.

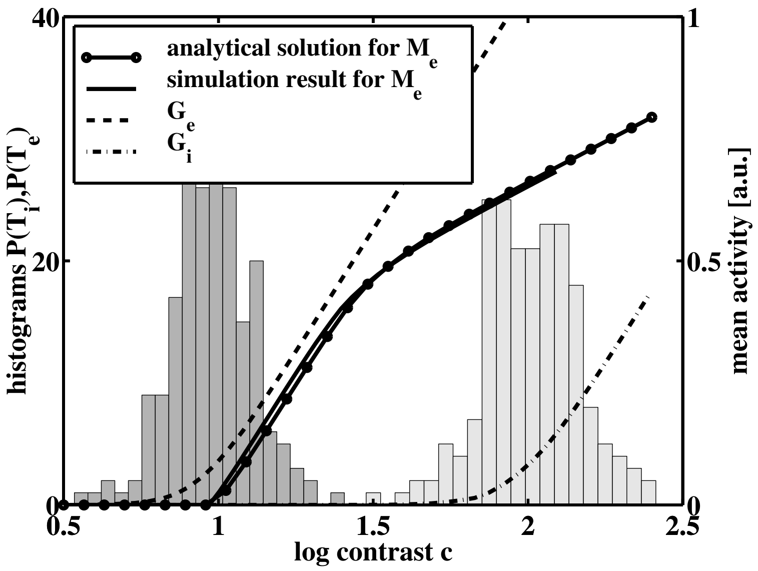


Figure 3.6.: Simulation of a contrast-response curve for a set of 400 coupled model neurons (200 exc., 200 inh.) in the linear phase ($S_e = 1, S_i = 1$) and for unimodal Gaussian threshold distributions $p_e(T_e) = \mathcal{N}(1, 0.1)$, $p_i(T_i) = \mathcal{N}(2, 0.1)$ (cf. threshold histograms in the diagram). *Solid line*: Numerical solution of the differential equation 3.8 on page 49. *Circles*: Evaluation of the analytical expression 3.18 on page 50. Both curves agree very well. Dashed and dash-dotted lines show G_e and G_i, respectively. Figure from (Bartsch et al., 2000a)

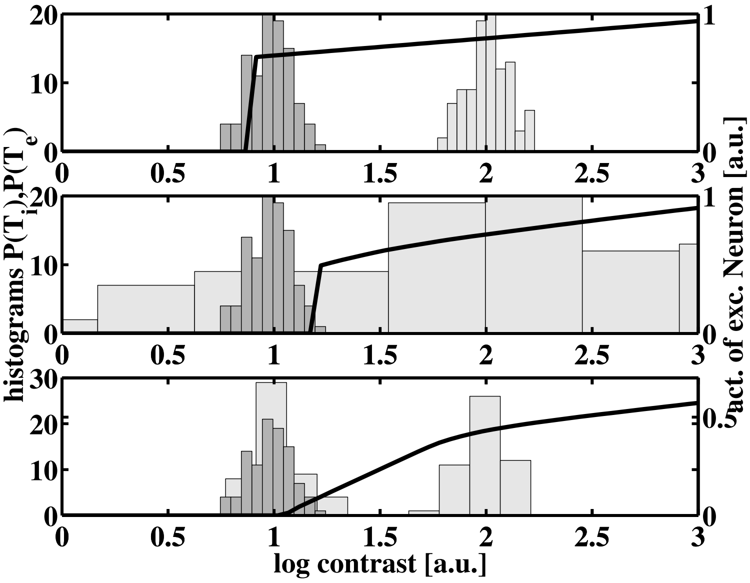


Figure 3.7.: Simulation for a set of 400 coupled model neurons (200 exc., 200 inh.) in the marginal phase ($S_e = S_i = 6$). *Solid lines*: Contrast-response curves; *dark gray*: Histograms of excitatory thresholds $p_e(T_e) = \mathcal{N}(1, 0.1)$; *light gray*: Histograms of inhibitory thresholds $p_i(T_i)$. *Top*: $p_i(T_i)$ unimodal, small variance ($\mathcal{N}(2, 0.1)$); *Middle*: $p_i(T_i) = \mathcal{N}(2, 1)$ unimodal, large variance. *Bottom*: A bimodal distribution of $p_i(T_i)$ is used ($\mu_{i,1} = 1, \mu_{i,2} = 2, \delta_{i,1} = \delta_{i,2} = 0.1$). A bimodal distribution p_i is necessary and sufficient for graded contrast-response and contrast saturation also in the marginal phase. Figure from (Bartsch et al., 2000a)

This relationship can be quantified by forming the 2nd derivative of the steady state activation Equation 3.18 on page 50 with respect to the LGN input. We arrive at the following relationship between the curvature of the contrast-response function and the distributions of activation thresholds $p_\alpha(T_\alpha)$:

$$\begin{aligned} \frac{d^2}{d(h^{\mathrm{LGN}})^2} M_e &= \beta_e \frac{S_e \beta_e p_e(H) - S_i \beta_i p_i(H)}{(-1 + S_e \beta_e G_e'(H) - S_i \beta_i G_i'(H))^3}, && (3.20) \\ H &= F^{-1}(h^{\mathrm{LGN}}) && (3.21) \end{aligned}$$

The denominator of Equation 3.20 is positive in the linear phase, because the gain of F has to be finite (invertibility of F). The contrast-response curve shows a negative curvature or saturation if more inhibitory than excitatory neurons are recruited by a small increase in the input, i.e. if $S_e \beta_e p_e(H) < S_i \beta_i p_i(H)$ holds. Otherwise, the contrast-response function increases its gain.

Besides a quantitative understanding of the structural origin of contrast gain in the linear phase, it might be even more important to see, whether a gradually increasing and finally saturating contrast-response can also be stabilized in the marginal phase by some threshold distribution. If this could be achieved, we would succeed in formulating necessary conditions for cortical circuitry to show a constant orientation tuning and contrast saturation for a single parameter setting.

Figure 3.7 on the facing page shows the contrast-response curve of an excitatory neuron with threshold $T_e = 1$ for different cases of the inhibitory threshold distribution in the marginal phase. If the threshold-distribution is small and unimodal (top), the contrast-response shows a pseudo-binary switch-on behavior as observed in the marginal phase with two neuron types (cf. Section 3.2.1 on page 45). This behavior remains stable as long as the distribution is unimodal, even if it is very wide (Figure 3.7 middle, on the facing page). As soon as the threshold distribution becomes bimodal (Figure 3.7 on the preceding page, bottom), the contrast-response first increases from zero and later saturates, as observed in biology. This demonstrates that two inhibitory neuron populations, one with low and the other with higher activation threshold, are necessary and sufficient to stabilize contrast saturation in the marginal phase.

Orientation and Contrast Response with Three Neuron Types

One can easily combine many structured orientation columns to a full hypercolumn. The orientation columns are mutually coupled by lateral connections with π-periodic Gaussian functions 3.25 on page 63, and are driven by weakly orientation biased input.

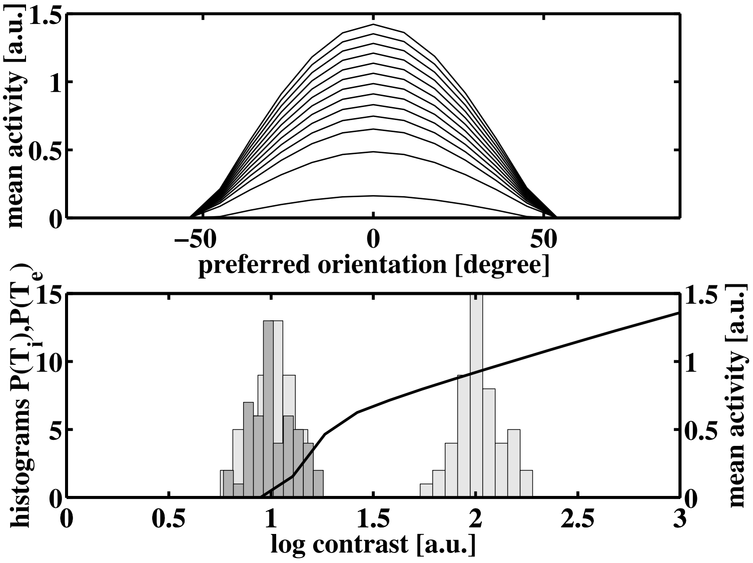


Figure 3.8.: *Top:* Orientation tuning curve and *Bottom* contrast-response curve of an excitatory neuron with preferred orientation 0^o for a hypercolumn with 21 orientation columns (50 excitatory neurons and 100 inhibitory neurons each) in the marginal phase. The system shows a graded and saturating contrast response, which is combined with a contrast-invariant orientation tuning width. Parameters: $S_e = 6, S_i = -6, \sigma_e = 34$ deg, $\sigma_i = \infty$, $\delta_e = \delta_{i,1} = \delta_{i,2} = 0.1$,$\mu_e = \mu_{i,1} = 1, \mu_{i,2} = 2$. Figure from (Bartsch et al., 2000a)

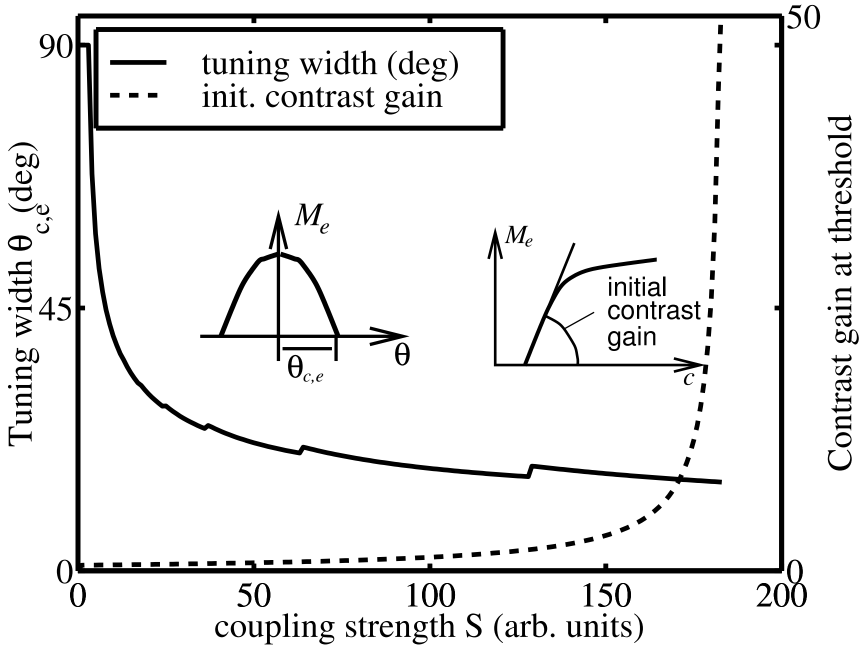
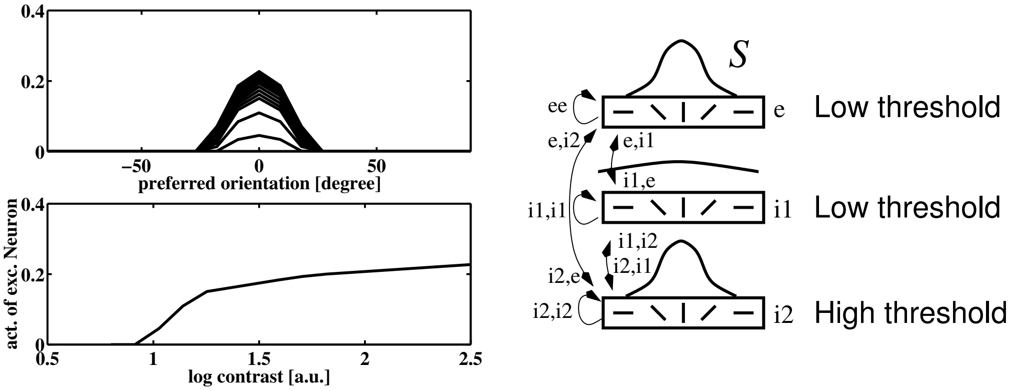


Figure 3.9.: The behavior of the contrast gain at activation threshold (*solid line*) and the orientation tuning width (*crosses*) as a function of the connection strength $S \equiv E_0 = E_2 = I_0$ $(I_2 = 0)$ for a hypercolumn with one excitatory and two inhibitory neuron populations. Other parameters were: $\beta_e = 0.5, \beta_i = 1, T_e = 1, T_{i1} = 1, T_{i2} = 2, \varepsilon = 0.01$, contrast for the orientation tuning width: $c = 2.0$. Insets illustrate criteria used for calculation of the curves. There is a wide range $(4 \leq S \leq 180)$, over which the linear and marginal phase coincide. Steps in the solid line are finite-size effect



(a) (b)

Figure 3.10.: *(a)* Orientation tuning curves (top) and the contrast-response function of the zero-deg. orientation column (bottom). *(b)* Schematic illustration of the corresponding wiring scheme: low-threshold lateral inhibitors (e.g., basket neurons) and high-threshold local inhibitors (e.g., chandelier cells). Parameters: $S_e = S_{i1} = S_{i2} = 50$; (marginal phase) *(b)* $\sigma_e = 34$ deg; $(\sigma_{i1}, \sigma_{i2}) = (\infty, 34)$ deg. $T_e = T_{i1} = 1$; $T_{i2} = 1.5$. The hypercolumn properly operates as in Figure 3.8 on page 56

Figure 3.8 on page 56 shows orientation tuning curves of the $m_{e,1}(\theta)$ excitatory populations of 21 orientation columns (top) and the contrast-response curves of a subset of 5 excitatory subpopulations of the $\theta = 0^o$ column (bottom) for unimodal $p_e(T_e)$ and bimodal $p_i(T_i)$. Even though the system operates in the marginal phase, where orientation tuning is independent of contrast, the contrast-response curve shows expressed saturation at the same time. This behavior is independent of the detailed shape of the threshold distributions, as long as it is bimodal.

A phase-diagram determined by the initial contrast gain and the orientation sharpening (cf. Figure 3.3 on page 47) for a hypercolumn with one excitatory and two inhibitory (low- and high-threshold) neuron types is plotted in Figure 3.9 on the page before. Due to low-threshold inhibition, the linear phase with finite initial contrast gain is stabilized up to very strong recurrent excitation strengths ($S \approx 180$ compared to $S = 2$ for two-neuron hypercolumns), and there is a wide range of coupling strengths, in which orientation tuning is invariant and the contrast-response saturates. In summary, this finding predicts that the experimentally observed cortical response properties require essentially two functionally distinct inhibitory neuron types to be present: Inhibitors

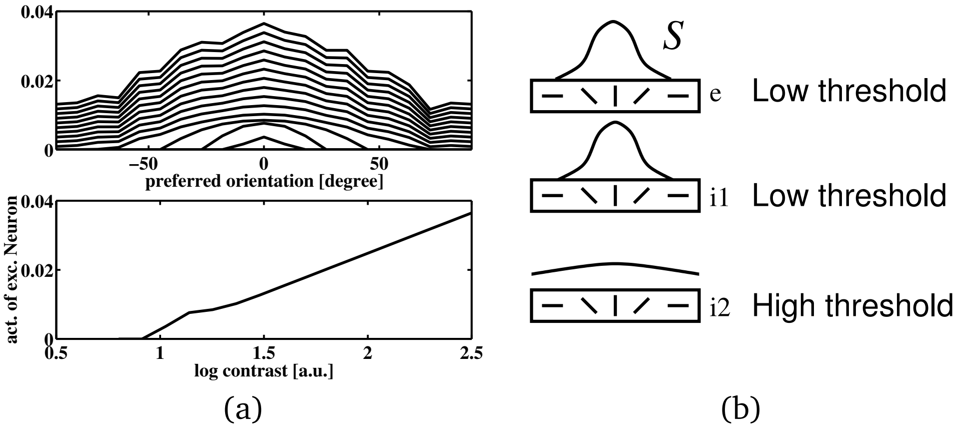
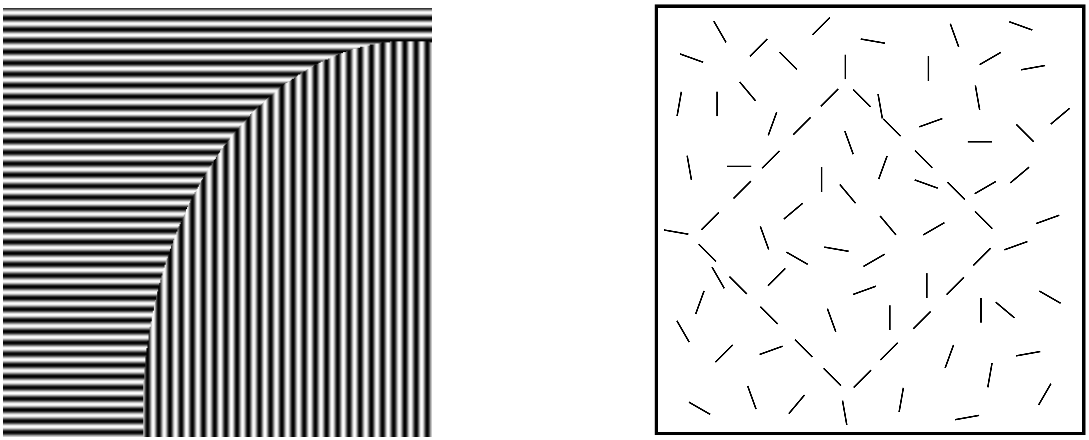


Figure 3.11.: *(a)* Orientation tuning curves (top) and the contrast-response function of the zero-deg. orientation column (bottom) for reverse properties of the inhibitory neurons: *(b)* low-threshold local inhibitors and high-threshold lateral inhibitors. Parameters: $S_e = S_{i1} = S_{i2} = 50$; (marginal phase) *(b)* $\sigma_e = 34$ deg.; $(\sigma_{i1}, \sigma_{i2}) = (34, \infty)$ deg. $T_e = T_{i1} = 1; T_{i2} = 1.5$

with a low activation threshold (or tonically active inhibitors) stabilize the contrast gain at near the contrast threshold to finite values, whereas inhibitors with high activation thresholds cause the saturation of the contrast-response curves at higher contrast levels.

In Figure 3.8 on page 56, both types of inhibitors were assumed to distribute lateral inhibition between different orientations. Possible candidates for such inhibitors are basket cells with axonal arborization up to $1200\mu m$ (Lund, 1987a), but it seems more reasonable to identify the two functionally different inhibitors with two anatomically distinguishable biological neuron types. Many inhibitors apart from basket cells are local companions, which contact only postsynaptic neurons within the same or closely adjacent orientation columns. One important local inhibitor is the chandelier cell. Therefore we may ask under which conditions a hypercolumn with pyramidal neurons as the excitatory population, basket cells as lateral inhibitors and chandelier cells as local inhibitors, still show the behavior seen in Figure 3.8 on page 56.

Orientation tuning and contrast response for a hypercolumn with three neuron types are provided in Figure 3.10 on the facing page, and in Figure 3.11 for two different combinations of wiring profiles and activation thresholds of inhibitory neurons. If the low-threshold cells mediate lateral inhibition and the high-threshold neurons local inhibition (Figure 3.10b on the facing page), ori-



(a) (b)

Figure 3.12.: *(a)* Example image for the demonstration of texture-based segmentation (a contour is defined by texture boundaries). *(b)* Example image for the demonstration of line-completion (aligned line segments are perceptually grouped as an interrupted diamond). Figure from (Bartsch et al., 2001)

entation tuning is sharp and constant with saturating contrast response function (Figure 3.10a on page 58). If the properties are reversed (low-threshold chandelier cells and high-threshold basket cells, Figure 3.11b on the preceding page), orientation sharpening is weak, unstable and contrast-dependent (Figure 3.11a on the page before). These simulations emphasize the important role of inhibition for the observed cortical representation of orientation and contrast of a stimulus (cf. Eysel, Shevelev, Lazareva and Sharaev (1998)), but additionally provide the following prediction: A hypercolumn needs two different inhibitors for the generation of experimentally observed contrast and orientation representation. At least one of the cell types must mediate lateral inhibition (e.g., basket cells), and this cell type must have a low activation threshold. If local inhibitors (e.g., chandelier cells) contribute to the recurrent circuit as modeled, they should have a high activation threshold.

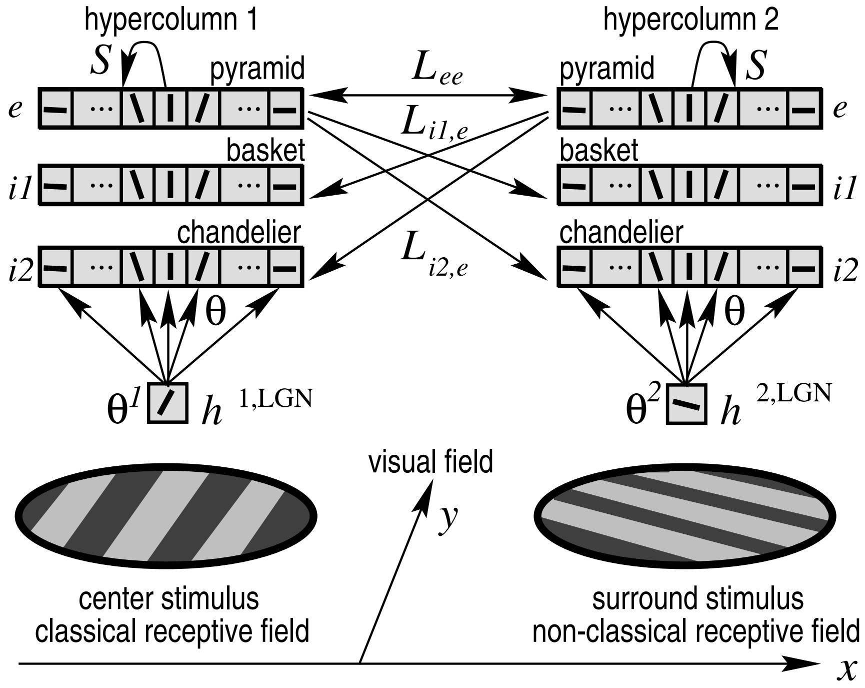


Figure 3.13.: Mean-field model of two coupled hypercolumns $a = 1, 2$, the orientation columns θ contains one excitatory ('e') and two inhibitory ('$i1$','$i2$') neuron populations. Both hypercolumns receive weakly orientation biased geniculocortical inputs $h^{a,\mathrm{LGN}}$, $a = 1, 2$, from adjacent but nonoverlapping patches of the visual scene, which correspond to the center and the nonclassical surround of hypercolumn 1. Orientation columns within each hypercolumn are densely interconnected by short range connections $S_{\alpha,\beta}(\theta-\theta')$, where α denotes the type of the target population and β the type of the source population ($\alpha, \beta =$ 'e', 'i'). In addition, both hypercolumns are mutually interconnected by symmetrical and excitatory long-range connections $L_{\alpha,\beta}(\theta - \theta')$

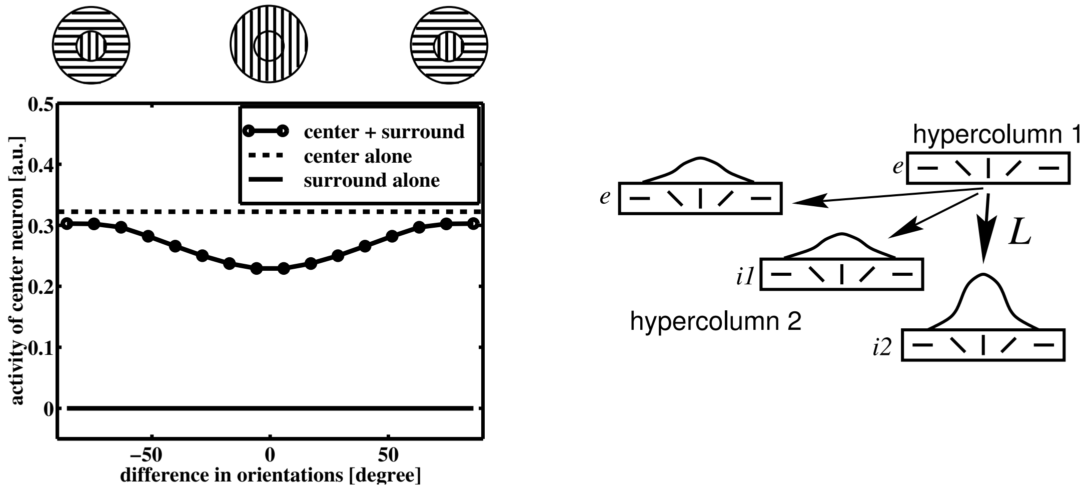
3.2.3. Hypercolumn Model Setup for Contextual Effects

Contextual effects are assumed to be mediated by lateral spreading signals or signaling cascades. These signals have to travel distances incorporating more than one hyper-column to connect neurons with non-overlapping receptive fields. One approach towards a model of contextual effects consists therefore naturely of two neighboring and coupled mean-field hypercolumns $a = 1, 2$ within the primary visual cortex. A different approach incorporating a lattice of model neurons is presented in Section 3.3 on page 69. Hypercolumn 1 is considered to process the visual input within the considered receptive field and is referred to as "center" hypercolumn. The aggregate field of hypercolumn 2 is assumed to be adjacent but still disjunct from the considered receptive field. It processes the nonclassical receptive field of the "center" hypercolumn and modulates it via their mutual couplings. Figure 3.13 on the page before schematically illustrates the model setup.

Again, each hypercolumn consists of a set of orientation columns, indexed by their preferred orientations θ, and each column consists of an excitatory (e) and two inhibitory neuron populations $(i1, i2)$. The activity of neuron (α), $\alpha = e, i1, i2$ in response to synaptic input h is given by a semi-linear activation function $g_\alpha(h) = \max(\beta_\alpha(h - T_\alpha), 0)$ where β_α denotes its slope and T_α its activation threshold. Similarly to Equation 3.8 on page 49, the dynamics of a neuron population α in hypercolumn a and column θ, $m_e^a(\theta, t)$, $m_{i1}^a(\theta, t)$ and $m_{i2}^a(\theta, t)$, are described by the following set of differential equations:

$$\begin{aligned}
\frac{d}{dt} m_\alpha^a(\theta, t) &= -m_\alpha^a(\theta, t) + g_\alpha\left(h^{a,\mathrm{lat}}(\theta, t) + h^{a,\mathrm{LGN}}(\theta, t)\right) && (3.22)\\
h^{a,\mathrm{lat}}(\theta, t) &= \sum_{\beta=e,i1,i2} \int\limits_{-\pi/2}^{\pi/2} d\theta' \left[S_{\alpha,\beta}(\theta - \theta') m_\beta^a(\theta', t) + \right. \\
&\qquad \left. L_{\alpha,\beta}(\theta - \theta') m_\beta^{b \neq a}(\theta', t)\right] && (3.23)\\
h^{a,\mathrm{LGN}}(\theta) &= c(1 - \varepsilon + \varepsilon \cos(2(\theta - \theta^a))), && (3.24)
\end{aligned}$$

where θ^a is the stimulus orientation presented to the center hypercolumn $(a = 1)$ or to the surround hypercolumn $(a = 2)$. Intracortical couplings are symmetric between both hypercolumns and all long-range connections are excitatory, i.e. $L_{\alpha,\beta} = L_{\alpha,e} =: L_\alpha$. They depend only on the difference in preferred



(a) (b)

Figure 3.14.: *(a)* Modulation of the center response by an oriented stimulus in the nonclassical surround. Compared to stimulation of the center hypercolumn alone (*dashed line*), the surround stimulus causes iso-orientation suppression (*circles*), but has only a weak impact in the cross-orientation stimulus condition. The surround stimulus alone cannot drive (*solid line*) but only modulate the center hypercolumn. *(b)* Connectivity needed for the behavior in *(a)*. Long-range connections must predominantly drive inhibitory interneurons for iso-orientation suppression. Parameters were $L_{e;i1;i2} = 0.5, 0.5, 3$, $\lambda_{\alpha\beta} \equiv \lambda_\beta = 34$ deg., $\beta = e, i1, i2$

orientations and are assumed as Gaussian functions in orientation space:

$$S_{\alpha,\beta}(\Delta\theta) = \operatorname{sign}_\beta S_{\alpha\beta} N_\sigma \exp\left(-\frac{\Phi^2(\Delta\theta)}{2\sigma^2_{\alpha\beta}}\right) \quad (3.25)$$

$$L_\alpha(\Delta\theta) = L_\alpha N_\lambda \exp\left(-\frac{\Phi^2(\Delta\theta)}{2\lambda^2_\alpha}\right), \quad (3.26)$$

where $L_\alpha \geq 0$ is the integral strength of the long-range connections to population α and N_λ is a normalization constant. It may be considered here to use the *von Mises* distribution which is the circular equivalent of the Gaussian function.

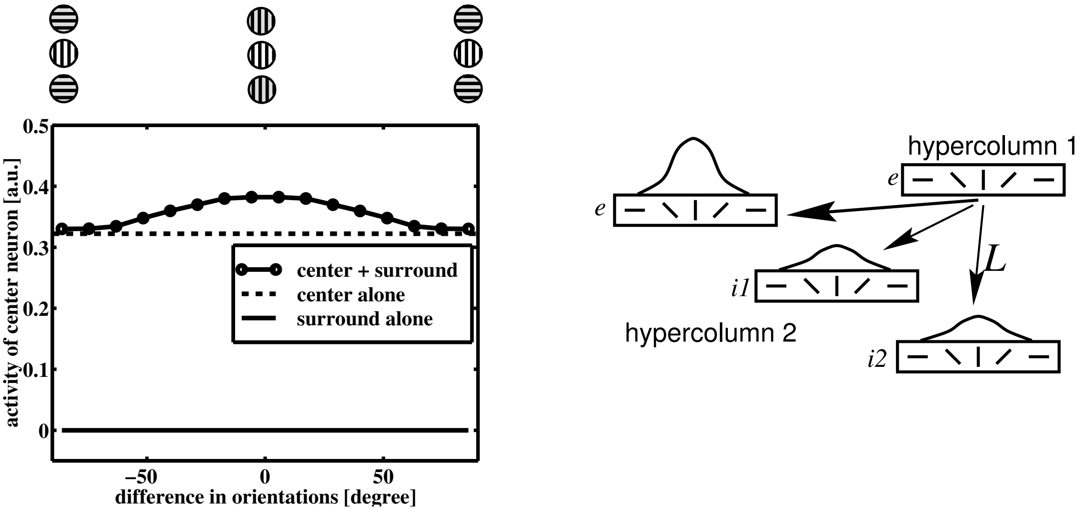
3.2.4. Numerical Simulations

Based on the coupled hypercolumn model we can now explore if and how long-range connections can modulate local cortical processing. For the following simulations we used a strong local recurrent connectivity with identical strengths $S_\alpha = 50$ and widths $\sigma_{\alpha\beta} \equiv \sigma_\beta$, $\sigma_e = \sigma_{i2} = 34$ deg, $\sigma_{i1} = \infty$ (cf. Figure 3.10 on page 58). Afferent input had intermediate orientation bias $\varepsilon = 0.3$ and $c = 2.5$, and the parameters for the activation functions were chosen as $\beta_e = 0.5$, $\beta_{i1} = \beta_{i2} = 1$, $T_e = T_{i1} = 1$ and $T_{i2} = 1.5$. The results reported do not qualitatively depend on these choices as long as the system operates in the overlap region of the linear and marginal regimes (central part in Figure 3.9 on page 57).

Figure 3.14a, on the page before demonstrates how a stimulus presented in the non-classical surround of hypercolumn 1 (the center hypercolumn) can modulate its response to a stimulus within the receptive field. Compared to the dashed line, which marks its response to center stimulation alone, the activity of the center column is reduced (circles), if an oriented stimulus is presented to the non-classical surround (see icons above plot). If center and surround orientations are identical or similar, the suppression is strongest, i.e., this system shows iso-orientation suppression. In contrast, if both stimuli are orthogonal to each other, only a weak suppressive effect is observed. In particular, over all orientation differences the sign of the modulatory effect is the same. The solid line in Figure 3.14a, on the preceding page shows the response of the center hypercolumn to surround stimulation alone and demonstrates that the surround stimulus cannot activate but only modulate the neurons in the center hypercolumn.

Modulatory suppression as shown in Figure 3.14a, on the page before requires a particular connection scheme for the long-range connections, which is summarized in Figure 3.14b on the preceding page: (i) Long-range connections should predominantly connect columns with similar preferred orientations, which is supported by experiments (Malach et al., 1993; Bosking, Zhang, Schofield and Fitzpatrick, 1997). (ii) The fibers must drive at least one inhibitory neuron type stronger than the excitatory populations which has been suggested by recent experiments (Das and Gilbert, 1999). For high contrast levels where all neuron populations are active, the effect does not depend on which inhibitory neuron type ($i1$ or $i2$) is driven strongest.

Figure 3.15 on the next page shows a parameter regime in which the surround stimulus facilitates the center response (circles vs. dashed line), but cannot drive the neurons of the center hypercolumn alone (solid line). Again, cross-orientation modulation is weak and has the same sign as the center modulation. Facilitation is observed if the long-range connections drive excitatory



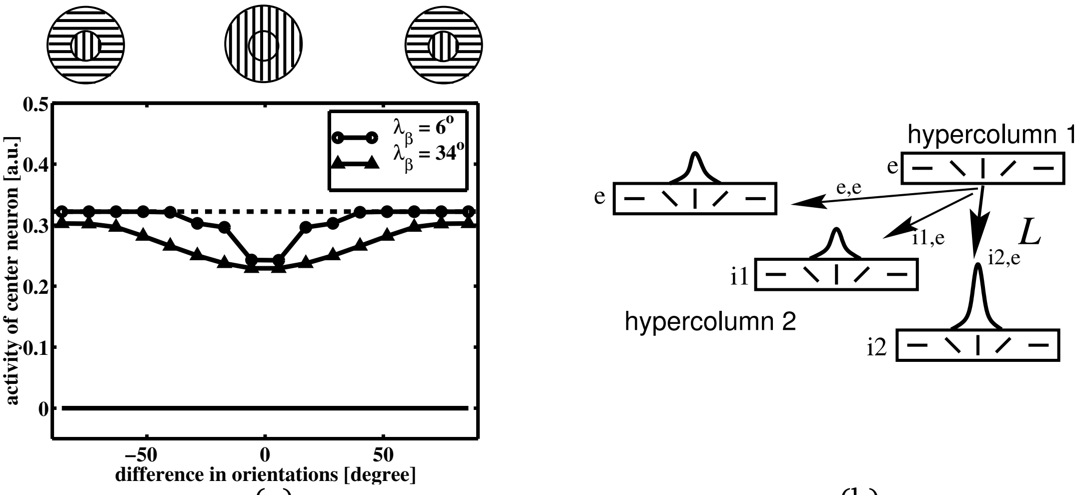
(a) (b)

Figure 3.15.: *(a)* Iso-orientation facilitation for the same circuit as in figure 3.14 on page 63, but this time the long-range connections drive excitatory target neurons stronger than the inhibitors *(b)*. Parameters: $L_{e;i1;i2} = 1, 0.5, 0.5$, $\lambda_\beta = 34$ deg, $\beta = e, i1, i2$

target neurons stronger than inhibitory ones (Figure 3.15b, on this page).

The angular profile of the nonclassical modulation is mostly determined by the orientation specificity of the long-range couplings $L_{\alpha,\beta}(\theta - \theta')$. This is demonstrated in figure 3.16a on the next page, which compares the suppressive modulation caused by strongly orientation specific long-range connections (circles) with suppression for more broadly tuned long-range connectivity (triangles). The orientation specificity of long-range connections is strongly correlated with the orientation tuning of the suppression. In contrast, the profile of non-classical modulation depends only weakly on the local connectivity within the hypercolumns, as long as they operate in the marginal phase (data not shown).

In order to understand why the modulatory effects are determined predominantly by the properties of the long-range connections, we have to realize that the activity pattern within each hypercolumn is determined by the local recurrent circuitry. Because the hypercolumns operate in the marginal regime, the local circuit forms a sharply tuned activity patch around the orientation column which matches the stimulus orientation (Figure 3.17 on page 67). The curve shape of the patch is relatively rigid and can only be weakly influenced by external (afferent or lateral) input. In particular, any synaptic input that is mediated by long-range connections can only modulate the activity level of



(a) (b)

Figure 3.16.: (*a*) Dependence of the suppression profile on the orientation-specificity of long-range connections. Circles: Strong orientation specific long-range couplings ($\lambda_\beta = 6$ deg) cause a narrowly tuned iso-orientation suppression. Triangles show the same curve as in Figure 3.14 on page 63 ($\lambda_\beta = 6$ deg.) for comparison. Again, the *dashed* and *solid* lines mark the response to center alone and surround alone stimulation. (*b*) Schematic illustration of the long-range connectivity used. Parameters: $L_{e;i1;i2} = 0.5, 0.5, 3$, $\lambda_\beta = 6$ deg, $\beta = e, i1, i2$

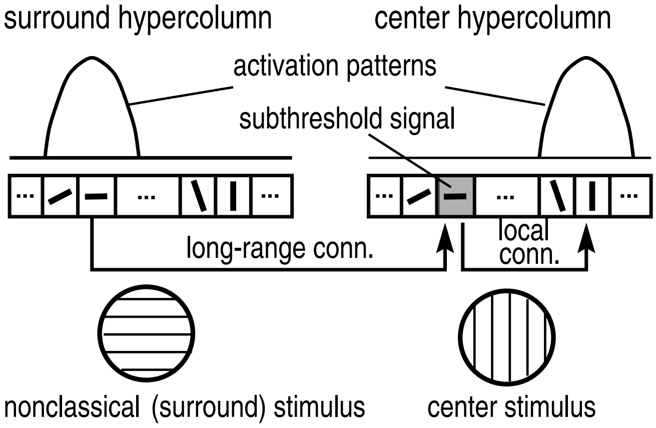
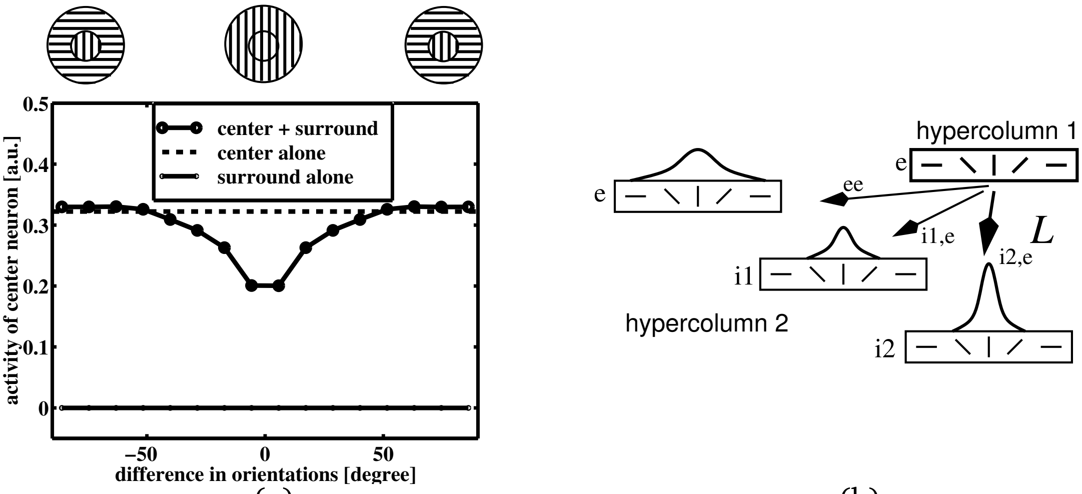


Figure 3.17.: Cross-orientation modulation in the marginal phase. Iso-orientation specific long-range connections cannot evoke any cross-orientation modulation, because all basket cells that signal across orientations are silenced by the local recurrent circuitry (shaded orientation column)

active neurons, but cannot activate silent neurons. This behavior is schematically illustrated in Figure 3.17 for cross-orientation stimulation. The shaded orientation column is driven by long-range connections, but cannot become active because its state is determined by the local recurrent dynamics. Consequently, we can only expect a non-classical modulation to occur if there are long-range fibers which connect active source neurons with active target neurons. In other words, the range of a surround modulation in orientation space is approximately given by the width λ of the long-range connection profile plus the width of the cortical activity pattern. Because the activity patterns have approximately constant shape, the angular profile of the surround modulation is determined by the angular profile of the long-range connections in orientation space.

Figure 3.17 also helps to understand why cross-orientation modulation is hard to achieve with iso-orientation specific patchy connections. It sketches the situation of cross-oriented stimuli and purely iso-orientation specific long-range fibers. The activity patterns of source and target neurons are disjunct in orientation space, and therefore no cross-orientation effect can be observed. In particular, orientation contrast sensitivity (iso-orientation suppression combined with cross-orientation facilitation) in the marginal phase cannot be caused by dis-inhibition as suggested earlier (Pawelzik et al., 1996).

Figure 3.18a, on the next page shows a simulation in which iso-orientation suppression is combined with a weak cross-orientation facilitation (sensitivity



(a) (b)

Figure 3.18.: *(a)* Iso-orientation suppression combined with a weak cross-orientation facilitation appears, as soon as long-range connections to excitatory target neurons are more broadly tuned than connections to inhibitory target neurons *(b)*. Parameters: $L_{e;i1;i2} = 0.5, 0.5, 3$, $\lambda_{i1} = \lambda_{i2} = 17$ deg, $\lambda_e = 34$ deg.

to orientation-contrast). This behavior is caused by long-range connections, which are more broadly tuned for excitatory target neurons than for inhibitory target neurons (Figure 3.18b on this page). As a consequence, long-range modulation via inhibitory interneurons dominates at small orientation differences between source and target orientation column, whereas for larger differences in orientation direct excitation dominates. In other words, the profile of the long-range connections implements an inverse mexican-hat in orientation space, which directly translates into orientation-contrast sensitivity.

3.3. A Lattice Model for Contextual Effects

Showing an iso–oriented annular surround stimulus outside the classical receptive field of a neuron, additionally to a centered stimulus, suppresses its response (Levitt and Lund, 1997). This can be seen as a possible mechanism for texture–based segmentation where contour is defined by a contrast orientation. But in a different stimulus paradigm, iso–oriented surround stimuli can also facilitate the response of a neuron (Polat et al., 1998; Kapadia, Sigman and Gilbert, 1999b). Long–range connections formed by excitatory pyramidal neurons were proposed to mediate these effects. Nevertheless, Das and Gilbert (1999) showed that contextual effects might also be mediated by short–range connections.

In Section 3.2.2 on page 48 we have shown that additionally to excitatory neurons, two groups of inhibitory neurons per orientation column (with low and high activation thresholds, respectively) are necessary and sufficient to generate both, contrast saturation and contrast invariant orientation tuning. We have also demonstrated, how long–range connectivity determines the characteristics of contextual effects. However, the model which acted solely in the orientation space neglected the geometrical arrangement of orientation columns, and consequently could not model poly-synaptic lateral signaling cascades.

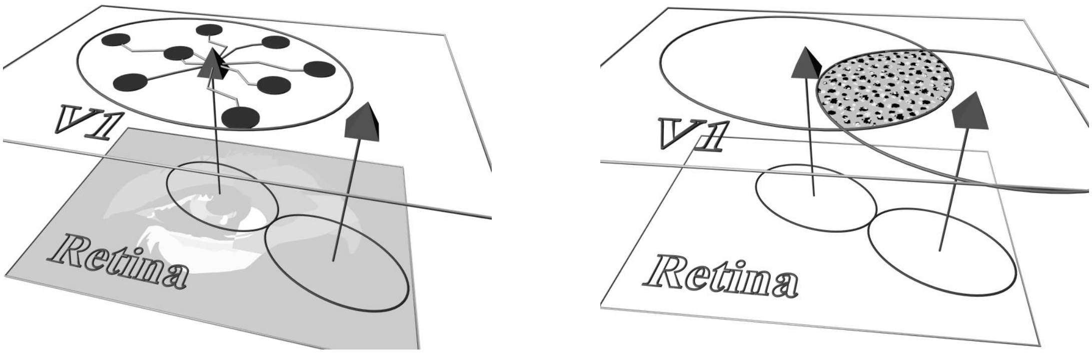


Figure 3.19.: *Left:* Intra-cortical relation between two neurons with non-overlapping receptive fields. Because of the large extend of the intra-cortical connections a neuron with a cortical position in the shaded area (*right*) can mediate information between the two neurons

In this section we set up a mean–field model of V1, which takes into account this geometric arrangement and systematically characterizes its consequences for local cortical processing. We take into account only local connections which are all unspecific: (i) excitatory neurons connecting to neurons up to $400\mu m$, (ii) inhibitory neurons connecting up to $750\mu m$ and (iii) inhibitory neurons connecting only to neurons $250\mu m$ away (Lund, 1987b; Fitzpatrick, Lund and Blasdel, 1985). However, we omit long–range excitatory connections (up to $3000\mu m$) because at this point we are interested in poly-synaptic signaling cascades by short-range connections only.

3.3.1. Model description

We model a cortical layer by a two–dimensional sheet of coupled orientation selective neuron populations (Bartsch et al., 2001). We describe their dynamics within a mean–field framework. A measured orientation map

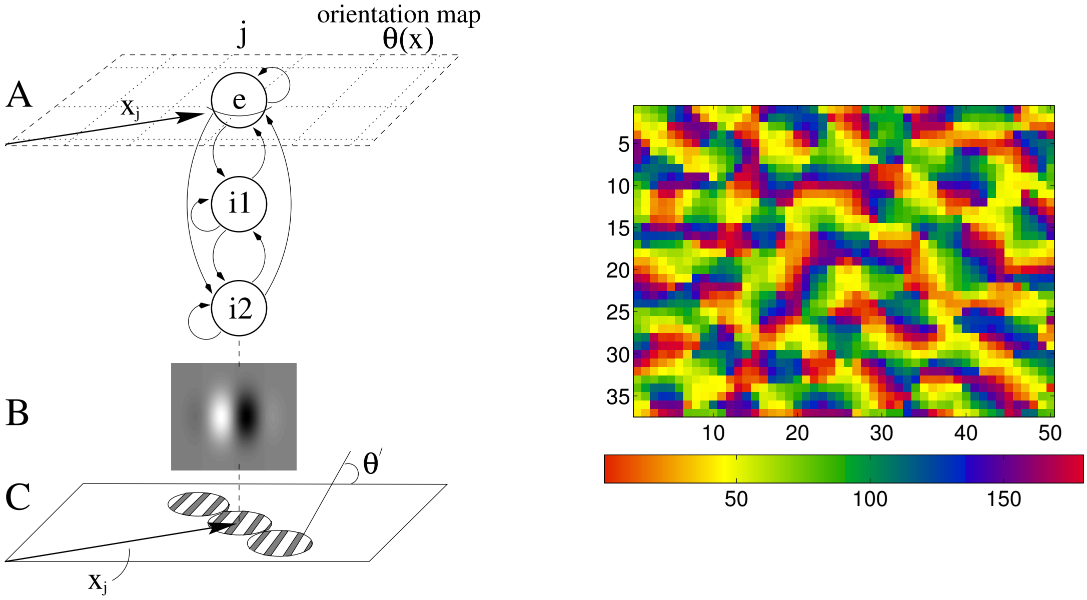


Figure 3.20.: Schematic sketch of the model setup *left*. (*A*) In a sheet of model neurons there are three types of neuron populations for each cortical position (one excitatory (e) and two inhibitory (i_1, i_2)). Neuron populations are interconnected following the lateral connectivity given in Equation 3.32 on page 72 (arrows). (*B*) Receptive fields of neuron populations are modeled as Gabor filters, resembling the basic properties of simple receptive field of neurons in V1. (*C*) The input consists of a sinus grating with orientation θ'. *Right:* color coded the preferred orientation of all neuron populations. Figures from (Bartsch et al., 2001)

(macaque monkey, Blasdel, Obermayer and Kiorpes (1995)), which represents

$4mm \times 3mm$ of the cortex, was divided into a grid of 50×37 orientation columns. Their positions are denoted by vectors $\mathbf{x}_j$, which also represent their retinal positions in the visual field. Each position hosts an excitatory (e) and two inhibitory (i_1, i_2) neuron populations. The preferred orientation θ of each population depends on its cortical position and is read out from the orientation map $\theta(x)$. Activations of model neurons are described by the mean firing rate $m_\alpha(x_j, t)$ of population α at position $\mathbf{x}_j$:

$$\begin{aligned} \frac{d}{dt} m_\alpha(x_j, t) &= -m_\alpha(x_j, t) + g_\alpha\left(h^{\mathrm{LGN}}(x_j, t) + h_\alpha^{\mathrm{lat}}(x_j, t)\right), \qquad (3.27) \\ & \alpha = e,\ i_1,\ i_2; \quad j = 1, \ldots, N. \end{aligned}$$

h^{LGN} and h_α^{lat} are the mean geniculate and intracortical synaptic inputs, respectively, and $g_\alpha(h) = \max(0, \beta_\alpha(h - T_\alpha))$ represents a thresholded nonlinear activation function. Parameters of the activation transfer function were chosen according to (Bartsch, Stetter and Obermayer, 1999c) to allow a wide parameter range with realistic contrast–independent orientation tuning and saturating contrast–response functions as $T_e = T_{i1} = 1$, $T_{i2} = 3$, $\beta_e = \beta_{i1} = 0.5$, $\beta_{i2} = 1$, and were fixed for all simulations.

Gabor filter

To model the geniculate input h^{LGN}, first the correlation of a two–dimensional stimulus $t'_\theta(\mathrm{x})$ and a Gabor filter $p_\theta(\mathrm{x})$, which models the receptive field of a cortical simple cell at x, is computed. Using absolute values of the synaptic input results in oriented π–periodic input that behaves like quadrature pairs in the neuron populations. This value is then fed through a logarithmic non–linear function mimicking the influence of the LGN,

$$h^{\mathrm{LGN}}(x_j) = \ln\left(c\left|\int t_{\theta'}(\mathbf{x}) p_\theta(\mathbf{x} - \mathbf{x}_j) d\mathbf{x}\right| + 1\right) \qquad (3.28)$$

and c being the contrast of the stimulus in arbitrary units.

$p_\theta(\mathbf{x})$ is formulated as a DC-free filter (cf. (Daugman, 1988)),

$$\begin{aligned} p_\theta(\mathbf{x}) &= \frac{k^2}{\sigma^2} \exp\left(-\frac{k^2}{2\sigma^2} x^2\right)\left(\exp(i\mathbf{k}_\theta \mathbf{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right), \qquad (3.29) \\ & ||p_\theta(\mathbf{x})||^2 \approx k^2 \end{aligned}$$

where $\mathbf{k}_\theta = \mathbf{R}(\theta) 2\pi/60$ is the center frequency vector, $\mathbf{R}(\theta)$ is a vector orthogonal to the preferred orientation θ. Center frequency $k = 2\pi/60$ was adjusted to yield a width of 60 pixel for one wave. $\sigma = 2$ ensures essentially two oscillations within the Gabor field. Artificial stimuli are modeled as combinations of circular regions of complex wave functions $t_{\theta'}(\mathbf{x})$

$$t_{\theta'}(\mathbf{x}) = \exp(i\kappa_{\theta'}\mathbf{x}) \qquad (3.30)$$

where $\kappa_{\theta'} = \mathbf{R}(\theta')2\pi/30$ is the wave vector of the sinusoidal grating and $\mathbf{R}(\theta')$ is a vector orthogonal to the direction of the grating.

The lateral input, h_α^{lat}, is given by

$$h_\alpha^{\text{lat}}(\mathbf{x}_j, t) \;=\; \sum_{\beta=e,i_1,i_2} S_\beta(\mathbf{x}_j - \mathbf{x}_l) m_{\beta,l}(x_l, t), \tag{3.31}$$

where $S_\beta(\mathbf{x}_j - \mathbf{x}_l)$ denotes the mean connection strength formed by a neuron population of type β at $\mathbf{x}_l$ to all neuron populations at point $\mathbf{x}_j$. Note that all three target neuron populations at a given location $\mathbf{x}_j$ receive identical input.

The intracortical connection strength was chosen independent of the distance from population β at $\mathbf{x}_l$ to the population at x_j.

$$S_\beta(\mathbf{x} - \mathbf{x}_l) \;=\; \begin{cases} S_\beta / N_\beta & : \; |\mathbf{x} - \mathbf{x}_l| < \sigma_\beta \\ 0 & : \; \text{else} \end{cases} \qquad \beta = e, i_1, i_2 \tag{3.32}$$

N_β is the overall number of connections formed by this particular population, and S_β is a scalar value describing the overall strength of the connections. σ_e refers to the size of the excitatory connections , σ_{i1} and σ_{i2} to the size of connections from inhibitors with low and high thresholds, respectively.

3.3.2. Results

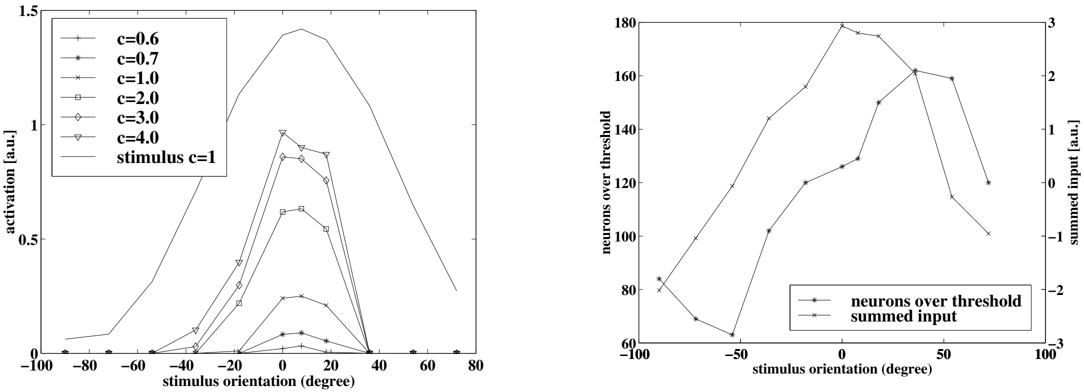
For high recurrent connectivity the model shows a hexagonal arrangement of activation blobs depending on the wavelength of coupling.

We fixed the principle wavelength of the intracortical couplings to fit the wavelength of the underlying orientation map ($550\mu m$, $\sigma_e = 160$, $\sigma_{i1} = 290\mu m$). The principal wavelength of intracortical couplings was computed by calculating the maximum of the Fourier–transform of

$$\begin{aligned} &H(x + \sigma_e) - H(x + \sigma_e - 2\sigma_e)/(2\sigma_e)S_e - \\ &[H(x + \sigma_{i1}) - H(x + \sigma_{i1} - 2\sigma_{i1})/(2\sigma_i)S_i]\,, \end{aligned} \tag{3.33}$$

where $H(x)$ denotes the Heaviside–function. Using larger wavelengths ($800\mu m$) leads to an inconsistent behavior of the model: Some regions show no activation though they receive excitatory input, because the cortical couplings enforce a roughly hexagonal arrangement of activity centers, which are not consistent with the underlying iso–orientation domains.

First we are interested in the effect of changing the strength of the recurrent excitation on the response properties of the neuron populations. We computed orientation–tuning curves for different levels of input contrast. The result (see Figure 3.21 on the facing page) shows that the model neurons produce saturating contrast–response curves. At the preferred orientation of the neuron



(a) (b)

Figure 3.21.: (*a*) Response of an excitatory neuron population (preferred orientation 7°) to a stimulus patch of different orientations (sinusoidal grating) for different stimulus contrasts. The solid line *without marker* shows the thalamo–cortical input for this neuron population for $c = 1$. The stimulus has an optimal phase for the shown cell. Coupling strength was chosen to be $S_\beta = 100$. The tuning is un-symmetric and its shapeness differs between the left and the right flank. (*b*) Numbers of activated neuron populations together with the global synaptic input for the neuron population $\mathbf{x}_j$ ($c = 4$) over stimulus orientation

population ($\approx$ 7°) the response change for equidistant contrast changes is reduced at hight contrast levels. The reason is that for strong recurrent excitation, high threshold inhibitory neuron populations are activated and suppress the response of the other neuron populations. Figure 3.21b, on the page before shows that the number of neuron populations, which take part in the cortical circuitry, depends on the stimulus orientation. At $\theta = 50$ degree, more neuron populations contribute to the recurrent feedback loop than at $\theta = -50$ degree, hence competition is stronger at $\theta = 50$ degree and leads to contrast invariant onset of the tuning curve as opposed to $\theta = -50$ degree. In other words, local inhomogeneities in the orientation map cause a varying number of neurons to participate in the local recurrent circuit. If the number of active neuron populations is higher, the circuit is shifted into a regime (the marginal phase), where orientation tuning is contrast–invariant (right half of the Figure 3.21 on the preceding pagea).

Now we address the issue of context effects in this model. Stimuli in the non-classical surround of a neuron population x_j have only modulatory effects according to experimental results. Figure 3.23a on page 76 shows the responses of all excitatory neuron populations to a stimulus of diameter $600\mu m$ centered above the neuron shown in Figure 3.21 on the preceding page. Figure 3.23b on page 76 shows the response to two flanking stimuli outside the CRF. In accordance with experiments these stimuli cannot elicit any response at the center location.

Figure 3.22 on the facing page shows that facilitation and suppression arise depending on the mutual configuration of the center and flanking surround stimuli. Figure 3.22 on the next page (solid polygon) plots the activity of the cell (radius encodes response level) versus the axis between the flanking stimuli (angle encodes the orientation of the axis). This particular cell is slightly facilitated if the flanking stimuli are spatially aligned with the center orientation (thick bar), whereas it is strongly suppressed for orthogonal flanking stimuli. The anisotropy of the contextual modulation is induced because the input to the center neuron differs depending on the orientation and the arrangements of the flanking stimuli.

Figure 3.22b, on the facing page, and Figure 3.22c, on the next page display gray–level plots of the cortical synaptic input evoked by a center stimulus, flanked by aligned (b) and orthogonal (c) stimuli. The inputs at the points marked $\#1$ and $\#2$ differ for changes in the alignment of the surround stimulation. Activity blobs $\#1$ and $\#2$ are more active for orthogonal aligned stimuli (Figure 3.22c, on the facing page) than for co–aligned stimuli (Figure 3.22b on the next page). Figure 3.23 on page 76 demonstrates how these changed inputs can evoke contextual modulation by local connections. The cell populations which are excited by the center and the flanking stimuli are

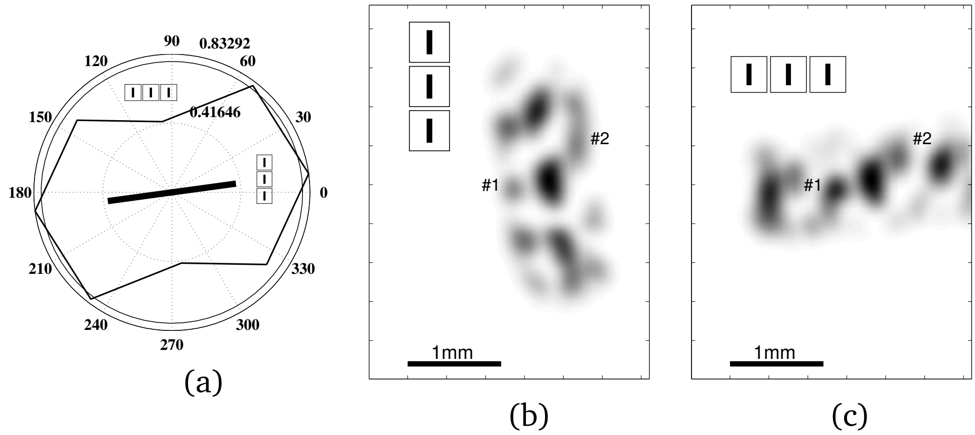


Figure 3.22.: Facilitation and suppression for different configurations of input patterns because of changes in the local input. *(a)* Response of the considered neuron population to different configurations of surround patches shows facilitation for aligned stimuli along the axis of preferred orientation (thick bar in the middle) and substantial suppression for orthogonal aligned stimuli. The activation of the center neuron population for center stimulus alone was 0.788. *(b)* Input pattern for co–aligned stimuli along the preferred orientation. *(c)* Input pattern for co–aligned stimuli orthogonal to the center stimulus. Parameters match that of Figure 3.23 on the next page. Insets show the used stimulus paradigm. Figures from (Bartsch et al., 2001)

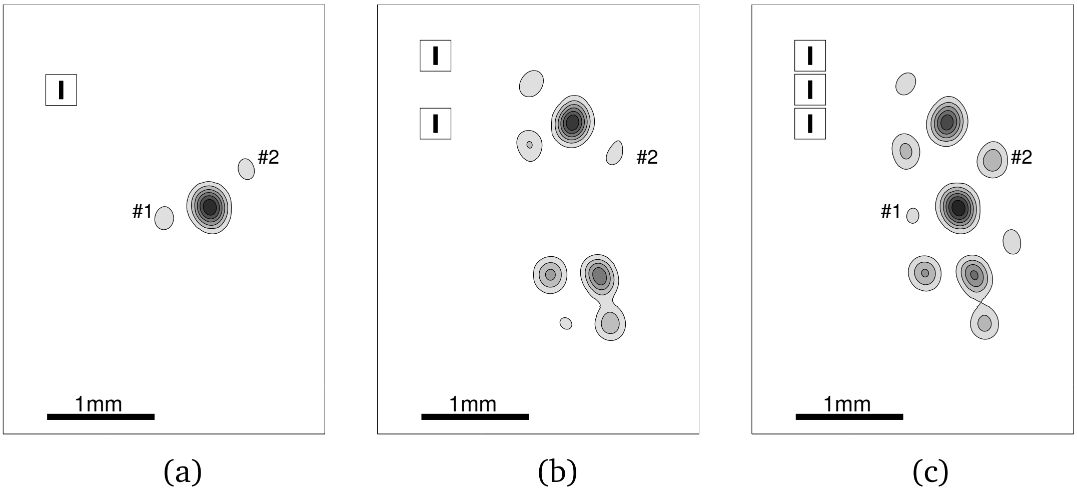


Figure 3.23.: Stimulation of center and surround changes the response of cortical neuron populations via short–range interactions. The responses of excitatory neuron populations for an oriented grating and two co–linear stimuli ((*a*) center stimulus alone, (*b*) surround stimulus alone and (*c*) center plus surround) are shown. Additionally two activation blobs are marked for further reference in the text ($\#1$, $\#2$). Surround stimulation alone cannot elicit response. Parameters are: radius of grating patches $300\mu m$, center to center distance of co–aligned patches (same radius) $\pm 600\mu m$. Insets show the used stimulus paradigm. Figures from (Bartsch et al., 2001)

not completely disjunct. The patches marked $\#1$ and $\#2$ are driven by both, center and surround, and communicate contextual effects via a bi–synaptic or poly-synaptic signaling cascade. Suppression is observed if a change in activity induces inhibition in their respective surround, that is at the center position of the stimulus. It reduces the response at the center for co–aligned stimuli in the surround, we get a suppressive effect.

Facilitation can be explained by the following: activity at intermediate positions $\#1$ and $\#2$ can be reduced by the surround stimuli in comparison to the activity induced by the center stimulus alone. This leads to an explanation of facilitatory effects by dis-inhibition.

We propose that two opposite effects contribute to the observed contextual modulation; (i) local inhibition that is induced by a local change in input (leads to suppression), and (ii) dis-inhibition. By changing the configuration of the stimulus different regions of the orientation map are activated. Changes in the local structure then define which is more prominent, suppression or facilitation.

3.4. Concluding Remarks

In the lattice model we observe a slight shift of the activity blobs depending on the stimulus (see Figure 3.23 on the preceding page). This leads to a divergent behavior of the neuron activation. Some populations may actually be facilitated but at the same time other neuron populations in the surround may show suppressed activity. This shift is induced by the enforced hexagonal arrangement of activity centers.

Our model shows strong interactions mediated by poly-synaptic connections. For the findings of Das and Gilbert (1999) that contextual effects might also be mediated by short-range connections we found two possible mechanisms: lateral inhibition and dis–inhibition. In addition, anisotropic contextual modulation can be an emergent property of a network with ideally isotropic and local connectivity. We suggest that it results from a different weighting of inhibition and dis–inhibition depending on the configuration of flanking stimuli.

In the last chapter we analysed the role of the intro-cortical networks in the generation, sharpening, and modulation of orientation preference as one of the main features of primary cortical neurons. The mathematical theory of interacting hypercolumns in primary visual cortex incorporated details concerning the arborization of three different neuron populations. As analytical methods we used the stationary state analysis. Whereas we took special care about not introducing time-periodic pattern the dynamic properties of the cortex are worthwhile to analyse. They can be compared, for example, to the findings of Volgushev et al. (1995) (see Figure 2.6 left, on page 23).

Bressloff and Cowan (2002) introduced bifurcation theory for the ring-model of the orientation hypercolumn. They derived non-linear equations for the amplitude and phase of the population tuning curves. It would be worthwhile to repeat this analysis for our model incorporating the third neuron population. Unfortunately both of the above projects could not been performed as part of this thesis.

We leave now the field of biological modeling in favor of a more theoretical ground which is concerned with the special nature of the input into the visual system.

4. Invariance and Symmetry

> *"They where masters of geometry.* 60 *stones build a circle of exactly* 360 *degree."*
>
> *TV documentation*

Biological Motivation

A popular rationale for the response characteristics of visual cortical neurons is that of convergent, weighted input from more specialized cells. For example the outputs of ON-OFF retinal ganglion cells converge to orientation selective simple cells in primary visual cortex. High up in the hierarchy, for example, in the inferior temporal visual cortex (IT) we find neurons that respond to objects or faces independent of their positions on the retina over many degrees (Gross, Desimone, Albright and Schwartz, 1985). But also neuron in lower visual areas respond well to complex stimuli (see Section 2.3.2 on page 26), sometimes even stronger than to the classical moving grating stimuli. Also the experiment of Földiàk (2001) on the non-linear reconstruction of complex cell receptive fields indicate that additional information is extracted by these neurons.

To explain the apparent contradiction of having neurons early in the visual pathway, showing strong responses to complex stimuli we hypothesize that neurons possess receptive fields characteristics that can be described in terms of second order correlations of image intensities. Here, we implicitly assume that the observed specificities of neural responses cannot fully be described by linear models which weigh single pixels only. This does not exclude that some neurons (e.g., simple cells) can be reasonably described by linear models, but, as we are going to higher levels of visual processing some non-linear processing strategies are likely to be used by the visual system, in the case of complex cells this may even happen at the very first stages of visual cortical processing. This seems reasonable also in the light of the complex nature of the visual input (see Section 1.3).

Statistical Motivation

Basically we have already seen in the Section 1.3 about natural images that there is a strong motivation in terms of statistics for symmetry detection. Natural images were characterized as containing a huge amount of redundancies.

To repeat some of that reasoning: A representation of the images by their non-redundant parts is preferable, because of its short description length. By using a short representation we select the correct representation in terms of (Kolmogorov) complexity (see Section 1). The non-redundant parts are also found in being highly structured, containing for example edges.

Kolmogorov complexity

Interestingly we demand here a search for structure in the data. Only the structured components hidden in the data will allow an efficient data representation. In this work we propose that symmetry is a measure of structure, thus highly symmetric components have a large amount of structure and can be used to efficiently represent the data. To avoid any confusion at the beginning the symmetry we are talking about differs from the geometrical abstract form of symmetry. We instead will introduce a symmetry measure that is smooth (more a scalar than a binary value) and can be evaluated locally in digitalized images.

4.1. Related Models

Response properties of visual cortical neurons are mostly described in terms of selectivity and invariance meaning that a neuron responds selectively to some stimulus feature, while at the same time does not respond to some other feature in the stimulus. A complete description based on a neurons selectivities is difficult because of the unknown number of possible features the neuron may respond to. Therefore, one generally intermixes the two statements; stating that complex cells are selective to the orientation of a grating stimulus but invariant to the spatial phase of the pattern.

4.1.1. Learning Invariance

Models for learning invariances in natural images try to cope with the transformations that images undergo if the position of the observer changes. If images shift, scale, and rotate, the networks in the brain have to learn these invariances in order to represent the information efficiently and to ensure stable representations. Models for learning invariances either rely on temporal sequences of input patterns undergoing transformations (Rao and Ballard, 1998; Földiàk, 1991; Wiskott and Sejnowski, 2002; Einhäuser, Kayser, König and Körding, 2002) or on modifications to the distance metric for comparing input images to stored templates (Simard, LeCun and Denker, 1993).

Rao and Ballard (1998), for example used a *transformation invariant coding* strategy to code images based on a first-order Taylor expansion, and ob-

transformation invariant coding

tained localized, oriented receptive fields from natural image inputs. Due to the first-order approximation only small transformations (e.g., small invariances) could be learned, the performance of the model was reduced for large transformations.

Rao and Ruderman (1999) introduced a Li-group approach (see Section 4.1.3) which could handle 1-D transformations and 2-D rotations. The model is based on the notion of continuous transformations and Lie group theory. In this approach a matrix G is learned that is called the *generator* of the transformation group. Applying this matrix to the input one assumes an infitesimal small change of the input according to the transformation. A macroscopic transformation can be produced by chaining together a number of these infitesimal transformations. This leads to an exponential based generative model of an image

$$I(x) \;=\; \exp^{xG} I_0 \tag{4.1}$$

where I_0 is the initial or 'reference' input. The model now learns the generator matrix G by a series of before and after images (before the transformation and after the transformation). It turns out to be problematic to learn different transformation generators G at once. Also the single transformations have to be small to ensure successful learning.

slow feature analysis

Another example for a model that uses the temporal sequence of the input is the *slow feature analysis* (Wiskott and Sejnowski, 2002). It is based on the assumption that a slowly varying representation can be considered to be of higher abstraction level than a quickly varying one. So we can observe a large change in illumination of pixels for a moving grating stimulus but the orientation of the grating is a more slowly varying feature thus of better use to describe the stimulus than the changes on the level of single pixels. The model use first and second degree monomials to describe the neurons' responds by a polynomial.

Another model that utilizes slowly varying features in order to learn invariances is the one by Eglen, Bray and Stone (1997). By jointly maximizing the long-term variance of the output and minimizing its short-term variance a network model could learn to discover stereo disparity and feature orientation.

The model of Einhäuser et al. (2002) also relies on temporal sequences. As a hierarchical network model it learns simple and complex cell receptive field properties based on a sparseness constrain (see (Olshausen and Field, 1996; Hyvärinen and Hoyer, 2000) for other models that use sparseness).

Crucial for models of temporal sequence learning is that either the type of invariance has to be predefined or the transformations between successive inputs have to be comparatively small in order to learn the right invariances.

It is questionable if the visual input coming from the eyes produces such a slowly varying stimulus transformation. For example, rapid eye movements should be rather common (cf. saccades) shifting our focus of interest over large proportions of the input image. Micro-saccades on the other hand occuring between saccades could introduce relatively small shifts.

A model that learns invariances (or a related feature class) for independently drawn input signals would be favorable. We need also to simultaneously learn more than one transformation.

4.1.2. Learning Symmetry

It is difficult to learn invariances without at least two closely connected patterns, one pattern before and one after the transformation in question was applied. In that sense a pattern is invariant if a mathematical or physical process has not changed the pattern. If we look for invariances we refer normally to processes like shift of gaze, micro-saccades or object movements, or rotations that causes the 'after transformation' pattern. For independently drawn pattern these processes are not directly observable.

There is a way around this. A related measure that is a feature of a single pattern, e.g., accessible from independent drawn samples is *symmetry* or self-similarity. Like invariance it reflects a property of a pattern not to be changed after some transformation. Normally the transformations we speak of in symmetry are rigid motions of geometrical figures, but we like to weaken this 'geometrical' symmetry in favor of a more general continuous property that is closer to invariance (Bartsch and Obermayer, 2002; Bartsch and Obermayer, 2001; Masuda, Yamamoto and Yamada, 1993; Zabrodsky, Peleg and Avnir, 1995).

symmetry

As we will see symmetry can be detected by local visual processes relying on correlations, so symmetry detection does not demand great computational resources. One example for invertebrates using symmetry as a visual feature are bees which according to Moller (1995) can detect the symmetry of flowers. In general, symmetry can signal the *presence* of objects in the visual scene and therefore is likely to be used to direct visual attention.

Evidence for Symmetry as a Visual Feature

Symmetry has been dealt with in both, art (architecture, sculpture, painting) and science (mathematics, physics, chemistry). Thus symmetry can be discussed in different aspects. In this section we will be mainly concerned with psychophysical findings of spatial visual symmetry detection.

Symmetry has been found to be a pre-attentive feature similar to size, brightness, color or movement. That means it is detected very fast, without eye movement, requiring less than one second (Corbalis and Roldan, 1974; Royer, 1981). In particular it is faster processed than form, shape, and structure, which all require registration in short term memory (Attneave, 1955), and is also independent of color processing (Morales and Pashler, 1999).

Different types of symmetry are known to be detected with different reaction times. The detection time for rotationally symmetries where found to be longer than for the detection of mirror symmetries (Royer, 1981; Corbalis and Roldan, 1974). For the mirror symmetries the vertical symmetry (axes) was found to be the most salient and easiest perceived symmetry.

mental rotation
template matching

In this context the question arises if the finding could be explained better by a mental rotation or by a template matching model. In particular, Royer (1981) found that the reaction times for detection of vertical symmetry are shorter than of horizontal and diagonal mirror symmetry, which in turn are shorter than the reaction time for the detection of rotational symmetry. In favor of the template model Corbalis and Roldan (1974) found that the reaction time to detect symmetry in a random dot cluster increased substantially as the angle to the vertical increases.

One of the earliest explanations of how the visual cortex detects symmetry was given by Julesz (1971). He pointed out that the saliency of vertical symmetry was due to interactions between the symmetry of the pattern and the bilateral symmetry of the visual system. Detection of symmetry could be performed by a point by point comparison based on neuroanatomy which is symmetric about the fovea. One concludes that the required symmetric projection to the visual system would be destroyed if the fixation point for symmetry detection was not on the symmetry axis. Nevertheless, Barlow and Reeves (1979) and Masame (1983) showed that although there is a decrease in performance, symmetry is also detected if the symmetry axis is displaced to the right or left of the fixation point.

Dakin and Herbert (1998) showed that symmetry detection exhibits *scale-invariance*. The size of the integration region to detect symmetry was measured with spatially band-pass filtered noise images in which symmetrical patches where embedded (see Figure 4.1 for an example stimulus). They showed that the size of the integration region varies in inverse proportion to spatial frequency and was elongated in the direction of the axis of symmetry, with an aspect ratio $\approx 2:1$. These results are compatible with a central role for spatial filtering in symmetry detection.

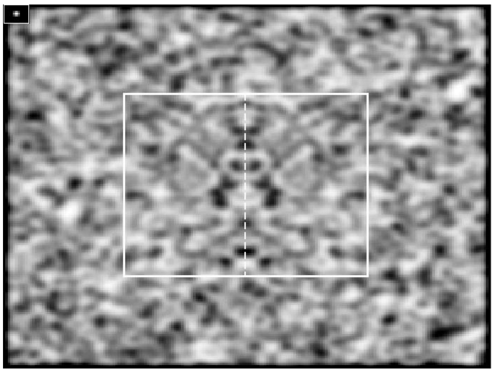


Figure 4.1.: Stimulus similar to the ones used by Dakin and Herbert (1998) for measuring the integration region of symmetry detection in humans. The inset top left shows the Gaussian filter used for bandpass filtering of the random dot image

Models for Symmetry Detection

There is rich literature on models for detecting symmetries in images primarily used in object recognition or shape representation (Gofman and Kiryati, 1996; Blum and Nagel, 1978; Brady and Asada, 1984; Pizer, Oliver and Bloomberg, 1987; Bruckstein and Shaked, 1995; Zabrodsky et al., 1995; Ponce, 1990). Here symmetry is thought of as a global feature of the image or the object displayed therein. Contrarily, later on we will assume that symmetry can also be defined as a local feature of parts of the image.

The input of some early algorithms of symmetry detection was assumed to be coming from a successful segmentation procedure. These shape based algorithms heavily rely on the preprocessing. Because general segmentation of images is a problem that is still not solved the applicability of shape-based symmetry analysis algorithms is limited. An example for problems with shape analysis is presented in Section 5.4.2 on page 110. There an algorithm based on gray values outperforms a shape based approach for image alignment.

Other algorithms compute symmetry by analyzing an edge or line rather than a binary images (Cham and Cipolla, 1995; Ogawa, 1991; Ylä-Jääski and Ade, 1996; Brady and Asada, 1984). Basically the same problem appears in this context. Extracting clean edges is as difficult as segmentation. Most pre-processing algorithm destroy some type of information in the data in order to highlight another. Pre-processing is bound to be task specific because what is neglectable information in one task is crucial in another. For this reason symmetry analysis is preferably done on grey level data. Our

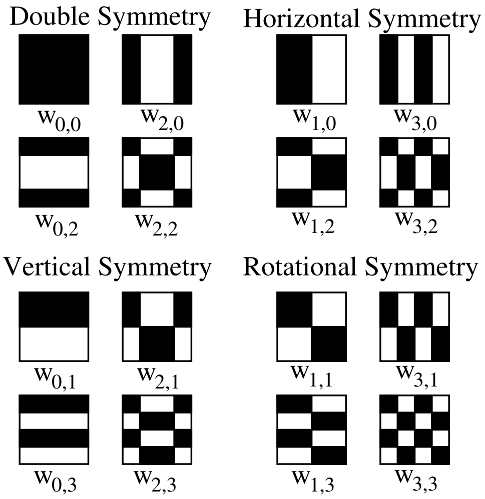


Figure 4.2.: Walsh basis functions used for symmetry detection

previous observation that symmetry detection is performed pre-attentive also indicates an early position for symmetry detection in the visual pathway.

Walsh basis functions

Some basis function approaches have been explicitly suggested for symmetry detection. One example are *Walsh basis functions* (see Figure 4.2). The Walsh functions form an orthogonal and complete set of functions representing a discretized function (Rao, 1983). There is also a direct relationship between Walsh functions and wavelet decomposition. Any wavelet component can be obtained as a superposition of all the Walsh components of the same order.

The basis functions denoted by $W_{n,m}$ (where n, m are integer values) can be separated into four different classes: (*i*) vertical mirror-symmetry (m-even, n-odd), (*ii*) horizontal mirror-symmetry (m-odd, n-even), (*iii*) double mirror symmetric (m-even, n-even) and (*iv*) rotational symmetric (m-odd, n-odd). A vector of four values can be calculated for each symmetry from a given image. For example, the entropy of these values represents the symmetry information of that image.

Another basis function approach was used by Bigün (1988) to detect rotationally symmetric images. Basis functions were defined as spirals with varying number of 'arms' and variable curvature (see Figure 4.3, left). Given an image of radius R the basis functions are given by

$$\phi_{m,n}(r,\theta) \quad = \quad \exp\left(i(m\omega r + n\theta)\right) \tag{4.2}$$

where $\omega = 2\pi/R$ and (r, θ) are polar coordinates. Note that m, n are natural numbers (sign of n defines the 'handiness' of the spiral). The above equation

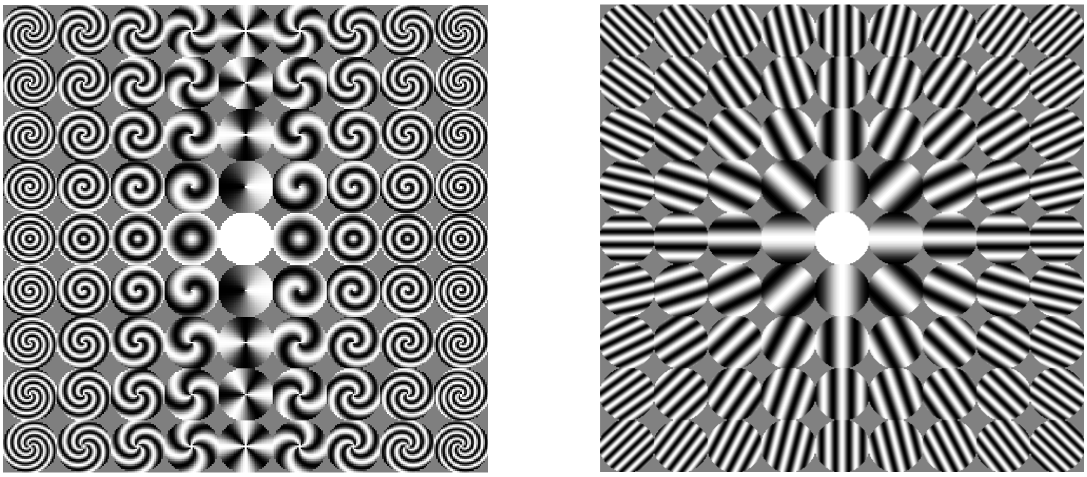


Figure 4.3.: *Left:* Basis functions to detect rotational and circular symmetry in images. *Right:* Basis function to detect orientations with different spatial frequencies

describes the local (in m, n) solutions of a differential equation obtained from a set of operators by the Lie Transformation Group model which is explained in more detail in the next section.

fractal coding

Note that there are more models connected to symmetry detection, for example *fractal coding* (Jacquin, 1990), a technique to find a transformation on an image, for which the result of its application upon the image will be to leave it unchanged. This transformation was used then in the context of image coding, e.g., image compression.

4.1.3. The Lie Transformation Group Model

Lie transformation group

Hoffman (1965) first formulated the concept of the Lie Transformation Group (LTG) in the context of visual perception. The model claims that it can explain how the locally smooth phenomenons that occur in the visual cortex lead to a model of human vision.

Our eyes receive constantly changing inputs due to our movement of body and head. The majority of this movement in our sensory input will be provided by transformations from the two-dimensional affine transform group. In order to have continuity in the perception of objects over time, human vision should account for such transformations.

The LTG model proposes a set of three basic pairs of operators, or transformations, which can be used to construct a model of the human vision. These three pairs come in the form of mutually orthogonal orbits. The first operator pair accounts for simple translation in the x respectively y direction

$$\mathcal{L}_x = \partial/\partial x \qquad \mathcal{L}_y = \partial/\partial y. \tag{4.3}$$

Examples for the corresponding orbits can be seen in Figure 4.3, right. Directions of equal gray value indicate the changes for which the pattern recognition process should be transparent. The patterns are generated as local[1] solutions of a differential equation calculated as

$$\phi_{m,n}(X,Y) = \exp\left(i(\omega/2mX - \omega/2nY)\right) \tag{4.4}$$

where ω was defined as in Equation 4.2 and X and Y are grid positions in x-respectively y-direction.

As well as lateral movement we can also include rotations and scalings

$$\mathcal{L}_S = x(\partial/\partial x) + y(\partial/\partial y) \quad \mathcal{L}_R = -y(\partial/\partial x) + x(\partial/\partial y). \tag{4.5}$$

An example of this operator pair and the corresponding orbits can be seen in Figure 4.3 on the page before, left. The corresponding solution of the differential equation can be found in Equation 4.2. It is apparent that these operator pairs are not unique. A third pair is usually defined as:

$$\mathcal{L}_\beta = x(\partial/\partial x) - y(\partial/\partial y) \quad \mathcal{L}_b = y(\partial/\partial x) + x(\partial/\partial y) \tag{4.6}$$

4.1.4. Symmetries of the Visual Cortex

Bressloff, Cowan, Golubitsky, Thomas and Wiener (2001) used the assumption that the anatomical connection structure of the visual cortex itself exhibits symmetries rendering it invariantly under the actions of the Euclidean group E(2) (see (Alexander, Sheridan and Bourke, 1997) for a similar ansatz to explain orientation selectivity).

If a system is constraint by some symmetries this defines the pattern that the system will generate spontaneously. For an unconstraint system generally a spatially constant solution emerges, which has by definition the largest number of symmetries. Each restriction introduced into the system now reduces (breaks) the maximum symmetry solution and leaves only specific symmetries as solutions. Bressloff et al. (2001) could show that the pattern created by a specific set of actions which were motivated by the anatomical structure of the visual cortex were used to predict common visual hallucinations. Using this model one can explain the pattern that appear due to abnormal states of the brain as they occur in the case of migraine or epilepsy.

[1] In this case also global solutions.

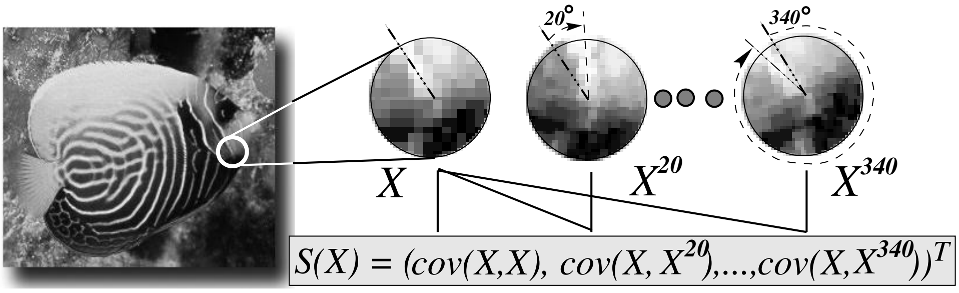


Figure 4.4.: Example image and sketch of the procedure of calculating $\mathcal{S}$

Whereas the latter model analyses the structure of the cortex in order to explain phenomena of human vision, we will reverse this ansatz in the following chapters. We will derive and analyse a model of the visual world in order to make predictions about the mechanisms of the visual cortex.

4.2. A Structure Preserving Transformation

In the light of symmetry as a pre-attentive feature to detect the presence of objects in a scene, we now introduce a measure of local rotational symmetry on gray valued digital images. Later on we will see that this model is compatible with a more general model of a quadratic form by which in the end we will be able to learn local features from natural images.

One basic idea to detect rotational objects is to calculate a *'product of the pattern with the rotated pattern'* (Bartsch and Obermayer, 2001). Correspondingly, a stimulus is interesting if the dot-products of the stimulus with a rotated version of the stimulus is large.

Operations on pixel coordinates (like rotations) work in general only for images in a continuous function space. It is difficult to find this space for digital images. One way is to define an ambiguous prior on the image structure (for example smoothness) for the re-calculation of coordinates. The idea of Ruderman and Bialek (1992) for example could also be used to construct a more educated guess about the sub-pixel structure of images based on the largely scale invariance of natural images.

Instead, we will here rely on the smoothness assumption. A simple interpolation techniques is used to approximate sub-pixel positions within the equally spaced (Cartesian) grid of pixel. This of course introduces discretization errors that will show up as prior structure if the smoothness assumption is not valid.

In the following we will refer to X as a stimulus or input in the receptive field of diameter $r_{\mathtt{max}}$ of a neuron implementing $\mathcal{S}$.

Let $\mathcal{S}$ be a transformation $\mathbb{R}^2 \rightarrow \mathbb{R}^{2(2\pi/\theta_{\mathtt{min}})}$ that is sensitive to structure:

$$\mathcal{S}(X) = (\mathcal{C}(\theta_{\mathtt{min}}), \ldots, \mathcal{C}(2\pi - \theta_{\mathtt{min}}))^T, \qquad \mathcal{C}(\theta) = \mathtt{cov}(X, X^\theta) \tag{4.7}$$

X^θ describes the rotation of the center of the receptive field of a neuron in $\mathbb{R}^2$ space by θ and $\mathtt{cov}(\cdot,\cdot)$ is the covariance function which is used to measure the similarity C of the original and the rotated image patch (see Figure 4.4). It is problematic to compute X^θ because the rotation in pixel coordinates is not well defined, however, we will assume for the moment that with certain tricks[2] by the computer vision community we can be pretty close to the correct solution (continuous pixels). We also do not need to apply our transformations (rotation) successively as in the generator model of Rao and Ruderman (1999). There are several ways of defining a scalar value for the symmetry. For example, we can use the $L1$-norm of $\mathcal{S}(X)$. This will result in a single value per pixel and thus it can be visualized as a *response* image. At this point this procedure is only used to illustrate the general process of calculating some 'symmetry' value. To analyse images this way we would rely on a more elaborated approach calculating, for example, the entropy of the symmetry vectors obtained for different images.

A set of 256 image patches, each with dimension 60×60, were drawn randomly from an image (Figure 4.5, left). Mirror symmetries were tested for a set of $\{\theta_0, \ldots, \theta_9\}$ orientations of the mirror axis which were equally spaced in the range of $[0, \pi]$. To avoid edge effects a circular region was used. Figure 4.5, right show the results of a re-sorting of the images according to the calculated symmetry values $||\mathcal{S}||_1$ for each random position (sorted from left to right, top to bottom). The values for $||\mathcal{S}||_1$ are shown in 4.6 and decrease exponentially which indicates a hight specificity of $\mathcal{S}$.

One sees that most symmetric images of this size contain mostly edges whereas the least symmetric ones contain mostly un-structured noise. This is not really surprising because at this scale one can basically classify image patches in the two categories 'noise' and 'contrast edge' (every more symmetric patch, for example a patch containing a circle, is extremely unlikely). The question remains why patches containing edges should be at any rate *more* rotationally symmetric than noisy patches. First of all, we could suspect that edges are better in having an overlap with (mildly) rotated versions of themselves plus 'negative' overlap with 180° rotated versions which is elevated by the $L1$-norm. Secondly, our transformation $\mathcal{S}$ uses covariances. Therefore

[2] Mainly we computed all rotated versions from one unchanged 'master' image, the original X, and in a backward way, e.g. finding for each pixel in X^θ the most likely original pixel in X.

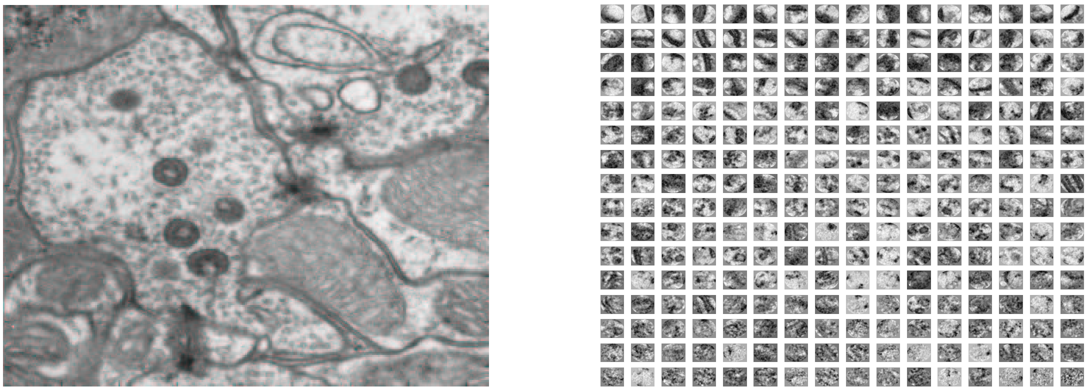


Figure 4.5.: Edges are the most rotationally symmetric image patches of natural images. *Left:* Microscopic image (1400×1400). *Right:* Resorted random samples from a single images according to the symmetry values $\mathcal{S}$ plotted in Figure 4.6

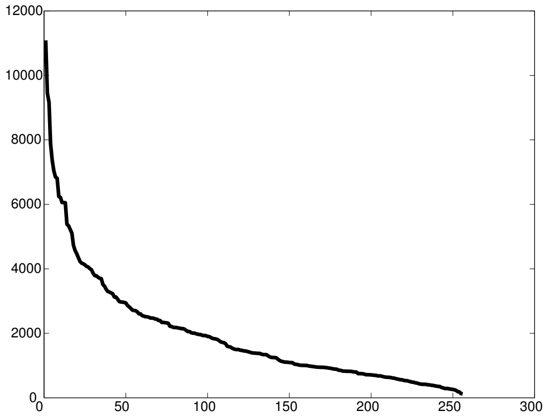


Figure 4.6.: The values of $\mathcal{S}$ for the 256 randomly drawn image patches shown in Figure 4.5

C depends heavily on the variance of the pixels in X. As was found by Reinagel and Zador (1999) high variance patches are most likely edges thus explaining our bias for edges in calculating rotational symmetries.

Although here we focus on detecting rotational symmetries, other types of symmetries can also be implemented in this framework. Detecting mirror symmetries, for example, requires only an additional mirroring of X^θ around its medial axis.

If we look in more detail into the model presented in Equation 4.7 on page 88, we notice that we have computed products of pixel intensities which were summed up to get an according C value. In order to present a single symmetry value for a region in Figure 4.6 we used the $L1$-norm of all C values which is also a sum of absolute values. So our whole procedure can be viewed, in first approximation, as a large sum of pixel products.

$$\mathcal{S}(X) \;=\; \sum_{\theta_{\mathtt{min}}}^{2\pi-\theta_{\mathtt{min}}} \left| \sum_{i,j} (x_i - \mu)(y_j - \mu) \right|, \quad x_i \in X, y_i \in X^\theta \tag{4.8}$$

If we assume that the absolute values are not essential to the function of the transformation the above equation can be represented by a single sum of products of specific pairs $\{s\}$ of pixel.

$$\mathcal{S}'(X) \;=\; \sum_{(i,j)\in\{s\}} x_i x_j \tag{4.9}$$

We emphasize here that the crucial ingredient is the choice of the pixel pairs and not the grouping of the products. In the attempt to calculate rotational symmetries we have selected a specific set s of all possible pixel pairs. If we like to detect mirror symmetries, we simply select another set $\{s'\}$ of pixel pairs but the overall structure of the algorithm remains. Thus the detection of rotational symmetry as done by Equation 4.9 is a special case of a model $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ that can be parameterized as a quadratic form

$$\mathcal{F}(X) \;=\; \sum_i a_i \phi_i(X) = \mathbf{x}^T A \mathbf{x}, \qquad \mathbf{x} = \mathrm{vec}(X). \tag{4.10}$$

where ϕ_i is a cross product of the coordinates of the input vector X and a_i is an entry a_{kl} of the matrix A which is 1 if $x_k x_l$ is in the corresponding set $\{s\}$ and 0 elsewhere. In general we are interested in matrices A which weight pairs of pixels. This is in contrast to basis function approaches that are interested in weights of single pixel ($\phi(X) = \mathbf{A}_i$). The above formulation relates the quadratic form to models of polynomial classifiers and notably to *sigma-pi units* or higher order nets.

sigma-pi units

From now on we will call A a parameterization and $\mathbf{x}^T A \mathbf{x}$ our model.

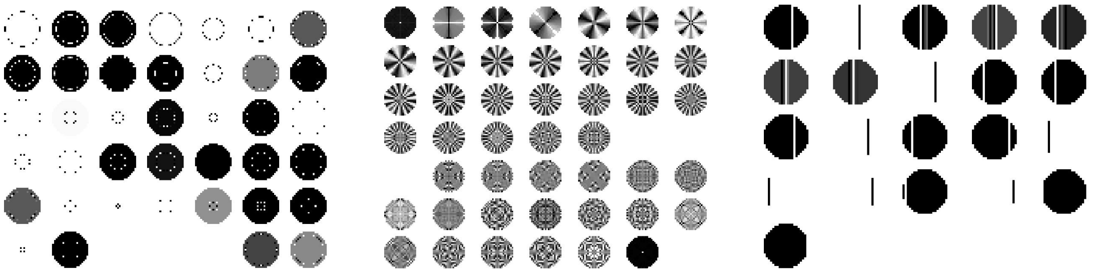


Figure 4.7.: Orthogonal basis functions for rotation (*left*), scaling (*center*) and shift (*right*) computed as eigenvectors of a quadratic form. (Figure from (Bartsch and Obermayer, 2003))

4.2.1. Co-variation and the Quadratic Form

Assuming an ensemble X of gray level images ($E(\mathbf{x} \in X) = 0$), what is the information about the image that is explored by A?

Using an eigenvalue decomposition of the matrix A

$$\mathbf{x}^T A \mathbf{x} \quad = \quad \mathbf{x}^T \sum_i^N \lambda_i \mathbf{n}_i (\mathbf{n}_i^T \mathbf{x}) = \sum_i^N \lambda_i (\mathbf{n}_i^T \mathbf{x})(\mathbf{x}^T \mathbf{n}_i) \tag{4.11}$$

it is simple to show that the expectation of the quadratic form equals

$$\langle x^T A x \rangle_x \quad = \quad \langle \sum_i^N \lambda_i \mathbf{n}_i^T \mathbf{x} \mathbf{x}_i^T \mathbf{n} \rangle_x = \sum_i^N \lambda_i \mathbf{n}_i^T \langle \mathbf{x}\mathbf{x}^T \rangle_x \mathbf{n}_i = \sum_i^N \lambda_i \mathbf{n}_i^T C \mathbf{n}_i . \tag{4.12}$$

Only the correlation matrix $C = \langle \mathbf{x}\mathbf{x}^T \rangle_x$ of the data will be of importance for defining A (statistics of pairs of pixel). A single entry $a_{i,j}$ in this model represents the covariance between the pixel i and the pixel j. By assuming shift invariance we arrive at the usual covariance function $C(x) = E(I_0 I_x)$ (assuming mean zero) where 0 is an arbitrary origin and x a position in the image ($E(.)$ is the expectation).

This interpretation of the quadratic form, the selection of a covariance structure points to a way to visualize the information, e.g., the symmetry coded in a specific matrix A. The covariance of a data distribution defines uniquely a multivariate normal distribution in that space. The principal axes of this hyperellipsoid can be calculated as the eigenvectors of the covariance matrix. In the next section we will define different matrices A^{rot}, A^{scal} and A^{shift} which code for the model equivalents of rotations, scalings, and shifts respectively. The resulting eigenvectors are shown in Figure 4.7 sorted according to their eigenvalues and reflect nicely the underlying transformations by pattern showing the respective invariances. Because the matrices are symmetric their eigenvalues are real. Nevertheless, negative eigenvalues occur (all basis functions after the gap) because they were not obtained from real images. Because of the orthogonality of the eigenvectors they build an orthogonal basis set.

Learning matrices A by selecting image sets and computing their covariance matrix one could in principle explore the respective symmetries and would arrive at positive definite or semi-definite matrices. The problem remains, how to build sets of images in a meaningful way. Preferably the algorithm itself learns different A matrices from one set of images. That will be the task in Section 6.

The next sections will look in more detail into possible settings for A, how to constrain A and how to build an energy functional that can be used to invert the symmetry transformation.

5. The Binary Valued Quadratic Form

In this chapter we analyse the model presented in the previous chapter in the case of a binary[1] valued matrix A. Each entry $a_{i,j}$ selects a pair of pixel (if it is set to 1). Different (supervised) choices for A are used to illustrate the information content of the quadratic form and its use in advanced methods of image analysis. This chapter is a pre-requisite to the next chapter in which real valued quadratic forms are learned from the data in an unsupervised manner.

Throughout this work we will assume that the matrix A is symmetric (which implies that its eigenvalues are real). We will refer to the transformation as

$$\mathcal{S}(\mathbf{x}) \;=\; \mathbf{x}^T A\mathbf{x}, \qquad a_{i,j} \in \{0,1\}. \tag{5.1}$$

where A is chosen in order to select the pixel pairs that occur for successive rotations of the original X. Alternatively, we will use the more intuitive formulation based on successive rotations from Equation 4.7 on page 88 (a scalar value is obtained from that form by the $L1$-norm).

Lets introduce a toy example in order to illustrate the information contained in the statistics of pairs of pixels. This is of importance because the models introduced to detect invariances or symmetries (see Section 4.1) rely either on a basis function approach, which is related to linear models, or on non-linear models. One of the simplest non-linear models can be expressed in the form of a quadratic model (introduced in Section 4.2 on page 87).

Lets build two long sequences of the symbols 0, 1, and 2 and analyse them in the context of a first order Markov chain:

Markov chain

$$\mathrm{Seq1} \;=\; 1\,1\,2\,1\,1\,1\,2\,1\,1\,1\,0\,1\,2\,0\,0\,2\,2\,0\ldots \tag{5.2}$$

$$\mathrm{Seq2} \;=\; 1\,0\,1\,1\,2\,1\,1\,0\,0\,0\,0\,2\,0\,2\,0\,0\,2\,1\ldots \tag{5.3}$$

Both sequences were built in order to let every symbol in sequence 1 and in sequence 2 appear with a probability of $p(0) = .39$ $p(1) = .39$ $p(2) = .21$. Thus by measuring probabilities of single events the sequences appear to be iid (i.e., sampled from the same probability density). But the probability of one symbol

[1] The terminus *binary quadratic form* usually defines a quadratic form in two variables. Here, we refer to the type of values in the matrix A, which is independent from the number of dimensions.

at position t occuring *after* some other symbol at position $t-1$ is found to be

$$p_{\text{Seq1}}(\alpha_{t+1}|\alpha_t) = \begin{pmatrix} .13 & .19 & .07 \\ .16 & .19 & .07 \\ .13 & .02 & .07 \end{pmatrix}, \quad p_{\text{Seq2}}(\alpha_{t+1}|\alpha_t) = \begin{pmatrix} .15 & .15 & .08 \\ .15 & .16 & .08 \\ .08 & .08 & .05 \end{pmatrix},$$

where $\alpha = 0, 1, 2$. A position $a_{1,2}$ in these (stochastic Markov[2]) matrices codes for the probability of $p(\alpha_{t+1} = 0|\alpha_t = 1)$. By this measure the two sequences can easily be distinguished from another. The probability of the occurrence of one symbol followed by a second symbol in the sequence is reflected by the covariance between successive symbols, which relates this toy example to the model proposed.

The first sequence was constructed by first using a probability of $1/3$ for each symbol. After that we arbitrarily deleted three times every symbol 2 which was followed by a symbol 1. In doing so we introduced correlations between consecutive symbols. The second sequence was constructed by an independent sampling according to the first order probabilities of the first sequence. The probabilities of each symbol in the first and second sequence are the same (by design), but the two sequences are generated from distinct processes.

We see that the choice of the statistic has to be adequate. By counting probabilities of single events we could not distinguish the two sequences[3]. This was only possible after measuring the statistic of the occurrence of pairs of events.

Obviously only the second matrix is symmetric whereas the first matrix indicates the asymmetric generation of the underlying sequence, there we removed only symbol 2 followed by symbol 1, and not symbol 1 followed by symbol 2.

Also we see that the probabilities of the single events show up in the column sums of the matrices ($p(0) = p(0|0) + p(0|1) + p(0|2)$). In other words measuring the complete statistics of pairs of events we measure additional and not just complementary information. This will be of importance because the second order model presented is able to detect features that are found by linear models also (namely edges in natural images). In doing this it reflects the generality of the ansatz compared to linear models and comparative performance to polynomial models.

Polynomial models of second order can be defined by

$$\mathcal{P}(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c. \tag{5.4}$$

[2] A Markov matrix is a stochastic matrix where the rows (or columns) sum to one.

[3] Choosing the appropriate statistic by measuring the probabilities of joint events we are also able to better predict a new symbol, if that is the goal.

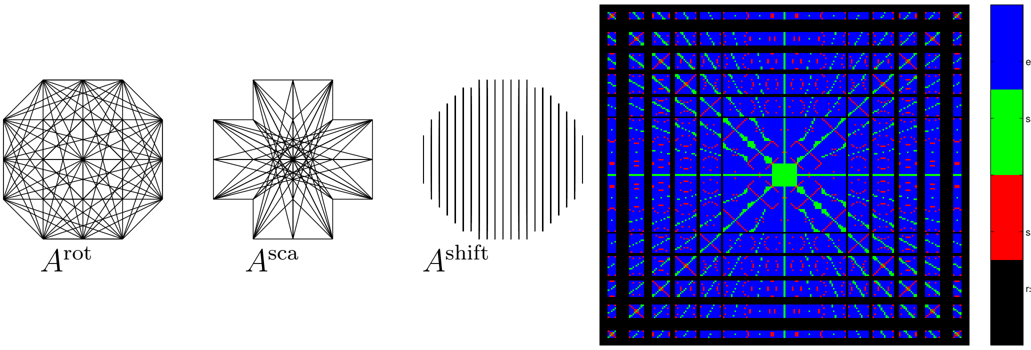


Figure 5.1.: Connection structure. *Right:* Three binary matrices A (color coded) for image patches of 15×15 pixels. The color code identifies the cases in which the pixel pairs are on the same radius to the center position (rotation) or on the same angle relative to the center position (scaling) or outside a circular receptive field. *Left:* For every matrix element $a_{ij} = 1$ the corresponding pixel pair (i, j) is connected by a line

Note that the quadratic form ($\mathbf{x}^T A \mathbf{x}$) appears in this formula next to a linear term ($\mathbf{b}^T \mathbf{x}$) and a constant c. The model produces a scalar output dependent on the values of A, $\mathbf{b}$, and c. A parameterization of this type was used by Wiskott and Sejnowski (2002) in the slow feature analysis model. Different from there model we will reduce the polynomial model by assuming $\mathbf{b} = \mathbf{0}$ and $c = 0$.

One advantage using the quadratic form only instead of the full polynomial model is the reduced number of parameters. A second advantage is that all the parameters $a_{i,j}$ (entries of the matrix A) are of one *type*, i.e., of comparative size/variance/meaning, which is favorable for algorithms exploring state space. Analysis of the state space will be done in more detail in Section 6.1 on page 114.

5.1. The Choice of a Binary Valued A

Before we go into more detail with respect to general properties of the transformation $\mathcal{S}$ we point out other 'natural' choices for a binary A matrix. In Section 4.2 on page 87 we have seen that the formalism of comparing an image with the transformed image boils down to a specific quadratic form (under certain assumptions). In the quadratic form each matrix element $a_{i,j}$ selects a specific pair of pixel. The knowledge which pairs of pixels we like to combine defines therefore different quadratic forms (matrices A). We have seen already that the choice we make here can be interpreted as designing a two-point correlation between pixel.

In the following, the coordinates of a pixel i will be denoted by π-periodic[4] polar coordinates (r_i, θ_i) in order to simplify the definitions.

Rotation: The goal is to let $A = A^{\text{rot}}$ encode the model equivalent of rotational symmetry of the image pixels around their mean position. This corresponds to the model in Equation 4.9. Again, we do not attempt to 'rotate' the image by A but code the general correlation structure found in perfect rotationally symmetric (centered) pattern.

We can incorporate our knowledge about the pixel pairs that appear if we make successive rotations of an image patch as described in Section 4.2. a_{ij}^{rot} $(i \neq j)$ will be set to one if $|r_i - r_j| < \varepsilon$ $(i \neq j)$ and otherwise to zero. ε can arbitrarily be chosen to be the distance of two neighboring pixels on the outer perimeter of the circular receptive field (see Figure. 5.1, left).

Scale: In a similar way A^{sca} is constructed to encode scaling. Entries a_{ij} of the matrix will be set to 1 only if the angular distance between i and j is sufficiently small, i.e., if $|\theta_i - \theta_j| < \varepsilon$ $(i \neq j)$ (see Figure 5.1, left). We are aware that we also code for symmetries that include inversions: scaling must not allow for pixel to cross the origin. We find that the choice of a specific pixel pair subset does not uniquely define a single symmetry operation. This points to a principle weakness of the ansatz of using the quadratic form. We may not uniquely define a symmetry transformation by using a binary choice for $a_{i,j}$.

Shift: A^{shift} is constructed to encode the pixel pairs that match for translations along the vertical direction. Correspondingly a_{ij} $(i \neq j)$ is set to 1 only if the x-coordinate of the pixel x_i equals the x-coordinate of the pixel x_j $(x_i = x_j)$ and otherwise to zero.

Mirror: A^{mirror} is constructed to encode the pixel pairs that occur for mirroring pixels along a chosen mirror axis (assumed to go through the image center). To illustrate the use of this transformation in Figure 5.2 on the next page we computed the resulting scalar value of the quadratic form for some (equally sampled) directions of the mirror axis and displayed them in a color code. The image shows an auto-radiogram of a gerbil brain slice (data courtesy by Andreas Hess).

5.2. The Properties of $\mathcal{S}$

To understand the properties of $\mathcal{S}$ it helps to show what information is lost or retained if X is projected into the space of $\mathcal{S}(X)$. For example, it is easy to see that $\mathcal{S}$ is invariant to changes in sign of the local contrast gradient. We already

[4] We do not distinguish a point $\mathbf{x}$ from a point $-\mathbf{x}$.

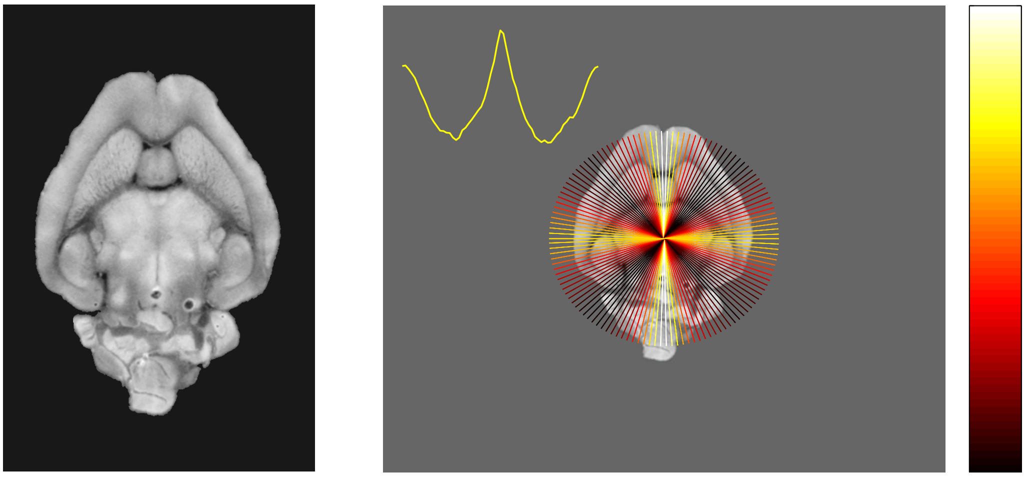


Figure 5.2.: Mirror axes detection example. The image displayed *left* is an auto-radiograph of a gerbil brain slice (data courtesy by Andreas Hess). *Right:* symmetry values are obtained by repeatedly selecting 5000 pixel pairs corresponding to a specific mirror axes (see Figure 5.5) and calculating the sum of their pixel gray value products. The inset shows the profile of the symmetry value over orientation

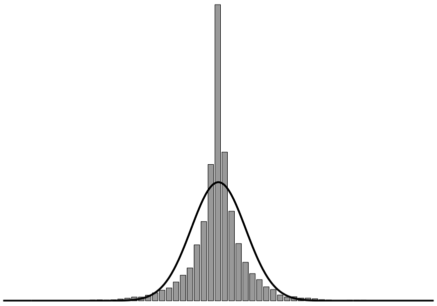


Figure 5.3.: Expressing data as dot-products yields a sparsely representation. Two independent (Gaussian) sources s_1 and s_2 are transferred into the space of dot-products. The histogram of the marginal distribution $s_1 s_2$ is shown. Whereas the underlying sources are Gaussian (*solid line*), in product space we observe a high forth order moment (kurtosis > 0)

have shown that $\mathcal{S}$ is sensitive to the correlations among particular input pairs or groups, which makes it a more powerful model than linear or threshold units (see Section 4.2.1 on page 91).

5.2.1. $\mathcal{S}$ is Linear in Contrast

Covariances in $\mathcal{S}$ (using Equation 4.7) can be expressed as correlations weighted by the non–normalized contrast of the image.

Assuming that X is a random variable with finite second moments it follows that $\texttt{cov}(X, X^\theta) = \sigma^2 \rho(X, X^\theta)$, where $\rho(\cdot, \cdot)$ is the correlation function and σ is the variance in the pixel intensities. σ is also a non-normalized measure of the contrast in X (compare with Section 2.1 on page 17). The covariance will therefore be large if the variance, e.g., the local contrast is large. Contrast sensitivity was found to be a crucial part in explaining the human active vision. In an active vision task humans prefer to look onto image regions with high spatial contrast (Reinagel and Zador, 1999). It would be of interest how the preference changes for contrast normalized images.

5.2.2. The Moments of $\mathcal{S}$

If we assume the input patch X to be a random variable then $\mathcal{S}(X)$ is also a random variable. For known input distributions X it is now of interest how the transformation changes the distribution $p(X)$. To verify this we analysed the transformation for one of the simplest distributions $p(X)$, that of a white noise distribution.

We found that the distribution $p(\mathcal{S})$ is *non*-white with high skewness and high kurtosis (see Appendix B.3 for detailed calculation and Figure 5.5). This 'generation' of higher order moments is not surprising if we keep in mind that in calculating pixel pairs we introduce correlations between pairs of pixel because a single pixel appears in more than one product.

Algorithms relying on distributions which are strongly non-Gaussian could therefore benefit from this data representation (we will use this property in later chapters in the context of feature extraction by ICA). On the other hand, it is clear that we have introduced redundancies (the number of pixel pairs is always much larger than the number of single pixels) which have to be sensible to be of use for the algorithm. Here we have to take care of two inter-weaved effects. One is the re-coding of information by changes of the coordinate system (like in PCA) and its use is clear on intuitive grounds. In our transformation the data is explicitly changed into the coordinate system of the two-point co-variances. The other effect is the introduction of new dimensions which is bound in our case to the goal of the re-coding.

5.2.3. Dependence on RFS and Preferred Orientation

Analytical solutions for $\mathcal{S}(X)$ can also be derived for specified structures. Only the results are repeated here for the derivation see the Appendix.

Given a horizontal edge (binary values) in the receptive field of size $r_{\mathtt{max}}$, we found (see Figure 5.4, A and Appendix B.2 on page 183 for derivation) that the symmetry value depending on the orientation θ of the edge is:

$$\mathcal{C}(\theta) \quad = \quad \begin{cases} \pi/2+\theta & : \quad -\pi\leq\theta< 0 \\ \pi/2-\theta & : \quad 0\leq\theta< \pi \end{cases} \tag{5.5}$$

$\mathcal{C}(\theta)$ will produce zero response if θ is orthogonal to the edge orientation or non-zero response otherwise. Note that this can be used to implement orientation selectivity in $\mathcal{S}$ if θ is restricted to a subset of all orientations. Changing the stimulus to a disc-like structure with radius r we found (see B.1 on page 182) that

$$\mathcal{S}(r, r_{\mathtt{max}}) \quad = \quad 2\pi r\left(1 - 2\frac{r^2}{r_{\mathrm{max}}^2} + \frac{r^4}{r_{\mathrm{max}}^4}\right). \tag{5.6}$$

This latter function (see Figure5.4, B) is zero at $r = 0$ and $r = r_{\mathtt{max}}$ and shows a maximum in between at $r = r_{\mathtt{max}}/\sqrt{5} \approx 0.48r_{\mathtt{max}}$. So a

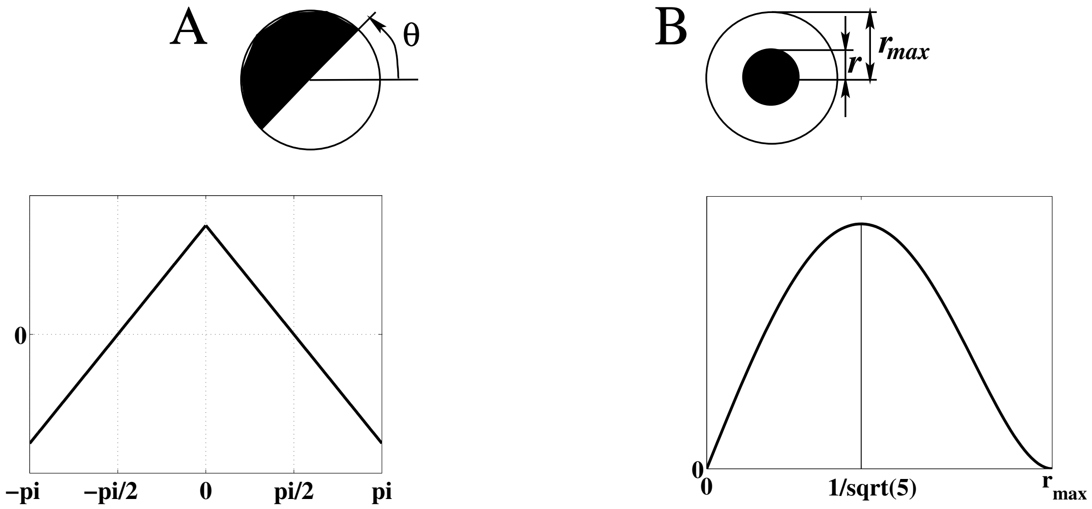


Figure 5.4.: Edge stimulus (*A*, top) and the corresponding analytical solution (*A*, bottom) of $\mathcal{C}(\theta)$ implementing orientation selectivity. Disc stimulus (*B*, top) and the corresponding analytical solution (*B*, bottom) of $\mathcal{C}(\theta)$. Non-zero response is restricted to the intermediate range of $0 \leq r \leq r_{\mathtt{max}}$

disc stimulus leads to maximum response if it fills nearly half the receptive field. This finding is interesting because it coincides with measurements of receptive fields of cortical neurons. One usually distinguishes between the stimulus size that elicits maximum response the receptive field and the extra-classical receptive field or integration region which is found as (*i*) being much larger and (*ii*) acting inhibitory (Solomon, White and Martin (2002), LGN, marmoset, Cleland, Lee and Vidyasagar (1983) LGN, cat, but see Felisberti and Derrington (2001) for contradicting results).

receptive field

integration region

5.2.4. Stability Analysis

The computational costs of computing $\mathcal{S}$ are in the order $O(n^2)$ where n is the number of pixels in our receptive field (because symmetry was based on pixel pair statistics). Compared to the filtering techniques usually used (e.g., linear filters) an algorithms using this may not be feasible in real time applications. To test its real time performance we implemented the rotational symmetry detection algorithm in C++ using DirectX8.1 on a laptop computer (600-MHz, PIII). Using a commercial web-cam (320×200 pixel, Phillips ToUCam Pro) we could calculate symmetry values at receptive field sizes of 5×5 pixel in the order of 1 frame per second.

One of the simplest methods that can be used in order to speed up this

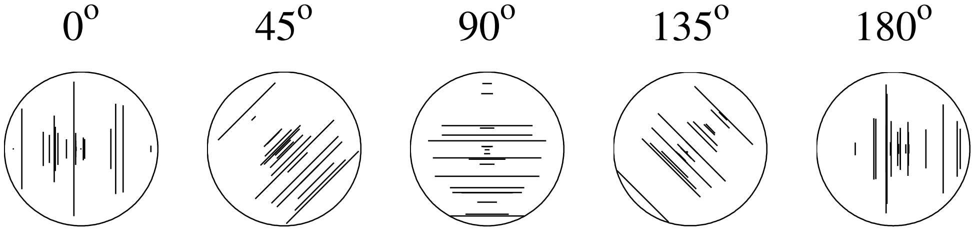


Figure 5.5.: Example of a sparse mirror selection filter for specific orientations. Pixel pairs are shown by connecting each pixel pair by a line

computation is to reduce the number of pixel pairs processed. To demonstrate the stability of $\mathcal{S}$ we used the quadratic form $\mathbf{x}^T A^{\text{mirror}}\mathbf{x}$ for the detection of mirror symmetries (see on page 96). We reduced the number of pixel pairs (entries $a_{ij} = 1$ in the matrix) by random deletion from the original set. The resulting structures for sparse A^{mirror} matrices are shown in Figure 5.5 sorted for different (approximative) mirror axes. Repeatedly the filters parameterized by their mirror axes where applied to the image in Figure 5.2 on page 97 (left).

We found that approximative symmetry values can reliable be computed over sub-populations of pixel pairs (see Figure 5.6 for results). In computer simulations the algorithm seems to be stable even if only very few pairs were used. This remarkable stability in face of large numbers of missing values is largely explained because of the relative smoothness of the image.

5.3. Inverting the Symmetry Detection

In the following section we will concentrate on the rotational symmetry case ($A = A^{\text{rot}}$). The stochastic algorithm proposed is not restricted in the choice to this specific A. It is used here as a general tool to visualize the information content of the 'naive' symmetry detection algorithm (binary valued quadratic form) for which A is known.

By inverting a transformation we can, in principle, obtain the original data. This idea is widely used to enhance the result of an image acquisition task (denoising). Lets assume that a process T corrupts an image $\mathbf{x}$ during measurement $T(\mathbf{x}) = \mathbf{y}, \mathbf{x} \in \mathbb{R}^n$. If the characteristics of the process, e.g., its transformation T are known by computing the inverse transformation $T^{-1} : \mathbb{R}^n \to X$ (provided that it exists) we can obtain the undisturbed data $T^{-1}(\mathbf{y}) = \tilde{\mathbf{x}} = \mathbf{x}$.

de-noising

The problem is always to invert T. A Gaussian low-pass filter for example will be generally not invertible, which states that we lost information about the image by applying the low-pass. If the process T acts in a linear way, in

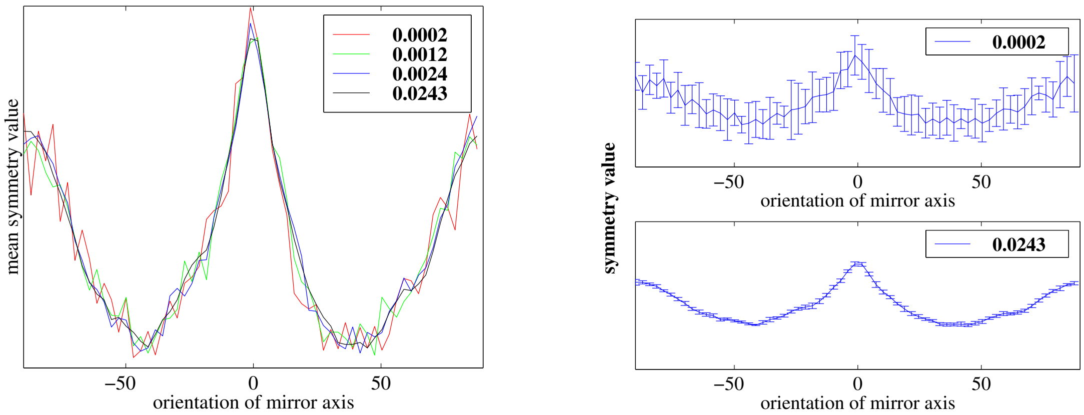


Figure 5.6.: *Left:* Mean symmetry for different numbers of pixel pairs of the Gerbil brain data of Figure 5.2 on page 97. Values in the legend are ratios of the number of pixel pairs with respect to the maximum number of pixel pairs for a given mirror axis (half of the number of pixels in the circle). *Right:* Low variance over 10 runs indicates that even with 2 percent of pixel pairs the mirror axis could be computed reliably

other words, T can be expressed as a (m, n)-matrix, this matrix may be either singular, or close to singular, or not square. In all three cases we cannot invert T. If T has full rank n we can use the pseudo-inverse $T^+ = (T^TT)^{-1}T^T$ as a 'workaround'. It can be used to describe the projection of the signals onto the subspace generated by a finite family of basis signals (lines of the matrix T).

pseudo-inverse

Apart from computing the pseudo-inverse of the transformation matrix, one can use the framework of an energy functional. In this framework additional knowledge about the data can easily be implemented in order to get a good approximation of $\mathbf{x}$. Let $T(\mathbf{x}) = S(\mathbf{x}) = \mathbf{y}$. A goal function is used to describe the quality of an estimated signal $\tilde{\mathbf{x}}$ which is sucessively improved during the optimization procedure. In our setup a $\tilde{\mathbf{x}}$ will be considered of having low energy or low error if $T(\tilde{\mathbf{x}})$ is sufficiently close to $T(\mathbf{x})$, that is we compare $\mathbf{x}$ and $\tilde{\mathbf{x}}$ in terms of their representation in feature space. Starting from an initial $\tilde{\mathbf{x}}$ (e.g., Gaussian white noise) we measure its likelihood by comparing the corresponding $\tilde{\mathbf{y}}$ with the known data of $\mathbf{y}$. Successively we change $\tilde{\mathbf{x}}$ in order to lower the corresponding energy (or error) function by an algorithm adopted from a Markov random field (MRF). $T(X) = S(X)$.

energy function

Markov random field

Now we like to present the used energy function. The values $\mathbf{y}$ obtained by $S(X)$ represent the local correlation of a sub-set of pixels in the image patch. The sub-set is defined by the choice of $A = A^{\text{rot}}$. Here all pixel pairs contain

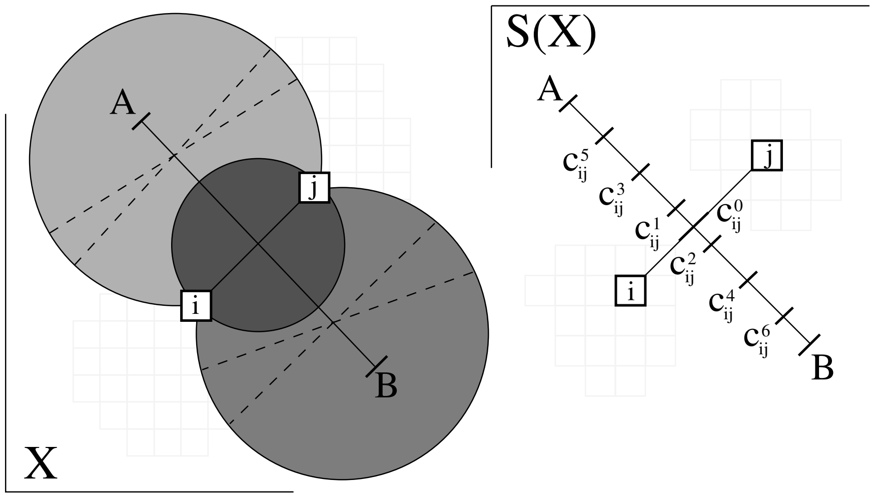


Figure 5.7.: The pixel-o-centric picture of the world. *Left:* For pixel pair (i, j) three example circles representing possible associated cliques are shown (living on the perimeter). Line $\overline{AB}$ is a perpendicular line to $\overline{ij}$ onto which all these cliques are centered. The connection between i and j is therefore defined as incorporating their respective pixel values but also information about the covariance structure on the line $\overline{AB}$. *Right:* The length of the perpendicular lines $\overline{AB}$ scale with the size of the receptive field of X and with the distance between the pixels ($|\overline{AB}| = \mathrm{rfs} - |\overline{ij}| + 1$)

pixel with equal distance to the center of the respective image patch. Because all pixel at a common distance to the center are connected to oneanother (are in product with each other) they form a *clique*. A pair of pixel $x_i x_j$ appears only ones if $S(\mathbf{x})$ is computed for a single image patch $\mathbf{x}$ only. But by extracting overlapping patches and reapeating the procedure of computing $S(\mathbf{x}(t))$ the pixel pair $x_i x_j$ will appear in more than one clique. In Figure 5.7 the line connection the points A and B connect the gravity centers of the corresponding cliques. The circles indicate the position of three examples cliques. Each clique is connected with one value $\mathbf{y}$ which we now re-name c_{ij}^k. Here k enumerates the clique in which the pixel pair $x_i x_j$ is contained. 5.7).

clique

In the framework of a Markov random field the probability of pixel i having a value x_i depends on the neighboring pixel j. In our case they depend on their respective covariances c_{ij}^k. Using the result $E(S(X)) = \texttt{cov}(x_m, x_n), m \neq n$ (see Equation B.43 on page 189) our energy function depends *(i)* on an initial guess of the image $\tilde{\mathbf{x}}_0$ and *(ii)* on the covariances c. It is defined as being zero, if each

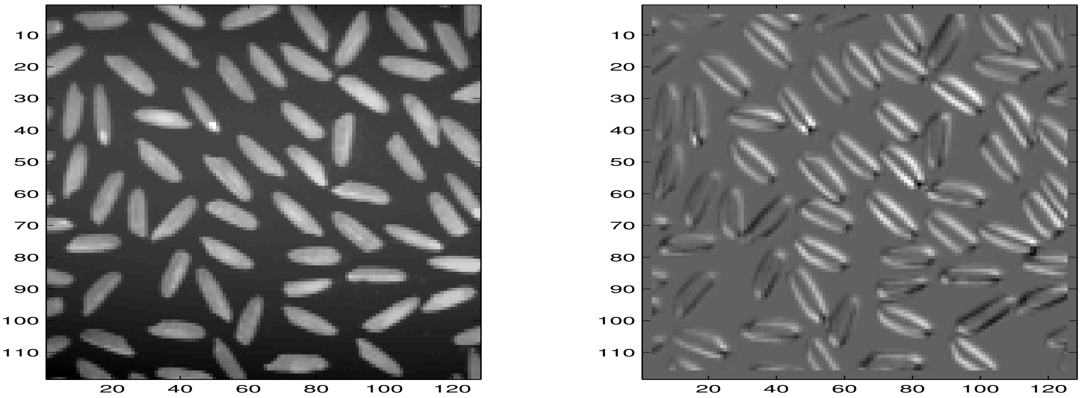


Figure 5.8.: *Right:* Original image X (rice.tif from Matlab). *Left:* The coefficients $c^k = ||S(\mathbf{x}_k)||_1$ for each corresponding pixel position $(\text{rfs} = 5)$. Edges elicit the highest contrast in the image

product of pixel values x_i and x_j equals their associated covariances c^k_{ij}:

$$F_i(\tilde{\mathbf{x}}, c) \quad = \quad \frac{1}{N}\sum_{j\in I_i}\left(\sum_k \left((\tilde{x}_i\tilde{x}_j) - c^k_{ij}\right)\right)^2 \qquad (5.7)$$

I_i is the index set of all pixels in a receptive field around i of size $2\,\text{rfs}$, N is the overall number of single operations of type $(xy) - c$. k counts a discreet number of centers of cliques (symmetry value of $\mathbf{x}_k$).

All values c^k are assumed to be equally important. Because they are obtained for cliques of changing radius and thus changing spatial distance between the pixel we herein assume implicitly that the covariance between any two pixels does not depend on their distance.

Minimizing F

There are several ways to minimize the above energy function. We use now a stochastic algorithm which is very simple to implement and furthermore a gradient descent algorithm which has some advantages in speed. For both cases initially $\tilde{\mathbf{x}}$ was set to (uncorrelated) Gaussian white noise. The c^k_{ij} were computed beforehand by the symmetry detection algorithm. The low energy state of the system corresponds therefore with the image that most likely causes the observed symmetry values. In Figure 5.8 an example image and the corresponding covariances $c = ||S(\mathbf{x})||_1$ are shown. High values are found predominantly at edges in the image.

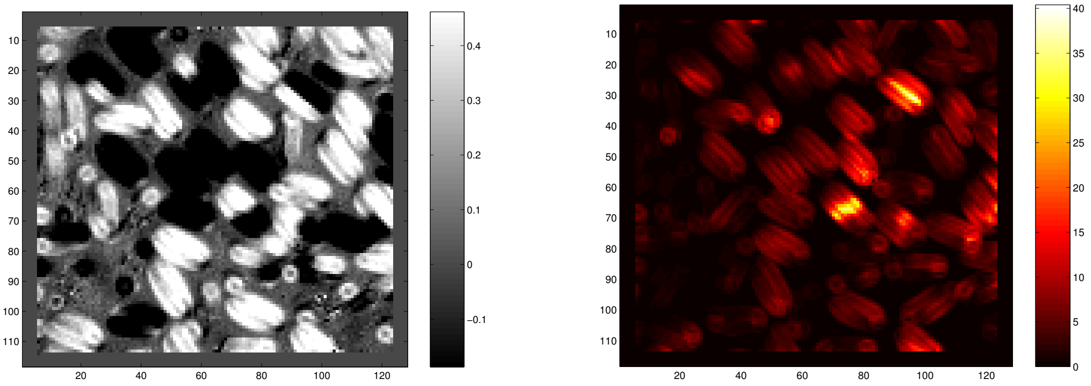


Figure 5.9.: *Left:* Reconstructed image $\tilde{x}$ with the algorithm of Metropolis et al. (1953). *Right:* High values of the energy function used (Equation 5.7) are shown in white and indicate errors remaining because of the deterministic annealing scheme

5.3.1. Minimizing F by Metropolis Algorithm

Minimizing our energy function is done by using the idea that a system in thermal equilibrium at temperature T has its energy probabilistically distributed among all different energy states F.

$$\mathtt{Prob}(F) \approx \exp\left(-\frac{F}{kT}\right) \tag{5.8}$$

The quantity k (Boltzmann's constant) is a constant of nature that relates temperature to energy. Using this formula there is a corresponding chance for the system to get out of a local minimum. Sometimes it goes uphill as well as downhill (in energy landscape); but the lower the temperature, the less likely is any significant uphill excursion.

This is implemented by the algorithm of Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953). We need to specify: (*i*) A description of the possible system configurations ($\tilde{x}$ initialized by white noise). (*ii*) A generator of random changes in the configuration; these changes are the 'options' present to the system (done by adding Gaussian noise to single pixel gray values). (*iii*) An objective function whose minimization is the goal of the procedure (Equation 5.7) and last (*iv*) a control parameter T (which works analog to temperature) and an annealing schedule which tells how it is lowered from high to low values, i.e., after how many random changes in configuration one downward step in T is taken (and how large that step is).

simulated annealing

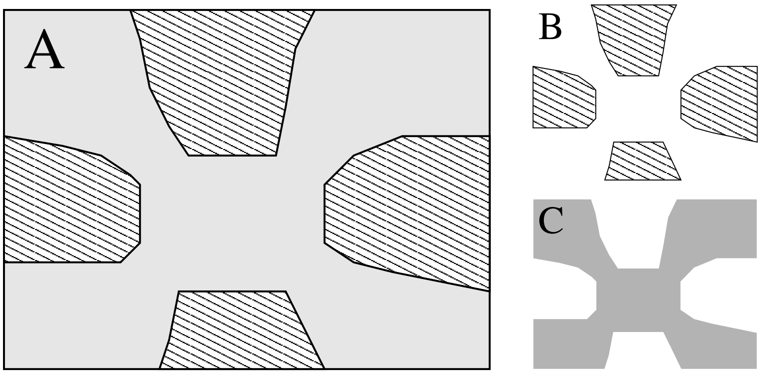


Figure 5.10.: Switching context for judging foreground versus background. The figure *A* resembles either four objects that are in close vicinity (*B*) or a single star like object (*C*)

The value of $\tilde{x}$ corresponding to the low energy state of the procedure can be seen in Figure 5.9 on the page before.

Due to the symmetric nature of the products, the algorithm cannot distinguish between different signs of the local contrast gradients (see Section 5.2 on page 96). This explains that the algorithm founds white as well as black representations for the elongated objects (compare with Figure 5.8 on page 104 left). Because the objects in the image are larger than the field size of our Markov Random Field initial solutions found at opposite ends of these objects are initially uncoppled and may contain by chance a change in sign. Whereas both solutions are optimal on their own they evolve together during the optimization and enforce an edge. This type of errors light up in Figure 5.9 right where the final energy is plotted.

Interestingly the algorithms repeatedly insert 'circles' in $\tilde{x}$ which have no correspondence in the original image (see Figure 5.8, left, and Figure 5.9, left). We can explain this behavior by the observation that in a close-up look the corresponding regions in the original image contain structured back-ground configurations. Because there is no clear distinction between the role of figure and background our algorithms interprets 'interesting' formations of backgrounds as structured objects which are represented in the reconstructed image by a proto-typical object, e.g., an open circle. Figure 5.10 illustrates this effect of ambiguous assignment of foreground and background that can also be found in many illusory figures. In this sense the representation of images in terms of symmetry lacks a prior about what is background and what foreground in the image. Additional information from outside the receptive field is needed to prefer one solution over the other. Information about the global distribution

figure-ground

of light and dark regions could provide such an bias, because objects in the foreground tend to be more compact, i.e., enclosed in background regions.

We observe also is a relative high amount of noise in the background. This is due to the initialization of $\tilde{\mathbf{x}}$ as Gaussian white noise which is preserved by the algorithm, because low symmetry values in the corresponding regions result in low overall energy in that regions, regardless of the pixel variance. In the next chapter we will deal with this problem by introducing an additional weight decay.

5.3.2. Minimizing F by Gradient Descent

An annealing algorithm finds (in principle) the global minimum of an energy function. But this holds only for a very time consuming and therefore non-practical annealing schedule. To speed up the computation we now introduce a gradient descent algorithm. It involves a measure of direction or gradient in parameter space along which the energy is becoming smaller. Given Equation 5.7 on page 104 we differentiate our energy $F(\tilde{\mathbf{x}}, c)$ with respect to a pixel $\tilde{x}_i$:

gradient descent

$$\frac{dF_j}{d\tilde{x}_i} = \frac{1}{N}\sum_{j\in I_i}\left(\sum_k \left(2(\tilde{x}_j\tilde{x}_i - c_{ij}^k)\tilde{x}_j\right)\right) \tag{5.9}$$

Using this formula we iteratively adjust the gray value of single pixel by

$$\tilde{x}_i^{new} = \tilde{x}_i^{old} + \nu\frac{dF_j}{d\tilde{x}_i} \tag{5.10}$$

where ν is the learning rate and is kept fixed at $\nu = 0.5$. Again we observe the same basic result as for the Metropolis algorithms (see Figure 5.9).

As a short extension to the previous algorithm we introduce now a weight decay term that should deal with the high variance in the background regions (see Figure 5.11 on the following page, left). We add a constrain for the variance of $\tilde{x}$ to the energy function 5.7:

$$\sum_i x_i^2 = 1. \tag{5.11}$$

Using the method of Lagrangian multipliers our new energy function F^2 is now $(\sum_i \tilde{x}_i^2 - 1 = 0)$

$$F_i^2(\tilde{x}, c) = \frac{1}{N}\sum_{j\in I_i}\left(\sum_k \left((\tilde{x}_i\tilde{x}_j)\mathbf{1} - c_{ij}^k\right)\right)^2 + \lambda\left(\frac{1}{M}\sum_m x_m^2 - 1\right) \tag{5.12}$$

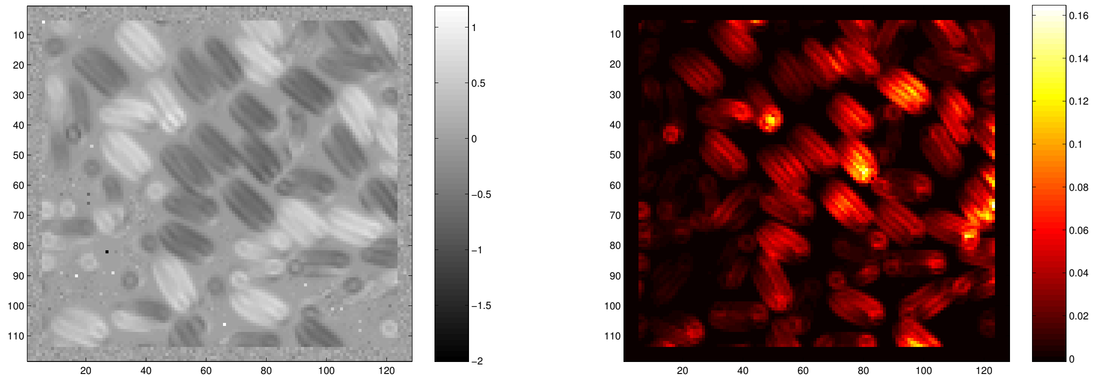


Figure 5.11.: *Left:* Reconstructed image $\tilde{X}$ by gradient descent. *Right:* High values of the energy function used (Equation 5.7) are shown in white. In background regions (bottom right) stable noise patterns can be observed

where m goes over all pixels M in the image. Our learning rule therefore is slightly changed compared to equation 5.9 to

$$\frac{dF_j^2}{dx_i} = \frac{1}{N}\sum_{j\in I_i}\left(\left|2(\tilde{x}_j\tilde{x}_i - c_{ij}^k)\tilde{x}_j\right|_1\right) + 2\tilde{x}_i \qquad (5.13)$$

λ was chosen arbitrarily, i.e., it was set to 1. The result of this operation can be seen in Figure 5.12 on the next page. As expected the constrain acts as a decay term removing the high pixel variance.

5.4. Applications

This section summarizes some attempts to use the methods derived so far in the context of object classification, image alignment and as a tool for automatic landmark detection. The work presented here about the binary valued quadratic form is merely a prelude to the next chapter. Correspondingly the applications presented are thought of as illustrations of the potential applicability of the method.

5.4.1. Object classification

incomplete data problem

Object classification can be formulated as an incomplete data problem. In the classification process we search for an unknown class label j associated with each pixel indicating which component density function was responsible for its generation.

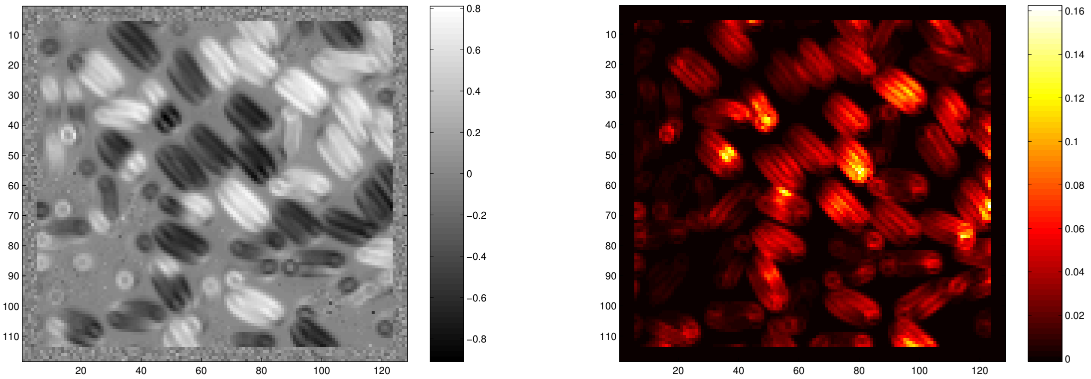


Figure 5.12.: *Left:* Reconstructed image $\tilde{x}$ by gradient descent and additional weight decay. *Right:* High values of the energy function 5.7 used are shown in white

Because we know that single pixel contain little information one generally uses a feature space representation of the image for classification of pixel. We are interested now in using the symmetry measure $\mathcal{S}(X)$ as a representation of the data in feature space. The classification algorithms is applied on toy data and outlines a particular direction of research which we think is worthwhile to follow (see Figure 5.13).

One of the best known algorithms for incomplete data problems is based on the probability density estimation with a Gaussian mixture model which can be trained with the Expectation Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977).

EM

A model of 9 multivariate Gaussian densities (diagonal covariance matrix) was trained on feature vectors extracted from overlapping image patches of size 32×32 according to $\mathcal{S}(X) = s_1(X), \ldots, s_{20}(X)$ (Equation 4.7 on page 88). The focus of the symmetry transformation was therefore just large enough to contain one circular respective triangular object in the toy data image. For learning we selected feature vectors in the upper 70% region of the $L1$-norm of $\mathcal{S}$. This acts in restricting the algorithm to non-background regions around the object. This heuristic can be understood as an first approach to active vision in which we change the relative frequency of occurrences of particular image regions according to their (pre-attentive) features. In our case this was nessesary because of the small number of training examples (only one image was used for training and testing).

active vision

In Figure 5.13 one out of nine clusters respectively its posterior probabilities $P(\mathbf{x}|j)$ are plotted in dark gray overlayed on the image (light gray). It turns out that for this image clusters in the feature space could be found that correspond to solemnly circular- or triangular objects (the cluster for the circular data is

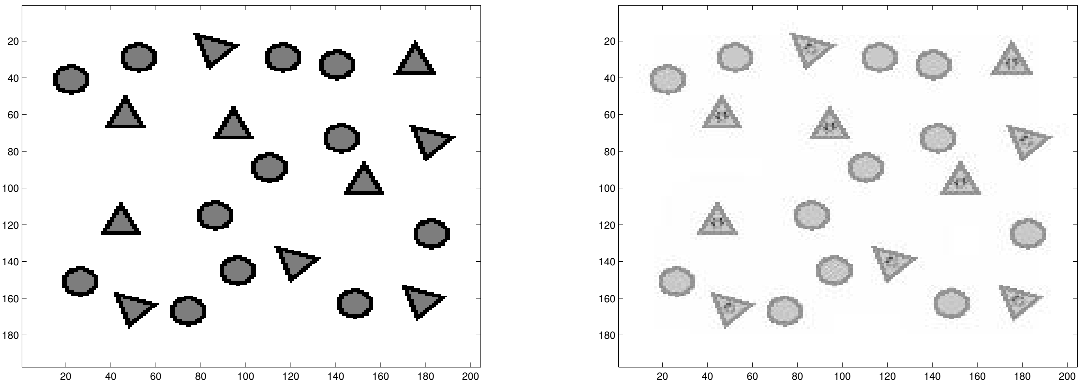


Figure 5.13.: Unsupervised image segmentation. *Left:* Test image containing triangular and circular shaped objects. *Right:* For cluster nr. 2 the posterior probabilities overlayed on the image are shown (selecting the central regions of the triangular objects). Note that the model is learned on only this single image. Two clusters (only one is shown) correspond to the two object classes

not shown).

5.4.2. Image Alignment

One case in which shape analysis is regularly used in practice is for the registration of digital images. Often single images from different measurements display connected information, images of consecutive brain slices for example, but due to the measurement procedure the single images are distorted. One of the first pre-processing steps before one can attempt a 3D segmentation or reconstruction is to realign (register) the images, in the simplest, and most often used case, by performing rigid transformations. In order to automatically recover the original axis of rotation of a stack of images for each image its major and minor axis are computed. Major and minor axis represent the directions in which the object shows its maximum respectively minimum elongation. Knowing this axis each image is rotated onto a common reference axis.

major-, minor axis

The procedure is the following: a binary map is computed which represents object and background regions. The eigenvectors of the pixel coordinates of the object pixel in the map represent now the major respectively minor axis of the shape of the object.

Obviously this procedure works only for images showing a clearly preferred axis of elongation. The brain slice in Figure 5.2 on page 97 shows such an axis. In effect the procedure is sensitive to the proximal parts of the frontal and of the occipital lobe of the cerebrum. A mismatch due to the measurement

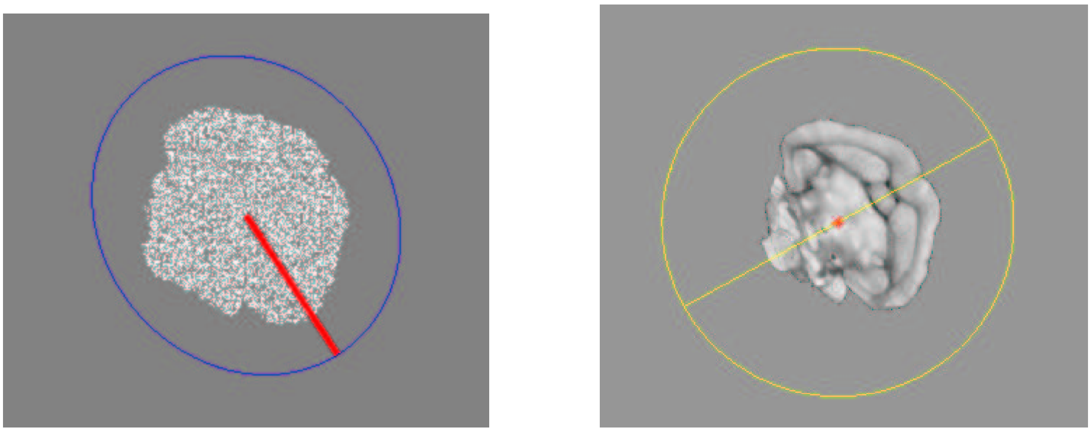


Figure 5.14.: *Left:* Shape analysis detects the largest variance direction. *Right:* Symmetry detection correctly find the direction of largest (mirror) symmetry

procedures is very likely in these regions. Because the method to extract the eigenvectors is very sensitive to the variance (that is what it measures) it may come up with directions that are largely influenced by errors in the measurement and not by the shape of the object. Principal component analysis is also not applicable if the shape of the object is nearly circular, or if the shape border is noisy, or, if the shape is mainly defined by the changes due to the measuring method.

In order to introduce our method we assume now that every image which is subject to re-alignment posses structural information in its gray values that are used in manual alignments.

Detection of the mirror axis is not solemnly based on the shape of the object but also on its gray value structure, thus it is predestined to explore the symmetrical structure of the two hemispheres visible in the horizontal section. In order to automatically recover the direction of the maximum mirror symmetry we used the Nelder-Mead simplex algorithm (Nelder and Mead, 1965) which adjusts parameters of a goal function. The goal function (which is minimized during the optimization) was calculated dependent on a direction (mirror axis), position, and size of the receptive field. Using the result from the stability analysis (Section 5.2.4 on page 100) we selected 5000 pixel pairs from the pool of possible pairs (positive correlated pixel pairs) for the given receptive field size. We summed over the products of the selected gray value pairs to obtain a mirror symmetry index $\mathrm{MI}(X)$. The gray values in the image are assumed to be positive, therefore $\mathrm{MI}(X)$ is positive (no negative entries in A^{mirror}). Its inverse $(1/\,\mathrm{MI}(.) \geq 0)$ was finally minimized by the Simplex algo-

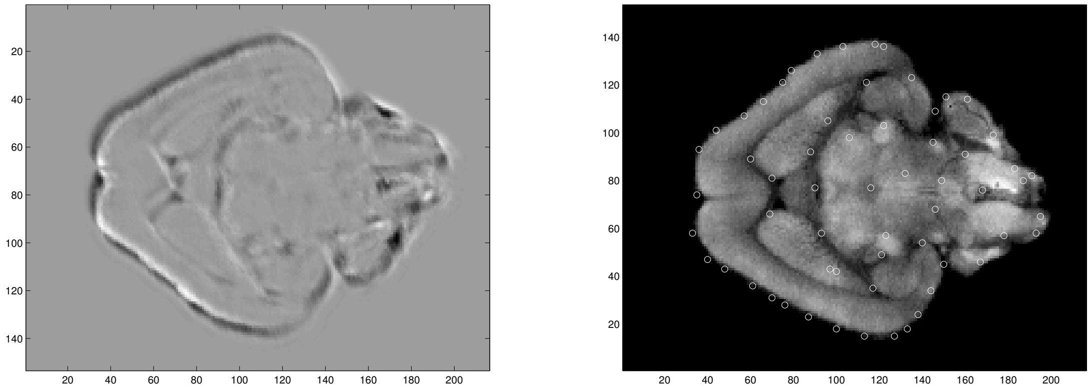


Figure 5.15.: *Left:* The $L1$-norm of $\mathcal{S}$ (rotational symmetry). *Right:* Found landmarks overlayed onto the original image

rithm. The obtained maximum mirror symmetry solution is shown in Figure 5.14 right. The algorithm converges fast and reliable, mainly due to the quality of the features and the apparently smooth energy landscape. For different orientations of the mirror axis the symmetry value (which is the inverse of the energy) is shown in the inset of Figure 5.2 on page 97. Notably, the size of the receptive field found by the mirror detection algorithm is greater than the outer perimeter of the brain slice. This reflects the analytical result in Section 5.2.3 on page 99 that the symmetry of an isolated object is maximal if the receptive field size is approximately twice as large as the object.

To construct a benchmark case we sub-sampled the brain slice of Figure 5.2 on page 97 reducing its vertical resolution by 2 and rotated the resulting image by an arbitrary angle mimicking an extreme measurement error. By sampling 5000 pixels from the inside of the figure the major axis of the object was computed by principal component analysis. Due to the sub-sampling procedure the resulting direction of highest variance was the one shown in Figure 5.14 left. The direction of largest mirror symmetry is shown right and corresponds to the expected direction.

5.4.3. Landmark Detection

We also tested if symmetry detection can be used to reliably detect local features in images. Again we used the brain slice data set introduced in the the last sections. If one can reliably find local positions (landmarks) in two images, one can use these points to perform warpings of one image into the other. This case of a usually non-rigid transformation of the image is used for example to align slices obtained in different experiments (from different animals). Later on one is interested, for example, in the improved statistics of the point-wise mean.

Landmark detection was started by computing the local rotational symmetry values of the image ($L1$-norm of $\mathcal{S}$). The first landmark was selected by finding the pixel position with maximum symmetry value. At this position an image with a Gaussian gray value profile was subtracted from the symmetry value image[5]. In the changed image the procedure of finding the maximum symmetry value was continued until a fixed number of landmarks could be found. Special care has to be taken for large numbers > 100 of landmarks because the iterative subtraction of the Gaussian profiles constantly lowers the mean gray value of the image. In effect the maxima selected late in the process are from regions where few Gaussian functions where subtracted, e.g., from the background of the image. Because the size of the Gaussian function influences the mean distance of detected landmarks it can be used beforehand to derive the maximum number of landmarks that can be detected by the method.

The resulting landmark set is relatively uniform distributed and prefers to select corresponding points in the two hemispheres (see Figure 5.15).

5.5. Concluding Remarks

In the last chapter we have analysed a special case of our model. Assuming that we select specific pairs of pixel for the statistics we have introduced a binary valued matrix A and suggested a design principle to implement symmetry transformations by that matrix. We found that the model could be used to code for rotations, scalings, and shifts. Applying symmetry detection on digital images can be done reliably and is noise-insensitive.

Symmetry detection was also shown to be selective to both foreground and background object configurations. In the application section it was shown that the binary quadratic form can outperform shape analysis and can be used for landmark detection.

The selection of the pixel pairs and by this of an associated symmetry group may sometimes appear to be *obvious*. But one has to be careful with intuition. The question remains: Is there a better symmetry group to describe our data? Apart from the settings in a specific problem class it may not be obvious which type of transformation is best. In the next chapter the transformation A is learned in an unsupervised manner from the data and so can be adjusted to an unknown problem class.

[5] This assumes that interesting points are modeled by singular points in the image that are subject to a Gaussian point spread function.

6. The Real Valued Quadratic Form

As we have seen in Section 4.2.1 the quadratic form can be analyzed in terms of a set of orthogonal basis functions. The basis functions appear as the eigenvectors of the matrix A in the quadratic form. They represent pattern that are invariant under the transformation. Each one is thought of as coding for one symmetry transformation. Using an ensemble of images and calculating the expectation of the quadratic form we found a non-trivial[1] solution by assigning to A the covariance matrix of the data. Image covariance calculated over many samples is a smooth function of the distance of the pixel (see Section 1.3), i.e., reflects that nearby pixels are correlated and distant pixel are not correlated or anti-correlated. Because of the extensive summing further aspects of the data are lost. That is, if there are underlying factors that produced the data and these factors are subject to different statistics only their common part survives.

It is reasonable to expect that natural images are composed from different factors. These factors can, for example, be the objects present in a scene or specific (non-accidental) parts of objects. We have to point out here that they are certainly not added linearly to produce the image. Because of occlusions of (mainly opaque) objects in a visual scene a logical operation would be more appropriate which reflects the presence of either one *or* the other object. But this is behind the scope of the current work. For now, lets assume that these factors are added linearly. Each factor is represented by a specific symmetry operation and a large number of symmetry transformations may constitute our visual input. To learn more than one symmetry transformation, in essence, we have to find a model that can cope with more than one matrix A.

6.1. A Tensor Form for Symmetry Detection

The polynomial model of Equation 5.4 on page 94 can be extended to incorporate more than one constant (now a vector), more than one linear weight vector (now a matrix) and more than one quadratic form (now a tensor of third order)

$$\mathcal{PN}(\mathbf{x}) \quad = \quad W^{ijk}\mathbf{x} + V\mathbf{x} + \mathbf{u}. \tag{6.1}$$

[1] Different from a diagonal matrix with $a_{i,i} = 1$ (identity matrix).

It is essential for our learning algorithm to keep from this model only the tensor W. We point here to the reasoning on page on page 95 and use the reduced model

$$\Phi(\mathbf{x}) \;=\; W^{ijk}\mathbf{x}. \tag{6.2}$$

We can arrive at the same model by the following, more intuitive reasoning: We want to find a number i of symmetries, each one expressed by its own matrix A^i. Let's define a vector valued function Φ which maps the data into the space of different symmetry operators:

$$\Phi(\mathbf{x}) \;=\; \left(\mathbf{x}^T A^1\mathbf{x},\; \mathbf{x}^T A^2\mathbf{x}, \ldots,\; \mathbf{x}^T A^N\mathbf{x}\right)^T \tag{6.3}$$

Additional assumptions can be made for the distribution of $p(\Phi)$ incorporating for example sparseness, independence or uncorrelateness. But before doing this, we first cast the problem into the notion of dot-products. This allows the analysis of a linear model in which assumptions can be justified more easily.

Any quadratic form can be written as a linear weighting in the space of dot-products:

$$\mathbf{x}^T A^i\mathbf{x} \;=\; (a^i_{11}, \ldots, a^i_{kk})(x_1x_1, \cdots, x_nx_m)^T = \mathbf{w}^i \operatorname{vec}(\mathbf{x}\mathbf{x}^T) = \mathbf{w}^i\mathbf{p}. \tag{6.4}$$

where $\mathbf{p}$ consists of monomials of constant order 2 and $\operatorname{vec}(.)$ describes the vector form of the entries of the matrix argument. We readily see that because of the commutativity of the multiplication ($x_ix_j = x_jx_i$) the full matrix A generates each basis function $j \neq i$ twice. The respective feature space would therefore be always degenerative (data points are concentrated in a linear subspace). By removing the obsolete basis functions we obtain the property of an *effective* transformation. We eliminate the respective entries from $\mathbf{w}$ by introducing the lower triangular form (vech) of the matrix A.

Adding more dimensions to the model Φ, i.e., more quadratic forms $\mathbf{x}^T A^i\mathbf{x}$, we arrive at

$$\Phi(\mathbf{p}) \;=\; W\mathbf{p}. \tag{6.5}$$

Here, W is a square matrix but Φ performs the same computation as in the tensor notation in Equation 6.2. It can be viewed as a linear mixture expressed by the mixing matrix W of the signals $\mathbf{p}$ in the space of dot-products. In the case of a single quadratic form $\mathbf{A}_0$, W is a row vector and the model reduces to the quadratic form. In order to learn this model we compute the correlation matrix $C = E_{\mathbf{x}}(\mathbf{x}^T\mathbf{x})$ of the data and set W to its vectorized, lower trinangular form $W = \mathrm{vech}(C)$ (see Section 4.2.1 on page 91). Consequently, for a square matrix W (with as many A matrices as components in $\mathrm{vech}(\mathbf{x}^T\mathbf{x})$) we need to learn a mixture of covariance matrices on the data.

To arrive at a learning algorithm for this non-linear model (overcomplete) independent component analysis is introduced. In solving this special case of ICA a novel learning algorithm will provide us with an efficient algorithm to learn the parameters of the model in Equation 6.5 on the page before.

6.2. Independent Component Analysis

Let us assume that the data $\mathbf{x}$ at a point in time t consists of a number m of observed variables. All together we have T different observations $x_i(t)$ where $i = 1, \ldots, m$ and $t = 1, \ldots, T$ $(T >> m)$. We are interested in a function that maps the m-dimensional space given by $\mathbf{x}(t)$ to an n-dimensional space, such that the transformed variables explicitly contain information on the data that is otherwise hidden. The transformed variables should capture the underlying *factors* that describe the essential structure of the data. Ideally the factors correspond to processes that have generated the data in the first place. But this correspondence is non-trivial and usually done afterwards, by looking onto found factors and proposing some ideas about a connected physical process. A simple model of the data is based on linear functions[2]. Every factor or component $s_i(t)$ can be expressed in this framework as a linear combination of the observed variables:

factor analysis

$$s_i(t) \quad = \quad \sum_j w_{ij} x_j(t). \tag{6.6}$$

Assuming that the observations $\mathbf{x}(t)$ are independent from each other, i.e., a re-ordering of the observations does not destroy essential information we can consider $\mathbf{x}$ a realization of a random process. In matrix notation the linear system is defined by

$$\mathbf{s} \quad = \quad W\mathbf{x}. \tag{6.7}$$

Because we know only the observations $\mathbf{x}$ and have to calculate the *de-mixing* or *separation* matrix W together with the underlying factors $\mathbf{s}$ this type of problem is termed *blind source separation*. The forward model for the generation of the data is a linear basis plus additive noise:

blind source separation

$$\mathbf{x} \quad = \quad W^{-1}\mathbf{s} + \nu, \tag{6.8}$$

[2] There is a (hidden) sign in every lab stating: 'Please Lord, let the world be linear, stationary, and Gaussian.'

where W^{-1} is an $M \times N$ mixing matrix with $N = M$ and ν is additive Gaussian white noise. The basis functions appear as columns of this mixing matrix. The corresponding data likelihood is

$$\log P(\mathbf{x}; W^{-1}, \mathbf{s}) \approx -\frac{1}{2\sigma^2}(\mathbf{x} - W^{-1}\mathbf{x})^2, \tag{6.9}$$

where σ^2 is the variance of the noise. Because the de-mixing matrix W is expensive to compute, there is strong incentive to find basis functions that are easy to invert.

Algorithms of principal component analysis are known to de-correlate the data thus learning an orthogonal basis set. Independent component analysis assumes the statistical independence of the sources which incorporates de-correlateness as a pre-requisite. By statistical independence we indicate that the value of one component x gives no information on the values of the other component y. We can define this mathematically in terms of the probability densities of the two involved random variables. They are statistically independent if and only if the joint density (or the cumulative distribution functions) factorizes into the product of the marginal densities $p_{x,y}(x, y) = p_x(x)p_y(y)$.

statistical independence

Only for Gaussian distributions both uncorrelateness and statistical independence are equivalent, i.e., uncorrelated Gaussian components are always independent. This is indicated by the fact that all higher order moments (higher than 2) of a Gaussian distribution can be expressed by its first two moments. Thus the higher order moments contain no additional information. This is used explicitely in the transformation of the moments of a function to the *cumulants* of the function (see Appendix). For a Gaussian distribution the cumulants of order higher than 2 are always 0. This is only approximately true for finite samples from Gaussian distributions. The higher the empirical moment the higher its sensitivity to outliers.

cumulants

An intuitive principle of estimating independent components is based on the measure of higher order moments or cumulants. One assumes that densities of mixtures of statistically independent sources are more similar to Gaussian densities than the densities of any single (unmixed) component. Thus we can jugde the quality of an estimated de-mixing matrix $\tilde{W}$ by the similarity of the densities $p(\tilde{s}_i)$ ($\tilde{W}\mathbf{x} = \tilde{\mathbf{s}}$) to the Gaussian distribution. Finding sources which are as non-Gaussian as possible we solve our problem. Non-Gaussianity is usually measured as non-zero third or fourth order cumulants.

It helps to think of W as a set of basis vectors $\mathbf{w}_i$ (rows of W) and the assumption of non-Gaussianity is expressed in the non-Gaussianity of the marginal distributions of the dot-products $\mathbf{w}_i^T\mathbf{x}$.

This fact relates ICA to a method called *projection pursuit* which is used mainly for visualization purposes and looks for maximally *interesting* distributions which are those of large higher order cumulants.

6.2.1. Overcomplete ICA

As we have seen in the previous section algorithms of independent component analysis seek to find the de-mixing matrix W and the underlying sources $\mathbf{s}$ given a hopefully large set of observations $\mathbf{x}(t)$. The model assumes that the observations are stationary and linearly mixed sources

$$\mathbf{x} \;=\; W^{-1}\mathbf{s} = A\mathbf{s}. \tag{6.10}$$

The W^{-1} is only defined if there are as many sensors (entries of $\mathbf{x}$) as sources. This is a strong assumption because in many applications one does not know the correct number of sources. Three cases are possible: (*i*) square ICA where we have as many sources as sensors, (*ii*) if we have more sensors than sources (in the case of no noise and linear mixtures this can be detected easily by decorrelation) and (*iii*) if we have more sources than sensors[3]. In (*iii*) the mixing matrix A is not square and we assume more sources than sensors. Because a complete set of basis functions[4] is given by as many sources as sensors the above case is termed *overcomplete* ICA. In Figure 6.1 left, we have symbolized the case of as many sensors as sources (a square mixing matrix A) and (right) the case of more sources than sensors on which we will focus our attention. In the literature this problem is also known as the estimation of an overcomplete dictionary. Analyzing natural images with (square) ICA one finds indications for an underlying overcomplete basis set. By increasing the number of basis functions Lewicki and Olshausen (1999) finds a better sampling of position, orientation, and scale. Starting ICA algorithms with different initializations one can find different sets of independent components.

Overcomplete bases are interesting because they admit (*i*) some advantages in terms of interpolation, for example an increased stability of the representation in response to small perturbations of the signal

wavelets

(Simoncelli, Freeman, Adelson and Heeger, 1992). Wavelet transforms on the other hand can be unstable with respect to translations and rotations. (*ii*) They also form a tight frame (Lee, 1996) that is, they allow stable reconstruction of images by linear superposition of the basis functions with their own projection coefficients (preservation of the signal energy). Through this they (*iii*) provide

sparse coding

representations of images using coarse neuronal responses, i.e., sparsity in the

[3] Another (The) problem for real world applications is that the model of a stationary linear mixtures is wrong.

[4] Given for example by the corresponding eigenvectors of the correlations matrix of the observations.

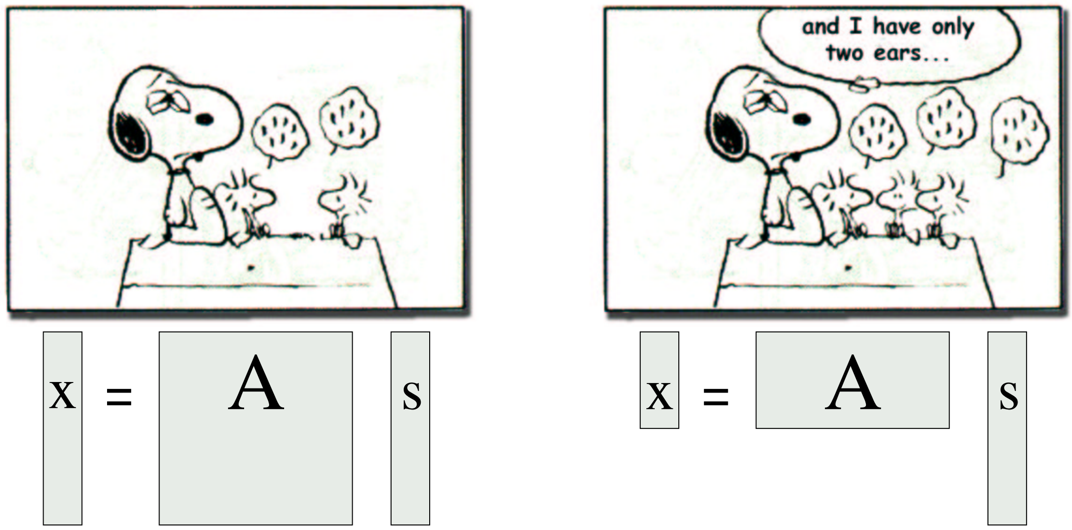


Figure 6.1.: Generative model. *Left:* As many sources as sensors, the mixing matrix A is square. *Right:* More sources than sensors, the mixing matrix A is non-square

representation (Olshausen and Field, 1997). A criticism of overcomplete representations is that they are redundant, i.e., a given data point can have many possible representations. Additional assumptions can select a representation by utilizing prior knowledge about the shape of the density functions of the single signals.

There is also a more technical reason to be interested in overcomplete dictionaries. If we look closely onto a particular model generating data (no noise, 6-dimensional observation and 12 sources, thus a two times overcomplete dictionary) we see that

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = A\mathrm{s} \quad = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,12} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,12} \\ a_{3,1} & a_{3,2} & \cdots & a_{3,12} \\ a_{4,1} & a_{4,2} & \cdots & a_{4,12} \\ a_{5,1} & a_{5,2} & \cdots & a_{5,12} \\ a_{6,1} & a_{6,2} & \cdots & a_{6,12} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ \vdots \\ s_{12}. \end{pmatrix} \tag{6.11}$$

The observations x are computed by the sources s and the mixing matrix A in a way that s_i is in product only with the i-th column of A.

$$\sum_{i=1}^{6} x_i = \sum_{i=1}^{12} s_i \sum_{j=1}^{6} a_{j,i} \tag{6.12}$$

ICA maps

Thus s_i acts as a weight of the vector $\mathbf{w}_{.,i}$. An observation $\mathbf{x}(t)$ is modeled by adding these weighted vectors. If a $s_i(t)$ is near zero its contribution to $\mathbf{x}(t)$ is small, but, if the source $s_i(t)$ is large (present) for one i only the corresponding $\mathbf{w}_{.,i}$ will solemny define the observation $\mathbf{x}$. Therefore we can view each column $\mathbf{w}_{.,i}$ of the mixing matrix A as a spatial (spatial[5]) *pattern* defining the structure of a source s_i. Sometimes the pattern are also called *maps* because they map the structure of the found source.

If we do not know the number of sources, how many possible pattern are there? Infinitely many, because we can also ask how many pairwise distinct columns can be defined for the matrix A (the entries of the matrix are scalar values). If we assume that the pattern contain binary values only the possible number of pattern is finite. The model would be still valuable in real world applications because a single binary value in the pattern can represent cortical activity and code for an either activated or inactivated area. The number of possible pattern (columns) for the mixing matrix is in this case $2^6 = 64$ (so A should be modeled by a 6×64 matrix). Even if not all 64 pattern are generated by the cortical circuitry it may be very likely that more than 6 pattern are present. With square ICA we can only detect 6 of them (as many pattern as sensors).

Introducing more sensors offers only a poor alternative. First of all the sensor signals will be highly correlated and thus the amount of additional information is very low. Second, adding more sensors gives the model the opportunity to explore the pattern between sub-groups of sensors (i.e. it will enhance the spatial resolution of the method). But it will not solve the original problem of obtaining 12 pattern from 6 sensors. Therefore we need more effective methods to solve overcomplete ICA.

A drawback connected with overcomplete bases is that the values of the independent components cannot be exactly recovered even if the mixing matrix is known. This is because the mixing matrix is not invertible. Therefore, even after estimating the mixing matrix A, the problem of optimal estimation of the realizations of the independent components needs to be solved. The naive approach of using the pseudo-inverse of the mixing matrix $\hat{\mathbf{s}} = A^{+}\mathbf{x} = A^T(AA^T)^{-1}\mathbf{x}$ does often not work well (Hyvärinen and Inki, 2002; Olshausen and Field, 1997). More advanced algorithms incorporate

[5] Spatial because it appears in a time step $\mathbf{x}(t)$.

some knowledge about the sources at this stage of the problem. For example, a smoothness assumption about the change of the active sources in time. In the context of image analysis we are interested in obtaining the basis functions as underlying structure of the images which consists out of the columns of the mixing matrix, thus we need not to cover the additional problem of obtaining the sources (Olshausen and Field, 1997). In summary we are interested in the recovering of the mixing matrix (ICA problem) which represents the pattern or maps of the sources and we ignore the additional problem of blind source separation (BSS).

Please keep in mind that we are still interested in solutions of the extended quadratic form in Equation 6.5. The connection of the two problems, that of learning features which describe symmetries and that of solving a non-square ICA problem is the following: Both will use the same non-linear transformation into a higher dimension feature space. Under the assumption that natural images can be explained by the model of ICA we can use the algorithm of non-square ICA to solve Equation 6.5 and find the independent factors of natural images.

6.2.2. Algorithms for Learning Overcomplete Dictionaries

Developing efficient algorithms to solve problems with overcomplete dictionaries is an active area of research. A first approach for learning an overcomplete basis set is, for example, to select the basis functions according to a low entropy description (Chen, Donoho and Sanders, 1996; Coifman and Wickerhauser, 1992; Mallat and Zhang, 1993) of a particular signal or a class of signals such as texture (Zhu, Wu and Mumford, 1997). These approaches pre-define a set of basis functions on intuitive criteria such as using 2D Gabor functions for modeling oriented structures in images. An entropy measure was then used to select from this set the best solution. Although overcomplete bases can be more flexible in terms of how the signal is represented, there is no guarantee that pre-designed basis vectors will be well matched to the structure of the data. Ideally, we would like the basis itself to be adapted to the data, so that for a signal class of interest, each basis function captures a maximal amount of structure.

An approach which learns the basis functions is motivated by the sparsity assumption of neuronal codes (Olshausen and Field, 1996; Olshausen and Field, 1997; Lewicki and Olshausen, 1998; Lewicki and Sejnowski, 1998; Lewicki and Sejnowski, 2000). The redundancy problem of the overcomplete basis set is solved by assuming a particular prior $P(\mathbf{s})$ on the activation of each component.

sparse coding

The principle states that only few neurons (s_i) should be active at any time to save metabolic costs, but if a neuron is active it should respond strongly in order to signal the presence of a feature[6]. The corresponding probability density of the firing of such a neuron should have a high peak at zero and 'heavy tails' which resembles the typical shape of a super-Gaussian distribution. One example for such a distribution is plotted in Figure 5.3 on page 98. It is obtained as the marginal distribution of a projected Gaussian distribution into the space of dot-products. Sparse coding favors non-Gaussian distributions, ergo an ICA solution.

independent factor analysis

In the previous model the prior on the source distributions (density of the activation of a unit) has to be chosen in advance. Usually the models work best if the prior used by the algorithm corresponds to the prior used by the generation of the data. In real world applications the correct priors are usually unknown. Thus it is preferable to learn the correct priors from the data at the time the mixture is estimated. This idea of adaptive source densities is implemented by an algorithm termed *independent factor analysis* (Attias, 1999). Here, the source distributions are modeled as factorial mixtures of Gaussians. Each source density i (prior distribution $P_i(\mathbf{x})$) is modeled by a mixture of n Gaussians. The term *factorial mixture of Gaussians* refers to the fact that the Gaussian distributions for one source are bound together.

quasi-orthogonality

All proposed algorithms are computationally very demanding. Hyvärinen (1999) introduce a fast algorithm for the estimation of an approximate overcomplete basis. The algorithm uses the interesting property of *quasi-orthogonality* (Kohonen, 1995). The observation shows that one can place a lot of vectors in higher dimensions that are close to orthogonal (quasi-orthogonal). In fact, if the dimension grows the angles between pairs of vectors can be made arbitrarily close to 90 degrees. The point here is that in two dimensions only two vectors can be orthogonal but in 50 dimensional many hundred vectors can be made 'quasi-orthogonal'. Algorithms like support vector machines rely (implicit) on this property, if they use projections into higher dimensional spaces. The probability to arrive due to an arbitrarily projection at an orthogonal set is high thus algorithms relying on linear separation work well in high dimensional feature spaces.

[6] A metabolically efficient code must balance between silent and spiking neurons. Laughlin and Attwell (2000) have produced estimates of the actual costs of dormancy and spiking in rat cerebral cortex neurons and they found that the maintenance of a neuron at rest absorbs a significant amount of energy. The generation of one action potential uses as much energy as maintaining a single neuron and its attendant glial cells at rest for only about $2s$.

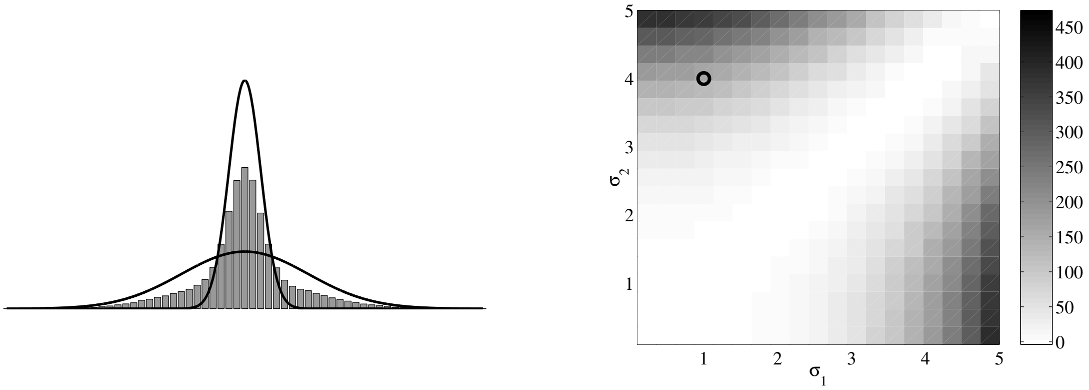


Figure 6.2.: Two Gaussian functions can model a sparse distribution. *Left:* A histogram of one-dimensional data x which is drawn from a mixture of two independent Gaussian distributions ($\mathcal{N}_1(0, \sigma_1 = 1), \mathcal{N}_2(0, \sigma_2 = 4)$). *Right:* the kurtosis ($k = E(x^4) - 3(E(x^2))^2$ for zero mean data) of the distribution dependent on the variances of the first and the second mixture component. The circle indicates the parameter settings for the two functions left

6.3. Solving Overcomplete ICA by Mixtures of Gaussians

We present now a solution to the problem of overcomplete ICA. We solve the source separation problem by a method of statistical data modeling. A source will be described by a multivariate distribution parameterized by A^i. Whereas density estimation is usually a harder problem then source separation we can use a crude model of the densities to reliably obtain the source directions of an additive linear mixture.

We like to describe the observed variables $\mathbf{x}_i$ (independently drawn, correlated sensor signals) by a set of L unobserved variables s_j which are mutually independent. The simplest description of the problem is given by the linear model

linear model

$$x_i \;=\; \sum_{j=1}^{L} w_{ij} s_j + \nu, \tag{6.13}$$

where w_{ij} describes an entry of the stationary mixing matrix W and ν is a random vector reflecting corrupted source, sensor, or mixing signals. We assume therein that the model is correct, thus the observed data can be produced by the model with some parameters.

Lets assume that our source densities are sparse and symmetric. The ICA solution we are looking for provides linear sources that explain our data as a factorizing distribution. In order to obtain the sources we now introduce a model of a mixture of Gaussian functions that is learned in order to estimate the factorizing distribution we are looking for.

We have to be sure that the sparsely distributed sources can be modeled sufficiently by a mixture of Gaussian functions. More specific we want to model n source directions by very few, e.g., $n+1$ Gaussians. Generally this cannot be achieved, we need at least two functions to model one sparse distribution. In Figure 6.2 on the page before two Gaussian functions are used to model sparse distributions. Given the density function of the data is symmetric and unimodal we can fix the mean of both Gaussian functions to zero. Only their respective variances σ^1 and σ^2 act as free parameters and produce distributions with arbitrary high kurtosis which we use here to indicate sparseness. For this reason the model is termed *centralized Gaussian mixture model*.

In two dimensions the restriction of zero mean symmetric basis function gives us a vantage point for analyzing sparse data. One source distribution can be efficiently modeled by two Gaussian functions, one with large and elongated and the other small and circular. For every additional (sparse) source present in the data the model needs only 1 additional (large and elongated) Gaussian function because the small circular shaped distribution can be utilized by both sources. Therefore we can model each sparse source distribution by 1 Gaussian function with an overhead of one additional Gaussian in the model. Note that we still have to know the number of sources.

6.3.1. Learning of the Mixture Model by EM

To find the parameters of the model we employ the the Expectation Maximization (EM) algorithm of Dempster et al. (1977). The EM algorithm is able to find a *locally* optimal solution for the estimation of the densities.

A common problem in learning mixtures of Gaussian densities with the EM algorithm is the occurrence of singularities due to source densities that concentrate onto a single data point. The corresponding covariance matrix eventually gets singular whereas the likelihood of the model is maximized. This can be seen by calculating the likelihood of a Gaussian distribution

$$\begin{aligned} L(x) &= -\log P(x) \approx -\log(p(x)\Delta) = -\log p(x) - \log\Delta && (6.14)\\ &= \frac{(x-\mu)^2}{2\sigma^2} + \frac{1}{2}\log 2\pi + \log\sigma - \log\Delta && (6.15) \end{aligned}$$

as Δ (scales the area under the function) goes to zero the likelihood goes to infinity.

Regularizing Σ can help in this case. We use an update rule for the co-variance matrices based on the Wishart density (Buntine, 1994;

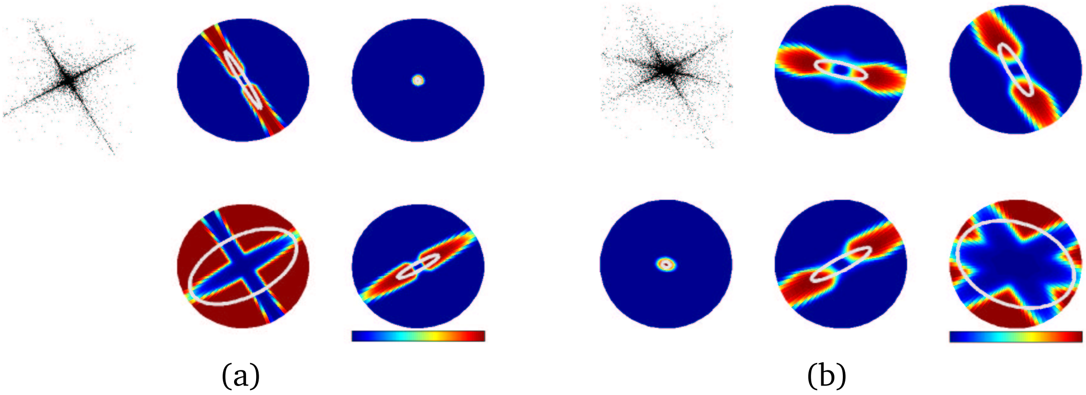


Figure 6.3.: Mixtures of two respectively three independent sparse sources learned by EM. (*a*) The first figure is a scatter plot of the data (ICA source model, $g(u) = u^3$). The other four figures display in color the posterior probabilities of the four Gaussian functions. The respective 2-dimensional Gaussian functions are shows as ellipses. (*b*) The corresponding result for three sources in two observations

Ormoneit and Tresp, 1996). We constrain the likelihood by adding the sum of the inverse of the eigenvalues of the covariance matrix defining the variances of the Gaussian functions. The respective change in the update rule compared to the standard M-step is

$$\Sigma^i_{\text{unr}} = \frac{\sum_{k=1}^m h_i^k (x^k - \mu_i')(x^k - \mu_i')^t}{\sum_{l=1}^m h_i^l} \quad \Rightarrow \quad \Sigma^i_{\text{reg}} = \frac{\sum_{k=1}^m h_i^k (x^k - \mu_i')(x^k - \mu_i')^t + 2\beta I}{\sum_{l=1}^m h_i^l + 1}. \tag{6.16}$$

where I is the unity matrix and β a regularization parameter. In effect this cannot prevent that for a Gaussian a eigenvalues converges to zero, if all other eigenvalues are sufficiently large. Nevertheless, it worked fine in all our experiments.

Observations on the un-regularized version of the original EM-algorithm suggest that in regions with one source density and two or more model densities the densities do not cooperate. Only one component is used to explain the data density whereas the other components converge to single data points. This is of importance because in the centralized Gaussian mixture model we fix all mixtures at zero mean, thus they strongly compete for the data. But this setting in fact helps, because there is only a single data point (at zero) for a singularity to occur (but see the regularization), thus by sacrificing a component with an approximative uniform distribution the other components (one for each sparse independent component) can explain the data. The direction of the first principle axis of each component is used as an indicator of the source direction.

In Figure 6.3 on the preceding page for two toy data problems (with 2 respectively 3 sources in 2 observations) the model of centralized Gaussian mixtures was trained with the proposed algorithm. The data are obtained as linear mixtures of independent super-Gaussian ($\mathcal{N}_{1,2,3}(\mu = 0, \sigma = 1)^3$) sources and are shown as scatter plots. We trained a mixture of $l = 1, \ldots, 4$ respective $l = 1, \ldots, 5$ Gaussian densities on a data set of size $10,000$. Classifiers of new inputs $\mathbf{x}$ are obtained by the maximum posterior class probability $p(\mathbf{x}|l)$ computed from the class-conditional data likelihood $p(\mathbf{x}|l)$, $p(\mathbf{x}|l)$ and $p(l)$ are estimated from the sample data by the EM algorithm of (Dempster et al., 1977) with regularization by Equation 6.16. In each maximization step in the algorithm the centers of the Gaussian functions are kept fixed to the center of mass of the data distribution.

In Figure 6.3 the posterior probabilities of each component l are coded in color (red for high probability of x to be generated by l). Overlayed is an ellipse illustrating the respective iso-probability curve of the two dimensional Gaussian function. In each case (a) and (b) two components could be learned with small respective large circular Gaussian-distributions. All other components correspond in their first principal axis to source directions in the data.

6.3.2. Conclusion

The proposed model estimates the source densities (parameterized by the covariance matrices) similar to the model presented in Xu, Cheung, Yang and Amari (1997) and the independent factor analysis of Attias (1999). Also Olshausen and Millman (2000) proposed a model in which a mixture of Gaussian priors is learned to characterize the (sparse) posterior distribution over the coefficients. Interestingly the last model also tried to model a single source direction by 3 Gaussian distributions and found that two large variance distributions converged to the same mean position. Subsequently the results in Olshausen and Millman (2000) are obtained in the case where two Gaussian distributions per source are used to model one (sparse) source distribution. In our approach each component independently models the data. Therefore, we cannot fully clarify which set of components is used to model one source direction. Nevertheless the results on the toy data set give us a clear intuition how to interpret the found solution. Basically one component at the center position is utilized by all other components to produce a factorizing distribution similar to the one dimensional case illustrated in Figure 6.2 on page 123. Before presenting first results for the GEM algorithm on data obtained from natural images we propose an alternative algorithm to learn a mixture model of quadratic forms. This algorithm allows us to extend the concept of higher order feature detectors to correlations of any order.

6.4. Solving Overcomplete ICA by Introducing New Dimensions

An intuitive reason for the problems of non-square ICA is that we do not have enough space to make our data statistically independent. A linear mixture of two distributions can be made statistically independent in two dimensions, regardless of the specific shape of the distributions. Wrong choices of the transfer function result in errors only in the scaling of the found sources. This is not true for the projection of a linear mixture of three or more distributions onto two dimensions. Introducing independence in the data is only possible if we can transform the data in such a way that every unique source direction corresponds to a unique space dimension. Therefore we like to introduce new dimensions. The idea is to project the data into a higher dimensional space.

In the next section we give a short introduction into the notion of features spaces.

6.4.1. Feature Spaces and Manifolds

The goal of data processing is to obtain a better representation of the information contained in the data. Classification is an extreme case were a high dimensional input is projected onto a single value describing a class label. Often the data is projected into a space of higher dimensions, for example by computing the response of each location in the image to a bank of filters. Selecting a specific bank of filters results in highlighting specific information which are also called the features in the data. For this reason the space defined by a projection is called *feature space* whereas the space of our original data we will call *input space*.

feature space

Projections may or may not destroy information in the data. This is expressed in the notion of (non-)invertible projections or functions. If a function is invertible no information is lost whereas if a function is non-invertible information about the data is lost during the transformation. Of some importance is the dimensionality of the feature space with respect to the dimensionality of the input space. If the feature space is of lower dimension information is certainly lost during the transformation[7] but often the dimension of the fea-

[7] It gets more complicated here if the information in input space fills only a sub-space. For data from this sub-space no information may be destroyed but this requires additional knowledge about the data and should be handled by pre-processing.

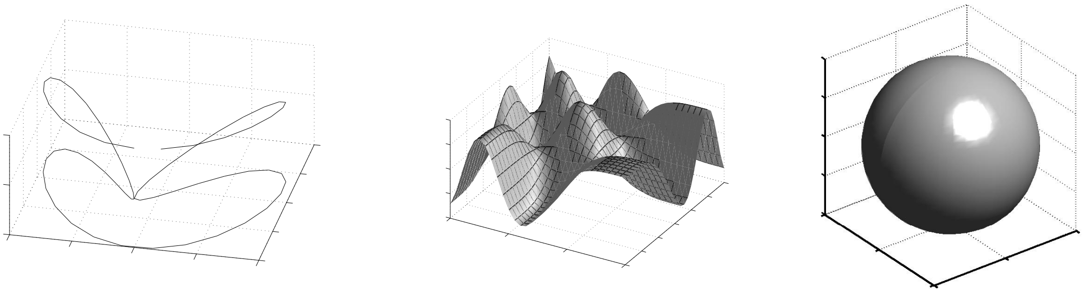


Figure 6.4.: Example spaces. *Left:* Intrinsic dimension 1, extrinsic dimension 3. *Center:* Intrinsic dimension 2, extrinsic dimension 3. *Right:* The space enclosed in the sphere is a manifold with intrinsic dimension 3 and extrinsic dimension 3

ture space equals or is larger than the dimension of the input space. A set of spatial filters $[\Phi^i]$ applied onto each position i in the image X is an example for a projection $f(i) : \mathbb{R} \rightarrow \mathbb{R}^m$ of a single pixel into a higher dimensional feature space (m filter coefficients). Because of the transformation there is no additional information which appears magically to fill the additional $(m-1)$ dimensions, during the transformation some information is only 'borrowed' from neighboring positions $j : |i-j| < \mathrm{rfs}$ in the image. So at best we can redistribute the information contained in the data evenly. It depends on this redistribution process if we (lose) retain information, i.e., if the transformation is (non-)invertible. Important is that in the case of no noise the transformation cannot effectively fill the higher dimensional space. The data will be restricted to a sub-space also called a manifold in the feature space.

manifold

A manifold is the generalization of the concept of surfaces or shapes to higher dimensions. It is defined a set of points with an associated coordinate system. In our above example a point of the manifold S is defined by the m filter coefficients obtained from a single image patch. By a coordinate system we mean a one-to-one mapping from S to $\mathbb{R}^n$, which allows to specify each point in S using a vector of n real numbers (the coordinates of the corresponding point). The natural number n is called the *intrinsic* dimension of S (writing $n = \dim S$). Let S be a manifold and $\phi : S \rightarrow \mathbb{R}^n$ be a coordinate system for S. Then ϕ maps each point p in S to n real numbers: $\phi(p) = [\chi^1(p), \ldots, \chi^n(p)] = [\chi^1, \ldots, \chi^n]$. These are the coordinates of the point p. The n functions $\chi^i : S \rightarrow \mathbb{R}(i = 1, \ldots, n)$ are called the *coordinate functions*.

intrinsic dimension

We are free in choosing a coordinate system. An interesting candidate is given by the coordinate system of our input space. In the case of a set of spatial filters the inverse of the transformation into the feature space $f^{-1}(p) = \phi$ maps a point p in S into $\mathbb{R}^n$, n being the dimension of an image patch i in

X. In linear filtering the coordinate functions χ^i are simply the lines of the inverse of the transformation matrix (provided the inverse exists). Therefore the introduction of a high dimensional feature space can be understood as the definition of manifolds in which the data is projected.

The dimension we are talking about in colloquial terms is often the intrinsic dimension, not the *extrinsic dimension*. Thus, a curve in 3-D space is a one-dimensional manifold (intrinsic dimension 1) of extrinsic dimension 3. The surface of a ball is a two-dimensional manifold (intrinsic dimension 2) of extrinsic dimension 3. Figure 6.4 shows three example spaces and the corresponding intrinsic and extrinsic dimensions.

extrinsic dimension

The important point is that a deterministic transformation of a n-dimensional space into a higher $m > n$ dimensional space can never enlarge the intrinsic dimension of the data. The data will always be in a lower-dimensional manifold. We can only enlarge its extrinsic dimension. A transformation which enlarges the extrinsic dimension we will call an *effective* transformation.

Lets do an example for a not-effective transformation. If we assume the input space as being $\mathbf{x} := (x_1, x_2) \in \mathbb{R}^2$ and we construct the feature space according to $f(\mathbf{x}) = \mathbf{y} := (x_1, x_2, x_1 + x_2)$ we do *not* enlarge the extrinsic dimension of the data from 2- to 3-dimensions. This can be seen by performing a change of variables $x'_1 = 2/3y_1 - 1/3y_2 + 1/3y_3$, $x'_2 = -1/3y_1 + 2/3y_2 + 1/3y_3$, and $x'_3 = 0$ (which inverts our feature space transformation, x'_i being the i's coordinate function). We can express the data in a coordinate system with only 2 non-zero dimensions. This indicates that all data in feature space are concentrated in a 2-dimensional plane.

In general, every linear transformation of the form $\mathbf{y} = A\mathbf{x}$ is not suitable to *effectively* construct a high dimensional feature space. In our example above $\mathbf{y}$ is constructed from

$$\mathbf{y} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = A\mathbf{x}. \tag{6.17}$$

The coordinate functions x'_i are computed from the pseudo-inverse of A,

$$A^+\mathbf{y} = \begin{pmatrix} 2/3 & -1/3 & 1/3 \\ -1/3 & 2/3 & 1/3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \tag{6.18}$$

6.4.2. A Feature Space for Overcomplete ICA

As seen in the last section the effective mapping into a higher dimensional feature space cannot be done by linear methods because there the *extrinsic*

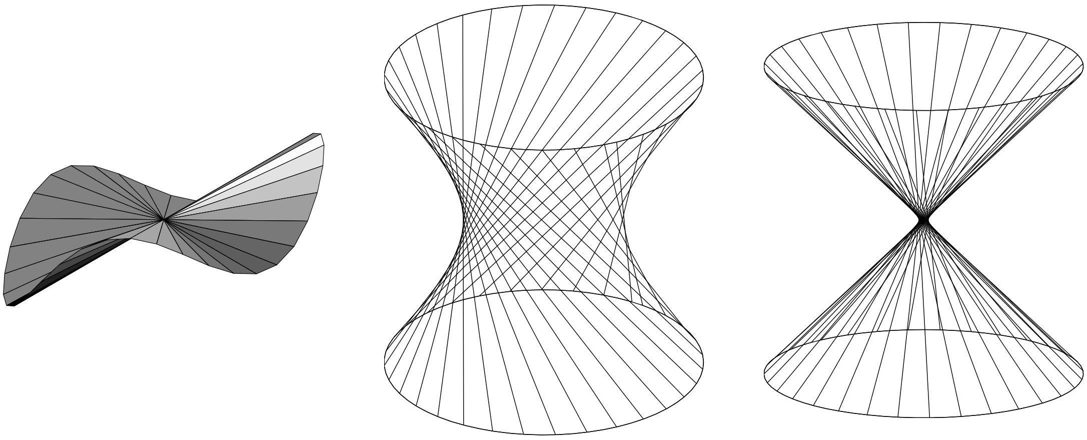


Figure 6.5.: Three examples for manifolds S (surfaces in 3 dimensions) defined by non-linear transformations where specific directions (depicted by lines) are lines, i.e., for some tangent vectors ψ each $a\psi$ is in S ($a \in \mathbb{R}$)

dimension of the data remains. The linear transformations can only construct rotated and scaled hyper-planes. The de-correlation applied as pre-processing step in many algorithms easily 'destroys' the additional dimensions. To obtain an effective enlarge representation of our data we need therefore non-linear transformations. Only they can provide an embedding of our data into a higher dimensional space that cannot be undone by linear methods. Whereas many non-linear transformations will enlarge the extrinsic dimension of the data not all of them can be used in our context of linear source separation.

Sorry, now its gets a little bit complicated: Our goal is to make the data in feature space statistically independent which solves the problem of linear source separation. If the data has extrinsic dimension m in a space of m dimensions than linear ICA provides an solution to our goal by estimating a de-mixing and a corresponding mixing matrix. Once this is done the source pattern are given simply as the columns of the mixing matrix. Using the algorithms of FastICA the columns of the mixing matrix are found as directions in which the data is sparsely distributed. But our data is in that space restricted to a lower dimensional manifold. We conclude that the transformation has to implement a property that directions, i.e., scaled vectors, are on the manifold (the manifold is closed under scalar multiplication). This may appear to be contradictory, because at the one hand we want to use non-linear transformations and on the other hand we want our data in feature space to be concentrated along directions in Euclidean space.

A solution to this problem is shown by the observation that we can combine linearity and non-linearity in orthogonal directions in the feature space. Three examples for manifolds with this property are shown in Figure 6.5 on the facing page. Every on of them displays a curved non-Euclidean 2D manifold in 3D space. Yet, in each of them some some of the tangent vectors (depicted as lines) are in the manifold. Only the first and the third example corresponds to feature spaces in which all tangent vectors also touch the mean of the data which is important because only those can be found by linear ICA.

A non-linear mapping $\mathcal{F} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ that projects scaled vectors $\alpha\vec{\gamma}$ in input space into scaled vectors $\vec{g}(\vec{\gamma})$ in feature space must obey the decomposition:

$$\mathcal{F}(\mathbf{x}) \;=\; f(\mathbf{x})\;\vec{g}\left(\frac{\mathbf{x}}{||\mathbf{x}||_2}\right) = f(r,\vec{\gamma})\;\vec{g}(\vec{\gamma}) \tag{6.19}$$

where the tuple $(r,\vec{\gamma})$ describes the data in polar coordinates. The non-linearity is banned explicitly to appear only in $\vec{g}$, i.e., in directions orthogonal to source directions[8].

$f(.)$ and $\vec{g}(.)$ are functions that separately describe either a *scaling* of the data in the direction $\vec{g}$ or the non-linear mapping (*bending*) of directions by $\vec{g}$. $\vec{g}$ is a vector valued function of dimension $1 \times m$, m being the dimension of the feature space. In ordinary ICA one would choose m as being the number of sources to separate.

scaling space

bending space

An assumption that can be placed onto the scaling function $f(r,\vec{\gamma})$ is that our problem is rotational symmetric, e.g., the scaling should not depend on direction.

$$f(r,\vec{\gamma}) \;=\; f(r) \tag{6.20}$$

The mapping $\mathcal{F}$ is therefore expressed in general form as

general form

$$\mathcal{F}(r,\gamma_1,\ldots,\gamma_{n-1}) \;=\; f(r)\begin{pmatrix} g_1(\gamma_1,\ldots,\gamma_{n-1}) \\ g_2(\gamma_1,\ldots,\gamma_{n-1}) \\ \vdots \\ g_m(\gamma_1,\ldots,\gamma_{n-1}) \end{pmatrix}. \tag{6.21}$$

Because we decoupled our class of functions it is sufficient to consider the *scaling* property implemented by f and the *bending* property implemented by $\vec{g}$ separately.

Desired Properties of the Transformation

Now we introduce some properties that further constrain the space of possible functions which up to now only obeys the form of Equation 6.21. In here

[8] Our manifold in feature space is like Flatland with its inhabitant A-Square exploring the curved nature his 2D space in 3D (*Sphereland: A Fantasy About Curved Spaces and an Expanding Universe*, Dionys Burger (1965)).

we borrow some methods of orthogonal functions, differential geometry, and information geometry (Amari and Nagaoka, 1993). We will mark properties that are needed in order to perform linear methods in feature space as *necessary* and explain also how for a given transformation the properties can be tested.

Properties of the function class $\mathcal{F}$ are:

1. *linear Decomposition (necessary):* The function $\mathcal{F}$ is of the form

   $$\mathcal{F}(r,\vec{\gamma}) \;=\; f(r,\vec{\gamma})\vec{g}(\vec{\gamma}).$$

   Test: by linear algebra

2. *Linear independence of basis functions (necessary):* The m functions $g_i(\vec{\gamma})$, $i = 1 \dots m$ are called *linear dependent* in a set G if there exist m constants $c_1, \dots, c_m$, not all zero, for which the function $c_1 g_1 + c_2 g_2 + \dots + c_m g_m$ is zero. If such constants do not exist, the m functions are called *linearly independent*. In our framework this property of a feature space assures that the data can effectively fill the feature space. In the context of projecting the data into higher dimensional feature spaces it enforces non-linear basis functions.

   Wronskian

   Test: Linear independence of functions is equivalent to a non-zero determinant (Wronskian)

   $$W(g_1, g_2, \dots, g_{m-1}) = \begin{vmatrix} g_1 & g_2 & \cdots & g_{m-1} \\ \frac{\partial g_1}{\partial \vec{\gamma}} & \frac{\partial g_2}{\partial \vec{\gamma}} & \cdots & \frac{\partial g_{m-1}}{\partial \vec{\gamma}} \\ \frac{\partial^2 g_1}{\partial^2 \vec{\gamma}} & \frac{\partial^2 g_2}{\partial^2 \vec{\gamma}} & \cdots & \frac{\partial^2 g_{m-1}}{\partial^2 \vec{\gamma}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^{m-2} g_1}{\partial^{m-2} \vec{\gamma}} & \frac{\partial^{m-2} g_2}{\partial^{m-2} \vec{\gamma}} & \cdots & \frac{\partial^{m-2} g_{m-1}}{\partial^{m-2} \vec{\gamma}} \end{vmatrix} \neq 0. \quad (6.22)$$

3. *General linear independence (necessary):* $\vec{g}$ should ensure that any m pairwise different vectors $\vec{g}(\vec{\gamma}_0), \vec{g}(\vec{\gamma}_1), \dots, \vec{g}(\vec{\gamma}_{m-1})$ are linear independent. This is connected with the goal of detecting arbitrarily, pairwise different combinations of source directions. If this property does not hold it is possible that m (highly super-Gaussian) sources populate a $m-k, 0 < k < m$ dimensional sub-space (by being linear- or nearly linear dependent).

   *Test:*General linear independence is equivalent to the matrix M having a non-zero determinant for all possible assignments of γ

   $$|M| = \begin{vmatrix} g_1(\vec{\gamma}_1) & g_2(\vec{\gamma}_1) & \cdots & g_m(\vec{\gamma}_1) \\ g_1(\vec{\gamma}_2) & g_2(\vec{\gamma}_2) & \cdots & g_m(\vec{\gamma}_2) \\ \vdots & \vdots & \ddots & \vdots \\ g_1(\vec{\gamma}_{n-1}) & g_2(\vec{\gamma}_{n-1}) & \cdots & g_m(\vec{\gamma}_{n-1}) \end{vmatrix} \neq 0 \qquad (6.23)$$

4. *Differentiable:* $\vec{g}$ should be rotational symmetric if problems are rotational symmetric. This favors differentiable, periodic functions.

5. *Factorizing distr. remain:* If we observe a factorizing distribution the projection in feature space should not introduce correlations in the data. This is a difficult problem, because the data is concentrated in a manifold (for example a bended plane in 3D). This will introduce correlations in the data. Because whitening is part of many ICA methods factorizing distributions will be spoiled.

6. *Invertible:* $\mathcal{F}$ should be invertible, so that back–projection of data is possible.

It remains to show (*i*) if there is such a feature space and (*ii*) if these considerations define a singular form of $\vec{g}$.

Because the linear methods applied in the feature space for them self change the space linearly they can cope with different linear transformations. Therefore any extended general form can be chosen

extended general form

$$\mathcal{F}(r, \gamma_1, \dots, \gamma_{n-1}) \quad = \quad f(r, \vec{\gamma}) P \vec{g}(\vec{\gamma}) \tag{6.24}$$

where P is a square, positive definite, and real valued matrix of size $(m-1) \times (m-1)$. By P one can produce any linear transformation on $\vec{g}$ which would correspond for example with rotations, scalings or translations of the manifold.

In the following we will examine some feature spaces for their properties regarding the restrictions we have placed so far onto $\mathcal{F}(r, \vec{\gamma})$ to reasonably perform linear methods as for example ICA.

Example 1: Polynomial Spaces

For polynomial spaces, which are most frequently used for defining non-linear feature spaces, directions in data space are mapped onto curves in feature space as can be seen for example by a polynomial space of order 2. The mapping of a data point $\mathbf{x}$ into feature space is done by

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \xrightarrow{\mathcal{FS}} \quad \left(1, x_1, x_2, x_1 x_2, x_1^2, x_2^2\right) \tag{6.25}$$

Consequentially, a direction $v = \binom{u}{au}$ is mapped onto a vector

$$\left(1, u, au, uua, u^2, (au)^2\right)^T \tag{6.26}$$

which is not compatible with the general form of Equation 6.21 on page 131. For example, in the sub-space spanned by u and u^2 a source direction is mapped

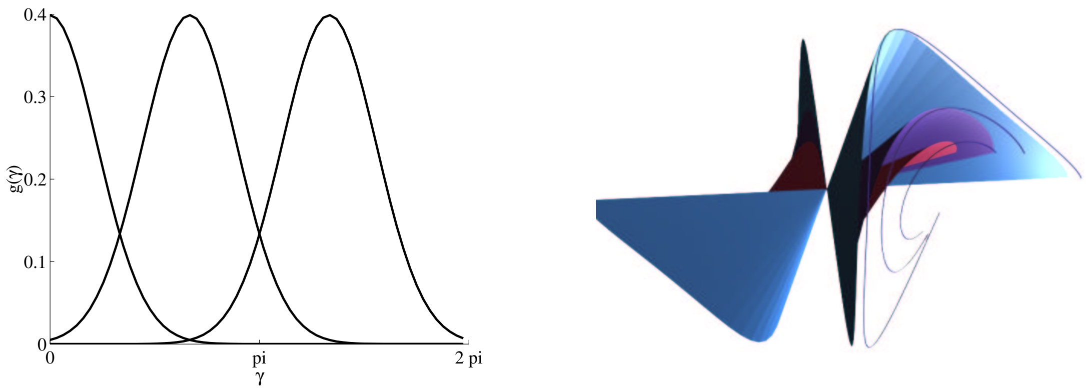


Figure 6.6.: Feature spaces defined by three radial basis functions. *Left:* The basis functions for one of the feature spaces, equally spaced in the interval $[0, 2\pi]$. *Right:* Each manifold is coded in a different color (blue, purple, red) with decreasing variances of respective basis functions

onto a parabolic function. Therefore a polynomial space is not compatible to linear methods.

One may suspect that the reason for this incompatibility is the different order of the single terms, later on in Section 27 on page 137 we will define a feature mapping with terms of constant order.

Example 2: RBF Spaces

radial basis functions

If we use radial basis functions' (RBF) as basis functions, our feature space is of the form

$$\vec{g}(\gamma) = \begin{pmatrix} G_1(\gamma_1, \ldots, \gamma_{m-1}) \\ G_2(\gamma_1, \ldots, \gamma_{m-1}) \\ \vdots \\ G_n(\gamma_1, \ldots, \gamma_{m-1}) \end{pmatrix} \tag{6.27}$$

where G_i is a Gaussian function with mean μ_i and variance σ_i. We can illustrate the mapping from data space into feature space in the case of $m = 2$ and $n = 3$ (see Figure 6.6). Doing ICA in this space one would expect to extract three independent components from two observations.

First of all, the feature space is not rotationally invariant (not a closed curve), which is due to the non-periodic defined basis functions. Second, directions can be linear dependent for small variances of the underlying Gaussian distributions. This property will be explain in more detail in the next section which is about circular radial basis functions which share this disadvantage.

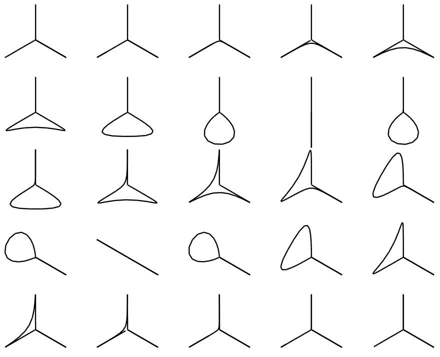


Figure 6.7.: Shape space for a moving first Gaussian distribution (shifted from $[0 \ldots 2\pi]$, left to right, top to bottom), the means of the second and third components are kept fixed. Variance was $\sigma = 0.3$

We summarize that equally spaced Gaussian radial basis functions may behave rather badly if the basis functions have small variances.

Usually one learns the parameters of the Gaussian functions from the data in order to get better results. Tuning the RBF's can be obtained by learning the means and variances. Clustering algorithms, for example, ensure that at best each single Gaussian function is placed at the direction of a component (because the data is concentrated in that direction). One can suspect that non-equal spacings are favorable, even though this assumes knowledge about the unknown components. To explore the utility of an adaptation of the center positions to the data in Figure 6.7 the shape space for a moving first component is shown (mean is varying in the interval $[0, 2\pi]$). We plot here not the 3D space but the curves obtained from the crossing points of the manifold with a plane that is tangent to the unit sphere. Each point on the curve represents a possible source directions. The variances of the basis functions is relatively small ($\sigma = 0.3$) and in all cases we can imaging three directions, i.e., three pairwise different points on the curve to be linear dependent.

Example 3: Circular Radial Basis Functions

Using ordinary Gaussians as radial basis functions for circular data is questionable because they cannot deal with the periodic nature of the data. In the next section we therefore move to a different form of basis function, the *von Mises* distribution, which is the equivalent circular to the normal distribution.

The *von Mises* Distribution

The circular equivalent of the normal distribution is the *von Mises* distribution

von Mises distribution

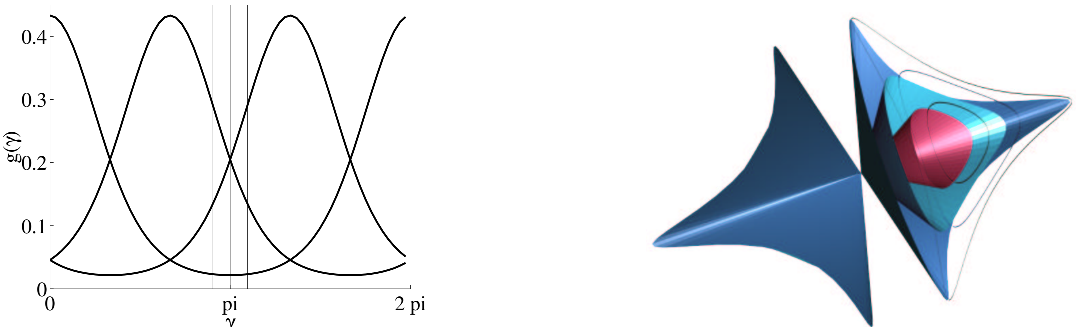


Figure 6.8.: Feature space defined by three *von Mises* distributions. with decreasing concentration parameter κ. *Left:* The three basis functions equally spaced with a concentration parameter of $\kappa = 0.3$. *Right:* Each manifold is coded in a different color (dark blue, light blue, red) with decreasing $\kappa = 1.5, 1, 0.3$ thus increasing variance

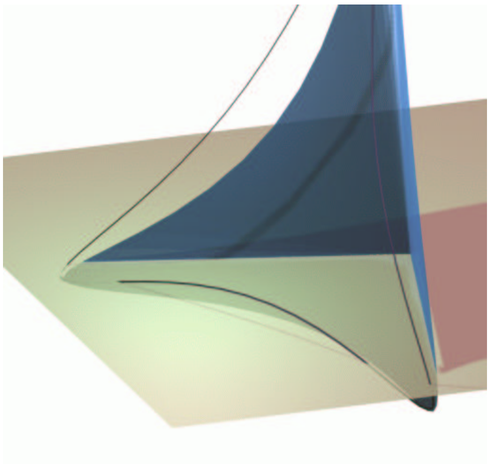


Figure 6.9.: Manifold in 3D (in blue) cut by a plane through the origin representing a 2D sub-space. It is possible to define 3 directions in feature space that all fall in the 2D sub-space. There is no guaranty for unknown source directions to fill the 3D space (be linearly independent)

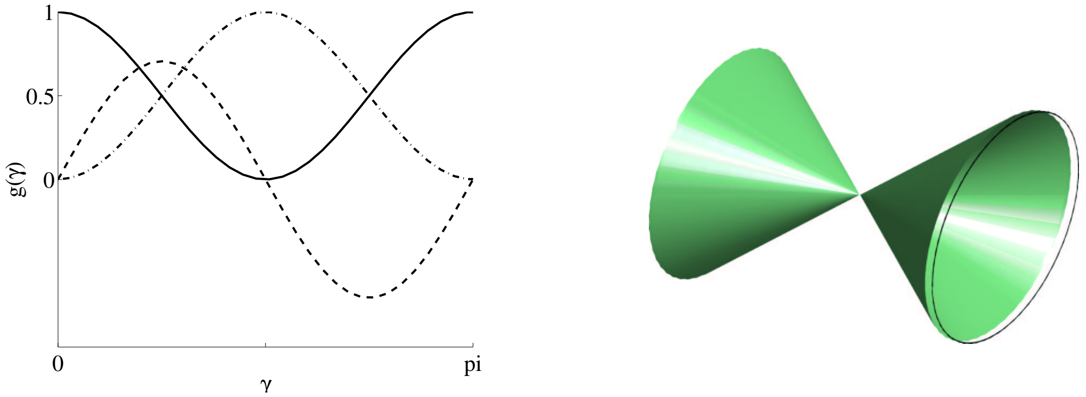


Figure 6.10.: Feature space defined by three monomials. *Left:* Three basis functions $\cos^2(\gamma), \sin^2(\gamma), \sqrt{2}\cos(\gamma)\sin(\gamma)$, $\gamma = 1 \ldots \pi$. *Right:* The 3-dimensional feature space which we termed due to its shape *Diabolo* space. A convex shape implies non-linearity for every combination of three different vectors and perfect circular symmetry

(Mardia and Jupp (2000), page 36),

$$M(\theta, \mu, \kappa) \;=\; \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \mu)) \tag{6.28}$$

where I_0 denotes the modified Bessel function of the first kind and order 0, which is defined as

$$I_0(\kappa) \;=\; \frac{1}{2\pi} \int_0^{2\pi} d\theta \;\, \exp(\kappa \cos\theta) \tag{6.29}$$

The parameter κ is known as the *concentration parameter* and behaves inverse to the variances of a normal distribution ($\kappa = 0$ equals a flat distribution). Note that $M(\mu + \pi, \kappa)$ and $M(\mu, -\kappa)$ are the same distributions. The shape of the function is very close to that of the normal distribution as can be seen in Figure 6.8 on the preceding page. But again for a large concentration parameter κ there are combinations of three directions that are linear dependent or nearly linear dependent. This is illustrated in more detail in Figure 6.9 on the facing page.

Example 4: Monomial Spaces of Constant Order

In the example about the polynomial spaces we suspected that the different order of the monomials provide the main reason for their incompatibility with linear methods. Based on this insight we now introduce monomial spaces of constant order.

Let each basis function $\vec{g}$ be defined in terms of a monomial in n variables of order d:

$$\vec{g}_n(d) \quad = \quad \{x_1^{e_1} x_2^{e_2} \cdots x_n^{e_n} \mid e_1 + e_2 + \cdots + e_n = d, e_i \geq 0\}\,. \tag{6.30}$$

It is easier to define the basis functions in Cartesian coordinates but the formulation in polar coordinates is straight forward. For example the monomials of order 2 in m variables ($\mathbf{x} = (x_1, \ldots, x_m)^T$ are defined by

$$\vec{g}(\mathbf{x}) \quad = \quad \mathbf{x}\mathbf{x}^T. \tag{6.31}$$

In order to increase the intrinsic dimension for each monomial term we only count on commutative monomials (vech = lower triangular form):

$$\vec{g}_2(x_1, \ldots, x_m) \quad = \quad \mathrm{vech}\begin{pmatrix} x_1^2 & & & \\ x_2x_1 & x_2^2 & & \\ \vdots & \vdots & \ddots & \\ x_mx_1 & x_mx_2 & \cdots & x_m^2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2x_1 \\ \vdots \\ x_m^2 \end{pmatrix} \tag{6.32}$$

Defining a feature space by monomials of constant order d ensures that the space is compatible with the general form of Equation 6.21 on page 131 because every dimension is of the form $r^d \,\mathrm{cossin}(\vec{\gamma})$, where $\mathrm{cossin}(.)$ is a product of sine and cosine terms.

We introduce now a linear scaling of the basis functions and show that all orthogonal directions in a 2 dimensional input space are mapped onto orthogonal directions in the space of monomials of constant order. Let d be the order of the monomials. If two vectors in input space are orthogonal their scalar product should be zero:

$$(x_1, y_1)\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \quad = \quad x_1x_2 + y_1y_2 = 0 \tag{6.33}$$

$$x_1x_2 \quad = \quad -y_1y_2 \tag{6.34}$$

This implies also that $x_1^e x_2^e = (-1)^e y_1^e y_2^e$. For $d = 2$ we have to test now if the scalar product of the projected vectors is zero as well:

$$(x_1^2, \alpha x_1y_1, y_1^2)(x_2^2, \alpha x_2y_2, y_2^2)^T \quad = \quad y_1^2y_2^2 - \alpha^2 y_1^2y_2^2 + y_1^2y_2^2 = 0 \tag{6.35}$$

here we used the fact that $x_1x_2 = -y_1y_2$ implies that $x_1^2x_2^2 = y_1^2y_2^2$. α is our unknown scaling factor which obviously has to be $\alpha = \sqrt{2}$. Lets do this again for an term of order $d = 3$:

$$\begin{aligned} (x_1^3, \alpha x_1^2y_1, \alpha x_1y_1^2, y_1^3)(x_2^3, \alpha x_2^2y_2, \alpha x_2y_2^2, y_2^3)^T \quad &= \\ -y_1^3y_2^3 + \alpha^2 y_1^3y_2^3 - \alpha^2 y_1^3y_2^3 + y_1^3y_2^3 \quad &= \quad 0. \end{aligned} \tag{6.36}$$

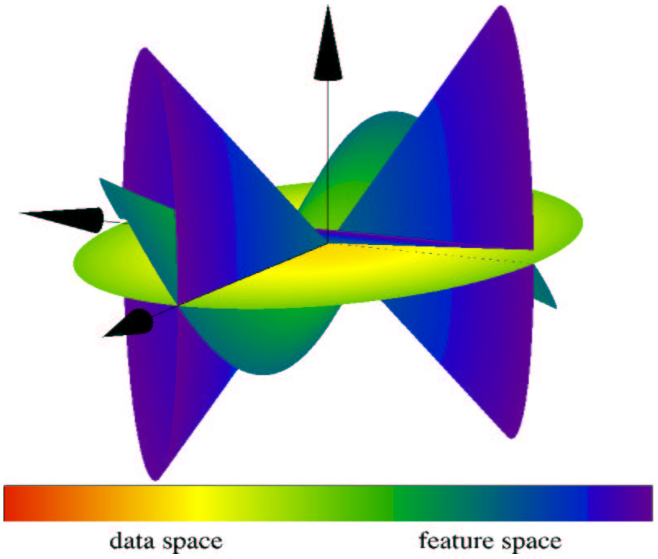


Figure 6.11.: Input space is folded into the space of monomials of constant order (Diabolo space). Shown is the procedure of folding a plane disc (in yellow) into the Diabolo space with one intermediate step in green

Here we see that α can be set to 1 and the scalar product is zero (e.g., the directions are orthogonal).

Depending on the order d the following pattern emerges:

$$(-1)^d y_1^d y_2^d + \alpha^2 \left((-1)^{d-1} y_1^d y_2^d + (-1)^{d-2} y_1^d y_2^d + \cdots + (-1)^1 y_1^d y_2^d\right) + y_1^d y_2^d = 0 \tag{6.37}$$

where

$$\alpha = \begin{cases} \sqrt{2} & : \text{if } d \text{ is even} \\ 1 & : \text{if } d \text{ is odd} \end{cases} . \tag{6.38}$$

The term is only needed for even orders of the monomials and for basis functions of the form $a_i a_j, i \neq j$.

Because of its length the proof of the general linear independence of monomials of constant order is presented in Section C on page 190.

To illustrate the mapping of directions onto Diabolo space in Figure 6.11 the two spaces, input- and Diabolo-Space are shown together with a linear interpolation ($.5$ * input space + $.5$ * Diabolo space). In this example the scaling function f was chosen to be $f(x) = |x|x$ ensuring that points in the positive quadrants are mapped onto the positive and points in the negative quadrant are mapped onto the negative quadrant. The function is therefore invertible.

The model is easily extended to monomials of higher degree or more variables. In the case that we have monomials of order n in d variables there are

$$\binom{n+d-1}{n} \tag{6.39}$$

commutative monomial (equals the number of feature dimensions). For monomials in $d = 2$ variables ($(u, v) = (r = \sqrt{u^2+v^2}, \gamma = \text{atan}(v/u))$) and order $n = 2, 3, 4$ the monomial feature spaces can be defined as $\mathcal{F}(r, \gamma) =$

$$r^2 \begin{pmatrix} \cos^2(\gamma) \\ \sin(\gamma)\cos(\gamma) \\ \sin^2(\gamma) \end{pmatrix}; r^3 \begin{pmatrix} \cos^3(\gamma) \\ \sin(\gamma)\cos^2(\gamma) \\ \sin^2(\gamma)\cos(\gamma) \\ \sin^3(\gamma) \end{pmatrix}; r^4 \begin{pmatrix} \cos^4(\gamma) \\ \sin(\gamma)\cos^3(\gamma) \\ \sin^2(\gamma)\cos^2(\gamma) \\ \sin^3(\gamma)\cos(\gamma) \\ \sin^4(\gamma) \end{pmatrix} \tag{6.40}$$

One may be worried about the periodicity of the used basis functions. Directions of linearly mixed centered distributions are π-periodic but that is not always true for our sine and cosine functions. It is easy to show that depending on the order $d = d_1 + d_2$

$$g(\gamma+\pi) \quad = \quad \cos^{d_1}(\gamma+\pi)\sin^{d_2}(\gamma+\pi) = (-1)^{d_1+d_2} g(\gamma). \tag{6.41}$$

For odd orders d this results in $g(\gamma+\pi) = g(\gamma)$ and $g(\gamma+\pi) = -g(\gamma)$ for even orders d. Either the data in the positive and negative direction are mapped onto the positive direction in feature space or the negative direction of the data is mapped onto the negative direction in feature space. To keep the distributions in feature space symmetric for even orders d, one can employ the scaling function f to map negative directions in data space onto negative directions in feature space (by, for example, the dot product $f(x) = |x|x$).

Correlations Introduced by the Transformation

It should be apparent right now that the transformation $\mathcal{F}$ of order 2 is up to the scaling identical to our symmetry transformation $\mathcal{S}$ expressed as a quadratic form in Equation 4.10 on page 90.

It remains to show how the transformation $\mathcal{S}$ behaves with respect to correlations that are introduced if the data is projected into feature space. We generated a data cloud from a circular two-dimensional Gaussian density and projected the data into the three-dimensional feature space defined by $\mathcal{S}(\mathbf{x}) = |\mathbf{x}|\mathbf{x}(x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$. Because the data was generated in input space

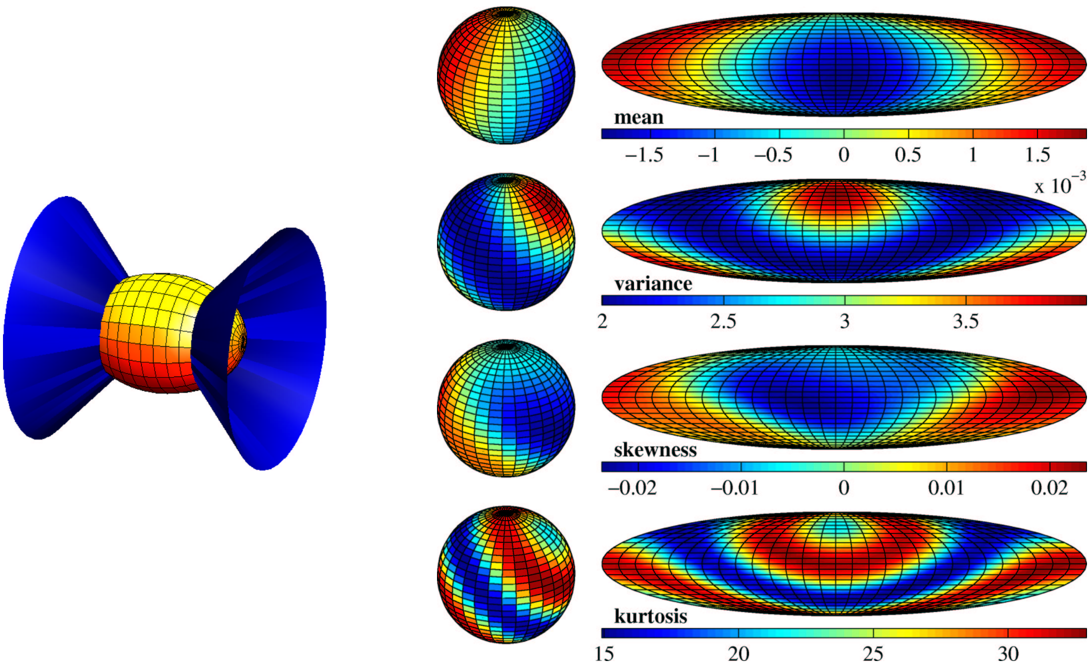


Figure 6.12.: Correlations introduced by the transformation of a circular Gaussian density. *Left:* Ellipse representing the iso-probability lines of a fitted Gaussian function to the data in feature space (Diabolo space indicated by manifold in blue). The data distribution is elongated in the direction of the cones. *Right:* The one-dimensional cumulants of the data in feature space projected onto vectors from the unit sphere

Table 6.1.: Table summarizing the properties of different feature spaces

	polynomials	RBF	v. Mises	monomials
invertible projection	√	√	√	√
linear decomposition	–	√	√	√
linear indep. of basis func.		√	√	√
differentiable basis func.		–	√	√
general linear independence		–	–	√
factorizing distributions remain		–	–	√

with variance one in all directions we can interpret directions of high variance in the data as directions in which new correlations are introduced because of the transformation.

In Figure 6.12 on the page before we analysed the above case of the circular Gaussian distribution in two dimensions after projection into the three-dimensional feature space. In the left figure, in blue the diabolo indicates the manifold in which the data is projected. Additionally we computed in the feature space the eigenvalue decomposition of the data. The resulting two-dimensional surface representing the iso-variance of the data is elongated in the direction of the mid-line of the two cones. This indicates that correlations are introduced into the data. Note, that this is a problem for all data which are not perfectly super-Gauss (sparse). Additionally we plotted in Figure 6.12 right, on the preceding page the first four cumulants of the data after it is projected onto linear sub-spaces to make explicit which information is used by the ICA algorithm. We sampled all linear sub-spaces in three-dimensional space (points on a the surface of a unit ball), projected each data point into that space (by a dot product) and indicated the cumulants of resulting distributions by a color code. To better visualize the structure of the cumulants dependent on the axes of projection we applied a Mollweide map transformation and obtained the ellipsoidal shaped two-dimensional mappings shown right. As already indicated by the eigen-ellipse left, the variance of the data is highest in the direction of the cones of the Diabolo. Notably, the kurtosis as an important measure of the higher order moments is found to be high in the sub-space defined by the manifold of the data in feature space.

6.4.3. Conclusion

As we have seen from the analysed transformations only the one by monomials of constant order fulfill the necessary properties proposed on page 131. In Table 6.1 on the facing page we have summarized the findings for the different transformations. It is obvious that the list does not cover all possible non-linear transformations. Our particular choice was motivated by the idea to illustrate the process of model selection, how to decide which transformation can be selected to effectively construct a non-linear feature space. Nevertheless, the properties obtained point to a single class of feature spaces obtained from a non-linear transformation based on monomials of constant order. Linear methods can be applied in this space. Note, that particular transformations from this set can be selected by defining the order of the monomials (which will depend on the number of sources we want to detect). Feature spaces are equivalent if they are built from different rotations of a single feature space expressed by an orthonormal matrix which is multiplied onto the vector of basis functions (this only affects the orientation of the diabolo in feature space).

6.5. Interpretation of the Directions in Diabolo Space

For solving an overcomplete dictionary problem by the means of the space expansion by commutative monomials of constant order the following steps have to be performed in sequence:

1. project the data $\mathrm{x}(t)$ into the space of monomial of constant order $d > 1$
2. apply a method of your choice (e.g., FastICA) in the feature space to obtain the components
3. back-project found components one-by-one into the input space.

For most projection methods it is better to visualize the results in the input space in order to judge the quality of the method. The question is now: How to project a direction, i.e., a row of the estimated mixing matrix found in feature space into a direction in input space.

Recovering of Source Directions from within the Manifold

Here we show that a source direction found by some method in feature space of monomial of constant order $d = 2$ can be used to obtain the corresponding source direction in input space.

Given that a direction $\tilde{\mathrm{s}}^i$ has been found in feature space which is a tangent vector *in* the manifold it corresponds to a direction x in the input space. From

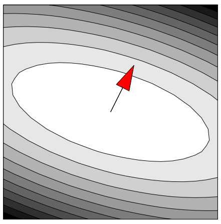


Figure 6.13.: Lines of constant Δ^2 are hyper-ellipsoids. The principal axes of hyper-ellipsoid are defined by Σ^{-1}'s eigenvectors u_i and eigenvalues λ_i as $s_i = \lambda_i^{-1/2} u_i$. Large λ_i result in small variance of $\Delta^2(x, y)$ thus a large gradient $\partial\Delta^2/\partial u_i$. For changes in the direction of $u_{\max}$ (red arrow) the neuron is most sensitive

Equation 6.32 on page 138 it follows that $\tilde{\mathbf{s}}^i = \mathrm{vech}(\mathbf{x}\mathbf{x}^T)$ for some unknown $\mathbf{x}$. Because of the symmetry of the matrix $\mathbf{x}\mathbf{x}^T$ we can construct from the source direction $\tilde{\mathbf{s}}^i$ the matrix $\mathcal{C}^i = \mathbf{x}\mathbf{x}^T$. It is easy to show that for $\sum_i x_i^2 > 0$ this matrix has a single non-zero eigenvalue with a corresponding eigenvector $\mathbf{u}_0 = \mathbf{x}$. Let λ_0 be a non-zero eigenvalue and $\mathbf{u}_0$ the corresponding eigenvector. We arrive at

$$\mathbf{x}\mathbf{x}^T\mathbf{u}_0 = \lambda_0\mathbf{u}_0, \tag{6.42}$$

which states that $\mathbf{u}_0$ is changed only in length by the multiplication with the matrix $\mathbf{x}\mathbf{x}^T$. Setting $\mathbf{u}_0 = \mathbf{x}$ we see that

$$\mathbf{x}\,\underbrace{\mathbf{x}^T\mathbf{x}}_{=\sum_i x_i^2} = \lambda_0\mathbf{x} \tag{6.43}$$

$$\sum_i (x_i^2)\ \mathbf{x} = \lambda_0\mathbf{x} \tag{6.44}$$

Thus for $\lambda_0 = \sum_i x_i^2$ the vector $\mathbf{u}_0 = \mathbf{x}$ is an eigenvector of $\mathcal{C}$.

$\mathcal{C}$ can be interpreted as the correlation matrix of a data distribution $p(y)$ which is restricted to a linear sub-space (a direction) in the input space. This implies that λ_0 is the only non-zero eigenvalue of $\mathcal{C}$ because for a linear sub-space C has rank 1. We can think of $\mathcal{C}$ as being a correlation matrix of some random vector $\mathbf{y} = \alpha\mathbf{x}$, $\alpha \in \mathbb{R}, \mathbf{x} = \mathrm{const}$, $E(\alpha^2) = 1$

$$E\left(\alpha_i\mathbf{x}(\alpha_i\mathbf{x})^T\right)_\alpha = E\left(\alpha_i^2\right)_\alpha \mathbf{x}\mathbf{x}^T = \mathbf{x}\mathbf{x}^T = \mathcal{C} \tag{6.45}$$

that is of data contained in a one dimensional sub-space in the direction $\mathbf{x} \in \mathbb{R}^n$. This directly indicates that the rank of $\mathcal{C}$ has to be 1 and all other eigenvalues $\lambda_{1,\ldots,n}$ are zero with corresponding eigenvectors $(x_1, 0, \ldots, x_{j\neq 1}, 0, \ldots, 0)$.

Multivariate Gaussians and Model Sensitivity

Here we show how the directions found by some method in the feature space of monomials of constant order can be interpreted as the response of the model that reflects its sensitivity to certain features of the input. The difference to the above back-projection technique is that we can now give an interpretation of directions which are not directly on the manifold.

In the case of commutative monomials of constant order 2 in two variables x_1, x_2 the model of a quadratic polynomial $\Delta^2 = a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_2^2 + a_4 x_1 + a_5 x_2 + c$ reduces to

$$\Delta^2 \;=\; a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_2^2 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \underbrace{\begin{pmatrix} a_1 & \frac{1}{2}a_2 \\ \frac{1}{2}a_2 & a_3 \end{pmatrix}}_{=:\Sigma^{-1}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad (6.46)$$

This reveals the structure of a 2-dimensional multivariate Gaussian distribution. In the case of more than two variables $x_1, x_2, \ldots, x_n$ the multivariate Gaussian is of dimension n. We have abbreviated this matrix as Σ^{-1} to make the connection to the distance measure called Mahalanobis distance. It defines the distance of a data point $\mathbf{x}$ to the center of a Gaussian function incorporating the quadratic distortion of the space expressed in the matrix Σ.

Mahalanobis distance

Its contours are defined by curves of constant density Δ^2. The equipotential lines of constant Δ^2 are hyper-ellipsoids[9]. The principal axes of the hyper-ellipsoids are given by the eigenvectors $\mathbf{u}_i$ of Σ which satisfy

$$\Sigma \mathbf{u}_i \;=\; \lambda_i \mathbf{u}_i \qquad (6.47)$$

and the corresponding eigenvalues λ_i give the variances along the respective principal axes. Immediately it becomes clear that in calculating the eigenvalues of the matrix Σ^{-1} we got the solution is $\Sigma^{-1}\mathbf{u}_i = \lambda_i^{-1}\mathbf{u}_i$ and thus the same eigenvectors are solutions but with reciprocal eigenvalues. The directions of largest λ_i point therefore in directions of smallest variance (see Figure 6.13 on the facing page). Drawing the corresponding eigenvector we look into a direction in which the data can be changed without much changing the response of the model, i.e., the model P is insensitive to that pattern. This idea of back-projecting directions in feature space by computing its most sensitive plus its most insensitive direction is similar to the one used by the model in Wiskott and Sejnowski (2002). There a polynomial function of second order is learned from successive drawn image patches undergoing transformations.

Sources for Errors

Because of the specific form of the transformation we can go one step further. We are interested in directions in the data, that is in points that are on the man-

[9] For $\mathrm{sign}(a_1) = \mathrm{sign}(a_3)$ it is an ellipsoid.

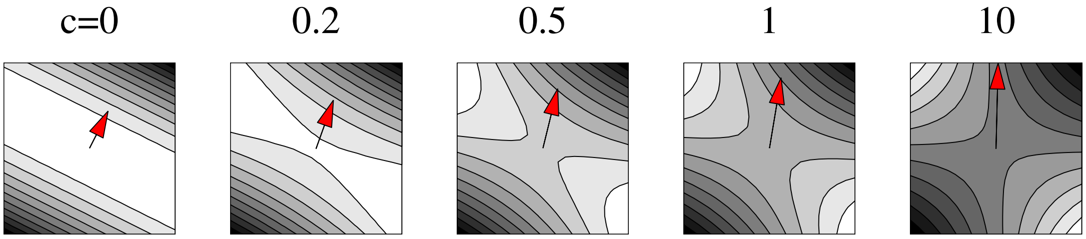


Figure 6.14.: Deviation from optimal solution (largest eigenvector by red arrow). In two dimensions the optimal solution (direction in diabolo space) is of the form $(\alpha^2, 2\alpha\beta + c, \beta^2), c = 0$ because that defines a direction that corresponds to a direction in input space. If $c \geq 0$ the solution is no longer elliptical

ifold defined by the feature space transformation. Because the ICA algorithm is applied in the feature space without using this knowledge it sometimes comes up with directions that are not in the manifold. In Figure 6.14 the effect of a wrongly found direction in feature space is shown for the solution in input space. There are several reasons for problems of this kind. One would be that the algorithm has not enough data to learn the correct source directions in the high dimensional input space. Another that assumptions about the source distributions are wrong (e.g., the linearity of the mixture).

The whitening used by many ICA methods may be a source for a methodical error. Whitening is performed in order to de-correlate the data (in feature space). Lets assume that the data is uncorrelated in input space. For the case of highly sparse sources (super-Gaussian) the projected data is also uncorrelated. But if the data is not that sparse or has a large noise component the transformation into the feature space will introduce additional correlations (see Figure 6.12 on page 141). This happens because the data is restricted to a manifold which does not cover the space between the sources equally. The whitening procedure in the feature space will attempt to undo these faulty correlations thus correlating the sources. As we can see for the working examples presented in the next section this may not be a practical problem.

Nevertheless, we can utilize this as a source of additional[10] information about the quality of our estimated sources. If we are only interested in components that correspond to a linear source direction in input space we would prefer sources $\tilde{\mathrm{s}}_i$ with rank one in the corresponding $\mathcal{C}_i$. The ratio near one of the largest eigenvalue to the sum of all other eigenvalues indicates a 'trustworthy' source direction.

[10] Information about the quality of a source is also contained in the amount of power in the signal explained by the source.

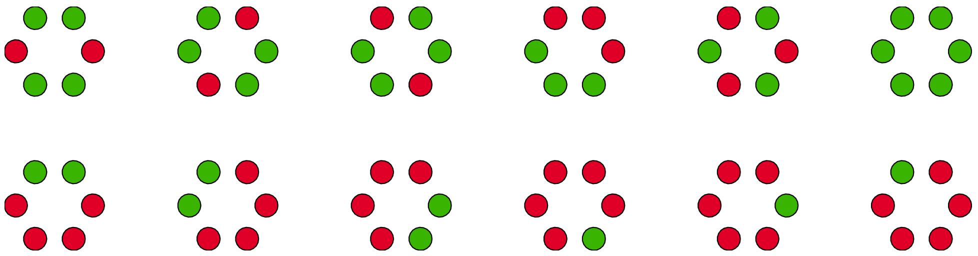


Figure 6.15.: The 12 columns of the mixing matrix $w_{.,i=1\dots12}$ represented as patterns associated with each source s_i

6.6. Applications

6.6.1. Toy-Example: 6 Observations 12 Sources

We now solve a two times overcomplete ICA problem with FastICA in the space of monomial of constant order.

Given are 12 statistically independent and sparsely distributed sources $\mathrm{s}_{1\dots12}$ each drawn according to the density function $p(s) = \exp(-s^2/2)^3$. The observations are 6 dimensional therefore the model generating the data is defined using a non-square (6×12) mixing matrix W:

$$\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots\cdots & w_{1,12} \\ w_{2,1} & w_{2,2} & \cdots\cdots & w_{2,12} \\ w_{3,1} & w_{3,2} & \cdots\cdots & w_{3,12} \\ w_{4,1} & w_{4,2} & \cdots\cdots & w_{4,12} \\ w_{5,1} & w_{5,2} & \cdots\cdots & w_{5,12} \\ w_{6,1} & w_{6,2} & \cdots\cdots & w_{6,12} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ \vdots \\ s_{12} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} . \tag{6.48}$$

An intuitive interpretation of the columns of the mixing matrix was given already in our section about overcomplete ICA on page 119. They contain the 'spatial' pattern (sometimes called maps) that can appear. As one can see in the above equation each column i is in product with only the corresponding s_i. If s_i is sparsely distributed also the pattern in the column i appear sparsely in the mixture (which is obtained by adding all the weighted columns). A pattern appears sparse if it is mostly inactive but sometimes it is expressed strongly.

For our toy problem we designed the mixing matrix in order to contain binary values only. This restricts us to 2^6 pairwise different pattern $\mathbf{p}_i$, if we count the pattern $\mathbf{p}_i = -\mathbf{p}_i$ as one $2^6/2 = 32$ patterns can be constructed. This

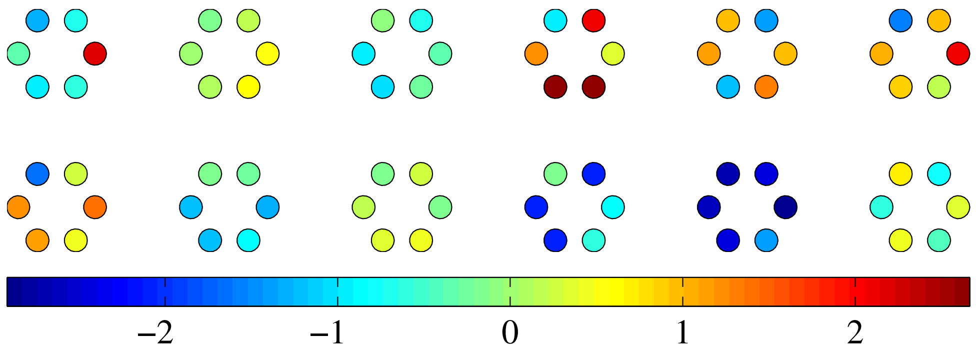


Figure 6.16.: Some of the mixtures

can be done because the ICA algorithm we are using is ambiguous in the sign of the solution. Notably this gives an upper bound for the number of independent *binary* patterns which is nevertheless ≈ 5 times larger as the number of pattern that can be extracted with a square ICA algorithm. We selected from this set (by hand) 12 pairwise different vectors based on good visual identification. The 12 pattern $w_{1\ldots 6,i}$ are shown in Figure 6.15 on the page before each represented by 6 color coded discs.

Some of the observations (mixtures) are displayed in Figure 6.16. Given many observations $\mathbf{x}(t)$ the task is now to find the pattern representing the hidden states of the system (the unknown columns of the mixing matrix).

Applying FastICA to Overcomplete Dictionary Problems

We now want to apply ICA to an overcomplete dictionary problem. Because for most real world applications we do not know the correct number of sources it is worthwhile to examine the algorithm in a case where the number of sources is twice as large as the number of observations. The hope is that the algorithm is, for example, able to find some of the correct sources, or in each run a different set. Both is in general not true.

To demonstrate this we applied the FastICA method (Hyvärinen, 1999) to the data. We chose the symmetric approach and as a non-linearity $g(u) = u^3$ which matches the density of the sources generating function. That is, the model fits the data in terms of being an additive linear mixture with the correct type of densities. Only the number of sources is not correct.

In Figure 6.17 the 6 found sources are shown. The algorithm converges to components in which mostly mixtures of the true sources are found. The problem is especially hard for the ICA algorithm because the frequency of a pattern to appear is the same for all 12 pattern and also the amplitude for all pattern was the same.

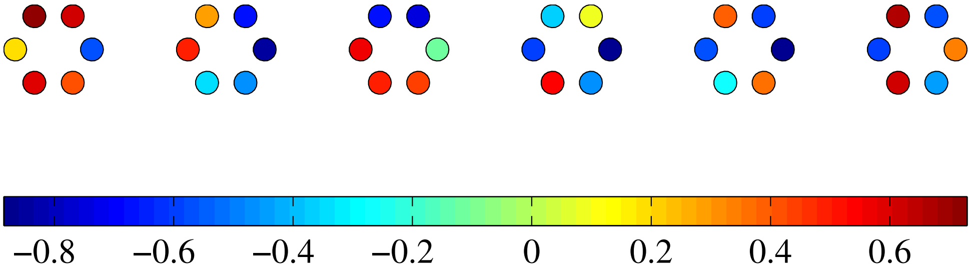


Figure 6.17.: Result of FastICA on the overcomplete mixtures

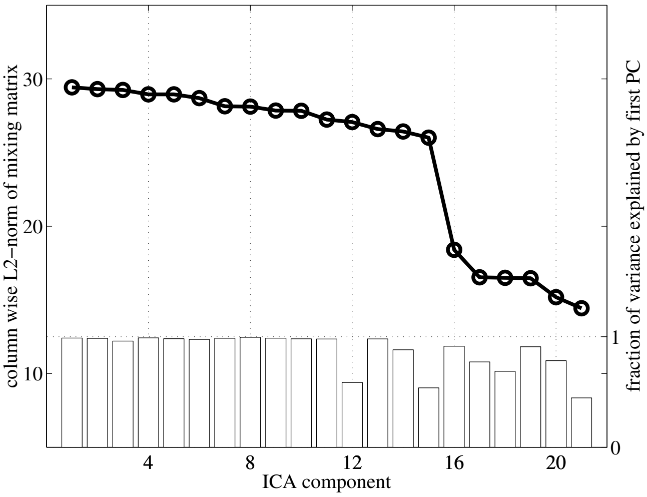


Figure 6.18.: $L2$-norm of the columns of the calculated mixing matrix. The bar plot displays a measure of the linearity of the found source. It is computed as the fraction of variance that is explained by the first eigenvector displayed in Figure 6.20

Applying ICA in Diabolo Space

Using the transformation in the space of monomials of constant order we represent our data in a 21-dimensional space which is built from the commutative monomials of order 2.

$$\begin{aligned}\mathbf{x} = (x_1, \ldots, x_6) \stackrel{\mathcal{FS}}{\longrightarrow} & (x_1^2, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_1x_6, \\ & x_2^2, x_2x_3, x_2x_4, x_2x_5, x_2x_6, \\ & x_3^2, x_3x_4, x_3x_5, x_3x_6, \\ & x_4^2, x_4x_5, x_4x_6, \\ & x_5^2, x_5x_6, x_6^2) = \mathbf{p} \qquad (6.49)\end{aligned}$$

To enforce the mean of the data to be zero half the data points were switched in sign. We also corrected each term of the form $x_ix_j, i \neq j$ by a factor of $\sqrt{2}$ to obtain rotational invariance (see Section 27 on page 137).

Linear ICA was performed on this data by the FastICA package in Matlab. Using the symmetric approach and the default non-linearity $g(u) = u^3$ the algorithms converged after 20 steps. The resulting independent components were sorted according to a decreasing norm of their corresponding row in the estimated de-mixing matrix (see Figure 6.18). This sorting preserves the relative importance of the found sources by selecting sources that contribute most to the power of the signal (Lee, Jung, Lee and Lee, 2000).

In Figure 6.19 on the facing page the first 12 components are shown. Because each ICA component is represented by a vector in feature space (see Equation 6.49) we display its entries as colored lines connecting the respective pixel positions for each dimension. In this way the second entry of an ICA component, for example is drawn as a line connecting the points x_1 and x_2. The color of the line indicates its respective value in the component. The result is best read using the interpretation of a weighting of a feature space dimension as a correlation of the corresponding two points. Similar colors represent similar fate of the points. Because the ICA detects only directions subject to a change in the sign of the component we can negate each component. All points connected by red lines in the first component can therefore assumed to have high correlations between their values whereas the top left point is different. This corresponds to the original source number 12 shown in Figure 6.19 on the next page. By this reasoning one can confirm that all 12 sources are found as the first 12 components.

The graphs in Figure 6.19 are hard to read but they contain the full information in the found directions. If we assume that the underlying sources are independent and the mixing is linear we can apply the the back-projection method presented in Section 6.5 on page 143 to visualize the found sources in input space. Constructing for each component the equivalent quadratic form

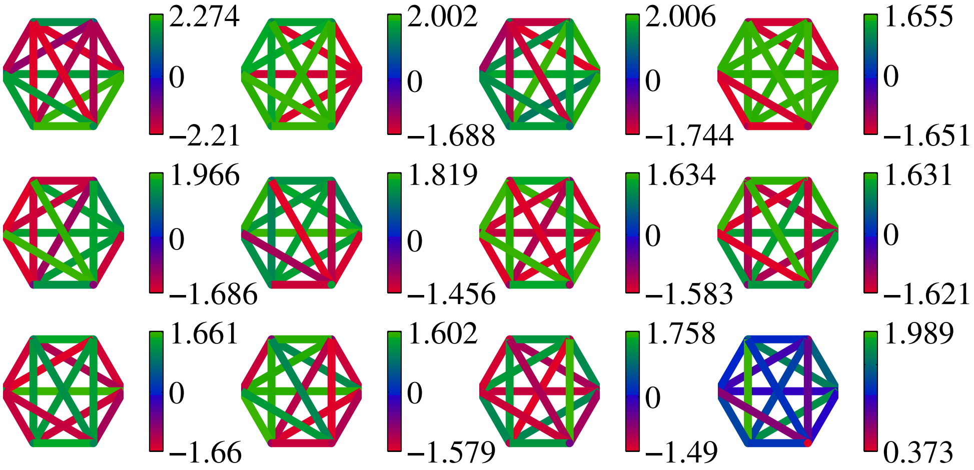


Figure 6.19.: Result of FastICA in Diabolo-space. The independent components are sorted according to the row norm of the calculated mixing matrix. It turns out that the first 7 components correspond to the original sources

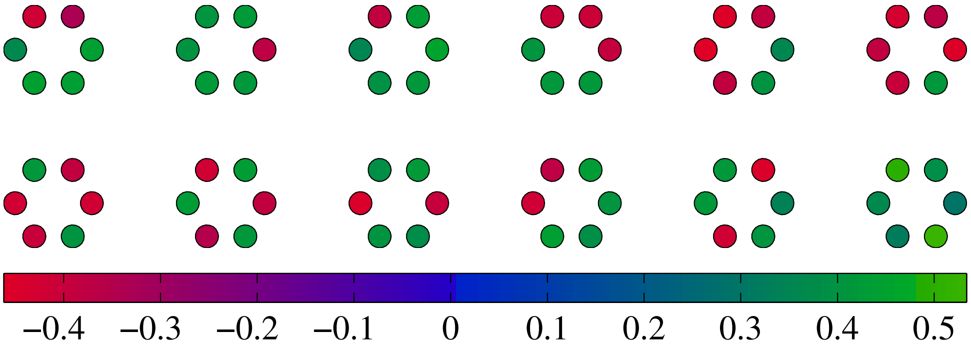


Figure 6.20.: Result of FastICA in Diabolo space. For the first 12 components the first eigenvector of the corresponding back-projection shows that all sources are correctly estimated. The components were sorted according to the $L2$-norm of the columns of the mixing matrix

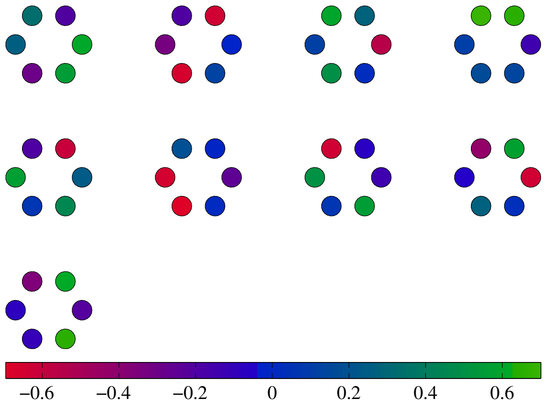


Figure 6.21.: ICA components $13 \dots 21$ are shown by their first principal component. Some of the eigenvectors explain only very few of the overall variance in the ICA component thus suggesting a direction found by ICA that is not a direction in input space

we display in Figure 6.20 on the page before its first eigenvector. A comparison with the true sources in Figure 6.15 on page 147 confirms that the overcomplete ICA problem was solved successfully.

We analysed also the remaining 9 sources. Especially the components $13, 14$ and 15 are of interest because they have similar explanation power as suggested by Figure 6.18 on page 149. It turns out that they can be interpreted as correcting for small mismatches in the first components to the true sources (Figure 6.21).

6.6.2. Application to Natural Images

We applied the algorithm of the centralized mixture of Gaussian functions to an ensemble of pre-whitened natural images obtained from the homepage of Bruno Olshausen (Olshausen and Field, 1996). 128 components where learned from $30,000$ image patches (64-dimensional). This results in a two times overcomplete dictionary. After 30 iterations of the EM algorithm for each cluster the direction of largest variance (vector of dimension 1×64) is shown in Figure 6.22. We assume here that the direction of largest variance corresponds to a source direction in the data (see Figure 6.3 on page 125). The obtained spatial filters resemble mostly localized Gabor like receptive fields but also a whole spectrum of curved edges.

Notably, the algorithm converges very quickly and the basic structure of the components is fixed after around 10 iterations. This is surprising, because

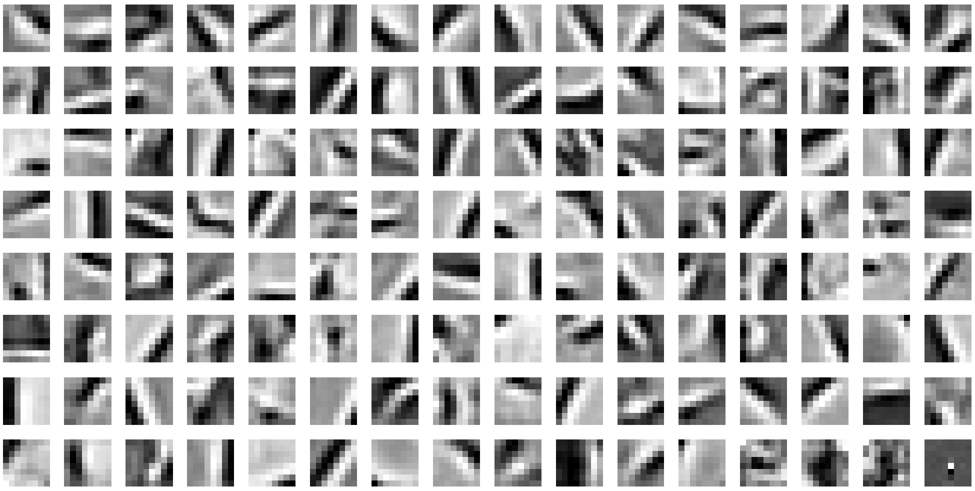


Figure 6.22.: Source directions learned by a centralized mixture of Gaussians (results were sorted left to right, top to bottom by decreasing sum of eigenvalues). Each image displays the eigenvector with largest eigenvalue and indicates a found source direction. In Figure 6.26 two more eigenvectors for the 16 mixtures in the first row are shown

of the large number of parameters in the model (we trained a full covariance matrix). We suspect that the nice convergence behavior is obtained because of the restriction of zero mean Gaussian distributions. The original Gaussian mixture model trained with EM in a first phase has to adapt the means of the components before fine-tuning the covariance matrices. By fixing the means the covariance matrices can be optimized beginning with the first iteration. Of course we cannot rule out the possibility of a plateau in the energy function (to balance this problem we trained the model for 30 generations).

For displaying purposes the components are sorted according to the sum of their respective eigenvalues. Thus we prefer 'large' Gaussians over ones that converge to the single data point at the origin. The singularity was prevented by the regularization described above. In Figure 6.23 the actual eigenvalues for all eigenvectors are plotted (note, that we plot the log of the data to elevate the exponential nature of the eigenvalue spectrum).

Indeed, we find that the eigenvector with the smallest sum of the eigenvalues has a flat eigenvalue spectrum. Also the next $2-3$ eigenvalues correspond to nearly flat spectra. Therefore an interpretation of the structure in the respective direction of the eigenvector with the largest eigenvalue is difficult (last $3-4$ source directions of Figure 6.22). The other source directions can

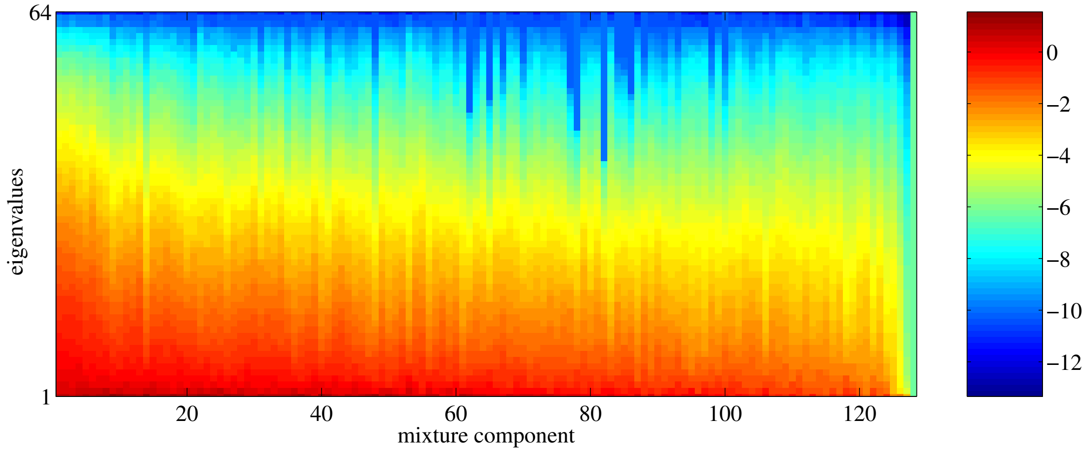


Figure 6.23.: A plot of the log eigenvalues corresponding to each of the first eigenvectors displayed in Figure 6.22. Only the last $3-4$ mixture components show a relative flat eigenvalue spectrum thus can be regarded as coding for the center distribution only (with vanishing covariance, but note the regularization by $\beta = 0.00001$)

be described as containing mostly edges but sometimes also curved- and more complex structures are found. Notably the edge pattern are not as localized as the edge pattern found, for example, in the model by Olshausen and Field (1996). We suspect that this is due to the regularization which counteracts the sparsest solution (which would correspond to singular Σ^i's with rank one).

At first glance the receptive fields appear to be localized in space. This is confirmed by calculating the mean correlation over the distance of pixel pairs (see Figure 6.25). Pixel pairs with small lateral displacement have predominantly high correlations. It is also interesting that the algorithms select very few negative correlations between pixel. But this conclusion may be misleading because the number of pixel pairs is not distributed equally over the distance between pairs (see the histogram of the number of pixel pairs per displacement in Figure 6.25 on page 157 right).

Because each component is learned as a multivariate Gaussian the description of one component by its direction of highest variance is insufficient. The density estimate obtained by the model assumed a full covariance matrix. To further analyse the found components we plotted in Figure 6.26 two more of the eigenvectors for the first 16 (first row in Figure 6.22) components. Most eigenvectors resemble stimuli with similar orientation and location but changes in phase and spatial frequency. The percentage values on top of the Figures represent the variance explained by the eigenvector and thus can be

Table 6.2.: Lifetime response kurtosis as given by Willmore and Tolhurst (2001) for pseudo-whitened images together with the values obtained for the learned codes (GEM) in Figure 6.22. The variances for the GEM codes are found to be ± 3.48 for the lifetime and ± 0.21 for the population response kurtosis

	Lifetime response kurtosis	Population response kurtosis
Gaussian	8.93	0.52
DoG	11.20	1.74
Olshausen	17.21	2.17
PCA	8.13	3.07
Walsh	10.91	4.01
Sinusoid	12.05	4.62
Gabor	18.47	5.37
GEM	8.35	8.83

interpreted as a relative weighting of the pattern. Neurons with fast decreasing eigenvalues can be assumed to respond to the stimulus in a linear way because they can be sufficiently described by the their first eigenvector only. Neurons with similar eigenvalues resemble more the complex type of cell. This indicates the remarkable property of the second-order neuron model to model independently from one another the response characteristics of simple and complex cells. Notably in the model we find complex cells which are not composed of simple cells.

Further eigenvectors with decreasing variances show the usual high frequency checker-board pattern (data not shown).

To evaluate the quality of the found sources we analysed the sparseness of the code. This is done in order to compare the non-linear filter with previously obtained filters by linear methods (e.g., ICA, PCA, Walsh, DoG, Gaussian, Olshausen and Fields, ...).

Sparseness in the Representation

We now analyse the code in terms of its sparseness . Let $p(\mathbf{x})$ describe the relative frequency of the occurrence of specific responses in our model. One source i is parameterized by a multivariate distribution A^i. The response of the source i to some data $\mathbf{x}$ is given by the length of that vector in a space distorted by A^i. This measure (or metric) is given by the Mahalanobis distance y between a vector $\mathbf{x}$ and a multivariate distribution described by its variance-

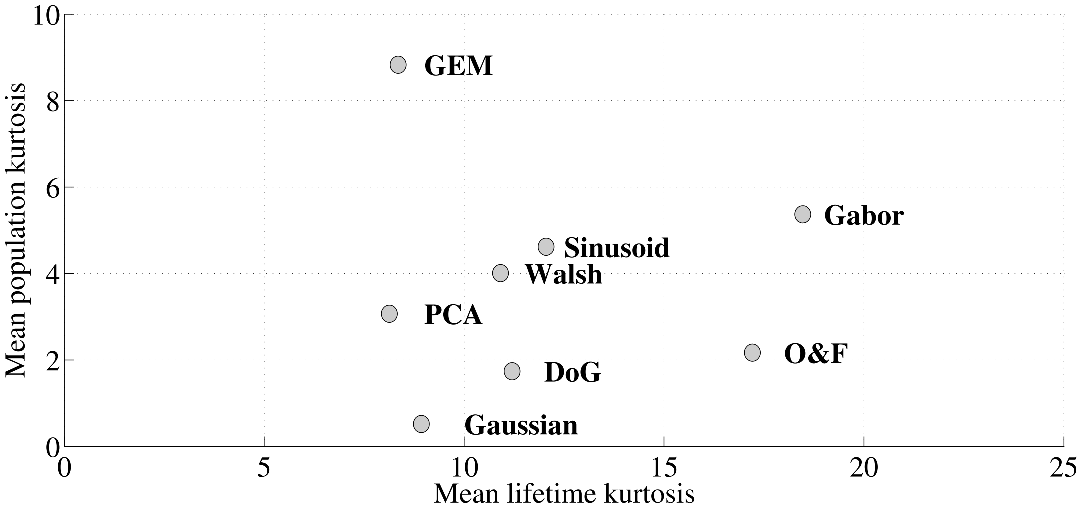


Figure 6.24.: Scatter plot of the data from table 6.2. The code obtained by centralized Gaussian mixture model trained by EM (GEM) outperforms the linear codes with respect to population sparseness and is comparable to the PCA code in its lifetime sparseness

covariance matrix A:

$$y = \mathbf{x}^T A^{-1} \mathbf{x}. \tag{6.50}$$

The inverse is used because the quadratic form $\mathbf{x}^T A \mathbf{x}$ only describes the variance along the direction $\mathbf{x}$. Directions with small variances correspond to directions with a steep gradient $\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial \mathbf{x}}$, thus the distance to a data point is large. The inverse of A solves the problem because it inverts the eigenvalues of A (the eigenvectors stay the same, see Section 6.5 on page 145).

If A is a covariance matrix of some data its eigenvalues will be real and positive (or zero) and therefore also the response of the model unit to the data calculated by the quadratic form is ≥ 0. Every distribution we obtain is therefore asymmetric with respect to an activation of 0. In assuming that both $\mathbf{x}$ and $-\mathbf{x}$ are equally probable we symmetrize our distribution by

$$y_{\text{sym}} = \epsilon\, y, \qquad P(\epsilon = 1) = .5, P(\epsilon = -1) = .5 \tag{6.51}$$

(flipping the sign of the response of each unit randomly). This is reasonable because by the quadratic form we measure variances of a multivariate Gaussian in the direction $\mathbf{x}$ which is symmetric around its center.

There are two major ways of obtaining distributions that describe the units responses. One is the distribution of the response of one unit over many data

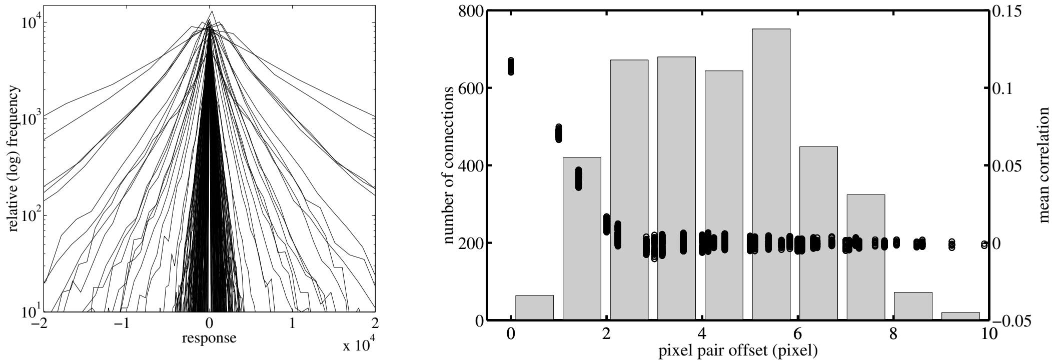


Figure 6.25.: *Left:* Distributions of the coefficients obtained from applying the 128 quadratic forms to the data (lifetime responses). Distributions are non-Gaussian with mean kurtosis of the data 8.35 (± 3.48). *Right:* Selected are predominately local correlations. Mean correlations over the distances between the respective pixel pairs. The histogram counts for 128 components the number of connections over the pixel pair distance

(lifetime measure), the other the distribution of the response of all units to a single image patch. Both can be used to describe the sparseness of the code by computing the kurtosis from the respective distribution. Usually the population response kurtosis is of higher importance because it describes a property of the whole population and not a avaraged property of single units. Willmore and Tolhurst (2001) found that both measures are largely uncorrelated. The first measure the *lifetime kurtosis* is usually defined by *life time kurtosis*

$$K_L = \left\{ \frac{1}{M} \sum_{i=1}^{M} \left[\frac{y^i - \bar{y}}{\sigma_y} \right]^4 \right\} - 3 \qquad (6.52)$$

where M is the number of image patches $(30,000)$ and y^i is the response of neuron i calculated by the quadratic form of Equation 6.51.

The *population kurtosis* is computed in the same manner but as a mean over the responses of the N neurons *population kurtosis*

$$K_P = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left[\frac{y^j - \bar{y}}{\sigma_y} \right]^4 \right\} - 3. \qquad (6.53)$$

In Table 6.2 the measurements of Willmore and Tolhurst (2001) are shwon together with the values obtained by our algorithm (last line). The lifetime kurtosis of the GEM code is comparable with the one observed for PCA. This

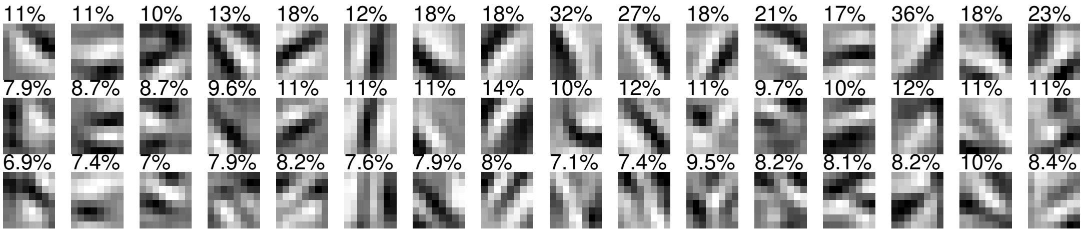


Figure 6.26.: Sub-space selected by each of the first 16 mixture components of Figure 6.22. Each time the three eigenvectors with the largest eigenvalues are shown. On top of each eigenvector the percentage indicates the variance explained by this component

simply reflects that each single neuron is parameterized by a single multivariate Gaussian distribution. In terms of population sparseness the mixture of Gaussian functions learned by the EM algorithms (GEM) outperforms all other codes as can also be seen from the scatter plot in Figure 6.24.

As a reason for the higher population kurtosis of Gabors, sinusoids, principle components, and Walsh functions Willmore and Tolhurst (2001) observed a large variability of the named codes in terms of representing spatial frequency. Since natural images are known to have amplitude spectra which are approximately proportional to $1/f$, there is a large amount of variance in the low-frequency Fourier coefficients. Thus the low-frequency filters can be expected to have larger response magnitudes than the high-frequency filters. As a result, the few low-frequency filters often produce responses that are large compared with the responses of the many high-frequency filters, and the resulting representation often have high population sparseness. This argument cannot fully be applied here because pseudo-whitened images were used for training and testing and pseudo-whitened images have had their amplitude spectra approximately flattened. So there is no longer a concentration of variance at low spatial frequencies. As we can observe in Figure 6.22 on page 153 our model indeed shows only a slight tendency to produce filters with varying spatial frequency.

Spread of Variance in the Representation

Another important factor is how evenly variance is spread amongst the population of coding units. Field (1994) discussed the idea that variance should be coded evenly amongst neurons (preference to evenly 'distributed' or 'dispersed' codes). In contrast to compact codes like PCA, a distributed code is more noise-insensitive because it does not depend on precise firing of some neurons. Based on the very low population response kurtosis variations in our model (± 0.21) the code found by GEM is indeed well distributed. This also

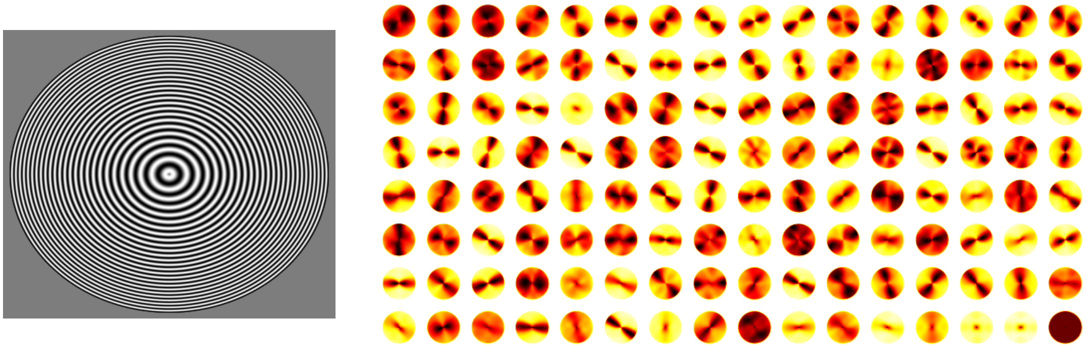


Figure 6.27.: *Left:* A pattern (400×400) used to test the receptive field properties orientation preference and spatial frequency. It consists of a circular sinusoid with increasing spatial frequency. *Right:* The response of all 128 neurons to the test pattern obtained by repeatedly applying the quadratic form to patches extracted from local positions

indicates that the population response distribution is non-Gaussian.

Taken together the Gaussian mixture model trained with EM produces a sparse and well distributed code which appears to be a useful strategy for encoding visual information.

Representation of Orientation Selectivity and Spatial Frequency

We analysed the quadratic form obtained from the Gaussian mixture model also in terms of their similarity to simple cells (that is how non-linear are the second order filters). This analysis ignores large parts of the structure of the receptive field of the neurons because it concentrates solemnly onto orientation preference and spatial frequency but in doing so it mimics the currently used methods for the analysis of cortical neurons.

We generated a spatial pattern resembling all orientations and a range of spatial frequencies (Figure 6.27 left). For each neuron i we successively extracted overlapping 8×8 image patches x from the circle pattern and computed the scalar valued quadratic form $\mathrm{x}^T C^i \mathrm{x}$. In Figure 6.27 right, the corresponding coefficients for each neuron are displayed in a color code, black indicating large responses of the neuron to this region in the pattern. The localized form of most solutions in angular- and radial direction indicates that a large part of the neurons resembles orientation selective neurons with a preferred spatial frequency (some neurons appear to be selective to two or three orientations in the stimulus).

We analysed also the linear response of each neuron so that it can be compared to the response obtained from the quadratic form. In Figure 6.28 the

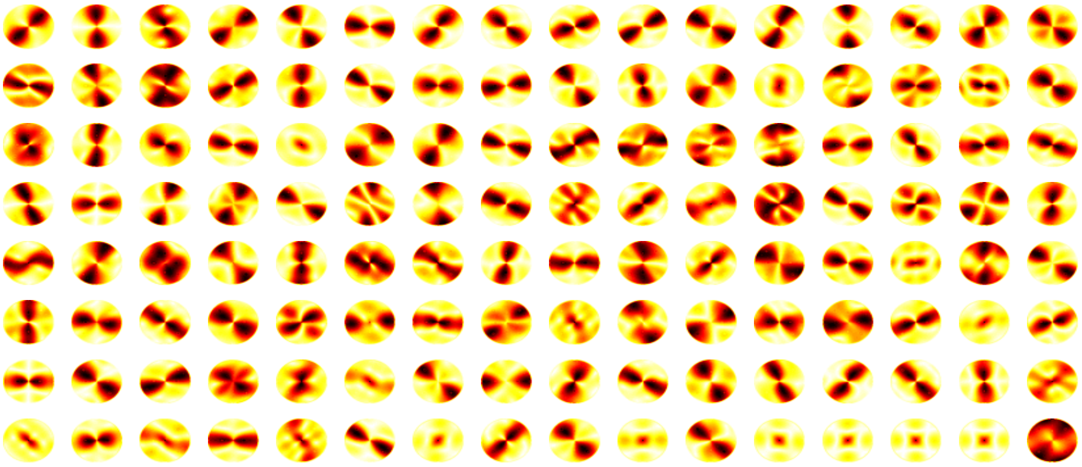


Figure 6.28.: The linear response part of the neurons obtained from spatial filtering of the first eigenvector with the test pattern in Figure 6.27 left. For visual presentation the results are smoothed to elevate the effects of interference pattern due to sub-sampling, the same smoothing was also applied to the images in Figure 6.27 right

linear part of the neuronal response is approximated by a linear filter defined from the largest eigenvector of the corresponding second-order neuron. Again this filter was applied to the test pattern in Figure 6.27 left. The degree of change between corresponding pattern in both figures can be used as an indicator for how well the neurons can be described as simple cells.

6.6.3. Application to Natural Images II

We applied the algorithm of FastICA in the space of monomials of constant order to the ensemble of images shown in Figure A.1 on page 175 (from the homepage of Patrick Hoyer). From these images we extracted $100,000$ circular patches with a diameter of 7 pixel. This was done in order to reduce the number of pixel, thus the number of pixel pairs and by this the number of dimensions in the feature space (≈ 700). Using the symmetric approach of FastICA we reduced the dimension of the data by using only the first 49 principal components in the whitened feature space. This results in 49 found independent components by the ICA procedure. For each component we displayed in Figure 6.30 on page 162 the eigenvectors to the two (absolut) largest eigenvalues. The ration between the eigenvalues is shown beneath each image pair. A negative value indicates that one eigenvalue is negative (we switched the

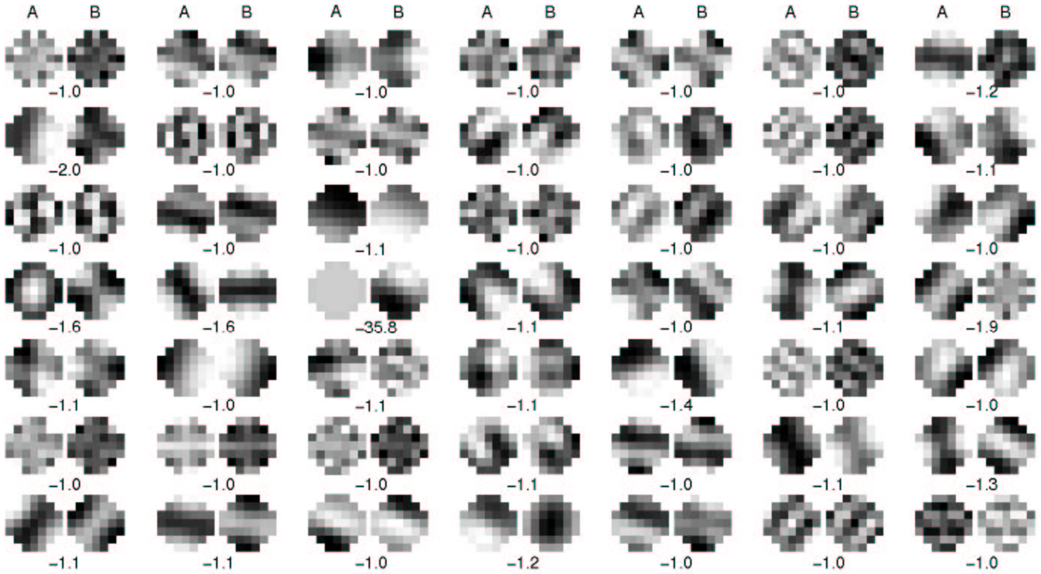


Figure 6.29.: Symmetry detectors from natural images learned by PCA. The first 49 principal components sorted by their eigenvalue are shown (explain 94% percent of the variance in the data). Each eigenvector is in dot-product space and represented by two images (rows *B, C*) representing the images maximizing or respectively minimizing the filter response (quadratic form)

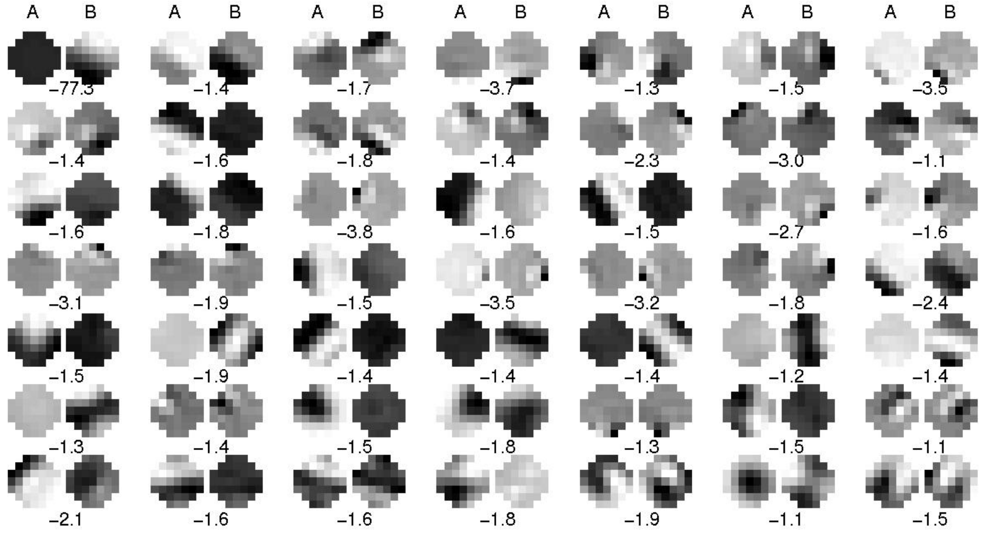


Figure 6.30.: Learned symmetry detectors by FastICA. For each of 49 found independent components two images (*A*, *B*) are shown. *A* is the eigenvector with the largest eigenvalue and maximizes the output of the filter. *B* is the eigenvector with the smallest eigenvalue and minimizes the output. The number below each independent component code for the ratio of the largest to the second largest eigenvalue, negative sign indicates that they have opposite sign. The (absolute) ratio can be used as an indicator for the linearity of the unit. (Figure from (Bartsch and Obermayer, 2003))

sign of the components so that the larger of the two has a positive sign). The pattern that corresponds to a negative eigenvalue will minimize our quadratic form thus we interpret the two patches for each components as antagonistic pairs describing pattern that either facilitate or inhibit the output of the unit.

As we already know, localized edge detectors form a set of independent sources for natural images (Olshausen and Field, 1996; Bell and Sejnowski, 1996). Differently from ours these models work with a representation of the sources in 'image-space' (linear models) rather than in a 'product-space' as our (non-linear) model.

Apparently edge like attributes are learned by the model. The edges are also, to some degree, localized in space. Further analysis of this model and comparisons of the properties of the units in terms of the sparseness are needed.

Two image pairs were analysed further. The first is the unit nr. 43 (in the

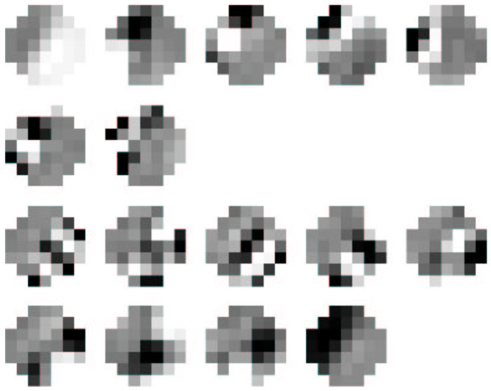


Figure 6.31.: A neuron selective to texture components in its sub-fields. All eigenvectors of the filter nr. 43 are orientation selective in the respective sub-fields. Neurons like this can be used to detect texture boundaries. (Figure from (Bartsch and Obermayer, 2003))

lower left corner). Both the most positive and the most negative eigenvector show the same basic structure, that of an edge. This is surprising because the corresponding neuron would be both activated (positive) and in-activated (negative) by an apparently similar pattern. To clarify this apparent contradiction we analysed some more of the eigenvectors of this unit.

The following eigenvectors depicted in Figure 6.31 indicate that the neuron responds differentiable to texture components in its on- and off-sub-fields. The units response as computed by the quadratic form is amplified for texture components in one sub-field and suppressed for texture components in the other sub-field. Because a change in texture is likely at object boundaries in images units of this type can be used to detect an important sub-population of edges.

Experiments in V1 show also that the activity levels near simple texture boundaries are increased $10-15ms$ after an initial cell response (Gallant et al., 1995).

Some of the components could not be sufficiently explained as edge detectors. Notably the components nr. 47 and nr. 49 are better described by a *yin-yang* type of pattern. Aware that we might interpret extensively we nevertheless assigned a function to these units. In Figure 6.32 on the next page we constructed examples for the orbit defined by the linearly weighted superposition of the respective two patterns. Approximately in the middle position both components resemble edge detectors. Therefore, an edge of the corresponding orientation will elicit a balanced amount of excitation and inhibition for these units. Only if the edge curvature is changed in the direction of either the left

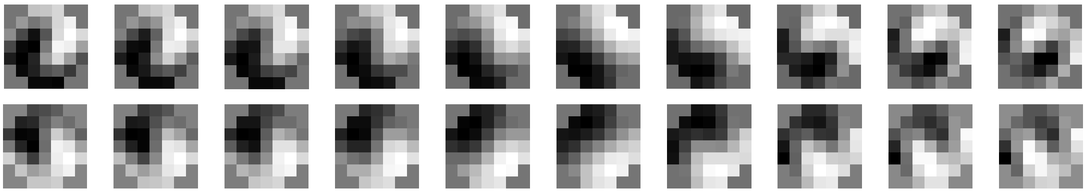


Figure 6.32.: The Yin-Yang type pattern of component nr. 49 and nr. 47 resemble bended edges thus can be regarded as coding edge curvature. For one row the pattern left shows the eigenvector $\mathbf{u}_0$ of the largest (positive) eigenvalue and the pattern right the smallest (negative) eigenvalues eigenvector $\mathbf{u}_{49}$. The pattern in between are obtained by the linear superposition according to $(1-\beta)\mathbf{u}_0 + \beta\mathbf{u}_{49}$, $\beta \in [0,1]$. (Figure from (Bartsch and Obermayer, 2003))

or the right pattern the respective unit will respond by being either facilitated or inhibited.

6.7. Biological Implementation by Dendritic Micro-circuits

Having shown that symmetry based on local spatial correlation is a 'useful' feature in terms of information processing, we now speculate how a neuronal module could implement this apparently abstract framework. Assuming a basically one-to-one (topographic) projection from the retinal space to the visual cortex, lateral connections in the cortex as well as converging projections from the LGN to the primary visual cortex are candidates for implementing structure detection at the level of V1. For now we only focus on the lateral connections.

Short-range lateral connections in contrast to the long–range steppy connections are known to be unspecific, connecting every neuron up to a certain range. Also the bar-shaped steppy-connections in the layers $4B$- upper $4C\alpha$ show only a slight tendency to connect similar orientations (Asi et al., in press). We speculate that these pattern can be thought of as an anatomical substrate for spatial information processing. We now point out how single neurons can detect local (spatial) correlation structure by micro-circuits implemented in their dendritic trees. This approach is tempting because of the availability of data about the morphology of single neurons from staining experiments. It is technically much more difficult to map the feed-forward connections to (complex) cells in visual cortex because of the large distance of the respective axons

Mel (1994) summarized some ideas of dendritic functions as:

- The spatially extended nature of a dendritic tree permits useful spatiotemporal interactions among active synapses.
- One dendritic tree can have multiple pseudo-independent processing sub-units.
- Nonlinear membrane mechanism appropriately deployed can allow the dendritic tree of a single neuron to act as a powerful multi-layer computational (e.g., logical) network.

For the 'biological' implementation of structure detection we need two ingredients, (*i*) a multiplicative comparison of many spatially extended input features and (*ii*) a summary of the result. This can be achieved by known properties of dendritic arbors: The delays from dendrites to soma are in the order of one membrane time constant (Agmon-Snir and Segev, 1993). In contrast, the local charging times on thin dendritic branches may be an order of magnitude faster than the membrane time constant τ. Therefore distal dendritic arbors may function more as coincidence detectors for local synaptic inputs whereas the function of the soma is more that of an integrator. By this functionality a dendritic tree can calculate a *sum of products*, i.e., a measure of structure. Which position in retinal coordinates is connected with which other position is in this framework defined by selective contacts on the dendritic tree.

coincidence detection

There are other models for biologically inspired multiplication. One example is multiplication based on coincidence detection (Bugmann, 1997). Neurons are sensitive to spike timing because the biological integration is leaky. If we suppose that for n inputs after $(n-1)$ spike increments the membrane potential is such that the $n-$th spike causes firing the probability of firing $P(n,\tau,\Delta t)$ is:

$$P(n,\tau,\Delta t) \;=\; \Delta t\tau^{n-1} n \prod_{i=1}^{n} f_i$$

where f_i is the firing rate of the i^{th} input. Hence the output firing rate is

$$f_{out} \;=\; \frac{P(n,\tau,\Delta t)}{\Delta t} = \tau^{n-1} n \prod_{i=1}^{n} f_i$$

where τ is the length of the time window.

Multiplication could also be generated by assuming that addition of voltages is accompanied by non-linearities that occur between spike activity and membrane potential. If (*i*) the synapse produces voltage from spike activity by a compressive non-linearity (logarithmic function) and (*ii*) the spike generation at the axon hillock is an accelerating process (exponential function) the output of the neuron would be the product of the inputs, $xy = \exp(\ln(x)+\ln(y))$ (see Payne and Peters (2002), page 376).

7. Summary and Outlook

Our understanding of the brain is still very fragmentary. Even in areas like early vision, where large amounts of data have been collected during the last decade reiteratively discoveries are being made that change our view on cortical information processing. Ideas about the function of cortical neurons are as long adequate as they comply to the known facts. If new experiments come up with conflicting data the models need to be modified to bring them back in line with the observations. For example, the basic description of cells in primary visual cortex as either simple or complex is known to be a *lie to children* that is a simplification for the sake of clarity. The cells are better defined by populating a continuum where we find a graded change from neurons acting solemnly as simple cells to neurons with complex cell responses. One the same line goes the explanation of complex cells as being build from converging simple cells with overlapping receptive fields[1].

During this thesis we first questioned the role of intra-cortical networks in the generation, sharpening, and modulation of orientation preference as one of the main features of primary cortical neurons. Step by step a model is derived and changed according to the subject under investigation. By modeling more explicit different populations of neurons in a single column we found that network effects can account for a large variety of phenomena like contrast invariant orientation tuning and contrast saturation.

The model was extended to predict response properties related to context effects that is to modulations of neuronal responses by stimuli applied outside the classical receptive field. First a full orientation hypercolumn and afterwards a system of two coupled orientation hypercolumns is used to show principle difficulties in having cross-orientation modulations by iso-orientation specific patchy connections.

Taking better into account the spatial layout of cortex we derive a model for analyzing the influence of local cortical connections on the activities of neurons in V1. A set of orientation columns is arranged according to a measured orientation map, and orientation columns are connected by local excitatory and slightly more distributed inhibitory fibers. We found that two opposite effects contribute to the observed contextual modulation; (i) local inhibition that is induced by a local change in input (leads to suppression), and (ii) dis-

[1] Some animals have complex cells but no simple cells.

inhibition. By changing the configuration of the stimulus different regions of the orientation map are activated. Changes in the local structure then define which is more prominent, suppression or facilitation.

In the second part of the thesis we analysed the input into the visual system. Starting from the observation that neurons in primary visual cortex already responde to a wide range of different stimuli we formulated the hypothesis that higher order features in spatial pattern can be described in terms of there intrinsic invariance and symmetry. A mathematical formulation of smooth local symmetry was given and led to the framework of polynomial functions.

In order to draw out and test its logical and empirical consequences we analysed the descriptive power of the model. We found that based on intuitive reasoning orthogonal basis functions for the detection of invariances to rotation, scaling and shifts could be defined. Also applications for object classification, image alignment, and landmark detection illustrate the principle advantage of structure analysis over methods of shape analysis.

One of the main points of this thesis is to introduce two new learning methods for high-order models. In the context of neuronal nets high-order models are known to have higher memory capacity (Poirazi and Mel, 2001) and to be computationally richer than linear or threshold units – just one can implement parity, exclusive-or, or lookup table functions. Furthermore, such models also better represent the operations of real neurons containing highly branched dendrites with voltage-dependent membrane conductances (Mel, 1994).

We showed that symmetry detection can be formulated as a linear model in the space of dot-products. The first presented algorithm for the second-order model extracted the underlying (sparsely represented) causes of the data by estimating the probability density of the data. The proposed algorithm is based on an algorithms for missing data, the Gaussian mixture model trained by EM. We showed how the estimated centralized densities can be used to extract a linear overcomplete basis set for natural images. The obtained non-linear code outperforms other known linear codes in being well distributed and having a high population sparseness.

Finally, based on geometrical considerations we have shown how correlations in the data can be learned by the means of linear methods. By this we extended the use of linear methods to an important group of non-linear transformations. In the context of independent component analysis the results indicate a distinguished transformation of the data into a feature space in which the independence assumption can be fulfilled for a set of overcomplete basis functions.

There are, of course, many open questions not been addressed in this thesis. In the following some ideas for further research are discussed, sometimes only briefly, sometimes more comprehensively.

▪ The proposed correlation space for statistical analysis of data in particular natural images was recently used in the context of classification (Schoelkopf, Simard, Smola and Vapnik, 1998) of handwritten characters. Here local correlations in the images where used to perform a linear classification task by the means of support vector machines. Decision boundaries are defined in the space of all possible products of d pixel by using a kernel $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d$. Notably, prior knowledge about the local correlation nature of natural images was incorporated by an image pyramidal sampling of the pixel pairs selecting more pixel pairs with short displacement. Our results in Figure 6.25 on page 157 also indicated that this strategy of reducing the dimensionality of the feature space is valid for our type of data. Intuitively this behavior may change for large receptive field sizes. The presence of objects in a visual scene implicitly defines long-range correlations in the range of the size of the object. Therefore it would be worthwhile to explicitly search for long-range correlations in images because they can be assumed to be rare events, i.e., be informative.

▪ It is attempting to perform the proposed algorithms for larger receptive field sizes which would comply with structure detection in higher visual areas. If we change the size of the receptive field we also change the scale on which structure is detected. Therefore different cortical areas parameterized by the mean size of the receptive fields of their neurons can independently 'work' on the same input. But do different areas have direct access to the visual input? For V2 and V3 it is known at least in cat, that these areas are also innervated by fibers coming from the LGN. For monkeys direct connections from the retina to visual areas apart from the connections to V1 are not known. Nevertheless, latency measurements have revealed higher latency for connections in one area compared to latencies between areas (Nowak and Bullier, 1997). Fast inter-area connections can therefore relay the visual input thus computations in the cortical areas can perform in quasi-parallel. An architecture that models the interplay between cortical areas based on these ideas could be analysed in terms of the interplay of latencies of responses related to intra-cortical connectivity and top-down connectivity from the higher visual cortical areas. This relates to attentional effects and in the context of vision to tasks of active vision.

▪ It is also interesting to follow the path of using even higher order correlations in the data. A model that incorporates third order correlations can be easily formulated using the mathematical background presented. By incorporating the fourth order tensor in the polynomial model we arrive at:

$$\mathcal{PPN}(\mathbf{x}) \quad = \quad X^{ijkl}\mathbf{x} + W^{ijk}\mathbf{x} + V\mathbf{x} + \mathbf{u}. \tag{7.1}$$

Again, reducing each polynomial model to its form of highest constant order monomials (three variables, $d = 3$) we can apply the linear methods in feature space. It is apparent that only the degree of the polynomial which is a free variable defines the form of highest constant order monomials. Another interesting extension of the method is connected with the notion of *canonical correlations*. There we want to find linear combinations of the variables which give us the maximum correlation between the combinations.

▪ If we assume the preferable solution to be on the manifold (that is, if we are interested in linear overcomplete ICA) we can constrain the search space of the algorithm to solutions *exactly* onto the manifold. This corresponds to a specific type of parameterized non-linear transformations in the input space. Basically we have to hold the property of linearity of source directions in feature space. First experiments indicate that models of this type can be learned by maximizing the entropy of the data distribution in feature space.

▪ Probability density functions as estimated in Appendix A.1 on page 174 are powerful tools that can be utilized, for example, to compare different sets of images or as a parameterization of generative models. The obtained form is appealing because it uses a very weak assumption about natural images.

▪ The computations performed by a cortical neuron may be expressed by the spatial layout of its connection pattern. We assume here that at least as much information is stored in the existence or non-existence of synaptic connections as in their strengths. Apart from the feed-forward pathway (converging LGN input produces an orientation bias) the lateral connection structure is an interesting candidate to explain the response property of cortical neurons.

To test implications of this we analysed the spatial structure imposed by our binary quadratic forms defined in Section 5.1 on page 95. The first two pattern (A^{rot} and A^{scal}) from Figure 5.1 on page 95 where selected for this purpose. We derived an additional connection pattern using a random subset of all possible connections (using 20% of all possible connections). Taking care of the hemi-retinal to V1 distortion[2] we computed density profiles indicating the mean density of connecting fibers (i.e., lines) between pixel at different distances to the center of the pattern. In Figure 7.1 on the facing page left, it can be seen that the profiles are notably different in all three cases. Especially the density near the center of the structure can be sufficiently used to distinguish between the artificial classes of neurons.

We used the method also to analyse lateral connection structures measured from macaque primary visual cortex. Lateral connections in this region exhibit a rich spatial structure in their intra-laminar connections. Short-range connections throughout the deeps of the cortex are found to be isotropic, con-

[2] The global retinotopic input-mapping from the hemi-retinal image to the primary visual cortex can be approximated by the function of complex variable supplied by (Schwartz, 1980): $G(z) := \log[(z + 0.333)/(z + 6.66)]$, where $z = x + iy$, $|z| \in [0, 1]$, $x > 0$.

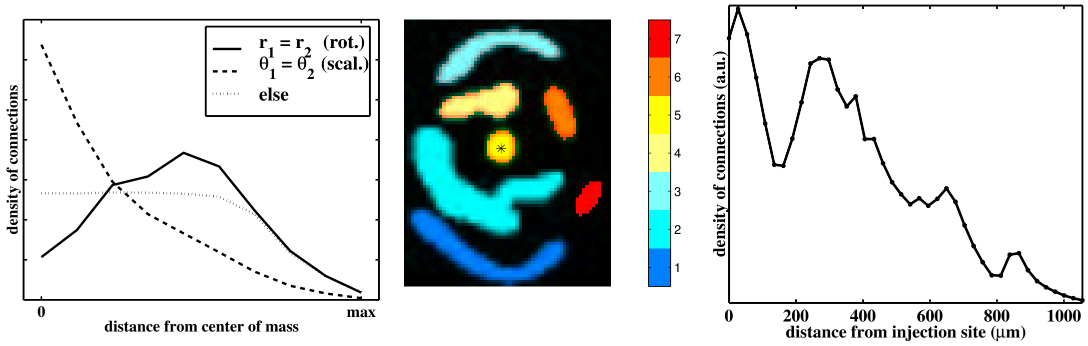


Figure 7.1.: *Left:* The type of operation performed by a neuron shows up in its lateral connection pattern. The density of the connections (mean over all orientations) was computed from the pixel image of Figure 5.1a on page 95 (artificial data) *Center:* The spatial layout of the cell in Figure 2.9 a (macaque monkey). In color coded are the local regions in which connections between spatial positions are assumed. *Right:* Profile of connection density over distance from the injection site

necting to all neurons in the local vicinity a the neuron. In the superficial layers $2/3$ long-range connections project mainly to patches of similar stimulus preference (see Section 2.3.3 on page 27). Interestingly in the layer $4B$ also long-range connections are found that indicate a different functional role from the iso-orientation biased layer $2/3$ circuitry (see Figure 2.9 on page 29).

As a first attempt we analysed the spatial layout of one layer $4B$- upper $4C\alpha$ cell (see Figure 2.9 a). The obtained profile of the connection pattern density is shown in Figure 7.1 center. Comparing it with the profiles for scaling and rotation (same figure, right) its multi-modal appearance indicates a more complex structure which is not surprising taking into account its bar-like structure.

If we assume that lateral connections influence the function of the neurons based on their selection of spatial correlations we can predict given the lateral layout of the connections the response properties of the corresponding neuron. To clarify the influence of this structure on the orientation preference and preferred spatial frequency of the neuron in question we reconstructed a quadratic form based on some assumptions about this particular neuron. (i) The neuron samples selectively spatial locations in the input and these locations are defined by the regions of dense terminal labeling in Figure 2.9 on page 29. (ii) The operation performed by the neuron in its dendritic tree is assumed to be dependent only on the structure found in the stimulus and can be modeled in the framework of second-order models (local multiplication fol-

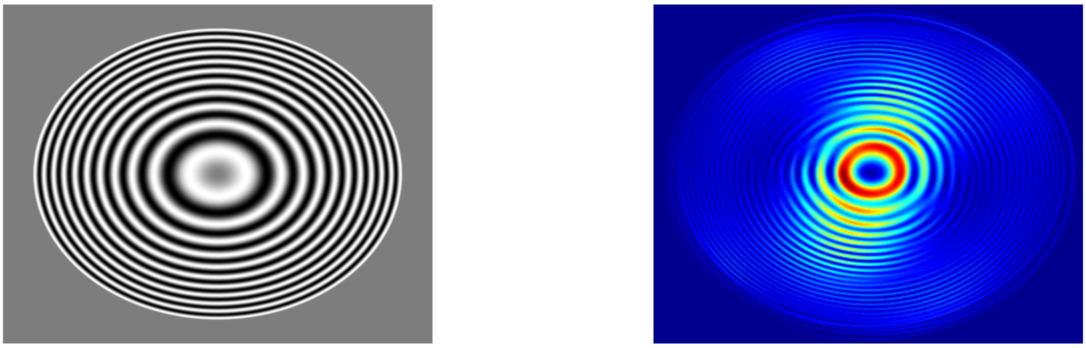


Figure 7.2.: *Left:* Test pattern (600×600) similar to the one used in Figure 6.27 on page 159 to test the orientation preference and preferred spatial frequency of a spatial filter (42×42). *Right:* The corresponding response image of the quadratic form defined by the lateral connection structure of the layer $4B$- upper $4C\alpha$ cell in Figure 2.9a on page 29

lowed by global summing). In particular we assume that apart from summing there is no dendritic micro-circuit between the bar-like pattern. Based on these assumptions we extracted from the shape of the lateral connections in Figure 2.9a, on page 29 $N = 7$ distinct regions (coded in color in Figure 7.1 center). In each region we assume that correlations between the terminal zones are computed. A 'correlation' matrix C was computed as the mean

$$C = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T \tag{7.2}$$

where $\mathbf{x}_i$ is obtained as the vectorized binary image selecting exclusively pixel corresponding to the i's colored region.

Using C we calculated a hypothetical response of the neuron in question by its quadratic form. Note that C is positive semi-definite and has rank 7 $(= N)$. In Figure 7.2 left, a test pattern is shown which we used already in Section 6.6.2 to estimate the orientation preference and preferred spatial frequency. In the corresponding response image red indicates a high response amplitude. As one can observe the neuron is selective for a particular orientation. This is remarkable because no particular orientation preference was assumed beforehand. Only the selective spatial sampling is sufficient to introduce a bias for a particular orientation. The way this is achieved is similar to the emergence of orientation selectivity in the local receptive fields by second order neurons. Inputs in the bar-like sub-fields implement a 'common-fate' mechanism producing large responses for uniform stimuli that fall into these regions.

By this predominantly the mean orientation of single bar's define the preferred orientation of the neuron.

This finding may serve as the end-point of this work. It relates the analysis of second order models starting from Section 4 to the role of lateral connections in primary visual cortex which was our subject in the first part of this thesis.

A. Measuring the Entropy of Natural Images

This chapter summaries some preliminary ideas about a framework to analyse natural images in terms of their probability distribution. Whereas previous attempts mostly started by defining a more or less arbitrary feature set and based on that analysed distributions of coefficients here we like to work with a more direct approach. We assume only the very weak assumption that images can be described by positions on a N-dimensional hyper-sphere (vectorized image patches are assumed to have length one). We show how one can estimate the probability distribution of natural images on the hyper-sphere and compute its entropy.

Measuring the entropy of a distribution is one way to measure the closeness of the distribution to the uniform distribution and by this its information content.

If images obtained from natural scenes (X_A) occupy only a small portion of the overall space of possible images (X_B) we expect that the entropies H of the two densities f_{X_A} and f_{X_B} differ significantly from each other. More precise we would expect that the entropy of X_B is the larger entropy because we can assume a uniform density for f_{X_B} whereas f_{X_A} should be more interesting.

The entropy or uncertainty of a random variable X is defined by the quantity

$$H(X) \quad = \quad -\sum_i f_X(x_i) \log f_X(x_i) = E_{f_X}[-\log f_X(X)] \tag{A.1}$$

where $f_X(x) \log f_X(x) = 0$ whenever $f_X(x) = 0$.

We will assume that the structure in natural images is well preserved if we assume $\mathbf{x} \in X$ as a vector $(x_1, \dots x_p)^T$ of length one. The set of all possible images lives therefore on the hyper-sphere $S^{p-1} = \{\mathbf{x} : \mathbf{x}^T\mathbf{x} = 1\}$ with radius one and dimension p. So, by Equation A.1 we need an estimate for the density f_{X_A} and a sampling scheme to get the expectation and by this the entropy of (a set of) natural images.

A.1. Kernel Density Estimation

To estimate the density of X_A on S^{p-1} we use a kernel density estimation approach. It is an efficient way to estimate the density in a case where we have

Figure A.1.: Sample ’natural images’ from Patrick Hoyers homepage

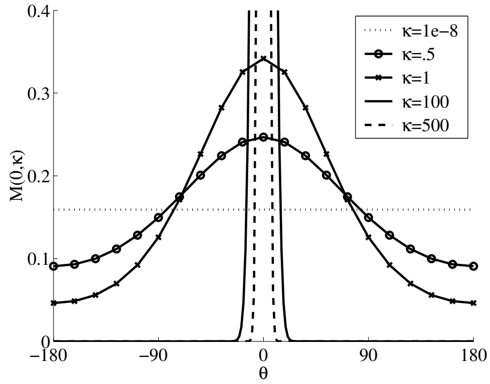


Figure A.2.: Density of the von Mises-Fisher distribution $M(0, \kappa)$

a small number of samples. For example for images patches of size 7×7 pixel the data space is the hyper-sphere S^{48}. All practical samples from this space are small samples.

Given observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we replace each data point $\mathbf{x}_i$ by a kernel function which assigns nearby (in S^{p-1}) data points non-zero probability. Note that it is not clear which distance measure (metric) to use. Two data points may be nearby given one metric. Given another metric the two data points may be far from each other. In our case image patches which are similar up to white noise of some amplitude are assumed to be similar e.g. having small distances from each other.

Because we have a periodic space it is natural to use the von Mises-Fisher distribution $M_p(\mathbf{x}_i, \kappa)$ as a kernel function with probability density function

$$\left(\frac{\kappa}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2) I_{p/2-1}(\kappa)} \exp\{\kappa \mu^T \mathbf{x}_i\} \tag{A.2}$$

as kernel function (Mardia and Jupp, 2000), p. 197. $\kappa \geq 0$ and $||\mu|| = 1$ are called the *concentration parameter* and *mean direction*, respectively. $I_p(.)$ is the modified Bessel function of the first kind and order p and $\Gamma(.)$ is the Gamma function. Note that the scalar product in the exponential function is over data in Cartesian coordinates as a measure of similarity of the mean μ and the data $\mathbf{x}$.

The corresponding kernel density estimate is given by (Mardia and Jupp (2000), p. 277)

$$\hat{f}_F(\mathbf{x}; \kappa) \;=\; n^{-1} a_p(\kappa) \sum_{i=1}^{n} \exp(\kappa \mathbf{x}^T \mathbf{x}_i), \tag{A.3}$$

where the constant κ determines the degree of smoothing (see Figure A.2 and Section A.2), and $a_p(\kappa)$ is the normalizing constant

$$a_p(\kappa) \quad = \quad \log\left(\left(\frac{\kappa}{2}\right)^{1-p/2}\Gamma\left(\frac{p}{2}\right)I_{p/2-1}(\kappa)\right). \tag{A.4}$$

A.2. Optimal Bandwidth for Kernel Density Estimation

Free parameters in kernel density estimation are the specific choice of the kernel and the used bandwidth of the kernel. The bandwidth changes the smoothness of the density estimate. For large bandwidth the resulting density estimate will be smooth, but generally the density will be overestimate (bias). If the bandwidth is small the resulting density estimate has a large variance so we need a large number of data points to estimate the density correctly. Because of this tradeoff we can assume that there is an optimal bandwidth for a given number of data points.

In general the kernel density estimate $\hat{f}_h(x)$ for $x_1, \ldots, x_n$ is

$$\hat{f}_h(x) \quad = \quad E_{x_i}\left[K_h(x - x_i)\right] = h^{-1}E_{x_i}\left[K\left(\frac{x - x_i}{h}\right)\right] \tag{A.5}$$

where h is the bandwidth of the kernel $K(x)$.

For the optimal bandwidth h_0 the mean integrated squared error (MISE) of the kernel density estimator should be minimal. It is composed of a variance and an bias term.

$$\begin{aligned}
\mathrm{MISE}(h) &= \int\limits_{-\infty}^{\infty} E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right]dx && \text{(A.6)}\\
&= \int\limits_{-\infty}^{\infty} E\left[\hat{f}_h(x) - E\left[\hat{f}_h(x)\right]\right]^2 + \left(E\left[\hat{f}_h(x)\right] - f(x)\right)^2 dx && \text{(A.7)}\\
&= \int\limits_{-\infty}^{\infty} \mathrm{Var}(\hat{f}_h(x))dx + \int\limits_{-\infty}^{\infty} \mathrm{Bias}^2(\hat{f}_h(x))dx && \text{(A.8)}
\end{aligned}$$

Asymptotically for the number of data points $n \to \infty$ the $\mathrm{MISE}(h)$ is according to Silverman (1986):

$$\mathrm{AMISE}(h) \quad = \quad \frac{1}{nh}\int K^2(x)dx + h^{2k}(\mu_k(K)/k!)^2\int\left(f^{(k)}(x)\right)^2 dx \tag{A.9}$$

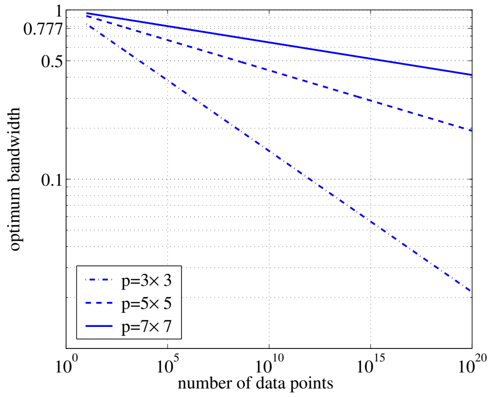


Figure A.3.: The order of the optimal bandwidth for different dimensions of image patches ($p = q + 1 = 3 \times 3, 5 \times 5, 7 \times 7$) over the number of samples. Note the log in both the x-and y-axes

where k is the order of the kernel K (normally $k = 2$) and $\mu_j(L) = \int x^j L(x)dx$ for any function L. Note that for this approach f has to be at least k times bounded or square integrable. The optimum kernel width h_0 is the minimizer of $\mathrm{AMISE}(h)$ and can be obtained by differentiating with respect to h and calculating the root of the derivative.

Härdle and Müller (2000) point out that for a multivariate kernel density estimator (Gaussian kernel $k = 2$) the optimal bandwidth depends strongly on the dimension q of the data and is in the order of:

$$h_0 \quad = \quad O(n^{-1/(4+q)}). \tag{A.10}$$

Figure A.3 displays this relationship for different dimensions $p = 3^2, 5^2, 7^2$ of the data (square image patches). The order of the bandwidth scales badly with the dimension. For image patches of size 7×7 we need 500000 samples to achieve a optimal bandwidth of $.777$, twice as many samples would lower the optimal bandwidth only slightly to $.767$. This indicates that using kernel density estimation in spaces of large dimensions is not feasible with kernels of small bandwidth. One would expect a large variance of the estimate.

Härdle and Müller (2000) derived a rule-of-thumb for the bandwidth in the case that we have a diagonal bandwidth matrix and a multivariate normal distribution as a reference distribution:

$$\hat{h}_j \quad = \quad \left(\frac{4}{q+2}\right)^{1/(q+4)} n^{-1/(q+4)} \sigma_j. \tag{A.11}$$

A.3. Entropy of the Density f_{X_A}

Using Equation A.3 we calculate the entropy as the expectation over random samples from the density $\hat{f}_{X_A}$

$$H(X_A) \quad = \quad E_{\hat{f}_{X_A}}(-\log(\hat{f}_{X_A})) \tag{A.12}$$

To calculate the expectation we have to sample from the distribution f_{X_A}. Because it is non-trivial[1] to sample according to the density in Equation A.3 we introduce a new distribution g and sample from this distribution because

$$\begin{aligned} E_f(-log(f(x))) &= -\int f(x)log(f(x))dx \\ &= -\int g(x)\frac{f(x)}{g(x)}log(f(x))dx \\ &= E_g\left(-\frac{f(x)}{g(x)}log(f(x))\right) \end{aligned} \tag{A.13}$$

will give us the desired expectation (Metropolis et al., 1953; Marshall, 1956). The expectation will converge faster, e.g. has smaller variance, for a distribution g that is similar to f. Fortunately there is a distribution g, the multivariate normal distribution, that is very similar to the von Mises distribution and there are efficient tools for sampling from that distribution. It has probability density function:

important sampling

$$g(\mathbf{x} - \mu, \Sigma) \quad = \quad \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu)\right\}, \tag{A.14}$$

where d is the dimension (in our case $d = p - 1$ because g is defined on the hyper-sphere S^{p-1}) and Σ is the variance-covariance matrix ($\Sigma_{ii} = \sigma$).

The normal distribution $\mathcal{N}(\mu, \sigma^2)$ wrapped around the circle gives the wrapped normal distribution $WN(\mu, \exp(-\sigma^2/2))$. The wrapped normal distribution is a close (first-order) approximation of the von Mises distribution

$$M(\mu, \kappa) \quad \simeq \quad WN\left(\mu, A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}\right) \tag{A.15}$$

for high and intermediate values of κ (Kent, 1978; Stephens, 1963). I_p is the modified Bessel function of the first kind and order p. Schou (1978) has given an approximation for $A(\kappa)$ as

$$A(\kappa) \quad = \quad 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} + O\left(\frac{1}{\kappa^3}\right) \tag{A.16}$$

[1] Sampling by the transformation method for example we would have to calculate the inverse function of the integral of $\hat{f}$.

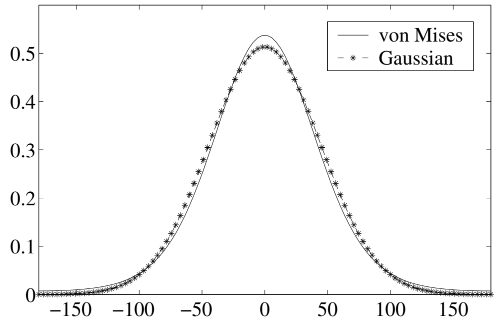


Figure A.4.: Comparison of a Gaussian function with $\mathcal{N}(\mu = 0, \sigma = .777)$ and a von Mises-Fisher distribution with $M(\mu = 0, \kappa = 2.1428)$

which is useful for large κ e.g., the small variance case in which the non-periodic function g can be reasonable compared to the (periodic) von Mises distribution function. Because of the above identities the normal distribution $\mathcal{N}(\mu, \sigma)$ that is closest to the von Mises distribution for large concentration parameter κ is[2]

$$\mathcal{N}\left(\mu, \sqrt{6\log(2) - 2\log\left(\frac{8\kappa^2 - 4\kappa - 1}{\kappa^2}\right)}\right) \simeq M(\mu, \kappa)\,. \qquad \text{(A.17)}$$

Using the above definition for the variance of g we can sample from g by drawing $p - 1$ values from a multivariate normal distribution. Calculation of the entropy in Equation A.1 can now be done by Equation A.13.

A.4. Results

We sampled 500000 image patches of size 7×7 from the 13 natural images shown in Figure A.1 on page 175. The bandwidth of the Gaussian g was chosen to be $.777$ (see Section A.2 and Figure A.4). This defines in turn a concentration parameter of $\kappa = 1.7722$ for the von Mises kernel (by Equation A.17). In Figure A.5 the result is plotted over increasing values of n and kernel sizes κ. The entropy for small κ (large variance) approaches the entropy of the uniform density f_{X_B}. Large negative entropy for small kernel variances indicate that the distribution of natural images on the unit sphere is highly non-uniform.

[2] For small κ a better approximation is $A(\kappa) = 1/2\kappa - 1/16\kappa^3 + O(\kappa^5)$.

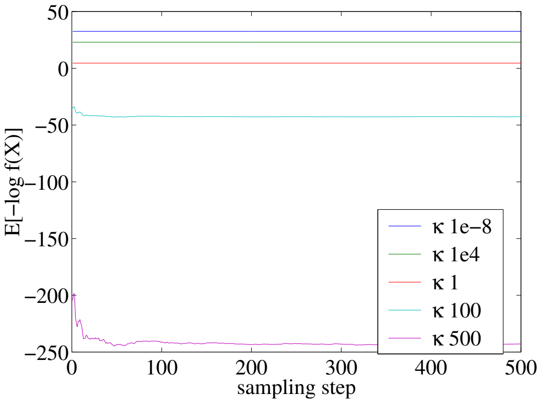


Figure A.5.: Entropy $E\left(-\log\left(f_{X_A}\right)\right) = 1/n \sum_i^n -\log(f_{X_A})$ of natural images for different concentration parameter κ of the kernel over growing sampling sizes n. Negative entropy for small kernels indicate a high non-uniformity of the density f_{X_A}

B. Derivations

B.1. Dependence of Symmetry on RFS

To calculate the (rotational) symmetry value for a given object in an image X we introduced the $L1$-norm of a vector calculated from the covariances between the original image and N successively rotated versions of that image (see Equation 4.8 on page 90). Lets analyse now, how this measure of the structure in X depends on the size of the object. For this reason we assume that the figure is perfectely symmetric,

$$x_{r,\theta'} = x_{r,(\theta'-\tau)}, \quad x \in X, \tau \in [0, 2\pi], \tag{B.1}$$

e.g., can be described by a disc with a certain radius α. In this case, the symmetry value for all τ will be the same.

$$\mathcal{S}(X) \;=\; N \operatorname{cov}_\tau(X,Y) = N \int\limits_0^{2\pi} \int\limits_0^{r_{\mathtt{max}}} (x_{r,\theta'} - \bar{x})^2 dr d\theta' \tag{B.2}$$

and let Y describe a by τ rotated version of X (because of rotational symmetry $Y = X$). We will omit the constant factor N for the next calculations.

A disc with radius α can be modeled by a single heavy-side function H (in polar coordinates):

$$\begin{aligned} x_r &= -H(r-\alpha) + 1 & \text{(B.3)} \\ \mathtt{cov}_\tau(X,Y) &= 2\pi \int\limits_0^{r_{\mathtt{max}}} \big(- H(r-\alpha) + 1 - E\left[-H(r-\alpha)+1\right] \big)^2 dr & \text{(B.4)} \end{aligned}$$

where $E(.)$ is the expectation of $(.)$.

$$E\left[-H(r-\alpha)+1\right] \;=\; \frac{1}{\pi r_{\mathtt{max}}^2} \pi\alpha^2 = \frac{\alpha^2}{r_{\mathtt{max}}^2} \tag{B.5}$$

By simple analysis it follows that:

$$\begin{aligned}
\mathrm{cov}_\tau(X,Y) &= 2\pi \int\limits_0^{r_\alpha} \left(-H(r-\alpha)+1-\frac{\alpha^2}{r_{\mathtt{max}}^2}\right)^2 dr && \text{(B.6)}\\
&= 2\pi \int\limits_0^{\alpha} \left(1-\frac{\alpha^2}{r_{\mathtt{max}}^2}\right)^2 dr \quad = 2\pi \int\limits_0^{\alpha} 1-2\frac{\alpha^2}{r_{\mathtt{max}}^2}+\frac{\alpha^4}{r_{\mathtt{max}}^4}\, dr && \text{(B.7)}\\
&= 2\pi\alpha \left(1-2\frac{\alpha^2}{r_{\mathtt{max}}^2}+\frac{\alpha^4}{r_{\mathtt{max}}^4}\right) && \text{(B.8)}
\end{aligned}$$

This later function is zero at $\alpha = 0$ and $\alpha = r_{\mathtt{max}}$ and shows a maximum in between at $\alpha = r_{\mathtt{max}}\frac{1}{\sqrt{5}} \approx 0.48 r_{\mathtt{max}}$. So a disc has maximum symmetry if it fills nearly half the receptive field.

A neuron implementing this strategy of integration over the input would have a measured receptive field (standard method) of half its ’real’ integration field and would show furthermore a strong inhibitory surround (for a disc like stimulus).

Nice: The result of the maximum symmetry seems (not) to be correlated with the golden ratio. The ratio $1-\frac{1}{\sqrt{5}} = \frac{\sqrt{5}-1}{\sqrt{5}}$ is near the value of the golden ratio $\frac{\sqrt{5}-1}{2}$ ($\sqrt{5} = 2.236 \approx 2$).

B.2. Dependence of Symmetry on Preferred Orientation

To calculate the (rotational) symmetry value for a given object in an image X we introduced the $L1$-norm of a vector calculated from the covariances between the original image and N successively rotated versions of that image (see Equation 4.8 on page 90).

$$\mathcal{S}(X,\tau) = \int\limits_0^{r_{\max}} \int\limits_0^{2\pi} (x_{r,\theta}-\bar{x})(x_{r,\theta-\tau}-\bar{x})d\theta dr \qquad \text{(B.9)}$$

Lets now analyse, how the measure of structure depends on the orientation of a contrast gradient in the image. This example is important because contrast gradients, or edges are common features in natural images.

Without restrictions we can assume that the edge is a horizontal one. This follows from the observation that we can rotate the original image and nevertheless obtain the same vector $\mathcal{S}$. Whereas this implies that $\mathcal{S}$ is not specific for a certain orientation (has no preferred orientation) we show now that the entries of $\mathcal{S}$ depend linearly on the difference between the edge orientation and the rotation angle τ.

The mean of an image containing a centered edge is $\bar{x} = (\mathtt{max}(X) - \mathtt{min}(X))/2$. For now we want to restrict ourselves to the case of a binary image for which $\bar{x} = 1/2$.

We first look onto the inner integral which describes for a given radius r the pixel pair products between the original image ($x_{r,\theta}$) and the by τ rotated image ($x_{r,\theta-\tau}$):

$$\int_0^{2\pi} (x_{r,\theta} - \bar{x})(x_{r,\theta-\tau} - \bar{x})d\theta = \int_0^{2\pi} (x_{r,\theta} - \frac{1}{2})(x_{r,\theta-\tau} - \frac{1}{2})d\theta \tag{B.10}$$

$$= \int_0^{2\pi} \underbrace{x_{r,\theta}x_{r,\theta-\tau} - \frac{1}{2}(x_{r,\theta} + x_{r,\theta-\tau}) + \frac{1}{4}}_{f_\tau(\theta)} d\theta \tag{B.11}$$

We can split the later equation into a sum of three integrals depending on $x_{r,\theta}$ and $x_{r,\theta-\tau}$:

$$= \int_{\substack{0 \\ x_{r,\theta}=x_{r,\theta-\tau}=1}}^{2\pi} f_\tau(\theta)d\theta + \int_{\substack{0 \\ x_{r,\theta}=x_{r,\theta-\tau}=0}}^{2\pi} f_\tau(\theta)d\theta + 2 \int_{\substack{0 \\ x_{r,\theta}\neq x_{r,\theta-\tau}}}^{2\pi} f_\tau(\theta)d\theta \tag{B.12}$$

$$= \int_{\substack{\tau \\ x_{r,\theta}=x_{r,\theta-\tau}=1}}^{\pi} f_\tau(\theta)d\theta + \int_{\substack{\pi+\tau \\ x_{r,\theta}=x_{r,\theta-\tau}=0}}^{2\pi} f_\tau(\theta)d\theta + 2 \int_{\substack{0 \\ x_{r,\theta}\neq x_{r,\theta-\tau}}}^{\tau} f_\tau(\theta)d\theta \tag{B.13}$$

$$= \int_\tau^\pi 1 - 1 + \frac{1}{4}\ d\theta + \int_{\pi+\tau}^{2\pi} \frac{1}{4}\ d\theta + 2\int_0^\tau -\frac{1}{4}\ d\theta = \frac{\pi}{2} - \tau \tag{B.14}$$

The outer integral and therefore $\mathcal{S}(X,\tau)$ is now simply:

$$\int_0^{r_{\mathtt{max}}} \frac{\pi}{2} - \tau\ dr \quad = \quad \left(\frac{\pi}{2} - \tau\right) r_{\mathtt{max}}. \tag{B.15}$$

So

$$\mathcal{S}_\tau(X) \quad = \quad \left\{ \left(\frac{\pi}{2} - \tau\right) r_{\mathtt{max}} \right\}_{0\leq\tau\leq 2\pi} \tag{B.16}$$

and $\mathcal{S}$ depends linearly on τ. In particular the symmetry value will be zero for an edge orthogonal oriented ($\tau = \frac{\pi}{2}$) to the rotation axis.

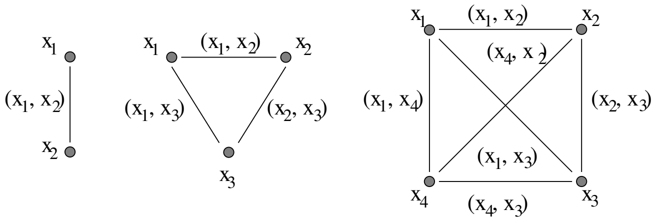


Figure B.1.: Example for a set of four pixels (x_1, x_2, x_3, x_4) and the corresponding 6 pixel pairs

B.3. Dependence of Symmetry on Noise

We show now that for white noise input the random variable defined by the symmetry detection procedure is characterized by large skewness and kurtosis as measures of higher order statistics. The distribution has mean 0 because we use only co-variances $(x_i x_j, i \neq j)$ and because of the independence assumption in white noise.

cliques

Let X be a random variable with $p(X) = \mathcal{N}(0, 1)$ and let Y be a sequence of the form $(x_1, x_2), (x_3, x_4), \ldots, (x_n, x_{n+1})$ where $x_i \in X$. This defines Y as a *clique*. A clique in graph theory is a collection of sites such that any two sites are neighbors (see Figure B.1). Here a clique defines a set of pixel such that any pixel in the clique has a connection weighted by 1 with all other pixel in the clique. The *order* of a clique refers to the number of distinct sites that appear multiplicatively. Note that the structure detection algorithm uses an average sum of cliques. For the detection of rotational symmetries all pixel at an 'equal' distance to the center of the patch are in on clique.

The random variable Y has a specific structure for symmetry pixel pairs that violates the independence assumption of x_i sampled iid. This is the consequence of using a single pixel in more than one element of Y.

More general, we look for the third and forth order central moments of the sum of a sequence X of $N = \frac{M(M-1)}{2}$ monomials of order 2. We can define a minimal non-trivial example of such a sequence for monomials of order 2 in 3 variables ($M = 3$). The sum of pixel pairs produces is in this case:

$$Y \quad : \quad (x_1 x_2) + (x_2 x_3) + (x_3 x_1). \tag{B.17}$$

Let each x_i be drawn independently from a normal distribution $\mathcal{N}(0, 1)$. It follows from definition that

$$E(x_i) \quad = \quad 0 \qquad E(x_i^2) = 1 \qquad E(x_i^3) = 0 \qquad E(x_i^4) = 3 \tag{B.18}$$

The corresponding cumulants are zero indicating no higher order information in X. Using the above results we explicitely calculate now the third and forth order central moments of Y. We assume that Y is a random variable which depend non-lineary on x_1, x_2, x_3 which in turn are sampled iid from a Gaussian distribution with mean 0 and variance 1. In the following $E(.)$ will always stand for the expection over all enclosed x_i.

$$Y := \sum_{i \neq j, i<j} x_i x_j \tag{B.19}$$

$$E(Y)_{x_1,x_2,x_3} = \underbrace{E(x_1x_2) + E(x_2x_3) + E(x_3x_1)}_{0} = 0 \tag{B.20}$$

because of x_i being iid and for all (x_ix_j), $i \neq j$ $E(x_ix_j) = E(x_i)E(x_j)$.

$$\begin{aligned} E(Y^2) &= E\left((x_1^2x_2^2) + (x_2^2x_3^2) + (x_3^2x_1^2) + 2\left(x_1x_2^2x_3 + x_1^2x_2x_3 + x_1x_3^2x_2\right)\right) && \text{(B.21)} \\ &= E(x_1^2)E(x_2^2) + E(x_2^2)E(x_3^2) + E(x_3^2)E(x_1^2) + 2\left(\underbrace{E(x_1)E(x_2^2)E(x_3)}_{=0} + \right. \\ &\quad \left. \underbrace{E(x_1^2)E(x_2)E(x_3)}_{=0} + \underbrace{E(x_1)E(x_2)E(x_3^2)}_{=0}\right) = 1 && \text{(B.22)} \end{aligned}$$

$$\begin{aligned} E(Y^3) &= [A+B+C]^3, \qquad (A = x_1x_2, B = x_1x_3, C = x_2x_3) && \text{(B.23)} \\ &= A^3 + 3A^2(B+C) + 3A(B+C)^2 + (B+C)^3 && \text{(B.24)} \\ &= x_1^3x_2^3 + 3x_1^2x_2^2(x_1x_3 + x_2x_3) + 3x_1x_2(x_1x_3 + x_2x_3)^2 + \\ &\quad (x_1x_3 + x_2x_3)^3 && \text{(B.25)} \\ &= x_1^3x_2^3 + 3x_1^3x_2^2x_3 + 3x_1^2x_2^3x_3 + 3x_1x_2(x_1^2x_3^2 + 2x_1x_2x_3^2 + x_2^2x_3^2) + \\ &\quad (x_1x_3 + x_2x_3)^3 && \text{(B.26)} \\ &= x_1^3x_2^3 + 3x_1^3x_2^2x_3 + 3x_1^2x_2^3x_3 + 3x_1^3x_2x_3^2 + \underline{6x_1^2x_2^2x_3^2} + 3x_1x_2^3x_3^2 + \\ &\quad x_1^3x_3^3 + 2x_1^2x_2x_3^3 + x_1x_2^2x_3^3 + x_1^2x_2x_3^3 + 2x_1x_2^2x_3^3 + x_2^3x_3^3 && \text{(B.27)} \end{aligned}$$

Only one term in the above sum is $\neq 0$, therefore

$$E(Y^3) = 6E(x_1^2)E(x_2^2)E(x_3^2) = 6 \tag{B.28}$$

The central moment of fourth order can be computed as:

$$\begin{aligned} E(Y^4) &= [A+B+C]^4, \qquad (A = x_1x_2, B = x_1x_3, C = x_2x_3) && \text{(B.29)} \\ &= x_1^4x_2^4 + x_2^4x_3^4 + 4x_1^4x_2x_3^3 + 12x_1^3x_2^2x_3^3 + 12x_1^2x_2^3x_3^3 + \\ &\quad 4x_1x_2^4x_3^3 + 4x_1^4x_2^3x_3 + 4x_1^3x_2^4x_3 + 6x_1^4x_2^2x_3^2 + 12x_1^3x_2^3x_3^2 + \\ &\quad 6x_1^2x_2^4x_3^2 + 4x_1^3x_3^4x_2 + 6x_1^2x_3^4x_2^2 + 4x_1 + x_1^4x_3^4 && \text{(B.30)} \end{aligned}$$

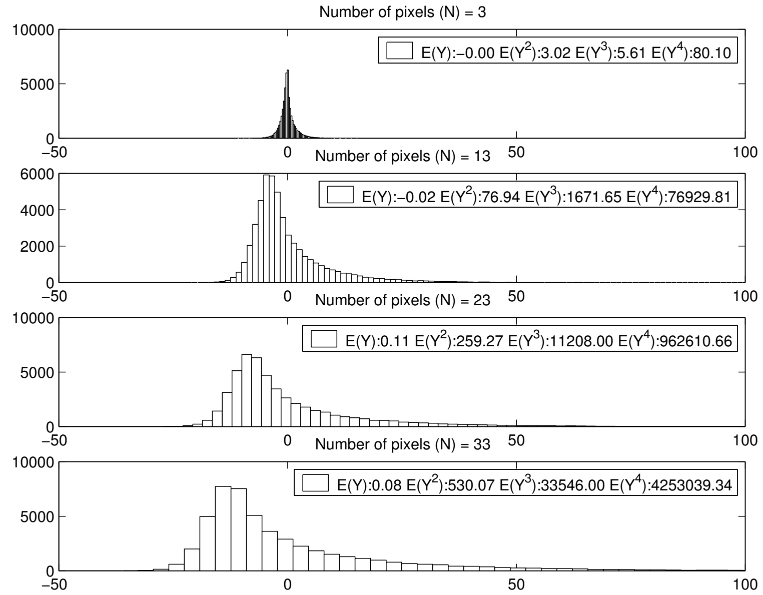


Figure B.2.: For the cases of $M = 3, 13, 23, 33$

Again we reduce this formula with the previously defined values for the mean, variance and third order moment of Y to

$$E(Y^4) = x_1^4x_2^4 + x_2^4x_3^4 + 6x_1^4x_2^2x_3^2 + 6x_1^2x_2^4x_3^2 + 6x_1^2x_3^4x_2^2 + x_1^4x_3^4 \tag{B.31}$$

$$\begin{aligned} &= E(x_1^4)E(x_2^4) + E(x_2^4)E(x_3^4) + 6E(x_1^4)E(x_2^2)E(x_3^2) + \\ &\quad 6E(x_1^2)E(x_2^4)E(x_3^2) + 6E(x_1^2)E(x_3^4)E(x_2^2) + E(x_1^4)E(x_3^4) \end{aligned} \tag{B.32}$$

$$E(Y^4) = 81 \tag{B.33}$$

To collect the above results we obtain for Y the first four central moments as:

$$E(Y) = 0 \quad E(Y^2) = 1 \quad E(Y^3) = 6 \quad E(Y^4) = 81 \tag{B.34}$$

To compare these numbers with the best fitting normal distribution we introduce now cumulants.

cumulants

Cumulants are normalized versions of the higher order central moments. They take into account that for a Gaussian distribution the higher order moments factorize, e.g. all higher order moments of a normal distribution can be expressed as combinations of its first two moments (mean and variance). The normalized versions of the central moments are corrected to yield zero for any higher than order 2 term.

In particular the third order cumulant is called *skewness* and is computed from the moment of third order as: *skewness*

$$s(Y) = \frac{1}{(N-1)\sigma^3}\sum_{i=1}^{N}(y_i-\bar{y})^3 = \frac{6}{(3-1)1^3} = 3 \qquad \text{(B.35)}$$

which incorporates the standard deviation $\sigma = E(Y^2)$ and the mean $\bar{x} = E(Y)$ of the sequence Y.

kurtosis The *kurtosis* is the cumulant of fourth order and defined as:

$$k(Y) = \frac{1}{(N-1)\sigma^4}E(Y^4) - 3 = \frac{1}{(3-1)1^4}81 - 3 = 37.5 \qquad \text{(B.36)}$$

As we can see Y is a skewed and highly kurtotic distribution. This is confirmed by the shape of the distributions in Figure B.2 on the page before. As it can be seen from the simulation there is a trend to introduce higher order moments by monomials in more variables (which correspond to more pixel for growing radii).

Multivariate Gaussian Distribution in Two Variables

For the model of a multivariate Gaussian distribution

$$p(X) = \mathcal{N}\left(\mu, \Sigma = \begin{pmatrix} \sigma_X & c_X \\ c_X & \sigma_X \end{pmatrix}\right) \qquad \text{(B.37)}$$

the expectation of Y can be expressed as:

$$\begin{aligned} E(Y) &= E\left(\sum_{i\neq j}^{N} x_i x_j\right) && \text{(B.38)} \\ &= \sum_{i\neq j}^{N} E(x_i x_j). && \text{(B.39)} \end{aligned}$$

Note that our covariance structure is restricted to be symmetric ($\Sigma_{12} = \Sigma_{21}$) and rotationally invariant ($\Sigma_{11} = \Sigma_{22}$). Due to the fact that $E(XY) = \mathtt{Cov}(X,Y) + E(X)E(Y)$:

$$\begin{aligned} E(Y) &= \sum_{i\neq j}(c_X + E(x_i)E(x_j)) \\ &= \sum_{i\neq j}(c_X + \mu^2) \end{aligned}$$

For a zero mean ($\mu = 0$) Gaussian distribution $P(X)$ we end up with an expectation for Y of

$$E(Y) = \frac{N(N-1)}{2} c_X \tag{B.40}$$

where c_X is the covariance between any two pixels $\mathrm{cov}(x_i, x_j), i \neq j$ in X and N is the order of the respective clique (number of pixels). If we build the expectation with respect to the number of pixel pairs we arrive at

$$E(Y) = c_X. \tag{B.41}$$

Lets combine the cliques of growing order so that we can compute the expectation of $S(X)$. Again we assume that X has a probability distribution that is multivariate Gaussian (Equation B.37 on the facing page). If N is the order of the clique Y_N, we can assume at each radius r arround the center position of X a number $N = 2\pi r$ of pixels forming a clique:

$$\begin{aligned} E(S(X)) &= E\left(\sum_{i=1}^{\texttt{rfs}} Y_{2\pi i}\right) = \frac{1}{\texttt{rfs}} \sum_{i=1}^{\texttt{rfs}} E(Y_{2\pi i}) && \text{(B.42)} \\ &= c_X. && \text{(B.43)} \end{aligned}$$

C. General Linear Independence of Monomial Spaces

Lets analyse a set of non-linear basis functions in order to prove that we can apply linear methods in this feature space.

Monomials of constant order d are of the form

$$\mathcal{F}_n(d) \quad = \quad \{x_1^{e_1}x_2^{e_2}\dots x_n^{e_n} | e_1 + e_2 + \dots + e_n = d, e_i \geq 0\}\,. \tag{C.1}$$

Linear methods have to identify directions in the feature space that correspond to directions in the input space, i.e., directions on the manifold. In order to reliably extract n independent source directions we have to ensure that the data projection in the feature space is effective, e.g., that the n unknown source directions span the whole space.

General linear independence of monomial spaces of constant order: We have to show that any n pairwise different vectors $\gamma_{1\dots n} \in [0\dots\pi)$ are linear independent if they are in the manifold defined by n monomials of constant order.

Commutative monomials of order n in two variables can be expressed in polar coordinates as ($\gamma \in [0\dots 2\pi)$)

$$\begin{aligned} &(y^0x^n, yx^{n-1}, \dots, y^{n-1}x, y^nx^0) = \\ r^n(\cos^n(\gamma), \sin(\gamma)&\cos^{n-1}(\gamma), \dots, \sin^{n-1}(\gamma)\cos(\gamma), \sin^n(\gamma)) \end{aligned} \tag{C.2}$$

Proving strong linear independence can be done by proofing that the matrix $M =$

$$\begin{pmatrix} \cos^n(\gamma_1) & \sin(\gamma_1)\cos^{n-1}(\gamma_1) & \dots & \sin^{n-1}(\gamma_1)\cos(\gamma_1) & \sin^n(\gamma_1) \\ \cos^n(\gamma_2) & \sin(\gamma_2)\cos^{n-1}(\gamma_2) & \dots & \sin^{n-1}(\gamma_2)\cos(\gamma_2) & \sin^n(\gamma_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \cos^n(\gamma_n) & \sin(\gamma_n)\cos^{n-1}(\gamma_n) & \dots & \sin^{n-1}(\gamma_n)\cos(\gamma_n) & \sin^n(\gamma_n) \end{pmatrix} \tag{C.3}$$

has a non-zero determinant in the range $\gamma_{1\dots n} \in [0\dots\pi]$. This matrix can be decomposed into a product of a diagonal matrix $V_{i,i} = \cos^n(\gamma_i)$ and a matrix

$$W \quad = \quad \begin{pmatrix} 1 & \tan(\gamma_1) & \tan^2(\gamma_1) & \dots & \tan^n(\gamma_1) \\ 1 & \tan(\gamma_2) & \tan^2(\gamma_2) & \dots & \tan^n(\gamma_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \tan(\gamma_n) & \tan^2(\gamma_n) & \dots & \tan^n(\gamma_n) \end{pmatrix} \tag{C.4}$$

ndermonde

because $\tan = \sin / \cos$. W is a *Vandermonde* matrix. The determinant of this matrix is known to be $\det(W) = \Pi_{i,j,i>j}(\tan(\gamma_i) - \tan(\gamma_j))$. Together with V the determinant of the matrix M in Equation C.3 on the preceding page is therefore

$$\det(M) \;=\; \prod_{i=1}^{n} \cos^n(\gamma_i) \prod_{i,j,i>j} (\tan(\gamma_i) - \tan(\gamma_j)) \tag{C.5}$$

Because $\tan$ is a strict monotonic function the only case in which the last term can be zero is, if for some $i \neq j$ $\gamma_i = \gamma_j$ which violates our requirement of having n different vectors $\gamma_{1 \dots n}$.

Because for n different directions the cosine function has only one zero-crossing between $[0 \dots \pi)$ and we require that all γ_i are different, there can only be a single γ_i for which $\cos(\gamma_i) = 0$. The only case in which the determinant can be zero is therefore if, without loss of generality, we assume that $\cos(\gamma_1) = 0$.

If $\cos(\gamma_1) = 0$ and $\cos(\gamma_{2 \dots n}) \neq 0$ it follows that in the first line of the matrix M only the last term $\sin^n(\gamma_1) \neq 0$. The determinant of the matrix M in Equation C.3 reduces to

$$\begin{aligned} \det(M) \;=\; & (-1)^{n+1} \sin^n(\gamma_1) \\ & \begin{vmatrix} \cos^n(\gamma_2) & \sin(\gamma_2)\cos^{n-1}(\gamma_2) & \dots & \sin^{n-1}(\gamma_2)\cos(\gamma_2) \\ \vdots & \vdots & \ddots & \vdots \\ \cos^n(\gamma_n) & \sin(\gamma_n)\cos^{n-1}(\gamma_n) & \dots & \sin^{n-1}(\gamma_n)\cos(\gamma_n) \end{vmatrix} . \end{aligned} \tag{C.6}$$

The later sub-determinant can again be expressed as the determinant of the product of a matrix $V_{i,i} = \cos^n(\gamma_i)$ and a Vandermonde matrix

$$\begin{pmatrix} 1 & \tan(\gamma_2) & \tan^2(\gamma_2) & \dots & \tan^{n-1}(\gamma_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \tan(\gamma_n) & \tan^2(\gamma_n) & \dots & \tan^{n-1}(\gamma_n) \end{pmatrix} . \tag{C.7}$$

Because of $\gamma_{2 \dots n}$ are different vectors and $\cos(\gamma_{2 \dots n}) \neq 0$ the determinant of the matrix M is non-zero if $\cos(\gamma_1) = 0$.

The presented proof uses as a crucial ingredient the strict monotonicity of the $\tan$ function in the interval $[0, \pi)$. This opens a path to define other feature spaces by introducing other strict monotonic functions. Important in this respect is that the property of linear separability has to be fulfilled for these feature spaces also. For example, with two variables x, y and $n = 2$ by exchanging $\tan$ by $\sin$ we arrive at basis functions (x^2, x^2y, x^2y^2) which are not separable into a scalar part and a vectorial part depending only on γ.

Of course we can also exchange the $\cos$ entries in the diagonal matrix. Exchanging $\cos^n$ by $\sin^n$ in the example above we arrive at separable basis functions of constant order $(y^2, y^3/x, y^4/x^2)$.

Bibliography

Adorján, P., Levitt, J. B., Lund, J. S. and Obermayer, K. (1999). A model for the intracortical origin of orientation preference and tuning in macaque striate cortex, *Visual Neuroscience* **16**: 303–318.

Adorján, P., Schwabe, L., Piepenbrock, C. and Obermayer, K. (2000). Recurrent cortical competition: Strengthen or weaken?, *in* S. A. Solla, T. K. Leen and K.-R. Müller (eds), *Advances in Neural Information Processing Systems*, Vol. 12, MIT Press, p. in press.

Agmon-Snir, H. and Segev, I. (1993). Signal delay and propagation velocity in passive dendritic trees, *J. Neurophys.* **70**: 2066–2085.

Albrecht, D. G. and Hamilton, D. B. (1982). Striate cortex of monkey and cat: contrast response function, *J. Neurophysiol.* **48**(1): 217–237.

Alexander, D. M., Sheridan, P. and Bourke, P. D. (1997). An algebraic-geometric model of the receptive field properties of the macaque striate cortex, *Proc. Aust. Neuroscience Soc.*

Allman, J. M. and Kaas, J. H. (1974). The organization of the second visual area (v ii) in the owl monkey: a second order transformation of the visual hemifield, *Brain Res.* **76**: 247–265.

Alonso, J. M., Usrey, W. M. and Reid, R. C. (2001). Rules of connectivity between geniculate cells and simple cells in cat primary visual cortex, **21**: 4002–4015.

Alvarez, L., Gousseau, Y. and Morel, J. (1999). The size of objects in natural images, *CMLA preprint, Ecole Normale Sup.- Cachan*.

Amari, S. and Nagaoka, H. (1993). *Methods of Information Geometry*, Vol. 191, American Mathematical Society, Oxford, University Press.

Arbib, M. A. (1998). *The handbook of brain theory and neural networks*, The MIT Press, Cambridge, Massachusetts, London, England.

Asi, H., Lund, J. S., Blasdel, G. G., Angelucci, A. and Levitt, J. B. (in press). Stripe-like patterns of lateral connections in layers $4b$ and upper $4c\alpha$ of macaque monkey primary visual cortex, area v1, *Journal of Comparative Neurology*.

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

Attias, H. (1999). Independent factor analysis, *Neural Computation* **11**(4): 803–851.

Attneave, F. (1955). Symmetry information and memory for patterns, *American Journal of Psychology* **68**: 209–222.

Azouz, R., Gray, C. M., Nowak, L. G. and McCormick, D. A. (1997). Physiological properties of inhibitory interneurons in cat striate cortex, *Cerebral Cortex* **7**: 534–545.

Baddeley, R. (1997). The correlational structure of natural images and the calibration of spatial representations, *Cognitive Science* **21**(3): 351–372.

Barlow, H. and Reeves, B. (1979). The versatility and absolute efficiency of detecting mirror symmetry in rd displays, *Vision Research* **19**: 783–793.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina, *J. Physiol., Lond.* **119**: 69–88.

Barlow, H. B. (1961). The coding of sensory messages, *in* W. H. Thorpe and O. L. Zangwill (eds), *Current Problems in Animal Behavior*, pp. 331–360.

Bartsch, H. and Obermayer, K. (2001). Detecting structure in images by detecting symmetry, *in* N. Elsner and G. W. Kreutzberg (eds), *Proceedings of the* 4^{th} *Meeting of the German Neuroscience Society 2001*, Vol. 1, Georg Thieme Verlag, p. 277.

Bartsch, H. and Obermayer, K. (2002). A structure preserving image transformation as the goal of visual sensory coding, *Neurocomputing* **44-46**: 729–734.

Bartsch, H. and Obermayer, K. (2003). Second order statistics of natural images, *Neurocomputing* p. in press.

Bartsch, H., Stetter, M. and Obermayer, K. (1997). A model for orientation tuning and contextual effects of orientation selective receptive fields., *in* W. Gerstner, A. Germond, M. Hasler and J.-D. Nicoud (eds), *Artificial Neural Networks - ICANN '97*, Vol. 1327 of *Lecture notes in computer science*, Springer Berlin.

Bartsch, H., Stetter, M. and Obermayer, K. (1999a). About the influence of neuronal variability in a mean-field model of the visual cortex., *in* N. Elsner and U. Eysel (eds), *From molecular neurobiology to clinical neuroscience.*, Thieme Stuttgart.

Bartsch, H., Stetter, M. and Obermayer, K. (1999b). On the influence of threshold variability in a model of the visual cortex., *Artificial Neural Networks – ICANN '99*, pp. 73–78.

Bartsch, H., Stetter, M. and Obermayer, K. (1999c). On the influence of threshold variability in a model of the visual cortex., *in* D. Willshaw and A. Murray (eds), *Artificial Neural Networks - ICANN '99*, Vol. 470 of *Conference Publication No.470*, Institution of Electrical Engineers, London, pp. 73–78.

Bartsch, H., Stetter, M. and Obermayer, K. (2000a). The influence of threshold variability on the response of visual cortical neurons, *Neurocomputing* **32-33**: 37–43.

Bartsch, H., Stetter, M. and Obermayer, K. (2001). Contextual effects by short range connections in a mean-field model of v1, *Neurocomputing* **38-40**: 475–481.

Bartsch, H., Stetter, M., Weber, C. and Obermayer, K. (2000b). Influence of the geometry of lateral connections in v1 on orientation selectivity: A model study, *Soc. Neurosci. Abstr.*, Vol. 26.

Bell, A. J. and Sejnowski, T. J. (1996). Edges are the 'independent components' of natural scenes, *Advances in Neural Information Processing Systems 9*.

Ben-Yishai, R., Bar-Or, R. L. and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex., *Proc. Natl. Acad. Sci. USA* **92**: 3844–3848.

Ben-Yishai, R., Hansel, D. and Sompolinsky, H. (1997). Traveling waves and the processing of weakly tuned inputs in a cortical network module., *J. Comput. Neurosci.* **4**: 57–77.

Bigün, J. (1988). Recognition of local symmetries in gray level images by harmonic functions, *International Conference on Pattern Recognition* pp. 345–347.

Billock, V. A. (2000). Neural acclimation to $1/f$ spatial frequency spectra in natural images transduced by the human visual system, *Physica D* pp. 379–391.

Blakemore, C. and Tobin, E. A. (1972). Lateral inhibition between orientation detectors in the cat's visual cortex., *Exp. Brain Res.* **15**: 439–440.

Blasdel, G. G., Obermayer, K. and Kiorpes, L. (1995). Organization of ocular dominance and orientation columns in the striate cortex of neonatal macaque monkeys., *Vis. Neurosci.* **12**: 589–603.

Blum, H. and Nagel, R. N. (1978). Shape description using weighted symmetric axis features, *Pattern Recognition* **10**: 167–180.

Bosking, W. H., Zhang, Y., Schofield, B. and Fitzpatrick, D. (1997). Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex, *J. Neurosci.* **17**(6): 2112–2127.

Brady, M. and Asada, H. (1984). Smoothed local symmetries and their implementation, *Int. J. Robotics Research* **3**: 36–61.

Bressloff, P. C. and Cowan, J. D. (2002). An amplitude equation approach to contextual effects in visual cortex, *Neural Computation* **14**: 493–525.

Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J. and Wiener, M. C. (2001). Geometric visual hallucinations, euclidean symmetry, and the functional architecture of striate cortex, *Phil. Trans. Royal Soc. London B* **356**: 299–330.

Bringuir, V., Chavane, F., Glaeser, L. and Frégnac, Y. (1999). Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons, *Science* **283**: 695–99.

Bruckstein, A. M. and Shaked, D. (1995). Skew symmetriy detection via invariant signatures, *Proc. 6th nt. conf. on Comput. Analysis of Images and Patterns (CAIP), Prague* pp. 17–24.

Bugmann, G. (1997). Biologically plausible neural computation, *Biosystems* **40**: 11–19.

Buntine, W. (1994). Operations for learning with graphical models, *Journal of Artificial Intelligence Research* **2**: 159–225.

Carandini, M. and Ferster, D. (2000). Membrane potential and firing rate in cat primary visual cortex, *J. Neuroscience* **20**: 470–84.

Carandini, M. and Rigach, D. L. (1997). Predictions of a recurrent model of orientation selectivity, *Vision Res.* **37**(21): 3061–3071.

Chaitin, G. J. (1966). On the length of programs for computing finite binary sequences by bounded-transfer turing machines, *AMS Notices* **13**: 133.

Cham, T. J. and Cipolla, R. (1995). Symmetry detection through local skewe symmetries, *Image and Vision Computing* **13**: 439–450.

Chapman, B. and Stryker, M. P. (1993). Development of orientation selectivity in ferret visual cortex and effects of deprivation, *J. Neurosci.* **13**: 5251–5262.

Chen, S., Donoho, D. L. and Sanders, M. A. (1996). Atomic decomposition by basis pursuit, *Technical report, Dept. Stat., Stanford Univ., Stanford, CA*.

Chung, S. and Ferster, D. (1998). Strength and orientation tuning of the thalamic input to simple cells revealed by electrially evoked cortical suppression, *Neuron* **20**: 1177–89.

Cleland, B. G., Lee, B. B. and Vidyasagar, T. R. (1983). Response of neurons in the cat's lateral geniculate nucleus to moving bars of different length, *J. Neurosci.* **3**: 108–116.

Coifman, R. R. and Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection, *IEEE Transactions on Information Theory* **38**: 713–718.

Connor, F. R. (1982). Noise, *Edward Arnold, London*.

Corbalis, C. and Roldan, E. (1974). On the perception of symmetrical and repeated patterns, *Perception and Psychophysics* **16**: 136–142.

Cormack, E. O. and Cormack, R. H. (1974). Stimulus configuration and line orientation in the horizontal-vertical illusion., *Perception and Psychophysics* **16**: 208–212.

Crair, M. C., Gillespie, D. C. and Stryker, M. P. (1998). The role of visual experience in the development of columns in cat visual cortex, *Science* **279**: 566–570.

Dakin, S. C. and Herbert, A. M. (1998). The spatial region of integration for visual symmetry detection, *Proceedings of the Royal Society of London, B265* pp. 659–664.

Das, A. and Gilbert, C. D. (1999). Topography of contextual modulations mediated by short-range interactions in primary visual cortex, *Nature*.

Daugman, J. D. (1988). Complete discrete 2-d gabor transforms by neural networks for image analysis and compression, *IEEE Trans. Acoustics, Speech, and Signal Processing* **36**: 1169–1179.

Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience*, MIT Press.

de Valois, R. L. and de Valois, K. K. (1988). Spatial vision, *New York Oxford Univ. Press*.

DeAngelis, G. C., Freeman, R. D. and Ohzawa, I. (1994). Length and width tuning of neurons in the cat's primary visual cortex, *J. Neurophysiol.* **71**: 347–74.

DeAngelis, G. C., Ghose, G. M., Ohzawa, I. and Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons, *The Journal of Neuroscience* **19**: 4046–4064.

DeAngelis, G. C., Ohzawa, I. and Freeman, R. D. (1995). Receptive-field dynamics in the central visual pathways, *Trends Neurosci.* **18**: 451–458.

DeAngelis, G. C., Robson, J. G., Ohzawa, I. and Freeman, R. D. (1992). Organization of suppression in receptive fields of neurons in cat visual cortex, *J. Neurophysiol.* **68**(1): 144–63.

deBoer, E. and Kuyper, P. (1968). Triggered correlation, *IEEE Transact. Biomed. Eng.* **15**: 169–179.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm, *J. Royal Star. Soc., Series B* **39**: 1–38.

Dobbins, A., Zucker, S. W. and Cynader, M. S. (1987). Endstopped neurons in the visual cortex as a substrate for calculating curvature, *Nature* **329**: 438–441.

Eglen, S., Bray, A. and Stone, J. (1997). Unsupervised discovery of invariances, *Network: Comput. Neural Syst.* **8**: 441–452.

Einhäuser, W., Kayser, C., König, P. and Körding, K. P. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli, *European Journal of Neuroscience* **15**: 475–486.

Ernst, U., Pawelzik, K., Tsodyks, M. and Sahar-Pikielny, C. (2000). Spontaneous emergence of orientation preference and direction selectivity through lateral intracortical interactions, *Neurocomputing*.

Eysel, U. T., Shevelev, I. A., Lazareva, N. A. and Sharaev, G. A. (1998). Orientation tuning and receptive field structure in cat striate neurons during local blockade of in tracortical inhibition., *Neuroscience* **84**: 25–36.

Felisberti, F. and Derrington, A. (2001). Long-range interactiosn in the lateral geniculate nucleus of the new-world monkey, callithrix jacchus, *Vis. Neurosci.* **18**: 209–218.

Ferster, D. and Miller, K. D. (2000). Neural mechanisms of orientation selectivity in the visual cortex, *Annual Reviews of Neuroscience*.

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**(12): 2379–2394.

Field, D. J. (1994). What is the goal of sensory coding, *Neural Computation* **6**: 559–601.

Fitzpatrick, D., Lund, J. S. and Blasdel, G. G. (1985). Intrinsic connections of macaque striate cortex: Afferent and efferent connections of layer 4C, *J. Neurosci.* **5**: 3329–3349.

Földiàk, P. (1991). Learning invariance from transformation sequences, *Neural Comput.* **3**: 194–200.

Földiàk, P. (2001). Stimulus optimisation in primary visual cortex, *Neurocomputing* **38-40**: 1217–1222.

Frégnac, Y. and Imbert, M. (1984). Development of neuronal selectivity in the primary visual cortex of the cat, *Physiol. Rev.* **64**: 325–434.

Gabor, D. (1946). Theory of communication, *J. IEE* **72**: 429–459.

Gallant, J. L., Essen, D. C. V. and Nothdurft, H. C. (1995). Two-dimensional and three-dimension texture processing in visual cortex of the macaque monkey, *in* T. Papathomas, C. Chubb, A. Gorea and E. Kowler (eds), *Early Vision and Beyond*, MIT Press, pp. 89–98.

Geman, D. and Koloydenko, A. (1998). Invariant statistics and coding of natural microimages, *CVPR 99, SCTV workshop.*

Gilbert, C. D. and Wiesel, T. N. (1990). The influence of contextul stimuli on the orientation selectivity of cells in primary visual cortex of the cat, *Vision-Res.* **30**(11): 1689–701.

Gisiger, T. (2001). Scale invariance in biology: coincidence or footprint of a universal mechanism?, *Biol. Rev.* **76**: 161–209.

Gödecke, I. and Bonhoeffer, T. (1996). Development of identical orientation maps for two eyes without common visual experience., *Nature* **379**: 251–254.

Gofman, Y. and Kiryati, N. (1996). Detecting symmetry in grey level images: The global optimization approach, *ICPR* p. A94.2.

Gross, C. G., Desimone, R., Albright, T. D. and Schwartz, E. L. (1985). Inferior temporal cortex and pattern recognition, *Experimental Brain Research* **11**: 179–201.

Hansel, D. and Sompolinsky, H. (1998). Modeling feature selectivity in local cortical circuits., *in* C. Koch and I. Segev (eds), *Methods in Neural Modeling*, MIT Press, Cambridge MA., chapter 13, pp. 499–567.

Härdle, W. and Müller, M. (2000). Multivariate and semiparametric kernel regression, *in* M. G. Schimek (ed.), *Smoothing and Regression: Approaches, Computation, and Application*, Whiley, Europe.

Hartline, H. K. (1940). The receptive fields of optic nerve fibers, *Am. J. Physiol.* **130**: 690–699.

Hegdé, J. and Essen, D. C. V. (1999). Selectivity for complex shapes in primate visual area v1, *Soc. Neurosci. Abstr.* **25**: 1548.

Hegdé, J. and Essen, D. C. V. (2000). Selectivity for complex shapes in primate visual area v2, *J. Neurosci.* **0**: RC61 (1–6).

Hoffman, W. C. (1965). The neuron as a lie group germ and a lie product, *Quarterly Journal of Applied Mathematics* **25**: 423–440.

Hubel, D. H. (1995). *Eye, brain, and vision*, Scientific American Library.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex., *J. Physiol.* **160**: 106–154.

Hubel, D. H. and Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex., *Proc. R. Soc. Lond. B* **198**: 1–59.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis., *IEEE Transactions on Neural Networks* **10**(3): 626–634.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces, *Neural Comput.* **12**: 1705–1720.

Hyvärinen, A. and Inki, M. (2002). Estimating overcomplete independent component bases for image windows, *Mathematical Imaging and Vision, in press*.

Jacquin, A. E. (1990). Fractal image coding based on a theory of iterated contractive image transformations, *SPIE: Visual Communications and Image Processing* **1360**: 227–239.

Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in the cat striate cortex, *J. Neurophysiology* **58**: 1,233–1,258.

Julesz, B. (1971). Foudations of cyclopean perception, *University of Chicago Press, Chicago*.

Kapadia, M. K., Ito, M., Gilbert, C. D. and Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys, *Neuron* **15**: 843–856.

Kapadia, M. K., Sigman, M. and Gilbert, C. D. (1999a). Lateral interactions in cortical area V1 and their role in perception., *Soc. Neurosci. Abstr.* **25**(1): 1049.

Kapadia, M., Sigman, M. and Gilbert, C. (1999b). Lateral interactions in cortical area v1 and their role in perception, *Soc. Neurosci. Abstr.*, Vol. 25, Part 1, p. 1049.

Kaschube, M., Wolf, F., Geisel, T. and Löwel, S. (2002). Genetic influence on quantitative features of neocortical architecture, *The Journal of Neuroscience* **22**: 7206–7217.

Kent, J. T. (1978). Limiting behaviour of the von mises-fisher distribution., *Math. Proc. Cambridge Phil. Soc.* **84**: 531–536.

Kohonen, T. (1995). *Self-Organizing Maps*, Springer Verlag, Berlin, Heidelberg, New York.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information, *Problems Inform. Transmission* **1**(1): 1–7.

Koulakov, A. A. and Chklovskii, D. B. (2001). Orientation preference patterns in mammalian visual cortex: A wire length minimization approach, *Neuron* **29**: 519–527.

Kuffler, S. W. (1953). *J. Neurophysiol.* **16**: 37–68.

Laughlin, S. B. and Attwell, D. (2000). An energy budget for glutamatergic signalling in grey matter of the rat cerebral cortex, *J. Physiol.* **525**: 61.

Lee, J.-H., Jung, H.-J., Lee, T.-W. and Lee, S.-Y. (2000). Speech feature extraction using independent component analysis, *IEEE International Conference on Acoustics, Speech and Signal Processing* **III**: 1631–4.

Lee, T. S. (1996). Image representation using 2d gabor wavelets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(10): 959–971.

Levitt, J. B. and Lund, J. S. (1997). Contrast dependence of contexual effects in primate visual cortex, *Nature* **387**: 73–76.

Lewicki, M. and Olshausen, B. (1998). Inferring sparse, overcomplete image codes using an efficient coding framework, *Advances in Neural Information Processing Systems 10, Proc. NIPS*97)*, pp. 815–821.

Lewicki, M. and Sejnowski, T. J. (1998). Learning overcomplete representations, *Advances in Neural Information Processing Systems 10, Proc. NIPS*97)*, pp. 556–562.

Lewicki, M. S. and Olshausen, B. A. (1999). A probabilistic framework for the adaption and comparison of image codes, *J. Opt. Soc. of Am. A: Optics, Image Science, and Vision (submitted)*.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations, *Neural Computation* **12**: 337–365.

Li, Z. (1998). A neural model of contour integration in the primary visual cortex, *Neural Compu.* **10**: 903–940.

Li, Z. (1999). Visual segmentation by contextual influences via intra–cortical interactions in the primary visual cortex, *Network: Comput. Neural Syst.* **10**: 187–212.

Lund, J. (1987a). Local circuit neurons of macaque monkey striate cortex: I. neurons of laminae 4c and 5a, *J. Comp. Neurol.* **257**: 60–92.

Lund, J. S. (1987b). Local circuit neurons of macaque monkey striate cortex: I. Neurons of laminae 4C and 5A, *J. Comp. Neurol.* **257**: 60–92.

Malach, R., Amirr, Y., Harel, M. and Grinvald, A. (1993). Relationship between intrinsic connections and functional architecture revealed by optical imagin and in vivo targeted biocytin injections in primate striate cortex, *Proc. Natl. Acad. Sci. USA* **90**(22): 10469–73.

Mallat, S. G. and Zhang, Z. F. (1993). Matching pursuits with time-frequency dictionaries, *IEEE Transactions on Signal Processing* **41**: 3397–3415.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*, Wiley series in probability and statistics.

Marshall, A. (1956). The use of multi-stage sampling schemes in monte carlo computations, *in* M. Meyer (ed.), *Symposium on Monte Carlo Methods*, New York: Wiley, pp. 123–140.

Masame, K. (1983). Detection of symmetry in complex patterns: Is symmetrical projections to the visual system necessary for the perception of symmetry, *Tohoku Psychologia Folia* **42**: 27–33.

Masuda, T., Yamamoto, K. and Yamada, H. (1993). Detection of partial symmetry using correlation with rotated-reflected images, *Pattern Recognition* **26**: 1245–1253.

Mato, G. and Sompolinsky, H. (1996). Neural network models of perceptual learning of angle discrimination, *Neural Computation* **8**: 270–299.

Mel, B. W. (1994). Information processing in dendritic trees, *Neural Computation* **6**: 1031–1085.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines, *Journal of Chemical Physics* **21**: 1087–1091.

Moller, A. P. (1995). Bumblebee preference for symmetrical flowers, *Proceedings of the National Academy of Science* **92**: 2288–2292.

Morales, D. and Pashler, H. (1999). No role for colour in symmetry perception, *Nature* **399**: 115–116.

Mundel, T., Dimitrov, A. and Cowan, J. D. (1996). Visual cortex circuitry and orientation tuning., *in* M. C. Mozer, M. I. Jordan and T. Petsche (eds), *Neural Information Processing Systems*, Vol. 9, MIT Press Cambridge, MA.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization, *Comput. J.* **7**: 308–313.

Nowak, L. G. and Bullier, J. (1997). The timing of information transfer in the visual system, *in* R. et al. (ed.), *Cerebral Cortex*, Plenum Press, New York, pp. 205–241.

Ogawa, H. (1991). Symmetry analysis of line drawings using the hough transform, *Pattern Recognition Letters* **12**: 9–12.

Ohzawa, I., Sclar, G. and Freeman, R. D. (1985). Contrast gain control in the cat's visual system, *J. Neurophysiol.* **54**(3): 651–667.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images., *Nature* **381**: 607–609.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1?, *Vision Research*.

Olshausen, B. A. and Millman, K. J. (2000). Learning sparse codes with a mixture-of-gaussians prior, *in* S. A. Solla, T. K. Leen and K. R. Müller (eds), *Advances in Neural Information Processing Systems*, Vol. 12, MIT Press, pp. 841–847.

Ormoneit, D. and Tresp, V. (1996). Improved gaussian mixture density estimates using bayesian penalty terms and network averaging, *in* D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds), *Advances in Neural Information Processing Systems*, Vol. 8, The MIT Press, pp. 542–548.

Owsley, C., Sekuler, R. and Siemens, D. (1983). Contrast sensitivity throughout adulthood, *Vision Res.* **23**: 689–699.

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of str iate cortex, *Biol. Cybern.* **58**: 35–49.

Pawelzik, K. R., Ernst, U., Wolf, F. and Geisel, T. (1996). Orientation contrast sensitivity from long-range interactions in visual cortex., *in* M. C. Mozer, M. I. Jordan and T. Petsche (eds), *Neural Information Processing Systems*, Vol. 9, MIT Press Cambridge MA.

Payne, B. R. and Peters, A. (2002). *The cat primary visual cortex*, Academic Press.

Pei, X., Vidyasagar, T. R., Volgushev, M. and Creutzfeldt, O. D. (1994). Receptive field analysis and orientation selectivity of postsynaptic potentials of simple cells in cat visual cortex, *J. Neurosci.* **14**(11): 7130–40.

Peters, A. and Sethares, C. (1996). Myelinated axons and the pyramical cell modules in monkey primary visual cortex, *J. comp. Neurol.* **365**: 232–255.

Pizer, S. M., Oliver, W. R. and Bloomberg, S. H. (1987). Hierarchical shape description via the multiresolution symmetric axis transform, *IEEE Trans. Pattern Anal. Machine Intell.* **9**: 505–511.

Poirazi, P. and Mel, B. W. (2001). Impact of active dendrites and structural plasticity on the memory capacity of neural tissue, *Neuron* **29**: 779–796.

Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T. and Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold, *Nature* **391**: 580–584.

Ponce, J. (1990). On characterizing ribbons and finding skewed symmetries, *Comput. Vision Graph. Image Process.* **52**: 328–340.

Rao, G. P. (1983). *Piecewise constant orthogonal functions and their application to systems and control*, Springer-Verlag, Berlin.

Rao, R. P. N. and Ballard, D. H. (1998). Development of localized oriented receptive fields by learning a translation-invariant code for natural images, *Network: Comput. Neural Syst.* **9**: 219–234.

Rao, R. P. N. and Ruderman, D. L. (1999). Learning lie groups for invariant visual perception, *in* M. S. Kearns, S. A. Solla and D. A. Cohn (eds), *Advances in Neural Info Processing Systems*, Vol. 11, MIT Press, pp. 810–816.

Reid, R. C. and Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex, *Nature* **378**: 281–284.

Reid, R. C. and Alonso, J. M. (1996). The processing and encoding of information in the visual cortex, *Current Opinion in Neurobiology* **6**: 475–480.

Reinagel, P. and Zador, A. M. (1999). Natural scene statistics at the center of gaze, *Network: Comput. Neural Syst.* **10**: 341–350.

Ringach, D. L., Sapiro, G. and Shapley, R. (1997). A subspace reverse-correlation technique for the study of visual neurons., *Vision Res.* **37**(17): 2455–64.

Rockland, K. S. and Lund, J. S. (1983). Intrinsic laminar lattice connections in primate visual cortex, *J. Comp. Neurol.* **216**(1): 303–318.

Roe, A. W., Pallas, S. L., Hahm, J. and Sur, M. (1990). A map of visual space induced in primary auditory cortex, *Science* **250**: 818–820.

Royer, F. (1981). Detection of symmetry, *Journal of Experimental Psychology: Human Perception and Performance* **7**: 1186–1210.

Ruderman, D. L. (1994). Designing receptive fields for highest fidelity, *Network* **5**: 147–155.

Ruderman, D. L. and Bialek, W. (1992). Seeing beyond the nyquist limit, *Neural Computation* **5**: 682–690.

Ruderman, D. L. and Bialek, W. (1994). Statistics of natural images – scaling in the woods, *Physical Review Letters* **73**(6): 814–817.

Sanchez-Vives, M. V., Nowak, L. G. and McCormick, D. A. (2000). Membrane mechanisms underlying contrast adaptation in cat area 17 *in vivo*, *J. Neuroscience* **20**: 4267–4285.

Schoelkopf, B., Simard, P. Y., Smola, A. and Vapnik, V. (1998). Prior knowledge in support vector kernels, *Neural Information Processing Systems Conference. MIT Press*.

Schou, G. (1978). Estimation of the concentration parameter in von mises-fisher distributions., *Biometrika* **65**: 369–377.

Schwabe, L., Adorján, P. and Obermayer, K. (2000). Spike–frequency adaptation as a mechanism for dynamic coding in v1, *Neurocomputing* **38-40**: 351–358.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding, *Vision Research* **20**: 645–669.

Sclar, G. and Freeman, R. D. (1982). Invariance of orientation tuning with stimulus contrast, *Exp. Brain Res.* **46**: 457–461.

Sengpiel, F., Sen, A. and Blakemore, C. (1997). Characteristics of surround inhibition in cat area 17., *Exp. Brain Res.* **116**: 216–228.

Sengpiel, F., Stawinski, P. and Bonhoeffer, T. (1999). Influence of experience on orientation maps in cat visual cortex, *Nat. Neurosci.* **2**: 727–732.

Shannon, C. and Weaver, W. (1948). A mathematical theory of communication, *Champaign, IL: University of Illinois Press* pp. 341–350.

Shapley, R. and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls, *in* N. N. Osborne and G. J. Chader (eds), *Progress in Retinal Research*, Vol. 3, Oxford: Pergamon Press, pp. 263–346.

Shevelev, I. A. (1999). What image characteristics are selected by neurons in the cat primary visual cortex?, *Ross Fiziol Zh Im I M Sechenova* **85**: 767–80.

Shouval, H. Z., Goldberg, D. H., Jones, J. P., Beckerman, M. and Cooper, L. N. (2000). Structured long–range connections can provide a scaffold for orientation maps, *J. Neurosci.* **20(3)**: 1119–1128.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudelro, J. and Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities, *Nature* **378**(378): 492–496.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York.

Simard, P., LeCun, Y. and Denker, J. (1993). Efficient pattern recognition using a new transformation distance, *Advances in Neural Information Processing Systems V*, pp. 50–58.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H. and Heeger, D. J. (1992). Shiftable multi-scale transforms, *IEEE transactions on informations theory*.

Singer, W. (1981). Topographic organization of orientation columns in the cat visual cortex, *Exp. brain Res.* **44**: 431–436.

Skottun, B. C., Freeman, R. D., Sclar, G., Ohzawa, I. and Freeman, R. D. (1987). The effects of contrast on visual orientation and spatial frequency discrimination: a comparison of single cells and behavior, *J. Neurophysiol.* **57**: 773–786.

Solomon, S. G., White, A. J. R. and Martin, P. R. (2002). Extraclassical receptive field properties of parvocellular, magnocellular, and koniocellular cells in the primate lateral geniculate nucleus, *The Journal of Neuroscience* **22**: 338–349.

Solomonoff, R. J. (1997). The discovery of algorithmic probability, *Journal of Computer and System Sciences* **55**(1): 73–88.

Somers, D. C., Nelson, S. B. and Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells, *J. Neurosci.*

Soodak, R. E., Shapley, R. M. and Kaplan, E. (1987). Linear mechanism of orientation tuning in the retina and lateral geniculate nucleus of the cat, *J. Neurophysiol.* **58**: 267–275.

Stephens, M. A. (1963). Random walk on a circle, *Biometrika* **50**: 385–390.

Stetter, M., Adorján, P., Bartsch, H. and Obermayer, K. (1998). Modeling contrast adaptation and contextual effects in primary visual cortex, *The Fifth International Conference on Neural Information Processing, ICONIP 98-Kitakyushu* **2**: 669–672.

Stetter, M., Bartsch, H. and Obermayer, K. (2000a). A mean field model for orientation tuning, contrast saturation and contextual effects in area 17, *Biol. Cybern.* **82**: 291–304.

Stetter, M., Bartsch, H. and Obermayer, K. (2000b). A mean field model for orientation tuning, contrast saturation and contextual effects in the primar y visual cortex., *Biol. Cybern.* **82**: 291–304.

Stratford, K. J., Tarczy-Hornoch, K., Martin, K. A. C., Bannister, N. J. and Jack, J. J. B. (1996). Excitatory synaptic inputs to spiny stellate cells in cat visual cortex, *Nature* **382**: 258–261.

Takeuchi, A. and Amari, S. (1979). Formation of topographic maps and columnar microstructures in nerve fields, *Biol. Cybern.* **35**: 63–72.

Tavazoie, S. F. and Reid, R. C. (2000). Diverse receptive fields in the lateral geniculate nucleus during thalamocortical development, *Nature Neurosci.* **3**: 608–616.

Todorov, E., Siapas, A. and Somers, D. (1996). A model of recurrent interactions in primary visual cortex., *in* M. C. Mozer, M. I. Jordan and T. Petsche (eds), *Advances in Neural Information Processing Systems*, Vol. 8, MIT Press Cambridge, Massachusetts.

Troyer, T. W., Krukowski, A. E. and Miller, K. D. (2002). Lgn input to simple cells and contrast-invariant orientation tuning: an analysis, *J. Neurophysiol* **87**: 2741–2752.

Tsodyks, M. V. and Sejnowski, T. (1995). Rapid state switching in balanced cortical network models., *Network* **6**: 111–124.

Turiel, A. and Parga, N. (2000). The multifractal structure of contrast changes in natural images: from sharp edges to textures, *Neural Computation* **12**: 763–793.

Valois, R. L. D., Albrecht, D. G. and Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex, *Vision Research* **22**: 545–559.

Versavel, M., Orban, G. A. and Lagae, L. (1990). Response of visual cortical neurons to curved stimuli and chevrons, *Vision Res.* **30**: 235–248.

Volgushev, M., Pernberg, J. and Eysel, U. T. (2000). Comparison of the selectivity of postsynaptic potentials and spike responses in cat visual cortex., *Eur. J. Neurosci.* **12**: 257–263.

Volgushev, M., Vidyasagar, T. R. and Pei, X. (1995). Dynamics of the orientation tuning of postsynaptic potentials in the cat visual cortex, *Visual Neuroscience* **12**: 621–28.

Walk, R. D. (1978). Perceptual lerning, *in* E. C. Carterette and M. P. Friedman (eds), *Handbook of Perception*, Vol. IX, Academic Press, New York, pp. 257–298.

Willmore, B. and Tolhurst, D. J. (2001). Characterizing the sparseness of neuronal code, *Network: Computations in Neural Systems* **12**: 255–270.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: unsupervised learning of invariances, *Neural Computation* **14**: 715–770.

Xu, L., Cheung, C., Yang, H. and Amari, S. I. (1997). Independent component analysis by the information-theoretic approach with mixture of densities, *in* P. B. Watta, M. Akkal and M. H. Hassoun (eds), *Proceedings of the IEEE International Conference on Neural Networks, ICNN'97*, Vol. 3, pp. 1821–1826.

Ylä-Jääski, A. and Ade, F. (1996). Grouping symmetrical structures for object segmentation and description, *Computer Vision and Image Understanding* **63**: 399–417.

Yoshioka, T., Blasdel, G. G., Levitt, J. B. and Lund, J. S. (1992). Patterns of lateral connections in macaque visual are v1 revealed by biocytin histochemistry and functional imaging, *Soc. Neurosci. Abstr.* **18**: 299.

Yoshioka, T., Blasdel, G. G., Levitt, J. B. and Lund, J. S. (1996). Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex, *Cerebral Cortex* **6**(6): 297–310.

Yousef, T., Bonhoeffer, T., Kim, D. S., Eysel, U. T., Tót, E. and Kisvárday, Z. F. (1999). Orientation topography of layer 4 lateral networks revealed by optical imaging in cat visual cortex (area 18), *Eur. J. Neurosci.* **11**: 4291–4308.

Zabrodsky, H., Peleg, S. and Avnir, D. (1995). Symmetry as a continuous feature, *IEEE Transactions on pattern analysis and machine intelligence*.

Zemel, R. S. and Pillow, J. (2000). Encoding multiple orientations in a recurrent network, *in* J. M. Bower (ed.), *Computational Neuroscience: Trends in Research*, Elsevier Science, pp. 609–616.

Zhu, S. C., Wu, Y. N. and Mumford, D. (1997). Minimax entropy principle and its application to texture modeling, *Neural Computation* **9**: 1627–1660.

Zohary, E. (1992). Population coding of visual stimuli by cortical neurons tuned to more than one dimension., *Biol. Cybern.* **66**: 265–272.

Index