Helsinki University of Technology
Dissertations in Computer and Information Science
Espoo 2004 Report D9

EXTENSIONS OF INDEPENDENT COMPONENT ANALYSIS FOR NATURAL IMAGE DATA

Mika Inki

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 10th of December, 2004, at 12 o'clock noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science
P.O.Box 5400
FI-02015 HUT
FINLAND

Distribution:
Helsinki University of Technology
Laboratory of Computer and Information Science
P.O.Box 5400
FI-02015 HUT
FINLAND
Tel. +358-9-451 3272
Fax +358-9-451 3277
http://www.cis.hut.fi

Available in pdf format at http://lib.hut.fi/Diss/2004/isbn9512273632/

ISBN 951-22-7362-4 (printed version)
ISBN 951-22-7363-2 (electronic version)
ISSN 1459-7020

Otamedia Oy
Espoo 2004

Inki, M. (2004): **Extensions of independent component analysis for natural image data**. Doctoral thesis, Helsinki University of Technology, Dissertations in Computer and Information Science, Report D9, Espoo, Finland.

Keywords: independent component analysis, latent variable models, natural image data, overcomplete models, topographic mapping, higher order structures.

ABSTRACT

An understanding of the statistical properties of natural images is useful for any kind of processing to be performed on them. Natural image statistics are, however, in many ways as complex as the world which they depict. Fortunately, the dominant low-level statistics of images are sufficient for many different image processing goals. A lot of research has been devoted to second order statistics of natural images over the years.

Independent component analysis is a statistical tool for analyzing higher than second order statistics of data sets. It attempts to describe the observed data as a linear combination of independent, latent sources. Despite its simplicity, it has provided valuable insights of many types of natural data. With natural image data, it gives a sparse basis useful for efficient description of the data. Connections between this description and early mammalian visual processing have been noticed.

The main focus of this work is to extend the known results of applying independent component analysis on natural images. We explore different imaging techniques, develop algorithms for overcomplete cases, and study the dependencies between the components by using a model that finds a topographic ordering for the components as well as by conditioning the statistics of a component on the activity of another. An overview is provided of the associated problem field, and it is discussed how these relatively small results may eventually be a part of a more complete solution to the problem of vision.

Acknowledgements

This work has been carried out in the Laboratory of Computer Science and Engineering at Helsinki University of Technology. The main funding for this work came from the Helsinki Graduate School in Computer Science and Engineering, and additional grants came from the Finnish Foundation of Technology (twice), Emil Aaltonen Foundation, and Jenny and Antti Wihuri Foundation.

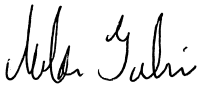
First of all, I would like to thank Docent Aapo Hyvärinen for his invaluable guidance and collaboration in the work presented in this thesis. I would also like to thank Professor Erkki Oja for his supervision, and for creating a pleasant working environment. Both of these men made time for my problems, big or small, when I needed it. I also thank Dr. Patrik Hoyer for the collaboration and friendly discussions.

I also thank Professors Heikki Hyötyniemi and Michael Lewicki for their comments on the manuscript. These comments improved the manuscript despite the brevity of time I had for incorporating them.

I also extend my gratitude towards Mr. Ville Könönen, Dr. Markus Koskela, and Dr. Jarmo Hurri for all the discussions and help in various matters, relating to work or not. Many other people from the laboratory also deserve big thanks. Furthermore, Mr. Mika Pollari deserves thanks for his help with MRI data.

I also need to thank all of my friends, my parents and brother, and everybody else who feels like they deserve gratitude. You know who you are, give yourselves a pat on the back.

Otaniemi, November 2004.



Mika Inki

Contents

Abbreviations

BSS	blind signal separation
cdf	cumulative distribution function
DCT	discrete cosine transformation
DFT	discrete Fourier transformation
GTM	generative topographic mapping
IC	independent component
ICA	independent component analysis
ISA	independent subspace analysis
JPG	JPEG, Joint Photographic Experts Group, an image compression method
LGN	lateral geniculate nucleus
MEG	magnetoencephalogram
MLP	multilayer perceptron
MRI	magnetic resonance imaging
PCA	principal component analysis
pdf	probability density function
SOM	self-organizing map
TICA	topographic independent component analysis
TIF	TIFF, Tagged Image Format File, an uncompressed image file format
V1	primary visual cortex

Some Mathematical Notations

non-boldface Latin or Greek letter	scalar, constant, scalar function, or random variable
boldface lowercase	column or row vector, vector valued function, or vector valued random variable
boldface uppercase	matrix, matrix valued function, or random variable
$\mathbf{A}$	m by n ICA mixing matrix
$\mathbf{a}_i$	i:th column of the mixing matrix, often referred to as basis vector or feature
a_{ij}	element on the i:th column and j:th row of $\mathbf{A}$
$\mathbf{C}$	covariance matrix
$\mathbf{D}$	diagonal matrix with the eigenvalues of $\mathbf{C}$
d_i	standard deviation of the i:th PCA component, i.e. d_i^2 is the i:th diagonal element of $\mathbf{D}$
$\mathbf{E}$	matrix of eigenvectors
$E\{.\}$	expected value over all parameters
$E_\alpha\{x\}$	expected value of x over α, when α is not the only parameter of x
I	index set or image
$\mathbf{I}$	identity matrix
$\mathbf{1}$	a matrix of all ones
kurt(.)	kurtosis of the input variable
mean(.)	mean value of the input variable, same as $E\{.\}$
m	number of observed signals in the ICA model
n	number of sources in the ICA model
$p(.)$	probability density function, $p(x) = p_x(x)$
$p_x(y)$	value of the probability density function for random variable x at y
$\mathbf{s}$	the sources in the ICA model, a vector valued random variable
$\mathbf{s}(t)$	source vector for the t:th data sample
s_i	i:th element of $\mathbf{s}$, i.e. the i:th source
T	number of samples in a data set
τ	indexes of a data set, $\tau = \{1, \ldots, T\}$
var(.)	variance of the input variable
$\mathbf{W}$	ICA separating matrix
$\mathbf{x}$	observed data
$\mathbf{y}$	linearly separated components, often estimates of the sources $\mathbf{s}$
$\mathbf{z}$	whitened or otherwise normalized data

Chapter 1

Introduction

1.1 Motivation and overview

This thesis discusses natural image statistics. An understanding of the statistical properties of images is important for any subsequent processing, such as efficient image coding, restoration and analysis. To put these tasks into context, one must understand what purpose they serve, and why the processing of images is useful. Simply put, we have vision, because it gives us information of our surroundings, helping us in our evolutionary struggle.

Images and image statistics as such are not the end products of a well-defined process. The statistics of images are defined by the properties of the observer, as well as those of the surrounding world. The objects that are important to identify from this visual data are linked to the objectives of the observer. Thus artificial intelligence may be necessary not only to solve many of the problems relating to image understanding, but also to fix the statistics and justify the problems themselves.

In this thesis, we will give a somewhat broad overview of the problem field relating to vision. We might have chosen to not discuss many of the viewpoints, and merely assume that goals such as finding sparse representations, efficient image coding, and noise removal are intuitively well enough motivated, simply because a need for efficiently storing and conveying image data exists in modern society.

However, this larger perspective opens interesting connections between natural image statistics and biological visual systems. It is possible to obtain information of the early processing of visual information from the brains of different mammalian species, and it has been found that this processing has clear links to the low-level statistics of natural images. It is thus apparent that, up to a point, the visual processing in a human brain does not need to be concerned with higher level objectives or actual objects in images, and is driven by the need to find efficient representations and dominant (statistical) image structures.

The work in this thesis explores some statistical models that have been found to be in relatively good correspondence with the early visual processing on the cortex. Especially we

concentrate on independent component analysis, and extending it to better match natural image statistics. It must be stressed that our aim with these extensions is not primarily biological, i.e. to build models that are in every way biologically sensible. Nevertheless, information known of biological visual systems can act as motivation for the usefulness of this analysis, and information extracted about image statistics can validate some of the assumptions made of biological visual systems.

1.2 Publications and structure of the thesis

This thesis consists of an introductory part and seven publications. Chapter 2 of the introductory part discusses natural image statistics. Chapter 3 introduces independent component analysis, discusses its properties and the basic results when it is applied to natural images. Chapter 4 presents extensions of independent component analysis for natural images, including the results obtained in the publications. Chapter 5 contains a summary of the main results in this thesis, as well as some perspective on the work and the research area in general.

The ordering of the publications is not chronological, but topical.

Publication I. Mika Inki. ICA features of image data in one, two and three dimensions. *Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 861–866, Nara, Japan, 2003.

In this work we explored the differences between image statistics, and ICA features in particular, for ordinary photographic natural images and magnetic resonance (MR) images. The features for MRI data were shown to be essentially three-dimensional extensions of the two-dimensional ICA features for natural photographic images.

Publication II. Mika Inki and Aapo Hyvärinen. Two approaches to estimation of overcomplete independent component bases. *Proceedings of the International Joint Conference on Neural Networks (IJCNN2002)*, pages 454–459, Honolulu, Hawaii, 2002.

In this work we presented two methods for overcomplete basis estimation. The first method used a probability distribution for the mixing matrix, that is related to a random distribution for the basis vectors. This quasi-orthogonality criterion replaces the classic criterion of orthogonality of the basis vectors in a whitened space. The second method was closely related to the method of gaussianization in projection pursuit literature. The iteration works iteratively by estimating one feature at a time using FastICA, after which the corresponding direction is gaussianized. The current author developed the methods with help from Dr. Hyvärinen, and did all the experiments. The current author also wrote the paper, while Dr. Hyvärinen provided help with editing of the paper.

Publication III. Aapo Hyvärinen and Mika Inki. Estimating overcomplete independent component bases from image windows. *Journal of Mathematical Imaging and Vision*, vol. 17, no. 2, pages 139–152, 2002.

This article expanded on Publication II, adding more experiments with image data, and with different levels of overcompleteness, all done by the current author. The article also included

a review of existing methods, done by Dr. Hyvärinen. The basic methods were the same as in Publication II. The rest of the paper was written jointly by Dr. Hyvärinen and the current author, mainly based on material by the current author.

Publication IV. Aapo Hyvärinen, Patrik O. Hoyer and Mika Inki. Topographic Independent Component Analysis. *Neural Computation*, vol. 13, no. 7, pages 1527–1558, 2001.

This article presented Topographic Independent Component Analysis (TICA), which extends independent component analysis by proposing a topographic ordering for the components by using the residual dependencies in the model. The basic ideas concerning TICA, the algorithm and derivations were by Dr. Hyvärinen and Dr. Hoyer was responsible for the experiments with natural image data. The current author was responsible for the experiments with simulated data, magnetoencephalographic (MEG) data, and estimating distributions in the generative model that produce source distributions matching those observed when using TICA on natural image data.

Publication V. Mika Inki. Examining the dependencies between ICA features of image data. *Proceedings of the International Conference on Artificial Neural Networks / International Conference on Neural Information Processing (ICANN/ICONIP 2003)*, pages 298–301, Istanbul, Turkey, 2003.

In this paper we continued exploring the dependencies between ICA components in natural image data. We studied what effect the activation of one independent component has on the statistics of other components. We found that when a new ICA basis is estimated in the data subset where a certain feature is highly active, the shapes of the features are essentially similar to those of the features in the original ICA basis. However, this conditioning does change their variances (activity levels).

Publication VI. Mika Inki. A model for analyzing dependencies between two ICA features in natural images. *Proceedings of the Fifth International Conference on Independent Component Analysis*, pages 914–921, Granada, Spain, 2004.

Similarly to Publication V, in this paper we explored the dependencies between ICA components in natural image data. We presented a model which was able to predict most of the low-order conditional statistics of a component when another component is highly active, by using information about the properties of the corresponding features.

Publication VII. Mika Inki. Natural image patch statistics conditioned on activity of an independent component. Publications in Computer and Information Science A79, Helsinki University of Technology, Espoo, Finland, 2004.

This article greatly expanded on the previous two papers, offering a somewhat comprehensive look at the dependencies between independent components. We examined how the (first and second order) statistics of the data change when a component is active, and the data has been properly normalized, and how much of the observed changes can be attributed to random variation. We also touched upon how to expand the features beyond the edges of the image patches.

Chapter 2

Image statistics

A digital image is a rectangular array of pixel intensities, mathematically equivalent to a matrix. Although any real-valued matrix can be understood as a grayscale image, all such images are not equally likely to represent a part of the natural, non-artificial world. Actually, the differences in probability are so great for any image larger than a hundred by a hundred pixels, that the probability of generating something resembling a natural scene by independently picking the values of the pixels from a (uniform) distribution is practically zero. These differences in likelihood indicate that images have structures, or redundancy. All processing of images uses these structures, and analysis of images has only limited usefulness without at least an implicit understanding of them, and of their underlying causes.

2.1 Problem definition

The problem field of vision, and image capturing in particular, has been discussed in many books on digital image processing [39, 107]. We will discuss the problem field from our own viewpoint here.

One can divide the problem field relating to images into five parts, see Figure 2.1. The first entity relevant to images is, naturally, the world from which the images are taken. The second is obtaining the images from the world, i.e. defining observer position and how the camera (eye) functions. These two parts are necessary to obtain images (image statistics), the third part of our segmentation. The fourth stage is low-level analysis of the images, obtaining a representation useful for further analysis. This stage may involve edge extraction, redundancy reduction (compression), etc. The final stage is high-level analysis of the images, i.e. image understanding, whereupon one obtains an understanding of the objects in the images as they relate to physical objects in the world. One can hardly expect to obtain such an understanding from a single image, so background information (image statistics) is needed.

We will have to assume here that the world is pretty much given, so we will only superficially

Figure 2.1: Problem setting. From left to right: Particle system known as the world. Observation device in the world (and part of the world). Image produced by the observation device. Low-level analysis of the image, taking advantage of image statistics. Image understanding, i.e. extracting information relevant to the observer from the image.

cover some essential aspects of it. However, the latter parts of this segmentation depend on the observer, and we will discuss them. The second part produces a set of numbers or a set of neural cell activations representing the image. It can be argued that the precise nature of the representation is irrelevant wrt. further processing, as long as it contains the same information in a related format.

As the world and image acquisition define image statistics, one (or both) of them is probabilistic. It may be argued whether or not the world is inherently probabilistic, but as probabilities are useful for describing deterministic processes that are only partially observed (as with pseudorandomness [101], although the process does not have to pass any tests for randomness here), or for simplifying a complex process, a probabilistic description of the world is convenient. We assume here that the position of the observer is random, given some constraints (or probabilities). If we attached information to each image on when, where, and how it was taken, speaking of probabilities would be more problematic as there could essentially be only one image associated with each parameter combination. Probabilities might then make sense in simplifying the structure (history of the world), or in predicting missing (or future) values. We will for the most part consider images without attaching information of the capturing process, and even the ordering of the images can be considered random. There is also measurement noise in the images.

The distinction between the different levels of image analysis can be somewhat vague. But the results and methods in this thesis belong for all intents and purposes to low-level analysis. In this low-level image analysis, the precise nature of the objects is often not relevant for the methods, which is more or less true for the results presented in this thesis. At such an early stage, all the structures in images have to be considered 'important'. That is, one can only 'throw away' redundancy, while transforming the data for further analysis.

In high-level image analysis, one basically attempts to build a usable model of the world depicted in the images. This goal is related in many ways to (artificial) intelligence, and solving it in a satisfactory way may require creating human-like intelligence and wisdom (including background information and a set of objectives for which the extracted information is to be used). However, in order to fully describe image statistics, one would probably have to describe (model) the entire world. Thus the task is as such infeasible. But less accurate models can still be useful for achieving any goal in biological or machine vision. Some useful approaches to various problems related to image understanding are discussed in [20, 28, 107].

An important task for artificial intelligence is to (only) extract relevant information from images, and not to model every grain of sand on the ground. This process of discarding superfluous information is generally referred to as attention [90].

We will here be discussing *natural images*. By natural images we mean that they depict natural scenes, i.e. they are taken of the real world. Sometimes the concept of natural images implies that image acquisition is also fixed to produce images somewhat comparable to what the human visual system receives or has received during human evolution. We conform to this definition in this thesis. We will, however, also explore different imaging processes in this work, and see in what way the results are comparable.

It must be noted that the five parts of the problem have interactions. The interaction between image acquisition, the image itself and subsequent analysis are quite obvious, as they all relate to the structure of the observer, and image acquisition and low-level analysis must serve high-level image analysis. Humans with higher-level image analysis capabilities can (and do) change the world even during the image capturing process, so the interaction between the observer and observations extends beyond parameters relating to observer position and camera (eye) functionality. But the images we use in this thesis do not have (obviously) man-made structures, so we will not explore this viewpoint further. It can be argued, however, that the effect of human activity on the low-level feature analysis techniques and results would be limited.

2.2 Image capturing

According to current understanding, the world consists of subatomic particles and related wave-functions, interacting with each other according to the (possibly undiscovered) laws of physics in a structure called spacetime. For the purpose of this thesis, it is useful to refer to groups of these particles as objects with different shapes, surface textures and materials, emitting, absorbing and reflecting light in different ways.

However, the concept of an object is somewhat vague, as objects can often be seen as consisting of even smaller objects, possibly down to the particle level, or of being parts of larger ones. For example, it is possible in some cases to consider a single strand of hair on the human head as an object, or the hair, or the head, or the entire human. Furthermore, two objects can partially overlap, and the division between them can be somewhat arbitrary. The concept of an object is therefore partially linked to the attention and objectives of the observer. Here it is useful to refer to a collection of particles as an object when their presence (location) and appearance is mostly independent of other objects. This definition is somewhat unsatisfactory, but as objects are only convenient abstractions, pursuing a better definition is not productive in the context of this thesis.

Part of the light emitted or reflected from these objects travels to the camera lens (human eye), where it is projected onto the film/photosensor (retina), on which it forms an image. This much of the normal photographic process (human sight) can be stated without going into details about sensor type (sensitivity, timing) and projection properties (lenses, focus), etc.

However, observer position and gaze direction alone can change image statistics considerably. If one were to assign coordinates for the camera randomly in the universe, nearly all images would be stars, nebulae, etc., i.e. astronomical. It is obvious that statistics for such images would be quite different from those important for biological organisms, and even for most machine vision tasks.

Therefore, observer position and orientation must be restricted somehow. One sensible way of doing this is to give different probabilities to different positions (in spacetime), orientations and velocities. These probabilities would have to be somewhat similar to the probabilities humans have of possessing these attributes. Highest probabilities would occur when the observer was within a couple of meters of the surface of the earth, yet above it, velocity pretty much matching the velocity of the planet surface, and orientation of the gaze most likely approximately towards the horizon. The 'important parts' of the images should also be in focus. Even if the purpose of image processing is purely 'technical', e.g. image compression, similar assumptions would be made at least implicitly. These assumptions guarantee that the images mostly consist of similar features, such as edges, contours, surfaces, etc. It can be argued (cf. Section 4.2) that many of the specifics of image capturing do not matter for low-level image analysis, but most of the above considerations are quite crucial. In practice, there are few digital images that do not satisfy these considerations as people operate the capturing devices.

We have so far only discussed still grayscale images, but the same assumptions hold even for color images, stereo images or image sequences. Adding these modalities adds information for later analysis, but the underlying objects in the world do not change. For example, stereo vision adds information that helps the observer infer the three-dimensional structure of the world. Image sequences can help an observer distinguish between the different objects in a scene, and additionally help to interpret the interactions between these objects.

After the observer position and orientation are fixed in the world, a description of how the image is projected (and possibly discretized) onto a surface (retina/film/sensor) is sufficient to fix the image statistics. One must also include a description of the suboptimalities in this process, i.e. noise and projection errors.

We used two sets of natural images in the Publications of this thesis. The first set consisted of 13 grayscale images of size 512 by 256 pixels, available at the web address http://www.cis.hut.fi/projects/ica/data/images/. The second and larger set consisted of 24 grayscale images of size 512 by 384 pixels, available at http://www.cis.hut.fi/inki/images/. More of the collection process and preprocessing of this larger set of images can be found in Publication VII. The first set was used in Publications I, II, III, IV and V; the larger set was used in Publications VI and VII, and for the experiments in this introductory part. Of the first set of images, we usually sampled randomly 50000 12 by 12 image patches, and from the larger set 200000 12 by 12 image patches. Note that two patches taken from the same image at different positions can be considered as different images from a similar setting. Nevertheless, obtaining more (high quality) images would always be beneficial to the subsequent analysis.

It must be noted, that even though the aforementioned considerations are necessary to obtain what we usually consider as natural images, one can have quite different imaging processes to produce images of the world. For example, magnetic resonance imaging (MRI [45]) does not project light onto a surface, but produces a three-dimensional image, where the intensities

of the voxels (volume elements) relate to hydrogen densities at the corresponding positions in the world. Also, a grid of measurements of air pressure (or any such quantity) could be understood as an image. Even though image formation and objects in these cases are very different, low-level analysis of such images does not necessarily produce completely different results.

2.3 Approaches to image analysis

There are several ways of approaching image analysis. One approach is to describe (speculate) how biological vision works, and then attempt to emulate it. The underlying assumption is that biological visual systems have evolved to be very good at analyzing their surroundings for the obvious reason that it has helped the associated organisms to survive and reproduce. Another approach is to concentrate on the data, analyzing what structures exist in them, and build models based on that analysis.

Fortunately, these approaches are linked in many ways, not only on the data being studied, but also on the apparently useful structures or methods. As it is very hard to build an optimal system for image analysis simply by analyzing image structures, biology can act as motivation on what methods are useful. On the other hand, analysis of structures in images can guide the search for and interpretation of structures in biological visual systems. Furthermore, when images are digitally processed for later human analysis, one often has to take into account the properties of the human visual system, especially in the choice of error measures, cf. psychovisual redundancy [39].

2.3.1 Human visual system

The full workings of the human visual system are still, and will be for some time, unknown. But a lot is known of the early stages of vision, through analysis of patients with localized brain damage [26, 72], psychophysical experiments [90, 96], advanced brain imaging methods including functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG) [41], and also more direct measurements from the brains of other mammalian species [40, 49, 50, 112]. The human visual system presumably works by progressively processing information, first by analyzing small local features, finding larger (invariant) features from these step-by-step, until at the very end the person has a mental image of the objects in his/her surroundings. There are, however, different types of feedback in the process, and it can be argued whether such discrete steps exist, especially at higher levels of processing.

Neural processing of visual input begins as light is projected onto the receptive cells on the retina. Information from these photoreceptors is combined already on the retinal ganglion cells [113], typically producing center-surround receptive fields, i.e. light striking one spot (small group of photoreceptors) on the retina activates the cell, while light at surrounding areas suppresses its activity, or vice versa, see Figure 2.2. These retinal ganglion cells usually synapse in the lateral geniculate nucleus (LGN) of the thalamus, whose cells typically also have center-surround receptive fields [49, 104, 128]. These LGN cells synapse in the primary visual cortex (area V1) at the back of the brain. From V1 the information is sent to

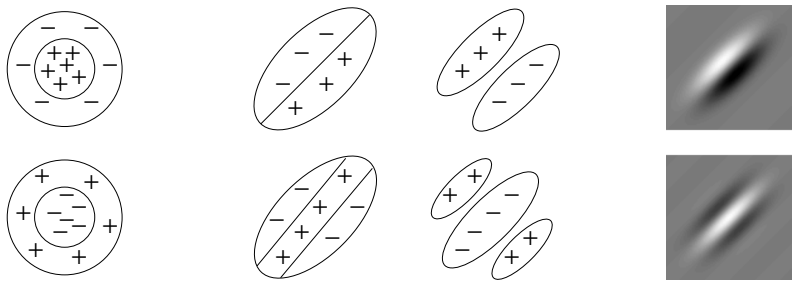


Figure 2.2: Different representations for cellural receptive fields. Left: Schematic representations of the receptive fields of retinal ganglion cells or cells in the LGN. Plus signs represent regions where light excites the cell, minuses inhibitory regions. Center: Schematic representations of cortical simple cell receptive fields. Right: Gabor functions, where zero corresponds to mid-grey, positive values to lighter areas, negative values to darker areas.

surrounding areas. A large portion of the brain is devoted solely to the processing of visual information, and an even larger portion of the cortex takes some part in its processing.

Cells in V1 have different kinds of receptive fields. *Simple cells* [10, 29] have localized, oriented receptive fields, which are best suited for edge-detection. These receptive fields are often described using two-dimensional Gabor functions [27], see Figure 2.2, or first or second order gaussian derivatives [108]. Gabor functions are able to describe many of the significant properties of receptive fields, even though the fit might not be perfect. Gabor functions (filters) are also useful in describing properties of many edge or line detector -type visual features, and we use them for various purposes in this thesis.

A two-dimensional real-valued Gabor function can be written as:

$$g(\mathbf{r}) \propto \exp(-\sum_{i=1}^{2} \frac{r_i^2}{2b_i^2}) \cos(2\pi\omega r_1 + \theta). \tag{2.1}$$

Any Gabor function can now be obtained with a rotation, translation and scaling of $g(\mathbf{r})$. One should note that only Gabor functions for which $b_2 > b_1$ and whose frequency ω is relatively small have properties similar to simple cell receptive fields, see Figure 2.2.

There are also so-called *complex cells* [50] on the V1, which no longer have receptive fields directly relating to a light pattern striking the retina. These complex cells are sometimes described using energy models, summing the squares of quadrature-phase simple cell outputs to produce a signal invariant to the absolute phase of the input [50].

It is natural that the retinal ganglion cells that process information from a group of photoreceptors is located near that group. However, also cells in V1 have *topographic ordering* [49], as information from neighboring areas on the visual field is processed in neighboring areas on the cortex. Actually processing of many different modalities of the input are topographically organized simultaneously on the cortex.

For example, the preferred orientation of simple cells is shifted gradually along the cortex,

and eye preference is organized in stripes. Higher brain areas devoted to the processing of visual information can also have topographic organization [112], and even input from other senses can be topographically organized on their dedicated areas on the cortex. This topographic ordering has been speculated to emerge partly because the length of wiring is a major concern in the brain [34, 109], and because cells tend to develop connections to nearby cells. As the connections on the brain are directly related to the processing it performs, two neurons that use the same input, and whose output is possibly needed as the input for a third neuron should tend to be located near each other on the cortex. A lot is still to be learned of the development of this topography [109, 125].

This has been a very much simplified glance at the early processing of visual information in the brain. We have not addressed how the information is transmitted between the cells (i.e. electropotential spike trains and the firing patterns of neurons), or what is known of color processing, time scales, moving images, noise properties, interconnectedness of surrounding brain areas (including feedback), how the connection strength between neurons is adjusted (see [13]) etc. For the scope of this thesis, this presentation is mostly sufficient. An accurate and complete presentation would be impossible, at least presently. More information can be found in for example [67].

2.4 Some interesting statistical properties of images and computer vision

Digital image processing (of natural images) begins when an array of numbers representing an image, obtained from either a scanner or digital camera, is inserted into the memory of a computer. A single number in this representation is called a pixel (picture element), whose value represents with some predetermined accuracy the average light intensity inside the corresponding area of the sensor plus measurement noise.

Low-level goals for digital image processing can include image compression/decompression [11, 99, 103, 123], image restoration (e.g. noise removal [58, 106]) and image enhancement (e.g. edge enhancement). Many books on digital image processing discuss these topics, e.g. [19, 39, 107]. These types of processing usually produce images that are comparable to the original, but somehow easier to interpret or store. Also, this processing usually does not throw away information, especially if the goals for further processing have not been specified. Even though one can see a link between these processing goals and early stages of biological vision, the actual methods used are often very different, at least partly for computational reasons.

2.4.1 Invariances and power spectrum

A solution for low-level problems does not need to involve descriptions of the actual objects in the world, taking advantage of low-level structures in images is often sufficient. Different types of invariances are examples of these low-level structures. Scale invariance implies that all structures in images could exist at any size [82]. For instance, if one were to capture

two sets of images from the same positions, but at different 'zoom levels', one would notice essentially similar statistical structures in both sets. Note, however, that even though a set of images may generally be scale invariant, a single image is often less so. Structures within a single image can be concentrated on only a couple of scales.

Translation (shift) and rotation invariances are also generally associated with images. Translation invariance implies that any structure at a certain position in an image could also be at any other position. Rotation invariance implies that any structure in an image can also be at any orientation.

These invariances can be seen to be related to the observer position. The observer can be at any distance from an object, so versions of every size of every object should be included in the statistics. Scale invariance can also emerge as similar objects can be at different distances from any observer position (i.e. relating to the properties of the world and not the observer). Similarly, translation and orientation invariances can be seen to be related to gaze orientation or observer position, or to actual positions of the objects in the world.

A lot of research has been devoted to second order statistics of images over the past decades. In fact, most of the basic techniques of compression (e.g. JPEG [39, 123], which we will return to later) and image restoration (e.g. Wiener filtering [39, 107]) rely on second order statistics. It has also been noted that invariances have some useful consequences for second order statistics, see for example [82]. We will discuss some of these consequences here.

Second order statistics can be described using the covariance function. For a continuous signal (image) $I(u)$, for which the expected value over all signals (images) is everywhere zero, i.e. $E_I\{I(u)\} = 0$, $\forall u$, the covariance function is $c(u, v) = E_I\{I(u)I(v)\}$. For the discretized, finite length version of the signal (image), the covariance function can be presented as a covariance matrix $\mathbf{C}_I$, where the elements are $c_{uv} = E_I\{I(u)I(v)\}$, $\{u, v\} \in \{1, \ldots, N\}$. For 2-D images, the covariance matrix is essentially four-dimensional, but the coordinates can be stacked so that it can be represented as a two-dimensional matrix.

To investigate, what translation invariance implies for the covariance of an image, we consider the eigenvalues of the continuous covariance function. Consider an operator C for which $Cf(v) = \int_{-\infty}^{\infty} c(u, v)f(u)du$. In the discretized case, the equivalent of this would be a matrix multiplication of a vector by the covariance matrix. (If we demand that $c(u, v) \in L^1$, this integral exists for all bounded f.) The eigenfunctions f of C have the property $Cf(v) = \lambda f(v)$, $\forall v$. Due to translation invariance, and symmetry of covariance, $c(u, v) = c(|u - v|)$. Let us test whether $f_{\omega,\alpha}(v) = \cos(\omega v + \alpha)$ are eigenfunctions for the corresponding operator.

$$
\begin{aligned}
Cf_{\omega,\alpha}(v) &= \int_{-\infty}^{\infty} c(|u - v|)\cos(\omega u + \alpha)du \\
&= \int_{0}^{\infty} c(x)\cos(\omega(x + v) + \alpha)dx + \int_{-\infty}^{0} c(x)\cos(\omega(v - x) + \alpha)(-dx)
\end{aligned}
$$

Here we made a substitution $x = |u - v|$. By using the trigonometric identity $\cos(s) + \cos(t) =$

$2\cos(s+t)\cos(s-t)$, we get:

$$\begin{aligned}
&= \int_0^\infty c(x)[\cos(\omega(x+v)+\alpha)+\cos(\omega(x-v)-\alpha)]dx \\
&= \int_0^\infty c(x)2\cos(\omega x)\cos(\omega v-\alpha)dx \\
&= \int_0^\infty 2c(x)\cos(\omega x)dx f_{\omega,\alpha}(v)
\end{aligned}$$

Thus $f_{\omega,\alpha}(v)$ are eigenfunctions for this translation invariant operator, and the eigenvalues do not depend on the phase α. Furthermore, it is easy to show that when considering a translation invariant covariance function of a two-dimensional signal $c(u_1,u_2,v_1,v_2) = c(|u_1-v_1|,|u_2-v_2|)$, e.g. radially symmetrically $c((u_1-v_1)^2+(u_2-v_2)^2)$, $f_{\omega_1,\alpha_1}(v_1)f_{\omega_2,\alpha_2}(v_2)$ are eigenfunctions for the corresponding operator. We will not analytically study, if all the eigenfunctions are of this form, or what effect discretization and bounds on the integration interval have for these functions.

For rectangular images, however, it appears that the eigenvectors of the correlation matrix are quite close to such products of two frequencies. In Figure 2.3 we have PCA basis vectors (eigenvectors of the correlation matrix) for image data, and discrete cosine transformation (DCT) basis vectors ordered by ascending variance. The similarity between these two bases is apparent, and explains why discrete cosine or discrete Fourier transformation (DFT) of images is so useful for many different purposes, cf. [39, 107]. However, by studying non-rectangular image patches, see Figure 2.4, we get also basis functions that appear to be products of a term depending on $\|\mathbf{u}-\mathbf{v}\|$, and a term depending on the orientation of $\mathbf{u}-\mathbf{v}$, plus possibly some other functions. This suggests that the (continuous, non-dicretized) covariance function has eigenfunctions not of the form $f_{\omega_1,\alpha_1}(v_1)f_{\omega_2,\alpha_2}(v_2)$. Note that the underlying grid is the same for both round and rectangular patches.

Another noteworthy aspect of image statistics is the power distribution of different frequencies, i.e. variances of the PCA components for rectangular images, also referred to as the power spectrum. Scale invariance implies that the average power of a frequency band is proportional to the inverted square of its spatial frequency [82]. Usually the empirical estimate for the exponent of the power law is close to minus two, but not exactly [82, 98, 117]. Even scale invariance is only an approximate feature of images, and it was shown in [98] that any power law for the distribution of object sizes leads to a power law for the power spectrum. There are some arguments that a power law for object size (e.g. scale invariance) is not necessary for a power law for the power spectrum [4], but it is nevertheless sufficient.

Note that one may study directly the shape of the covariance function instead of the power spectrum. Scale invariance (with rotational symmetry and translation invariance) implies that the covariance between two pixels depends linearly on the logarithm of the distance between them [82], which in turn implies the inverse square law for the power spectrum. Note that this type of covariance explodes when distance goes to infinity or zero. However, as a pixel essentially represents the average value in an area of nonzero measure, and images are not infinitely large, this does not have practical implications, especially if the average value is removed from each image, cf. "infra-red" blow-up [82].

The structure of the power spectrum is often used in image compression, for example in

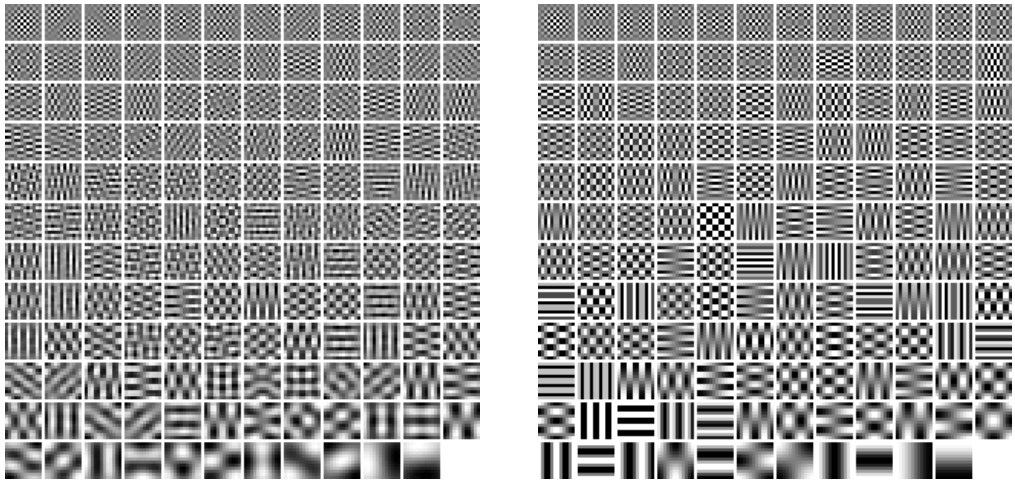


Figure 2.3: Bases convenient for the description of the power spectrum in images. Left: PCA basis obtained from 200000 12 by 12 pixel image patches. Right: Two-dimensional DCT basis for 12 by 12 image patches. Components in ascending order by variance. Note that the bottom right feature in both corresponds to patch mean.

JPEG (or JPG) [123]. It can also be shown that if the data is represented in a linear subspace, the best subspace for a large set of images in terms of squared error contains the lowest frequencies, i.e. the largest principal components. However, in compression methods such as JPG, the optimal subspace is chosen for each small image patch separately, so it uses spatial locality of image structures as well as frequency structure.

A problem with describing images by their frequencies is, indeed, that frequencies are not spatially localized. Images consist of edges (sharp changes in intensity that separate two areas of relatively small intensity variation), that are very much localized. After a DFT or DCT transformation of a single straight edge, almost all the coefficients may be nonzero. Furthermore, there are seldom such 'wavy' areas in images that DCT or DFT components would be best suited to represent.

A better transformation would use localized (in space and possibly frequency) basis vectors, only few of which would be needed to describe a single edge. This means that the distribution of such basis vectors would be sparse, i.e. most of the time they are close to zero, and only occasionally significantly nonzero. A certain subset of wavelets and Gabor functions [39, 75, 89] are such basis vectors, somewhat similarly to simple cell receptive fields. In addition to being sparse for single edges, they are also sparser for image data in general than DFT or DCT coefficients. We will discuss sparse bases for images in more detail in Section 3.5.

It has been argued that the existence of edges is sufficient for a power law in the power spectrum [4] similar to what has been observed in natural images. It can be conjectured that different invariances, and the existence of localized edges are fundamental properties of images, and the covariance structures as well as the optimal basis for representing images

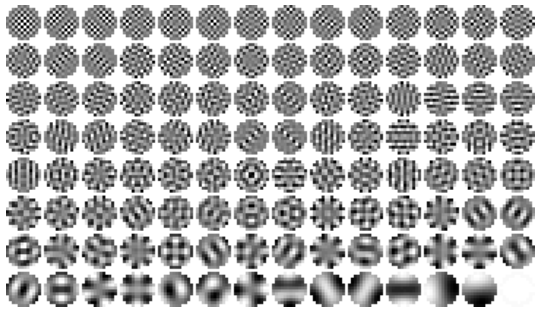


Figure 2.4: PCA basis obtained from 200000 round image patches. Center point of each patch is in the middle of four adjacent pixels. Radius is six pixels. Components in ascending order by variance. Bottom right feature corresponds to patch mean.

are mere consequences of these and possibly other fundamental properties.

An optimal linear transformation of images only reveals the most simple structures, and more complicated models are needed for higher-level analysis. Such an advanced analysis can possibly use linear transformations as a first stage. A more complete description of the edges in the scene can possibly be the next stage in the analysis. At the end of such an analysis lies an understanding of the objects in the image. Note that edges are not directly related to the objects in the world, as lighting in the scene can produce edges on flat surfaces, and there is not necessarily any visible edge between two objects.

2.4.2 Generative models

Generative models basically describe how the observed signals have been formed from a set of irreducible (probabilistic) sources. As has become apparent from our discussion of image statistics, a perfect generative model for images is quite infeasible, and less perfect models have to be used. If the generative model only attempts to describe the data as a linear transformation of independent sources (as in independent component analysis, see Chapter 3), the features are surprisingly close to simple cell receptive fields. However, generative models are often not the most computationally efficient solution to low-level image processing or analysis, even though such models have been used for these purposes [58, 77].

By using generative models one could basically generate (synthesize) images, even though these images would be far from realistic with low-level models. However, more complex models could be able to capture more of the structures in images, or successfully synthesize some special cases of images, e.g. textures [94]. As the complexity of the generative model increases, the model loses its elegance and comprehensibility. But this is arguably a fundamental property of any complex system, biological or artificial. Functionality of the system can only be described with a system (model) of comparable functional complexity (cf. Principle of Computational Equivalence, [126]), which is not to say that simplified models could not retain some of the interesting properties of a complex system.

A significant problem then is, how complex should the models be when used for image data. Again, the complexity of the 'true' model is for all intents and purposes infinite, but simpler models may be (depending on the criterion and amount of data) nearly indistinguishable from it. More complex models should be able to fit into more complex data sets or more complex structures in infinitely complex data sets, but the amount of data required to satisfactorily estimate the model increases with complexity. This is referred to as the *bias-variance dilemma*, and often there is a model of certain finite complexity that provides (in some sense) the best generalizability to unobserved data samples.

There are well-known analyses of such trade-offs between model complexity and estimation error, using such concepts as minimum description length [97], statistical learning theory [43, 120], and Bayesian inference [95]. All these methods rely on being able to quantify the complexities of a set of models using coding length, VC-dimension or prior probabilities. However, especially with image data, useful models and methods already incorporate unquantifiable amounts of prior knowledge about image structures, and it is hard to get as good results with generic models. However, this 'knowledge' is usually only approximative (e.g. scale invariance) and may result in a bias when building more complex models. Therefore, no definite answers exist about the appropriate model complexity for image data. If something useful can be learned from the model, or the model is by itself useful for some end, while still being simple enough to be manageable, it can be said to be of appropriate complexity.

Independent component analysis (ICA) is one generic technique of unsupervised learning, in which a generative model for the data is estimated. ICA is also a technique in the field of blind signal separation (BSS), as one does not essentially need to assume (know) anything else of the problem than that independent sources are mixed linearly. The model is often simple enough to identify, but not very expressive. The main point of this work is to investigate how one can extend the model for image data to better account for image structures, while maintaining comprehensibility. Consideration is given not to bias the model to find structures that are 'known' to exist in images, although any extension implicitly introduces some bias, or knowledge of preferred structures, cf. "no free lunches" [127]. Thus it may be possible to identify underlying image structures and to compare the model to what is known of biological visual systems.

Chapter 3

Independent component analysis

Independent component analysis (ICA) is a generative model, expressing how observed signals are mixed from underlying independent sources. We can express the model as

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_i \mathbf{a}_i s_i, \tag{3.1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is the vector of latent variables called the independent components (ICs) or source signals, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix [59]. The columns of $\mathbf{A}$, denoted by $\mathbf{a}_i$, are often called features or basis vectors. Usually only $\mathbf{x}$ has been observed, although some properties of the source distributions are usually assumed. The problem is to identify the mixing matrix, and possibly to identify the source distributions p_{s_i} or even to separate (recover) the source signals s_i. The problem is related to factor analysis, where one wants to find latent variables that are the sources of variability in the data. However, only second order statistics are considered in factor analysis [42].

Nongaussianity is an important concept with the ICA model, as (under some mild conditions) the sum of independent variables tends towards gaussianity by the central limit theorem [100], and the sum of (independent) gaussian variables is gaussian. Thus it is intuitive to assume that the sources are at least as nongaussian as any of their linear combinations. The model has connections to projection pursuit methods [36], where one wants to find interesting (for example nongaussian) projections of the data. Note that projection pursuit was developed originally to find directions that separate data clusters, whereas it is usually assumed with ICA that the source distributions are unimodal.

Usually one cannot find independent sources from natural data sets by a linear transformation, but one can find a linear transformation that produces the most nongaussian signals. Therefore, even though it can be argued that independence is the more fundamental concept of the two, one usually has to settle with maximal nongaussianity. This maximally nongaussian description can be useful for coding purposes, or for the information it reveals about the data structures, even if the components are not independent.

There are some fundamental indeterminacies in the ICA model. The model is identifiable

only up to a (nonzero, possibly negative) scaling of the sources and basis vectors, and a permutation of the components. Assuming that the probability densities can be estimated with sufficient accuracy (in practice, if sufficiently large sample sizes are available), the ICA mixing matrix $\mathbf{A}$ is identifiable [24, 35] up to these indeterminacies, if all the sources are nongaussian, or if $\mathbf{A}$ has full column rank and at most one source variable is gaussian. The sources are recoverable (the model is separable, see [35]) if and only if $\mathbf{A}$ is of full column rank and at most one source variable is gaussian. The distributions of the sources are identifiable in even some overcomplete cases [35] when all the sources are nongaussian.

3.1 Independence, entropy and nongaussianity

Independence of two events in everyday language means that the events are not causally dependent on each other. However, the independence of two events (processes) probabilistically means that the occurrence (value) of one does not change the probabilities for the occurrence (value) of the other. Therefore, two events that appear to be probabilistically independent can be causally dependent. It could be argued that building a statistical model of images has nothing to do with artificial intelligence, but identifying and understanding the causation behind the images and extracting relevant information from them is by most definitions intelligent. Causation can also (at least partly) explain why the different variables have the distributions they do, instead of just modeling them.

Statistical independence of two variables means by definition that their joint probability distribution can be factorized using the marginal distributions:

$$p(x_1, x_2) = p(x_1)p(x_2). \tag{3.2}$$

If the joint distribution of several variables can be factorized, it means that the maximal complexity of modeling it is essentially reduced from c^n to cn, where c is the complexity of modeling one variable (marginal distribution) and n is the number of variables. This partly explains the appeal of the ICA model, even though it might be far from an accurate description of the data.

The information content of a single component (which is related to the complexity of modeling it) is most often measured by entropy. Entropy gives a lower limit for the minimum coding length of a variable without a loss of information. (The difference between the minimum coding length and the actual coding length is called redundancy.) Entropy can be calculated also for several variables, but as its estimation essentially requires estimating the joint probability distribution, accurate estimates for complex, high-dimensional distributions are infeasible. But, if the probability distribution is factorizable, so too is entropy.

Entropy for a (possibly multidimensional) continuous variable $\mathbf{x}$ is defined as

$$H_{\mathbf{x}} = -\int p_{\mathbf{x}}(\mathbf{z}) \log(p_{\mathbf{x}}(\mathbf{z}))\mathrm{d}\mathbf{z}. \tag{3.3}$$

It is relatively easy to show [25] using calculus of variations that when the variance of a variable has been constrained to unity, the one-dimensional distribution with maximal entropy is the unit variance gaussian (aka normal) distribution. If the covariance matrix of

a multidimensional variable has been constrained to be the identity matrix, the maximizing distribution is again the multidimensional gaussian distribution.

If one constrains some statistic other than variance, the distribution with maximal entropy generally changes [25]. For instance, if one were to constrain the average absolute value, the maximizing distribution would be a Laplacian distribution (with appropriate scaling), i.e. $p(x) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x|)$. If the minimum and maximum values for the distribution are fixed, the maximizing distribution is a uniform distribution. For our purposes, it is in many ways natural to constrain variance, as it can usually be associated with signal energy, and it is easy to identify the effect of a linear transformation on variance structure, even in multiple dimensions. However, finding minimal coding length in a computer essentially means that the minimum and maximum values have been constrained, and therefore the entropy of a continuous variable and that of its discrete representation are somewhat different. However, a sparse distribution that is far from both a uniform distribution and a gaussian distribution can be a minimizing distribution in both cases.

The difference between the sum of entropies of a group of variables, and the entropy of their joint distribution tells of the redundancies of the variables (as a group) or of their mutual dependencies. This difference is called mutual information:

$$I_{\mathbf{x}} = \sum_{i=1}^{m} H_{x_i} - H_{\mathbf{x}}. \tag{3.4}$$

Mutual information is always non-negative, and zero only if the variables are independent. Intuitively, by finding variables with minimal mutual information, one should also get maximally independent variables. In ICA we search for a linear transformation that produces maximally independent variables, and we have for an invertible linear transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$:

$$I_{\mathbf{y}} = \sum_{i=1}^{m} H_{y_i} - H_{\mathbf{x}} - \log|\det \mathbf{W}|. \tag{3.5}$$

Here the final term on the right side comes from the Jacobian of the transformation. The complete derivation can be found in [59]. The second term on the right side (entropy of the original observed vector) does not depend on the transformation, and is usually ignored. Therefore we have two terms, one that depends on the entropies of the separated variables (which can be approximated using any measure for nongaussianity as entropy is maximal for a gaussian variable), and on the determinant of the separating matrix. If mutual information is used as an ICA objective function, $\mathbf{W}$ is a separating matrix, and at the correct solution $\mathbf{W} = \mathbf{A}^{-1}$.

One can derive ICA objective functions from different starting points and criteria [6, 16, 38, 44, 76, 83], but often they can be shown to be equivalent to the criterion in Equation (3.5) under some assumptions. When the data has been generated according to Equation (3.1), $p(\mathbf{x}) = p(\mathbf{s})|\det \mathbf{A}|^{-1}$. If we do not know $\mathbf{A}$, but know the distributions of the sources, we can test how likely the data is given a candidate $\mathbf{W}$ for the separation matrix, $p(\mathbf{x}|\mathbf{W}) = p_{\mathbf{s}}(\mathbf{y})|\det \mathbf{W}|$, where the distribution of $\mathbf{s}$ is separable due to independence. Thus the log-likelihood of the data $\mathbf{x}(\tau)$, $\tau = \{1, \ldots, T\}$, given a candidate separating matrix $\mathbf{W}$ is:

$$\log L(\mathbf{W}) = \log \prod_{t \in \tau} p(\mathbf{x}(t)|\mathbf{W}) = \sum_{i=1}^{m} \sum_{t \in \tau} \log p_{s_i}(y_i(t)) + T \log|\det \mathbf{W}|. \tag{3.6}$$

Using the logarithm of the likelihood is sensible, because otherwise the numeric range of the numbers for a large number of samples is not easily representable in a computer. Also, logarithm converts the product into a sum, usually producing simpler iteration formulae with techniques involving gradients.

The form of the objective function in Equation (3.6) is (up to irrelevant scaling and sign) identical to the one in Equation (3.5) at the solution, where the source distribution matches the distribution of $\mathbf{y}$. However, at other points entropy is calculated using the distribution of $\mathbf{y}$, whereas the above objective function calculates the likelihood using the probability distributions of the sources. On the other hand, entropy is usually approximated in various ways, sometimes even by fixing the distributions (inside the logarithm) to those of the sources, in which case there is no difference between the objective functions in Equations (3.5) and (3.6).

3.2 Whitening and cumulants

One consequence of the independence is that the sources are uncorrelated. It can be easily shown that if there are at most as many sources as sensors ($m \geq n$), linearly separated independent components are uncorrelated as well. As we mentioned earlier, the ICA model contains indeterminacy relating to the variance of the sources and/or to the length of the basis vectors. Therefore the variance of the sources is in most cases fixed to unity. As a result of these considerations, the data is in most cases prewhitened, and the separating matrix is constrained to be a rotation matrix (orthogonal matrix) in the whitened space. This also means that the matrix is not restricted to be orthogonal in the original space. These steps are necessary and sufficient for the separated components to be uncorrelated and of unit variance. The mixing matrix $\mathbf{A}$ can then be obtained from the separating matrix as its transpose. As we saw with mutual information in Equation (3.5), objective functions usually have a term that depends on the determinant of the separation/mixing matrix, and with this preprocessing that term disappears, as the logarithm of the determinant of an orthogonal matrix is zero. We use the term 'orthogonal matrix' interchangeably with 'orthonormal matrix'.

The difference between the original and whitened space is illustrated in Figure 3.1 in a 2-D toy example of ICA (the source distributions are piecewise constant). In the original space, the basis vectors are not necessarily orthogonal, and therefore an orthogonal basis in the original space is not the best to separate the components, and neither are the basis vectors (columns of $\mathbf{A}$). The basis vectors and separation vectors are connected by the equation $\mathbf{WA} = \mathbf{I}$, and therefore each separation vector $\mathbf{w}_i$ (row of $\mathbf{W}$) is orthogonal to all other basis vectors $\mathbf{a}_j$, $j \neq i$. Basis vectors are arguably the more fundamental ones, corresponding to actual directions of the components in the data, and separation vectors (filters) correspond to the best way of observing these components. In the whitened space, the basis vectors and separation vectors are the same $\mathbf{W} = \mathbf{A}^T$, which naturally simplifies things considerably. It is easy to show that in the original space, the separation vectors can be obtained by multiplying the basis vectors by the inverse of the covariance matrix. One can extend this procedure even to overcomplete cases, but then one cannot obtain the uncorrupted sources using these separation vectors, which are thus less meaningful.

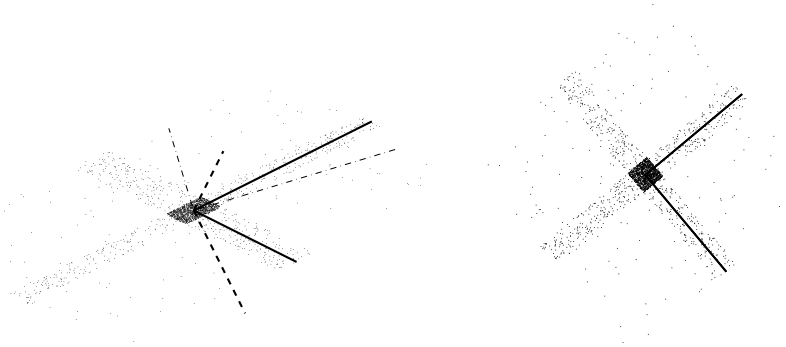


Figure 3.1: Example of whitening. Left: Original space. Right: Whitened space. Basis vectors (solid lines) are not orthogonal in the original space, so the separation vectors (dashed lines) do not point to the same directions as the corresponding basis vectors (each is orthogonal to the other basis vector). However, the basis and separation vectors match in the whitened space. Also illustrated with a thinner, dash-dotted line are the PCA directions that are orthogonal in the original space as well as uncorrelated, but not very useful as such.

Whitening can be achieved by calculating the eigenvectors and eigenvalues of the covariance matrix for the data $\mathbf{C_x} = E\{\mathbf{x}\mathbf{x}^T\} - E\{\mathbf{x}\}E\{\mathbf{x}\}^T$, and multiplying the data by a symmetric matrix $\mathbf{V}$ whose eigenvectors are the same as for $\mathbf{C_x}$ and whose (nonzero) eigenvalues are obtained as the inverse square root of the (nonzero) eigenvalues for $\mathbf{C_x}$. We will denote the i:th eigenvalue of $\mathbf{C_x}$ by $\lambda_i = d_i^2$, as it correspond to the variance of the i:th PCA component ($d_i \geq 0$ is the standard deviation). Thus the eigenvalues of $\mathbf{V}$ are $\lambda_i^{\mathbf{V}} = 1/d_i$. This whitening can now be written as:

$$\mathbf{V} = \mathbf{C_x}^{-\frac{1}{2}}, \qquad \mathbf{z} = \mathbf{V}\mathbf{x}, \tag{3.7}$$

where $\mathbf{z}$ now contains the whitened version of the data. This type of whitening is sometimes called zero-phase whitening [7], due to its symmetry. If one wants to reduce the dimensionality of the data, this type of a whitening is no longer possible. However, if some eigenvalues are zero or very close to zero, one can use pseudo-inverse, where one only inverts (and takes a square root of) the nonzero eigenvalues. Some consideration should then be given to these directions in the algorithms. We will now prove that the components produced by zero-phase whitening are as close as possible to the nonwhitened components in mean square error sense.

Theorem 1
Assume that $E\{\mathbf{x}\} = 0$ *and that* $\mathbf{C_x} = E\{\mathbf{x}\mathbf{x}^T\}$ *has full rank. Then* $R = E\{\|\mathbf{x} - \mathbf{W}\mathbf{x}\|^2\}$ *is minimal for a* $\mathbf{W}$ *s.t.* $\mathbf{W}$ *whitens the data, i.e.* $\mathbf{C_{Wx}} = \mathbf{I}$*, if and only if* $\mathbf{W} = \mathbf{V}$ *as in Equation (3.7).*

Proof. Any whitening matrix can be written as $\mathbf{W} = \mathbf{O}\mathbf{V}$, where $\mathbf{O}$ is orthogonal. Note that $\mathbf{V}\mathbf{C_x} = \mathbf{C_x}^{\frac{1}{2}}$. Denote the trace of a matrix $\mathbf{B}$ by tr$\mathbf{B}$. Note that tr$\mathbf{B}^T$ = tr$\mathbf{B}$ and tr$\mathbf{B}_1\mathbf{B}_2$ = tr$\mathbf{B}_2\mathbf{B}_1$.

$$R = E\{(\mathbf{x} - \mathbf{W}\mathbf{x})^T(\mathbf{x} - \mathbf{W}\mathbf{x})\} = \mathrm{tr}E\{(\mathbf{x} - \mathbf{O}\mathbf{V}\mathbf{x})(\mathbf{x} - \mathbf{O}\mathbf{V}\mathbf{x})^T\}$$
$$= \mathrm{tr}(\mathbf{C}) - \mathrm{tr}\mathbf{O}\mathbf{C_x}^{\frac{1}{2}} - \mathrm{tr}\mathbf{C_x}^{\frac{1}{2}}\mathbf{O}^T + \mathrm{tr}\mathbf{I} = \sum_{i=1}^{m} d_i^2 + m - 2\mathrm{tr}\mathbf{O}\mathbf{C_x}^{\frac{1}{2}}$$

The first two terms do not depend on $\mathbf{O}$. For the last term we obtain:

$$\mathrm{tr}\mathbf{O}\mathbf{C}_\mathbf{x}^{\frac{1}{2}} = \mathrm{tr}\mathbf{O}\mathbf{E}\mathbf{D}^{\frac{1}{2}}\mathbf{E}^T = \mathrm{tr}\mathbf{E}^T\mathbf{O}\mathbf{E}\mathbf{D}^{\frac{1}{2}} = \mathrm{tr}\mathbf{Q}\mathbf{D}^{\frac{1}{2}} = \sum_{i=1}^{m} q_{ii}d_i \leq \sum_{i=1}^{m} d_i$$

Here $\mathbf{E}$ has the eigenvectors of $\mathbf{C}_\mathbf{x}$ and $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{C}_\mathbf{x}$. The last inequality holds because $\mathbf{Q} = \mathbf{E}^T\mathbf{O}\mathbf{E}$ is orthogonal $\implies -1 \leq q_{ii} \leq 1\ \forall i$. Due to full rank of $\mathbf{C}_\mathbf{x}$, $d_i > 0$, $\forall i$, and the limit is attained when $q_{ii} = 1$, $\forall i$. Thus R is minimized iff $\mathbf{Q} = \mathbf{I} \iff \mathbf{O} = \mathbf{I}$.

□

The components corresponding to the largest eigenvalues of the covariance matrix are also the first (largest) principal components. The description using the principal components is also uncorrelated, but the components are often quite far from the original components. If one wants to reduce the dimensionality of the data, picking the desired number of principal components is usually the best way of doing it. This kind of a dimensionality reduction preserves maximum amount of data variance, i.e. a description of the data using only the first principal components is as close as possible to the original in terms of mean square error. Principal component analysis (PCA) is often used in similar problems as ICA, but as it only considers the second-order statistics, it is often not capable of separating the components, cf. Figure 3.1.

Zero-phase whitening can also be used to find an orthogonal (orthonormal) matrix $\mathbf{W}$ that produces the nearest (in mean square error sense) components to a square separating matrix $\mathbf{B}$ in a whitened space. The covariance of $\mathbf{y} = \mathbf{B}\mathbf{z}$ is $E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{B}\mathbf{B}^T$. Denote $\mathbf{W} = \mathbf{W}_2\mathbf{B}$. When $\mathbf{W}$ is orthogonal, $\mathbf{W}_2$ whitens $\mathbf{y}$, and Theorem 1 implies that $\mathbf{W}_2 = (\mathbf{B}\mathbf{B}^T)^{-\frac{1}{2}}$ minimizes $E\{\|\mathbf{y} - \mathbf{W}_2\mathbf{y}\|^2\}$. It is easy to show that

$$\mathbf{W} = (\mathbf{B}\mathbf{B}^T)^{-\frac{1}{2}}\mathbf{B} \tag{3.8}$$

is orthogonal. This method of orthogonalization is used by many ICA algorithms, for example by FastICA [54] in symmetric mode. Orthogonalization (and associated whitening) reduce the number of free parameters to be estimated from n^2 to $n(n-1)/2$. Note that while the precise form of whitening should not matter to ICA algorithms, the form of orthogonalization used does matter. For instance, if we would use Gram-Schmidt orthogonalization, the first component would only be normalized, and subsequent components would have to be orthogonal to the already estimated components, thus accumulating 'errors'. When estimating components in deflation mode, i.e. one-by-one, the Gram-Schmidt procedure is essentially the only possibility for orthogonalization. It is also possible to develop so-called natural gradient algorithms that maintain orthogonality of the separating matrix [1, 3].

One approach to ICA is to jointly diagonalize cumulant tensors. In addition to the sources being uncorrelated, independence dictates that all the cumulant tensors should be diagonal, i.e. all cross-cumulants are zero for independent variables. Fourth-order cumulant tensors are four-dimensional matrices, whose elements are $\mathrm{cum}(x_i, x_j, x_k, x_l) = E\{x_ix_jx_kx_l\} - E\{x_ix_j\}E\{x_kx_l\} - E\{x_ix_k\}E\{x_jx_l\} - E\{x_ix_l\}E\{x_jx_k\}$, when $E\{\mathbf{x}\} = 0$. The kurtosis, $E\{y^4\} - 3E\{y^2\}^2$, of any projection y of $\mathbf{x}$ can be calculated from the fourth-order cumulant tensor. As the normalized kurtosis,

$$\mathrm{kurt}(y) = \frac{E\{y^4\}}{E\{y^2\}^2} - 3, \tag{3.9}$$

is often as a measure of nongaussianity in ICA, the usefulness of this cumulant matrix in ICA is apparent. Note that kurtosis is zero for a gaussian variable, negative for subgaussian ('flat') variables, and positive for supergaussian ('peaked' or sparse) variables. Fourth-order blind identification (FOBI) was an early ICA algorithm using fourth-order statistics by calculating the eigenvectors of a weighted correlation matrix [14, 15]. JADE (Joint Approximate Diagonalization of Eigenmatrices) is a popular algorithm that uses fourth-order cumulant matrices [18, 17]. A review of existing methods can be found in [59].

3.3 FastICA

In addition to JADE, there are several algorithms for performing independent component analysis [1, 6, 66, 83]. The one we mostly use in this thesis is FastICA [54]. The algorithm is essentially an approximative Newton's method for finding the maximum likelihood (mutual information) estimate of the separation matrix. The Hessian matrix used in the estimation of a single component is approximated by a diagonal matrix, which is then easy to invert. There are also other ways of deriving the FastICA algorithm, and in fact it was originally derived as a fixed point algorithm [60]. The iteration formula for a single component is:

$$\mathbf{w} \leftarrow E\{\mathbf{z}g(\mathbf{w}^T\mathbf{z})\} - E\{g'(\mathbf{w}^T\mathbf{z})\}\mathbf{w}, \tag{3.10}$$

where the expected value of the integral of nonlinearity $g(\mathbf{y})$ approximates entropy (or log probability density) of $\mathbf{y}$, g' is the derivative of g, and the expected values are estimated by sample means. The data is usually prewhitened in FastICA, and the norm of $\mathbf{w}$ is normalized to unity on each step. Note that if the sign of the cost function or nonlinearity is changed in FastICA, the signs of the components change each iteration, everything else stays the same.

The algorithm can work in either deflation mode or in symmetric mode. In deflation mode, the components are estimated one-by-one, and the estimated component is constrained to be orthogonal to the already estimated components (due to prewhitening, $\mathbf{W}$ can be fixed to be orthogonal). In symmetric mode, all the components are estimated simultaneously using the same update procedure, while making sure they are uncorrelated by orthogonalizing as in Equation (3.8).

It has been shown that FastICA in deflation mode has cubic convergence [54, 60], when the distributions of the sources are symmetric and nongaussian. Convergence in symmetric mode has not been studied as much, but it appears that if the data has been generated according to the ICA model, the algorithm nearly always converges. If the data does not conform to the ICA model, e.g. the sources are (close to) gaussian or the data has no independent components, convergence does not necessarily occur, or is only linear (i.e. exponential) or sublinear. Assuming that the model holds, it was argued in [32] that the initial convergence of FastICA is linear using a cubic nonlinearity, with the inter-channel interference falling on average on each step to a fraction (about one third) of its value on the previous step. The convergence of symmetrical FastICA algorithm for the cubic nonlinearity was proven in [84].

Another aspect that has to be considered when using FastICA (or any similar ICA algorithm), is the choice of nonlinearity g. It appears that the nonlinearity does not have to be precisely related to the actual distribution of the sources (so the source distributions do not

have to be precisely known, which is one reason why ICA is a successful BSS technique), and therefore other considerations can guide the choice, such as robustness against outliers. Hyperbolic tangent nonlinearity, relating to source distribution of $p(u) = \frac{1}{a_1 \cosh(a_1 u)}$, has been found to be a good choice for most types of sources. Another popular choice is the cubic nonlinearity relating to the kurtosis, Equation (3.9), of the separated variables. However, the cubic nonlinearity corresponds to (the likelihood of) a subgaussian distribution $p(u) = a_1 \exp(-a_2 u^4)$ and it is sensitive to outliers.

3.4 Limitations of the model

ICA is a simple model, and in practice it is almost never accurate with real data. It is actually very hard to imagine situations (in real life) where the model would describe the data generation process exactly. All the properties of the ICA model can also be, depending on the application, its limitations. The properties of the model are generally instantaneousness (i.e. all the samples have been drawn from the same distribution, and their ordering has no relevance), linearity, and that the sources are assumed to be independent.

Probably the most often cited example of an ICA problem is the cocktail party -problem. Assume there are numerous (static) speakers in a room, and an equal number of microphones. By omitting the delays and echos in the problem, this would indeed be a nearly perfect case for ICA, where one wants to find the positions of the speakers in relation to the microphones, and the source signals (speeches). However, by not considering the time structure, the separation results are generally unsatisfactory. The exact content of the source signals may also depend on the angle between the orientation of the speaker and the vector from the speaker to the microphone. Nevertheless, a lot of work has been done in extending the ICA model to cases where there is time structure (see e.g. [8, 23, 130]), and for audio data especially (for a review see [114]).

The linear mixing of the sources is another limitation, as it can be argued that every type of natural data is generated from a set of independent sources combined in a nonlinear fashion; at least such a model can be used to generate any joint probability distribution. Linearity also means that the model has difficulties describing data with (several) distinct clusters. Also, with image data the dominant rule for combination of several objects is not addition but occlusion. That is, the value of a pixel is generally not determined by a linear combination of the light intensities emitted or reflected by several objects, as the closest object occludes others behind it. It must be noted, however, that nonlinear mixing models are generally not identifiable, so additional constraints on the sources and the mixing are necessary [61], and some choices on the objectives or model structure have to be made, see e.g. [111, 116].

Independence of the sources becomes a limitation, when it is combined with linearity of the mixture. Even a linear mixture of sources can be a sufficient model for many types of data, but generally the sources then depend on each other. It can be argued that if the ICA model is useful for a particular type of (natural) data, the best way to extend that model is to explore the residual dependencies in the model. In a sense, ICA can be seen as basis pursuit method [21], producing an interesting basis and corresponding set of source signals useful for further analysis.

Also (computational or algorithmic) limitations on the number of sources can reduce the usefulness of ICA on a problem. Both objective functions in Equations (3.5) and (3.6) work on the premise that there are (at most) as many sources as sensors. However, for example image data can be seen to have an infinite number of sources, corresponding to different objects at different positions, orientations and scales.

There are no extensions of ICA that would be appropriate for all situations and data sets. One can argue that certain extensions are useful for certain groups of problems, but it can also be argued that the application has to guide the nature of the extension, and not vice versa. It should be noted, however, that despite the limitations of the basic ICA model, it can be useful with many types of real data, either as a preprocessing tool or as something akin to a projection (basis) pursuit method. By eliminating the most obvious dependencies from the data, one can concentrate on more interesting structures.

3.5 Independent component analysis for image data

ICA was first used in image data by Bell and Sejnowski [7], although Olshausen and Field [87, 88] basically obtained the same results from a different starting point a bit earlier. In both cases the main results involve feature extraction (basis vector estimation), even though the signals (feature activations) can be used in for example compression and analysis.

Due to computational reasons, images are usually studied in small patches, of size 8 by 8, 12 by 12, or 16 by 16 pixels. Even then, tens of thousands of image patches are required for good estimation of the independent components (features). Usually a sufficiently large number of patches is randomly sampled from a collection of images. Due to this process, all the samples are from the same distribution (i.e. there is no time structure in the signals) defined by the collection of images. Each two-dimensional patch is stacked into a vector, thus a 12 by 12 image patch corresponds to a vector with 144 elements.

Image data is usually relatively sparse or supergaussian in any direction in the whitened space (see Publication I), with a possible exception being the mean value. This suggests that image data has not been generated according to the ICA model, because the projection in a random direction would have to be nearly gaussian compared to the source distributions, see Figure 3.2. At least, not with a complete ($n = m$) ICA model, although overcompleteness should only make the distribution in a random direction more gaussian. Even so, with ICA one can find the most sparse (supergaussian) basis, which is useful in redundancy reduction (compression) and noise removal. In fact, similar results have been obtained when sparsity is explicitly used as an objective, see [88, 89].

Usually, the mean is subtracted from each image patch, and the data is whitened. If the mean is not subtracted, one of the ICA features most likely corresponds to it, but as the mean value component can be quite strong, it is often preferable to completely remove it. As discussed earlier in Section 3.2, after whitening the basis vectors and separation vectors are obtained from each other by taking a transpose. If the data has been whitened as in Equation (3.7), and one wants to visualize these vectors in the original (non-whitened) space, the basis vectors must be multiplied by the square root of the covariance matrix, and the separation vectors by the inverse of the square root of the covariance matrix. As we argued earlier, basis

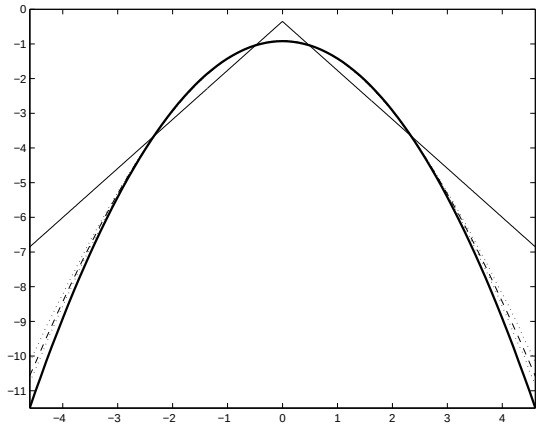


Figure 3.2: Distribution in a random direction (in log-scale) in the whitened space, when the data has been generated according to the complete ($n = m = 100$) ICA model. Thin solid line: Distributions of the sources (Laplacian). Thick solid line: Gaussian distribution. Dashed line: Average distribution in a random direction, calculated from 100 random directions. Dotted lines: Least and most kurtotic distributions for 100 random directions. We used large sample sizes (tens of millions of samples), so the distributions are accurately estimated.

vectors are the more fundamental ones, corresponding to the directions of the actual features (edge segments) in the patches. The separation filters are much 'wavier' and 'sharper' than the basis features, because the inverse of the covariance matrix amplifies higher frequencies very much. We will only illustrate basis vectors in this work. An ICA basis obtained from image data using FastICA with hyperbolic tangent nonlinearity is presented in Figure 3.3.

Sometimes, PCA is used to remove the smallest components from the data before using ICA, to remove 'noise' or to lower computational demands. However, there is usually little actual noise in digital images, and this procedure can make the basis vectors 'wavy' as well, because the smallest principal components correspond to the highest frequencies. This procedure does also produce less of the smallest components that are the least sparse and whose properties (such as orientation and width) are somewhat ambiguous. However, if possible, it could be argued that it makes more sense to exclude the smallest components from further consideration after a complete basis has been found with ICA, as the reduction in dimension does not then interfere with feature estimation.

The type of image data used in ICA affects the results somewhat. If the images contain man-made structures with clear straight edges, for example walls, the features are generally longer and narrower than with 'natural' images, see e.g. [77]. Also by choosing certain subsets of natural images, e.g. different textures or different 'scenery', one can slightly change the appearance of the features [119], or their statistics [81].

ICA description of image data is useful in many respects, even though the model is not perfect. If noise is assumed to be gaussian and white in the original space, then it is useful to transform the data into a basis where the actual data is maximally nongaussian. An optimal

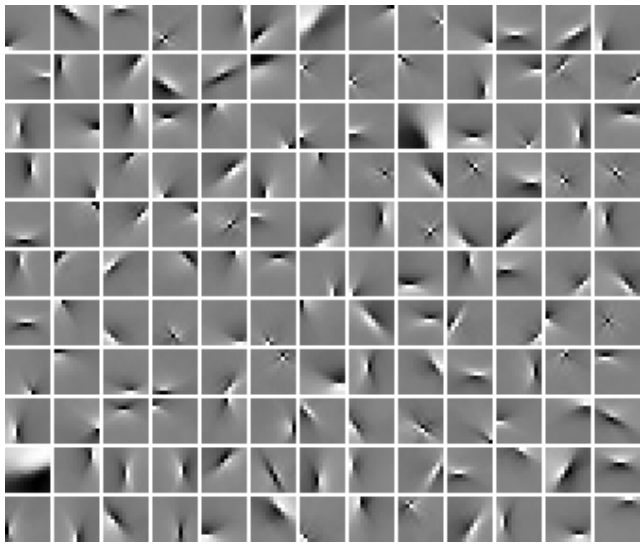


Figure 3.3: ICA basis for image data.

(component-wise) transformation in this basis removes the noise better than a (component-wise) transformation in any other basis, as with sparse code shrinkage [55, 58]. However, sparse code shrinkage uses a basis that is orthogonal in the original space, as otherwise the noise is no longer white, which complicates matters. Also, the components in a maximally nongaussian basis are as sparse as possible, meaning that they are close to zero most of the time, and only occasionally have large values. Therefore this basis also provides the best (linear) basis for component-wise compression, which is actually what minimizing mutual information implies.

Already the earliest adapters of ICA on image data noted the similarity of the features to cortical simple cell receptive fields [7, 88, 119]. This can be considered as support for the notion that the primary visual cortex (and early visual system in general) employs a strategy of sparse coding or redundancy reduction [5]. In any case, the features obtained by ICA, and especially their maximal independence (nongaussianity) and relation to edges, can reasonably be argued to be useful for consequent processing related to image analysis or understanding.

It should be noted that by using only statistical properties of image patches, and no constraints on the shapes of the features, it appears not to be easy to find (linear) basis vectors that are qualitatively distinctly different from both PCA and ICA basis vectors. Simple changes in the cost functions usually result in only slight changes in the properties of the features. This suggests that the ICA description for image data is quite robust, and does not depend too much on hard-to-quantify factors. It also means that from a statistical standpoint, there are no obvious arguments for choosing radically different linear features. Note that nonlinear or higher-order features cannot usually be represented as simple activity patterns similar to those in Figure 3.3.

However, there are many limitations on the usefulness of the ICA description of images. As was mentioned earlier, a basic limitation is that ICA considers the components to be

independent, which they are not in any sense with image data. Also, it can be argued that every scaling, translation and rotation of every ICA feature should also be in the basis, resulting in very highly overcomplete description, computationally infeasible to estimate. Another computational hindrance is the small window size. It is necessary to generalize the results for bigger windows to obtain a good description for images.

These limitations partially motivate the different types of extensions and the analysis we present in the next Chapter. However, we can only touch upon many of these limitations, and our solutions are often only one possible approach to dealing with these problems. Especially when constructing higher-order models, one cannot really even speak of 'solving' the associated problems. A true solution would have to be demonstrated to work as a part of a larger visual system. This being hard, one can partially validate the models by searching for the connections to biological visual systems, or by constructing smaller tasks (e.g. image segmentation or coding) and by demonstrating the performance of the models in these. This would always be desirable, but in the author's view, simply exploring the existing 'residual' dependencies between the ICA components can be of value, even if such an analysis would not immediately translate into working algorithms and methods.

Chapter 4

Extending the ICA model for images

As we shortly described earlier, there are many ways of extending the ICA model, but we will concentrate here on some extensions applicable to images. We will first very shortly discuss adding constraints to the model, then describe results for adding modalities to the data. Furthermore, we will describe two methods for finding overcomplete bases, and finally examine the residual dependencies between the 'independent' components, either by a model that finds a topographic arrangement between the components, or by studying how the activation of one component affects the statistics of other components.

4.1 Additional constraints

Although possibly counterintuitive, models can be extended by adding further constraints to them. When using ICA on images, one sensible constraint can be nonnegativity of the sources. When comparing ICA features to cortical simple cells, a nonnegativity constraint makes the model more sensible biologically, as neurons tend to have low background firing rates and the firing rates cannot be negative. It was shown in [46] that a model conforming to the nonnegativity constraint still produces features with essentially similar shapes. It has also been argued [74] that nonnegative models can be useful for learning parts of objects, but similar results have been obtained with standard ICA models as well [33]. Nonnegative models were examined also in [85, 86].

Nonnegativity implies that the source distributions are nonsymmetric and that the sources do not have zero mean, which has to be noted in the algorithms. As it is not our primary purpose here to produce a model that is in every aspect biologically consistent, we will not use nonnegativity constraints.

4.2 Additional modalities

Additional modalities in the input images can either be biologically reasonable, or otherwise interesting. Biologically sensible modalities can include three-dimensional color (i.e. observing three wavelengths of the light spectrum, corresponding to red, green and blue), stereo images (having two images observed at the same time instant at slightly different positions, somewhat similar to what is observed by two eyes), or image sequences (images successively captured by the same observer, with short and equal delays). Sometimes additional modalities result from a fundamentally different image capturing process, as with three-dimensional images or to some extent with image sequences. Inclusion of additional modalities makes the estimation of the probability distribution for a given number of pixels more difficult, yet it should significantly aid in identifying the underlying objects in the scene. We will for the most part study two-dimensional still images in this work, as our aim is to study only some of the basic statistical structures.

Adding color and stereo properties to images basically multiplies the amount of observed parameters, in the case of color by a factor of three, and in the case of stereo images by a factor of two. Including both modalities simultaneously therefore increases the parameters by a factor of six. ICA has been used on color images by several researchers (e.g. [47, 110, 121]) and stereo images first by Hoyer and Hyvärinen [47]. The results had similarities to observations made about retinal and cortical cells [12, 115]. It is not surprising that ICA features of color images have similar overall shapes as those of grayscale images, and color opponency, as a change of color is often associated with an edge. After removing the mean of each patch, the color transition occurs in a linear fashion in the color space over gray (zero), at least assuming there is only one prominent edge in a patch. However, the results also suggest that some color transitions are more likely than others.

When studying image sequences, the simplest way of transforming the problem for ICA is to consider the image windows at the same position in a certain small number of consecutive frames as a sample point, stacking all the dimensions (spatial and time) into a data vector. In [118], it was shown that the spatiotemporal properties of the features obtained from such data resemble those measured from simple cells, cf. [30, 31]. These features were essentially ICA features of still images moving linearly over time. However, it may be reasonable to treat the time dimension differently from the spatial dimensions, and develop models tuned to the particulars the problem, see [52, 53, 62]. Such models can be more sensible even from a biological viewpoint.

One can also add modalities that are not biologically sensible. For instance, in [121] ICA was used on hyperspectral images with information from 31 different wavebands (colors). It was shown that ICA features for individual pixels corresponded to different natural illuminants, e.g. leaves. Such analyses can yield interesting new insights into the statistics of image data, even if no immediate practical applications can be imagined.

In Publication I, we used ICA to calculate features of three-dimensional MR (magnetic resonance) images. These naturally no longer have any direct link to simple cell receptive fields, but the results were still even surprisingly similar. The obtained features were still Gabor-like, with the sum in Equation (2.1) going up to three. However, most likely due to the fact that MR images have large 'flat' intensity areas, the features were more step-

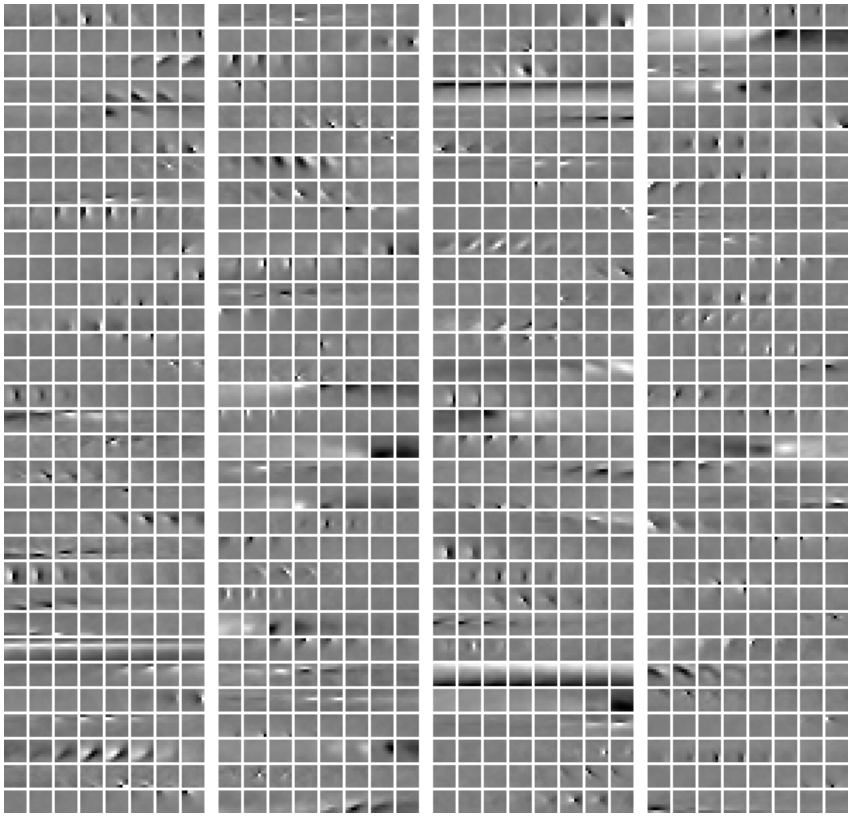


Figure 4.1: A quarter of the $8 \times 8 \times 8$ features found in MRI data. Eight consecutive slices represent one feature.

function like than Gabor functions, or features obtained from traditional natural images. The same was visible even in two- or one-dimensional restrictions of the data. Some of the three-dimensional features are shown in Figure 4.1. The corresponding source distributions were sparse, similarly to ordinary natural images.

As we noted in Chapter 2, the image capturing process for MR images is quite different than the one used for normal photographic images. The intensity values in MR images relate to hydrogen density, not to light emitted or reflected by an object. Additionally, there is no perspective or occlusion in the images. And finally, most available three-dimensional 'natural' images are medical, explaining why we used an image of a human head as the data in Publication I. Therefore, only an insignificantly small percentage of possible three-dimensional objects in the natural world were sampled in the data set.

Note also that the MRI data set can be assumed to be translation invariant, simply as a result of the sampling process. We picked some tens of thousands of three-dimensional $8 \times 8 \times 8$ patches from random positions in the image. Therefore, as a result of the analysis in Section 2.4.1, we may expect the PCA features for this data set to be products of different

frequencies, which appears to be the case. In Publication I we showed that the correlation function depended linearly on the distance between the voxels, which further implies a power law for the power spectrum, and more generally suggests scale invariance.

The similarity between the ICA features for ordinary photographic and MR images strongly suggests that most considerations relating to image capturing, while important to identify, do not have a very significant impact on low-level image analysis methods. There have actually been attempts to create synthetic data sets that have similar statistical properties to natural images, and the results would suggest that combining circles of different sizes with an occlusion rule can yield scale invariant images [73, 98]. Our results would further suggest that even occlusion is not necessary to obtain a correlation structure corresponding to scale invariance, or ICA features that are similar to features obtained from natural images.

One should note that there is demand for noise removal from 3-D images [129], and the results in Publication I suggest that ICA-based techniques such as sparse code shrinkage [58] for noise removal from 2-D images can (relatively straight-forwardly) be adapted for MR images. Moreover, the results suggest that most statistical techniques for image processing and analysis could be adapted for 3-D images. Confirming such conjectures, however, would require considerable work.

4.3 Overcomplete models

As we noted in Chapter 3, the ICA mixing matrix is identifiable in overcomplete cases when all the sources are nongaussian. It is nevertheless easier to develop algorithms when there are at most as many sources as observations, as the sources are then separable. When the sources are not separable, tasks like compression (i.e. finding a maximally sparse representation in an overcomplete basis [88, 21, 80, 92]) become significantly more complicated. However, it may be argued that image data has practically infinitely many sources, and assuming a complete model unnecessarily restricts the search. Indeed, when ICA is used on the same image data starting from different initial values, the basis vectors generally change completely, yet they can nearly always be described using translations, scalings and rotations of a small number of basic features.

Our aim was not to develop methods aimed at separating the components optimally, but to find overcomplete bases. Most existing methods for overcomplete ICA work by simultaneously estimating both signals and basis vectors, for example by demanding that the error of the representation is minimized while the sparsity of the sources is maximized [87, 88], or by alternately maximizing the likelihood of the basis vectors and posterior probability of the sources, assuming the other [78, 79, 80]. By not estimating the source signals explicitly, our methods should be computationally lighter. (Note that it has been argued that sparse coding is harder than sparse code learning [92].) More information about how to best separate the components in overcomplete cases can be found in Publication III.

As we saw with mutual information and likelihood (Equations 3.5 and 3.6), in complete cases most ICA objective functions are essentially divided into two parts: One measuring the nongaussianities of the estimated sources, and the other their orthogonality. If orthogonality was not ensured, all the basis vectors would tend to point to the most nongaussian direction.

However, orthogonality cannot be used in overcomplete cases, as there are at most as many orthogonal directions as there are dimensions. Therefore methods have to be devised with overcomplete situations in mind.

In Publication II, we formulated two methods for overcomplete cases. The first method, which we will refer to by estimation using quasi-orthogonalizing prior, uses a probability distribution for the dot products of the basis vectors. The second method, which we will refer to by estimation using gaussianization, works in deflation mode by removing the already found components by gaussianizing the data in the corresponding direction.

In Publication III we provided numerous experiments with image data, and in Publication II we also showed experiments that suggest that the methods are equally well suited to finding subgaussian sources. The methods are perhaps more elegantly presented in Publication II.

4.3.1 Estimation using quasi-orthogonalizing prior

A crucial assumption for the quasi-orthogonalizing method in Publications II and III is that the basis vectors are randomly distributed in the whitened space, and of unit length. This way one can formulate a probability distribution for the dot product between two unit-length vectors. By assuming that the dot products between the vectors $\mathbf{a_i}$ in a basis are independent, one gets a probability for a basis matrix $\mathbf{A}$:

$$p(\mathbf{A}) = \prod_{i<j} c_m (1 - (\mathbf{a}_i^T \mathbf{a}_j)^2)^{\frac{m-3}{2}}, \tag{4.1}$$

which holds when $m > 3$. Here $c_m = \frac{m-1}{m} \frac{\Gamma[\frac{m}{2}+1]}{\sqrt{\pi}\Gamma[\frac{m-1}{2}+1]}$, where $\Gamma[.]$ is the Gamma function. A more detailed explanation and derivation can be found in Publication II. Note that this probability is invariant under orthogonal transformations of the coordinate system, as the angles between the basis vectors do not change. This choice is thus uninformative wrt. the shapes of the features. Priors on the mixing matrix that assume something about the shapes of the features can also be useful in some cases [63].

Although testing overcomplete methods in low-dimensional spaces is computationally (and possibly algorithmically) easier, there are reasons why the higher-dimensional problems can actually be easier. Random vectors in high-dimensional spaces are highly quasi-orthogonal [69], i.e. the dot product between unit-length random vectors approaches zero as the dimensionality grows, and therefore the interference between two sources approaches zero. Actually, even when one assumes a fixed level of overcompleteness (two times overcomplete, say), the minimum angle between all pairs of vectors in such a random basis approaches zero as the dimensionality increases. Therefore there should be 'room' for higher levels of overcompleteness in high-dimensional spaces.

As there are also more components in these high-dimensional cases (again assuming a fixed level of overcompleteness), the effect of other components in the direction of an estimated component can be approximated by a gaussian variable that is independent of the estimated component, cf. Figure 3.2. Furthermore, by assuming that the basis vectors are randomly distributed in the space, the change in the size of a volume element due to $\mathbf{A}$, cf. $|\det\mathbf{A}|$, can be approximated by a constant. Thus, we approximate the likelihood of the data in the

whitened space (given $\mathbf{A}$) by a product of the likelihoods of components in the directions of the basis vectors (plus a constant term), as explained in Publication II. The gaussian 'noise' caused by the other components would require a modification of the nonlinearity, but as the nonlinearity in ICA methods does not have to be exactly related to the source distribution, the same nonlinearities g work even in overcomplete cases.

Multiplying the likelihood with the aforementioned probability for the matrix, one gets a probability for the separating matrix when the data is observed. Again, it is more sensible to use the logarithm of this:

$$\log \prod_{t\in\tau} p(\mathbf{A}|\mathbf{z}(t)) = \log \prod_{t\in\tau} \frac{p(\mathbf{z}(t)|\mathbf{A})p(\mathbf{A})}{p(\mathbf{z}(t))} \approx \sum_{t\in\tau}\sum_{i=1}^{n} \log p_{y_i}(\mathbf{a}_i^T\mathbf{z}(t)) + \alpha T \sum_{i<j} \log(1-(\mathbf{a}_i^T\mathbf{a}_j)^2) + \text{const.}, \quad (4.2)$$

where $\tau = \{1,\ldots,T\}$. This now differs (ignoring constant terms) from the usual maximum likelihood objective function (Equation 3.6) simply by the replacement of $|\det \mathbf{A}|^{-1}$ by $p(\mathbf{A})$. The parameter α weights the probabilities, and its value could be calculated analytically, but we adjust it empirically.

We earlier presented a slightly different derivation of the algorithm [65], where the dot-product $\mathbf{a}_i^T\mathbf{a}_j$, $i \neq j$, was approximated (with good reason) as gaussian, leading to an algorithm depending on the squares of the dot-products. However, minimizing the squares of dot-products is not enough to prevent at least some of the basis vectors from pointing in the same directions. Note that logarithm of the probability in Equation (4.1) explodes as any dot-product grows toward unity. Minimizing higher powers of the dot-products does work, and the results in [65] were valid, even though the justification was lacking.

Previously some extensions of classic orthogonalization methods were presented for overcomplete cases [56], but it appears such extensions are insufficient to keep the vectors apart. Natural gradient learning for overcomplete cases without estimating the sources was studied in [2], but it appears that demanding $\mathbf{A}\mathbf{A}^T = \mathbf{I}$, while essentially maximizing sparsities of $\mathbf{a}_i^T\mathbf{z}(t)$, is not enough to prevent multiples of the basis vectors.

This method and the underlying assumptions work best in high-dimensional spaces with highly overcomplete problems. Then the vectors are most likely somewhat randomly distributed in the space. Furthermore, as the distribution of the basis vectors is not known in advance, it is sensible to have a prior that is an uninformative one. Due to the inherent high-dimensionality of image data, it should be most well suited for this algorithm, with the caveats associated with ICA and image data in general. One should note that whitening does change the distribution of the basis vectors in the space away from a completely random distribution, as it spreads the vectors out somewhat even in overcomplete cases.

In figure 4.2, we have a four times overcomplete basis for image data found with this quasi-orthogonalizing method. We used gradient ascent in the maximization of the objective function in maximization of Equation (4.2), the value of α was set to 0.17. More experiments can be found in Publication III, and tests even with subgaussian data in Publication II. Note that even with this level of overcompleteness, the smallest angle between the basis vectors in the whitened space was close to 60 degrees.

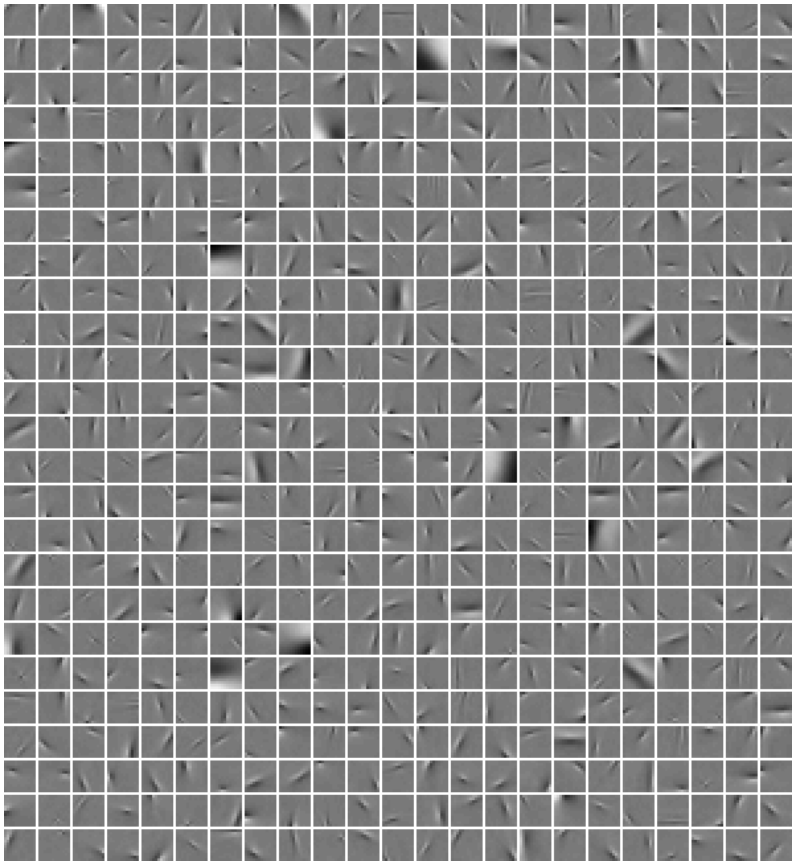


Figure 4.2: A 4 times overcomplete basis obtained using the quasi-orthogonalizing prior.

Some things not discussed in Publications II and III should be pointed out. First, as the orthogonality measure uses the dot products between the components and not the determinant of a square matrix, it does not necessarily force the basis vectors to 'fill' the space. The determinant explodes, if any dimension does not get used, whereas the angles between the basis vectors are hardly affected by the omission of a dimension in a high-dimensional space. As some directions in the space have distributions much closer to gaussian than others, it is (in principle) possible that none of the vectors use these dimensions. In practice, by giving the parameter α in Equation (4.2) a small value, one can get bases that sample mainly only the most nongaussian directions, and by giving α a large value, one can get bases that sample all the directions quite well. Large values of α can also lead to slower convergence with gradient ascent.

Note that in addition to using a probability for the dot-products between the basis vectors, we might also require at each step 'one-sided' orthogonalization $\mathbf{A}\mathbf{A}^T = \frac{n}{m}\mathbf{I}$, so the vectors would use all the dimensions 'evenly'. However, enforcing this the same way as with

standard orthogonalization in Equation (3.8) can change the norms of the basis vectors, and normalizing the basis vectors can change $\mathbf{A}\mathbf{A}^T$ away from $\frac{n}{m}\mathbf{I}$. Both restrictions could be achieved approximately by repeating this 'one-sided' orthogonalization $\mathbf{A} \leftarrow \frac{n}{m}(\mathbf{A}\mathbf{A}^T)^{-\frac{1}{2}}\mathbf{A}$ and basis vector normalization alternately enough times. The approximation that the change in the size of a volume element is approximately constant may be more accurate with this 'one-sided' orthogonalization.

Despite the similarity between Equations (4.2) and (3.6), using similar approximations as for FastICA does not yield an algorithm with cubic or even quadratic convergence. It appears this is true even in complete cases, so the addition of the quasi-orthogonalizing term (or the removal of orthogonalization) affects the convergence, and slightly better approximations do not appear to improve convergence. Note that the full Hessian matrix needed in the Newton's method for the problem has n^4 elements, and inversion of it is not sensible when there are more than a couple of sources, e.g. with image data. However, it should be noted that even FastICA does not converge quadratically (let alone cubically) with image data. We used gradient ascent in our experiments.

4.3.2 Estimation using gaussianization

The gaussianization method presented in Publications II and III, originally in [65], proceeds by searching for the most nongaussian direction using single component FastICA, removing the nongaussianity associated with it by gaussianizing the distribution in that direction, i.e. transforming the distribution to gaussian, and repeating the search. By denoting the cumulative distribution function (cdf) of the component $y_i(t) = \mathbf{a}_i^T\mathbf{z}$ by F (in the whitened space y_i is the linearly separated component corresponding to $\mathbf{a}_i$, or the component in the 'direction' of $\mathbf{a}_i$), and the cumulative distribution of the standardized gaussian variable by Φ, the gaussianized component is $h(t) = \Phi^{-1}(F(y_i(t)))$. The reconstructed data from which basis vector $\mathbf{a}_{i+1}$ can then be estimated is obtained as:

$$\mathbf{z}(t) \leftarrow \mathbf{a}_i h(t) + (\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^T)\mathbf{z}(t). \tag{4.3}$$

A data sample version of this gaussianization is easily implemented, where one only needs to sort the samples $y_i(t)$ instead of calculating cdf F, see [36].

This procedure makes sure that the same direction is not found (at least) in the next couple of iterations, but does not remove the information associated with that direction completely. This method, like the previous quasi-orthogonalizing method, uses the observation that in high-dimensional spaces with data that has been generated according to the ICA model, the distribution in a random direction is very close to gaussian compared to the distributions of the sources, see Figure 3.2. Thus, assuming that the basis vectors are randomly distributed in the space, gaussianization transforms the distribution close to what it would be without the component. As our algorithm finds directions that maximize nongaussianity, gaussianization also guarantees that the found direction is maximally 'uninteresting' after the transformation.

However, the distribution in a random direction in image data is not gaussian, but closer to Laplacian, as observed in Publication I. Therefore, with image data, the distribution of the component should possibly not be transformed to a gaussian distribution as this may produce artifacts, cf. Figure 4.3. However, the average distribution in a random direction changes

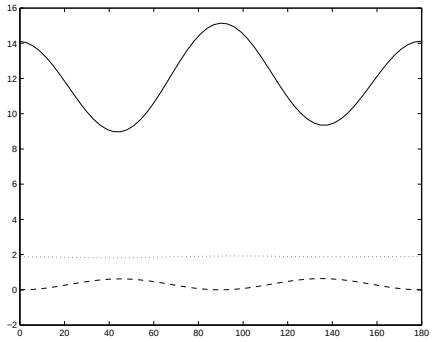


Figure 4.3: A two-dimensional example of gaussianization, where the data has not been generated according to the ICA model. Independent Laplacian sources s_1, s_2 are aligned with the axis (0 and 90 degrees), and the samples have been post-processed $\mathbf{x}(t) = \mathbf{s}(t)\sqrt{\|\mathbf{s}(t)\|}$ to produce positive energy-correlations somewhat similar to those existing in image data. Solid line: Kurtosis (Equation 3.9) of the data as a function of angle a, i.e. $\text{kurt}\{\cos(a)x_1(t) + \sin(a)x_2(t)\}$. Dashed line: Kurtosis of gaussianized data $\mathbf{u}$. Note that kurtosis at the gaussianization angles, 0 and 90 degrees, is zero, but spurious maxima have appeared at angles 45 and 135 degrees. Dotted line: Kurtosis when data is not completely gaussianized, $\mathbf{u}' = \frac{1}{7}\mathbf{x} + \frac{6}{7}\mathbf{u}$. It can be argued that the procedure resulting in $\mathbf{u}'$ is better than gaussianization (producing $\mathbf{u}$) for our purposes, as we do not want to gaussianize the joint distribution, but find directions of maximal nongaussianity.

as more components are removed, so the target distribution should change as well, adding technical difficulties. Furthermore, as more and more components are gaussianized, the joint distribution converges weakly (pointwise) to a gaussian distribution [22], if the directions chosen for gaussianization are sufficiently close to the most nongaussian directions. (Note that according to the Cramér-Wold theorem, any distribution is uniquely characterized by the family of its marginals in all possible directions [91].) Thus all statistical structures are removed, and after some point it is reasonable that the basis vectors obtained with this method no longer resemble normal ICA features. With images, the usability of the method ended after about a 2 times overcomplete basis.

In Figure 4.4, we have a slightly over four times overcomplete basis. As the search progresses in deflation mode, there is an ordering for the components, and smaller levels of overcompleteness correspond to taking only a certain number of components from the top. Indeed, the basis vectors become more and more noisy after a 2 times overcomplete basis.

As the method uses FastICA, it also inherits FastICA's convergence properties. The gaussianization takes a negligible amount of processing time compared to the FastICA iteration. As more and more components are estimated, the data becomes more and more gaussian, which may affect convergence speed.

This gaussianization procedure is almost identical to the procedure presented in [36] for finding interesting directions in data, e.g. for separating clusters, although the measures for interestingness (nongaussianity) are different, and the method in [36] was not used for

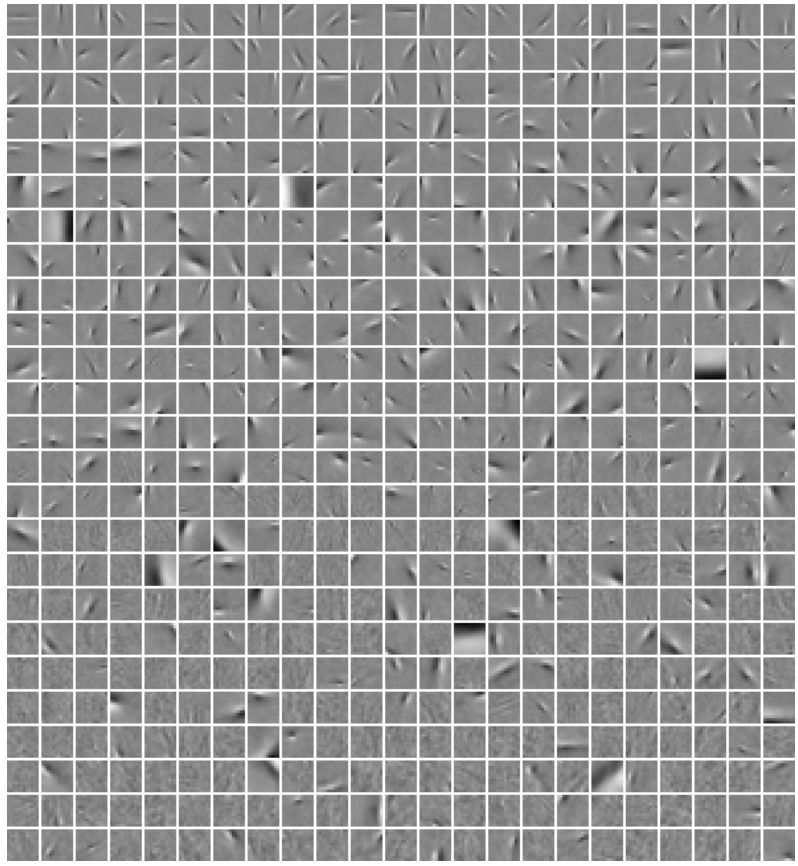


Figure 4.4: A 4 times overcomplete basis obtained using the gaussianization method. Features ordered from left to right, top to bottom with the feature found first at the top left corner.

ICA, or for finding bases, let alone overcomplete ones. Similarities exist also with [22], but they estimated complete ICA bases that were good for gaussianization in the context of density estimation, an approach similar to what was described already in earlier works [51, 36]. As the probability density becomes more and more gaussian, after a sufficient number of gaussianizations, one can basically invert all the gaussianizations, and generate samples of the original distribution from a gaussian distribution. One should note that this inverse transformation may be almost noncontinuous (if each individual gaussianization transformation has not been sufficiently smoothed, which in turn may reduce the flexibility of the model), and storing data of all gaussianizations may require as much (if not more) space as the original data set. The procedure in [22] was not used for finding overcomplete ICA bases. We would argue that for finding overcomplete bases, it is more sensible to estimate and gaussianize the components one-by-one, and not complete (uncorrelated) bases.

4.4 Topographic ICA

As explained in Chapter 3, the components in ICA are assumed to be completely independent, and therefore do not necessarily have any meaningful order relationships. However, natural data in its various forms is hardly ever exactly described by the ICA model, and the components found using ICA are far from independent. Topographic independent component analysis (TICA) presented in Publication IV uses this residual dependence structure to define a topographic order for the components. More information on topographic ICA can be found in [64].

The motivation for the TICA model arises partly from the observation that input from most senses is ordered topographically on the cortex, as discussed in Chapter 2. This two-dimensional mapping often exists on the cortex even when the features can no longer be described using only their position, e.g. a description of their orientation and size is necessary as well. In order to minimize wiring length, two neurons that correspond to visual features with similar properties need to be close together, especially as further processing may require input from both of them. Therefore the topographic ordering on the cortex may reveal different aspects of the dependency structures in the data.

Another motivation for topographic ordering is visualization, as it is easier for a human to (visually) analyze dependencies in data sets if the dependencies can be (with reasonable fidelity) mapped onto a one-, two- or three-dimensional space. Even in machine learning applications this kind of a mapping can be useful to reduce the dimensionality of the data, see e.g. [71]. It has been argued, that even the topography on the cortex can be understood in terms of dimension reduction [34].

The topographic ICA model is related to independent component analysis for subspaces (independent subspace analysis, ISA) introduced in [57]. In ISA, one searches for the most independent subspaces, measured by the sparseness of their energies, i.e. expected squared sums of the components in the subspace S, $E\{\sum_{i \in S \subset \{1,\ldots,n\}} (\mathbf{w}_i^T \mathbf{x}(t))^2\}$. An independent subspace can have energy correlations between the constituting signals, but the energy correlations between different subspaces should be zero. This means that signals that tend to be active (not zero) simultaneously should end up in the same subspace. It was found that for image data, the subspaces often had properties similar to those of complex cells [50, 93].

Other often used topographic mappings include the self-organizing map (SOM, see [70, 71]) and generative topographic mapping (GTM, see [9]). These are, however, quite different models. SOM is not a generative model for the data (does not define an explicit probability density), and it is most often used as a visualization tool, or as a tool for finding data structures in some machine learning applications. GTM is a generative model for the data, and it describes the probability distribution using a constrained gaussian mixture model. Both SOM and GTM use Euclidean distance when fitting to the data, basically finding a two-dimensional surface (when used for visualization) in the data space close to which most of the data lies. In contrast, ICA and TICA use dot products, somewhat similarly to "dot-product SOM" [71]. However, in ICA and TICA, the components are uncorrelated and maximally independent, and the topography in TICA can be expected to be influenced by only higher order dependencies between these components.

Independent subspace analysis can be seen as a special case of the topographic ICA model. In TICA, a nonnegative neighborhood function $h(i, j)$ is defined to describe the dependency between components i and j, similarly to the SOM [71]. There is considerable latitude in the choice of neighborhood function. One can choose a neighborhood that is one- or two-dimensional, i.e. the values of the neighborhood function relate to the distance of the components on a one- or two-dimensional grid. One can also choose something more exotic. If one chooses a neighborhood function, where groups of components are equally connected, and the different groups are not connected, one has the ISA model. In ISA, the sizes of the subspaces have to be predetermined, but in TICA one does not need to set strict boundaries for the groups, allowing the data to drive the search for dependencies. On the other hand, if one chooses a neighborhood function where $h(i, j) > 0$ iff $i = j$, the model is essentially equivalent to the classic ICA model.

One should note that as a rule of thumb, as long as $h(i, j)$ has a maximum when $i = j$, the TICA model still attempts to find maximally independent sources. That is, it does not generate topography where dependencies do not exist, and even though the components may be slightly different from the ones found by ICA, these differences arise from trying to fit the existing dependencies on the grid while maximizing the independence of the sources.

The log-likelihood for the TICA model can be approximated as it was in Publication IV:

$$\log L(\mathbf{W}) = \sum_{t \in \Gamma} \sum_{j=1}^{n} G(\sum_{i=1}^{n} h(i, j)(\mathbf{w}_i^T \mathbf{x}(t))^2) + T \log |\det \mathbf{W}|, \tag{4.4}$$

where function G has a similar role as the log probability density of the independent components in classic ICA. This function depends on $\sum_{i=1}^{n} h(i, j)(\mathbf{w}_i^T \mathbf{x}(t))^2$, which can be viewed as the energy of the neighborhood around component j for the current sample point $\mathbf{x}(t)$. Indeed, the basic type of dependency in TICA and independent subspace analysis is energy correlation. This means that signals in the same subspace in ISA or close to each other on the TICA grid are assumed to be active and inactive simultaneously.

It has been shown that this kind of a dependency structure is quite natural for images [102, 122, 68], and can be found in many other types of natural data [64]. In image data, it would appear that the most obvious dependencies between the components are such energy correlations, or dependencies between the activity levels of the components. The strongest dependencies appear to be between features with similar properties, e.g. features with similar prevailing orientations, positions, and sizes (or frequencies), regardless of the specific shapes of the features.

However, these energy correlations are insufficient to describe all the dependencies existing in images and other types of natural data, and the fact that the models have been designed to find such dependencies may highlight these warrantlessly, while ignoring some other, possibly even stronger dependencies between the components. Therefore it is valuable to examine what types of dependencies exist between ICA features in image data without first restricting the search to such dependencies one a priori expects to find. Even though we used a two-dimensional neighborhood function when using TICA on image data in publication 4, there is no reason to assume that the dependencies would actually be structured in such a way, and the same can be said of the topographic ordering on the visual cortex. The ordering is just in some sense the best way of describing the structures on a 2-D surface.

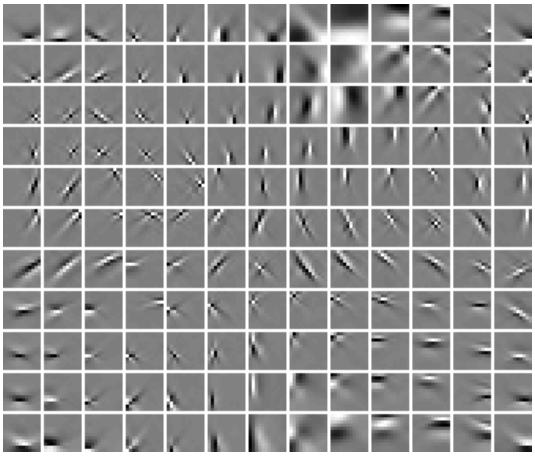


Figure 4.5: TICA basis of image data.

In Figure 4.5 we have a TICA basis for 12 by 12 image patches (with the mean value removed). We used a torus-shaped 2-D topography where the neighborhood function was one between 8-neighbors, otherwise zero. Being torus-shaped means that the units on the bottom of the map are neighbors of the units on the top of the map, and similarly for left and right, i.e. the map had no edges. As one can see, features with similar orientations, sizes and positions are clustered together on this map, confirming what we discussed about the existing energy correlations in image data.

In Publication IV, we also had experiments of using TICA on magnetoencephalographic (MEG) data, and in [64] on audio data, to test the usefulness of TICA as a tool for finding dependencies in other types of natural data. Publication IV also includes derivation of a generative model for TICA, where the sources are obtained by multiplying independent (zero mean, unit variance) gaussian variables by positive variance variables that depend on the neighborhood function. These non-independent sources are then mixed with the usual ICA model. Using this generative model, we tested TICA with simulated data, revealing that it was consistent in finding the structure that was used when generating the data. We also had experiments showing that the generative model could produce such source distributions as observed in natural image data.

4.4.1 TICA revisited

By comparing the forms of the TICA objective function in Equation (4.4), the maximum likelihood for the classic ICA model in Equation (3.6), and the objective function for overcomplete cases using the quasi-orthogonalizing prior in Equation (4.2), one may wonder if the TICA model can be extended into overcomplete cases by replacing $|\det \mathbf{W}|$ by $p(\mathbf{A})$. The short answer to this is positive, the differences between the models do not significantly interfere with the usability of this approach. In Figure 4.6 we have a more than 3 times overcomplete TICA basis. The angle between the closest two basis vectors in this basis is still almost 57 degrees.

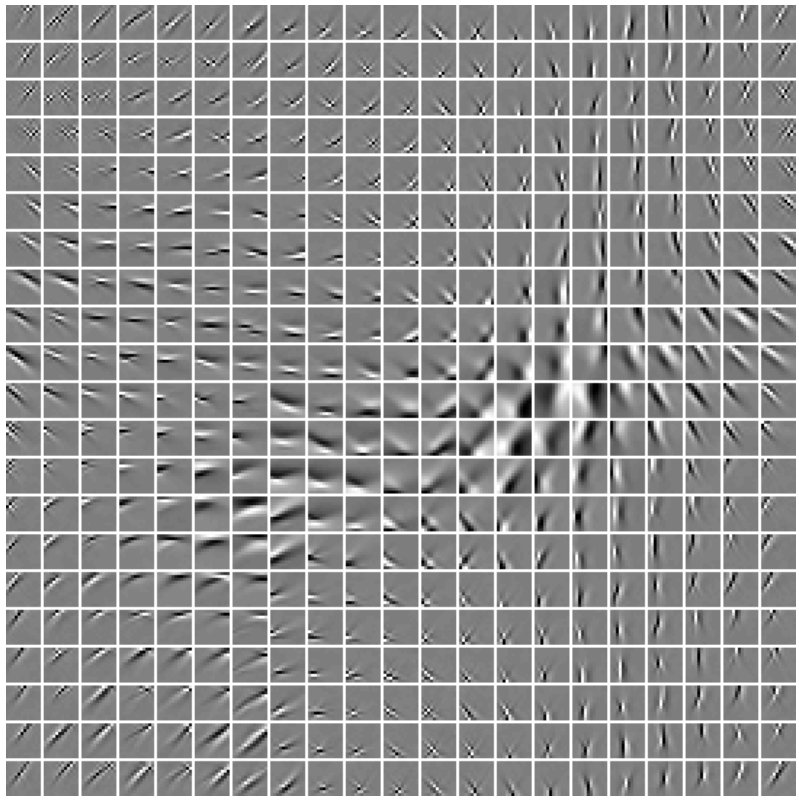


Figure 4.6: Overcomplete TICA basis of image data.

It can be convincingly argued that such an overcomplete model is more sensible biologically, as the description on the cortex is essentially overcomplete, and is definitely not confined to be orthogonal in a whitened space. However, it should be noted that the signals are no longer uncorrelated, and it should be analyzed how these correlations affect the ordering, and the validity of the approximations and assumptions used in deriving TICA should be re-examined. It is likely that the neighboring features have the strongest correlations, i.e. the basis vectors are least separated in the whitened space.

It should be noted that despite the similarity of the objective functions, one cannot derive an algorithm with quadratic or cubic convergence for the TICA objective function in Equation (4.4) using similar approximations as for FastICA. Actually with image data, the convergence of the algorithm is sublinear.

What was left unclear by the results in Publication IV was what image structures are necessary to produce such structures as observed in the topographic independent component analysis of the data. To study this, we generated data consisting of independently sampled Gabor functions, Equation (2.1). Each image patch consisted of six Gabor functions of random size, orientation, position and amplitude. All the functions had phase of $\pi/2$, and a very low frequency, so the Gabor had essentially only two lobes, one positive and one negative. Also the length of the Gabors (along the edge) was fixed to three times the width.

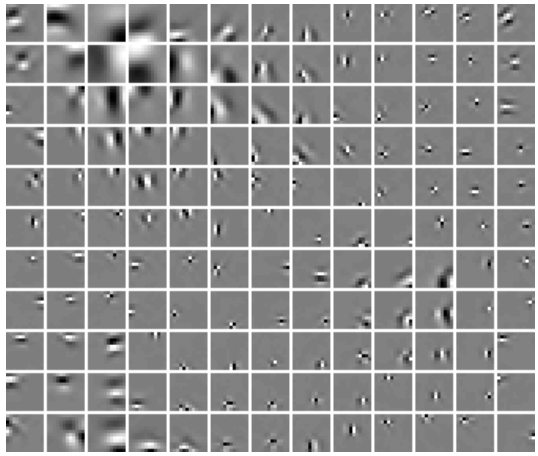


Figure 4.7: TICA basis of synthetic data consisting of Gabor filters. Compare with Figure 4.5.

As our results in Figure 4.7 show, even this type of data generation process is enough to generate structures somewhat similar to those identified by TICA. This can be explained by the fact that the basis consists of only a limited number of different Gabor-type basis vectors, so when a Gabor in a 'synthetic' image patch does not match any of these, it activates several different basis vectors, typically in a limited range of positions, orientations and sizes. Note also that some of the basis vectors even exhibit phases near zero, i.e. their shape is that of line detectors, even though the data had no such lines.

Even though this result does not show that the structures visualized by TICA are not at least partly affected by more interesting image structures, according to the famous Occam's razor, we should assume a simpler reasoning is also a better one in the absence of further evidence. The dependencies in natural image data appear to be somewhat stronger, although it is difficult to identify the reasons for this. However, by what we know a priori about image data, it is reasonable to assume that more interesting image structures do have an effect on the TICA results.

These results point out how difficult it is to identify the actual generative (causal) model by fitting a model to the data. The results of using ICA on image data suggest that images consist of Gabor-like functions, yet one cannot see such features in images. Such Gabor-like features give only the most sparse transformation of image data. Similarly, TICA results suggest that similar features are active simultaneously, even that the presence of a Gabor in an image patch implies that other Gabors with similar properties are also likely to be present. However, the reasoning above suggests that this may not be the case, and most of the observed dependencies might be due to certain types of elementary misfits.

Nevertheless, a TICA-type representation can be useful in that it summarizes the energy correlations for a fixed, complete basis in a simple way. Also, regardless of the causal origin of these energy correlations, their apparent importance for the topography on the cortex suggests that they are dominant low-level statistical structures in image data.

In order to examine the statistics of images further, it is useful to relax many of the assumptions implicitly or explicitly made in models such as TICA, even though these assumptions may be sensible biologically or otherwise. However, no method of analysis is truly unbiased, one can only observe such dependencies as the method or model allows, and interesting results can be obtained from other perspectives as well.

In addition to TICA, there are other examples where energy-correlations have been used to build higher-order ICA (or ICA-type) models. For example in [48] the model learned larger contour features, and in [68] the model learned variance features. In both cases, these features often still had properties similar to normal ICA features, including orientation selectivity, edge detection properties, and similar spatial arrangement, although there were some interesting exceptions. Both models (as well as the neighborhood energies in TICA, see Publication IV) are usually tuned to a greater set of orientations, positions, and sizes than individual ICA features, i.e. slight changes in the input data do not produce as different activity patterns as they can in ICA. For example in [68] some of the features were orientation selective, but not localized, and some were not localized nor orientation selective, but scale selective. Such observations are interesting, even if the model structures are in some aspects restrictive, and tuned to find certain types of dependencies.

4.5 Examining the dependencies in the ICA model

In Publications V, VI and VII we examined the dependencies between ICA features of image data. We took the approach of examining how high activation of a component affects other components. In these cases the highly active ICA feature can be expected to describe something essential appearing in a patch, i.e. part of a line or an edge. Even though the ICA basis is the most sparse basis for the whole data, it still consists of only a limited number of basis vectors, meaning that even the simplest features in images, such as lines and edges, usually require the activation of numerous ICA components. If the activation of a feature is low in a patch, it is probable that it only describes the residual 'error' of fitting other features to the content of the patch, and not any actual object in it. On the other hand, high activation of a feature means that it most likely represents a significant portion of a prominent edge appearing in the image patch. Even in such cases the edge most likely cannot be described using only the active feature, and the residual errors are not completely random. Thus by only examining the patches where a component is highly active, we have a more sensible starting point for examining the dependencies between it and others.

Normalization is important for observing the statistics. By normalizing variances of individual patches (i.e. fixing patch norms), contrast differences between different areas in the images are eliminated, and a certain level of activation of a feature means that it contributes a certain amount of the patch variance (norm of an individual patch), or patch 'content'. This normalization has similarities to the normalization used in [102, 122] for neighboring Gabor filters. Whitening normalizes the second-order statistics so that changes due to the activation of a component are easier to observe. Note that the mean value has been subtracted from all the patches. Basis functions obtained from this normalized data are illustrated on the left side in Figure 4.8.

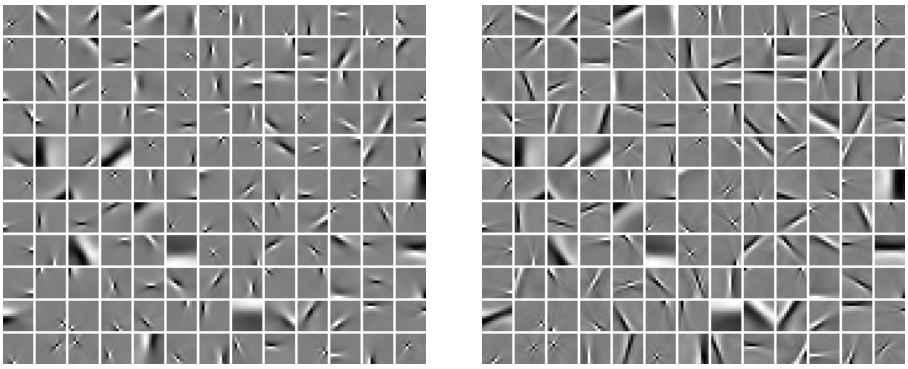


Figure 4.8: ICA basis for normalized image data. Left: Using hyperbolic tangent nonlinearity (supergaussian density). Right: Using third power nonlinearity (kurtosis) and starting from the basis on the left.

Furthermore, normalizing by the sign of the active component allows one to study the changes in first order statistics. Independence suggests that the signs of two components are independent of each other. However, two collinear features can be expected to have matching signs more often than not when either is highly active, as they are then likely to describe different parts of the same edge.

With these steps, first and second order statistics are normalized and changes in them are easy to observe. It is harder to normalize higher order statistics, and to identify which portion of them is changed due to the same factors observable in the changes of first and second order statistics. First and second order statistics do, however, account for the bulk of the dependencies. One should not forget that the difficulties in exploring higher order statistics is one of the reasons for the popularity of ICA and projection pursuit methods.

In Publication V we obtained results indicating that when a new ICA basis is estimated for the data subset where a specific independent component is highly active, the features are essentially similar to the features in the original ICA basis, but the variances of the components change. Thus, there are more components close to the active feature (high variance) and less further away from it (low variance), as a consequence of estimating the components in a whitened space.

In Publication VII we argued that the changes in mean value for high activations show how the feature extends into orthogonal dimensions, or what usually accompanies the feature in the patches, relating to what is known as a 'good continuation' of a (visual) feature in psychology [37, 105, 124]. The fact that apparently the mean value of the patches where a feature is highly active is not restricted to the subspace spanned by that feature should translate into coding advantages.

These first order statistics can be represented using image patches, similar to the original features. It was observed that when the threshold about equaled the standard deviation of the feature (one), these mean value features were very close to the original features. When the threshold exceeded one, the mean value features were longer and more step-function

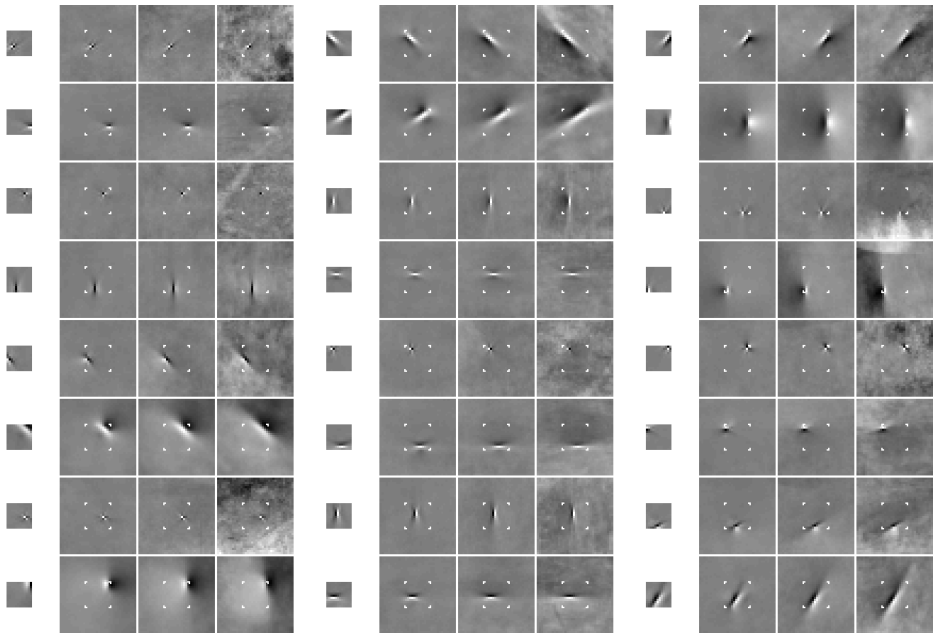


Figure 4.9: Extended ICA features. On the left of each group is the original feature, and on the right mean value features for different thresholds (one, three, and five times the standard deviation). Each mean value feature is 36 by 36 pixels, and the centermost 12 by 12 pixels correspond to the original patch. The area corresponding to the original feature has been explicitly marked for clarity. Compare with Figure 4.8.

like than the original features. Note that if the ICA model were valid for the data, these features would match the original features. The proportion of the mean value directed at other dimensions rose about linearly with the threshold after a threshold of one. These mean value features also present a way of extending the original features outside patch edges, by calculating mean values for larger patches and by using appropriate normalizations. In Figure 4.9, we have illustrated these extended mean value features for different thresholds.

There is a link between the shapes of the mean value features at different thresholds, and features obtained with different nonlinearities. High thresholds correspond to nonlinearities that are insensitive to low activities, i.e. nonlinearities g (cf. Section 3.3) that are flat near zero and jump at the positive and negative thresholds. However, such nonlinearities are not differentiable at the threshold. Furthermore, it is difficult to test the effect of a nonlinearity as it is likely that different thresholds may produce quite different basis vectors (with different positions, orientations and sizes) despite using similar initial values. On the right side in Figure 4.8, we have a basis obtained using the cubic nonlinearity, relating to the kurtosis of the variable, emphasizing large values much more. Again, the features are longer and more step-function like. Note that normalization limits the maximum values of the components, so the outliers cannot be arbitrarily large.

Our analysis of second order statistics in Publications V and VII showed that the PCA dimensions that increased in activity (i.e. had variance of more than unity) due to the activation of an IC were positioned similarly, and had similar general orientations and sizes as the active feature. The very smallest principal components (minor components) that correspond to dimensions that decreased in activity (i.e. had variance of less than unity) had properties as far as possible from the active feature. The shapes of many of the remaining features in the PCA basis were mostly due to estimation error (i.e. finite sample size), and in Publication VII we analyzed precisely how much of the observed statistics can be accounted for by the finite sample size. However, even the first principal component features (that had the smallest estimation error) were often far from Gabor- or ICA-type features, making description of their properties difficult and thus limiting the generalizability of the results.

In Publication VII we also conjectured that the changes in second order statistics relate to different types of 'errors' made in describing an object in the image by the active feature. Similarly to our experiments with TICA and synthetic data, the orientations, sizes and positions of the objects do not perfectly match those of the active feature, and therefore other features with similar properties are also activated. Note that most of these errors are such that they do not show up in first order statistics. For example, the orientation and position of the actual object (edge) can vary to either side of those of the active feature, and therefore such errors 'cancel out' in the mean value. It is difficult to identify all the factors that affect these conditional statistics, especially as it is difficult to remove the effects of these simple 'errors' in the fit. Additional difficulties come from the amount of data required in this analysis.

4.5.1 A parametric model for analyzing the dependencies

In order to better examine the specifics of the changes in conditional statistics, we fitted a model to the statistics in Publication VI, described in more detail in Publication VII, where the input to the model consisted of the properties of the conditioning (active) and conditioned (reactive) feature, and the desired output was either the mean value or variance of the conditioned feature. This way the dependency between the conditioning and conditioned feature was reduced to two parameters, which naturally reduces the amount of data required in the analysis. Also, we chose the structure of our model to be such that the contributions of different parameters were easily identifiable. We made some control experiments with multilayer perceptron (MLP) networks to identify, how much of the remaining variance was explainable with a more flexible model, which was also harder to interpret.

Our model was of the type

$$R'_{i,j} = f(f_{WI}(G^{WI}_{i,j}) f_{OR}(G^{OR}_{i,j}) f_{CO}(G^{CO}_{i,j}) f_{OL}(G^{OL}_{i,j})), \tag{4.5}$$

where the functions on the right side are adjusted so that the squared error between $R'_{i,j}$ and the observed statistic $R_{i,j}$ is minimized over all conditioning ICs i and conditioned ICs j. Function f was required to be monotonically increasing and positive, and $G_{i,j}$:s were properties calculated from Gabor functions fitted to the features i and j. $G^{WI}_{i,j}$ related to the logarithmic difference between the sizes of the features, $G^{OR}_{i,j}$ to the difference between feature orientations, $G^{CO}_{i,j}$ to collinearity between the features, and $G^{OL}_{i,j}$ to the overlap of the features. More details on the procedure can be found in Publications VI and VII.

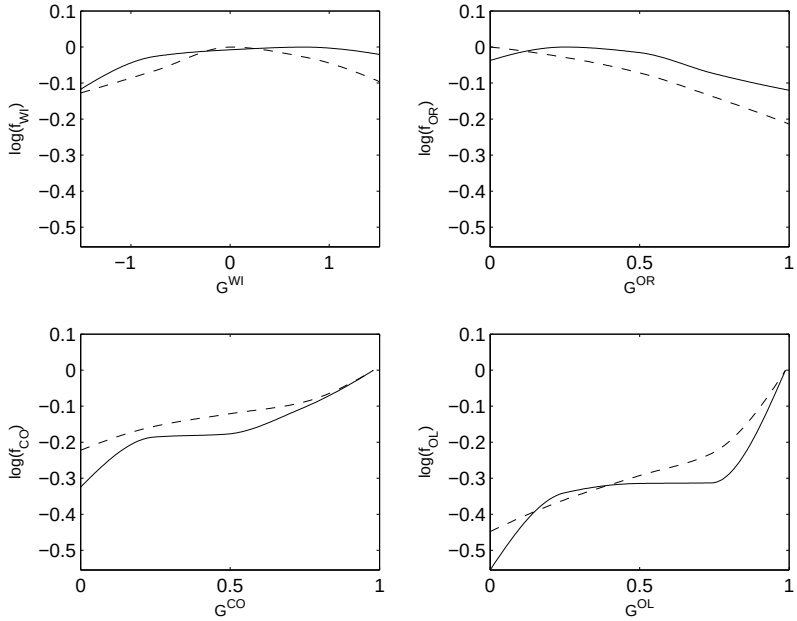


Figure 4.10: Plots of the logarithms of f_{WI}, f_{OR}, f_{CO}, and f_{OL}. Top left: Logarithm difference in the sizes of the components. Top right: Difference in angle. Bottom left: Collinearity measure. Bottom right: Measure of how much the features overlap. Solid lines: $R_{i,j}$ is the absolute value of mean of the conditioned component. Solid lines: $R_{i,j}$ is the variance of the conditioned component.

In Figure 4.10, we have plotted f_{WI}, f_{OR}, f_{CO}, and f_{OL}, when using a threshold of three, i.e. the absolute value of the active feature exceeded three times its standard deviation. You can see that f_{WI} has a maximum when the features are of the same size, f_{OR} has a maximum when the features point in the same direction, f_{CO} has a maximum when the features are maximally collinear, and f_{OL} has a maximum when the features overlap the most. If function f had linear shape, the whole model would be separable. In our experiments f had a slightly sigmoidal shape.

We identified that overlap of the conditioning and conditioned feature was the most important factor in deciding how much conditioning changes the statistics. However, other factors weighed in as well. Collinearity of the features was found to be important for the first order statistics, whereas second order statistics depended more on the difference in angle between the features. These findings offer some support to our hypotheses that first order statistics essentially 'extend' the features, and second order statistics are related to the many types of 'errors' in the fit. Our model did not explain all of the changes in the statistics, at least partly because we used Gabor functions to estimate the properties of the features, but the features were not exactly Gabor functions, cf. Publication I. However, according to control experiments, our model did capture most of dependency explainable by the parameters we used. A much more complete analysis of our results can be found in Publication VII.

In order to examine the dependencies in image data further, one could examine how the statistics change when several features are active. However, the amount of data required in such an analysis would explode with all the different pairings and activation levels. One might also entertain the thought of searching for a best fitting Gabor (from a highly over-complete basis) for each edge segment, and then conditioning the statistics of the patches on the activation of these, essentially eliminating some of the elementary errors in fitting the features.

Our analysis also pointed out how analyzing even some of the most basic statistics of 12 by 12 patches requires very large sample sizes. Thus it is questionable, how far the analysis can be taken by blindly exploring the statistical structures. In the end, to build useful models, one may have little choice but to use prior knowledge of useful image structures. Such knowledge restricts the space of possible models, and can reduce the amount of data required for solving the problem [63], or enabling solving problems for much larger window sizes [89].

Chapter 5

Discussion

This thesis discussed natural image statistics and independent component analysis. We discussed the generation process and some of the dominant low-level statistical properties of natural images. Independent component analysis, and its connections to sparse, efficient representations for images were covered. The main focus of the work was concentrated on extending the known results of using ICA on image data.

We explored how some image statistics change when a drastically different imaging technique is used. We presented methods for estimating overcomplete ICA bases for images as well as methods for estimating sparse 'independent' bases, whose components are topographically arranged according to their mutual dependencies. We also examined the dependencies between the 'independent' components without fixing such a model. A detailed presentation of the results was provided in the Publications, and an overview and discussion of the main results was given in this introductory part. We also discussed some of the connections to biological visual systems and some ways of further improving the results or extending the methods.

Nevertheless, it must be said that a lot remains to be learned of the statistics of natural images, and one has to wonder how much can be achieved by collecting image data and blindly analyzing the statistical structures of such collections. The scope of the problem is huge, and some parts of it are not well-defined. It is possible to study aspects of image statistics to develop solutions to well-defined, small problems. Larger problems related to image analysis and understanding probably require defining the eventual goals of the processing, so one can decide which statistical structures are relevant, and which are not. Therefore the problem field of computational vision is tied in with artificial intelligence in general. Precisely which statistical properties of images are not dependent on these higher level goals is unclear.

One should not underestimate what millions upon millions of years of evolution has achieved in a mammalian (human) brain. It is known that some grazing animals are able to see, as well as walk and follow their mother, within minutes of their birth, implying that many of the brain structures related to vision have been coded into the genes of the beast. This is no doubt the case even with other mammals, including man. The task of learning all significant

image statistics without a good starting point (prior knowledge) might be too demanding to achieve within the lifespan of an animal, especially as the rest of the brain would have to adapt accordingly. However, the current author believes that it should be possible to simulate all the necessary processes, whatever their specifics may be, on a sufficiently capable computer.

Until we can actually mimic all the processes relevant to the development and functionality of the human brain, we are left to speculate on which statistical properties of natural images are relevant for which purposes, and to speculate on the workings of the human visual system, guided by experiments that illuminate only very small aspects of the overall behavior. In computational neuroscience, we make the assumption that it is possible to build a solution to the overall problem of artificial intelligence piece by piece, or at least that the insights obtained along the way are useful as such. The small results in this thesis can be useful for many purposes from image compression to analysis and understanding, as well as for the possible parallels that can be drawn with the behavior of biological visual systems. The usefulness of the results is eventually decided by whether they are conducive to the solution of larger problems, which may or may not be the case.

Bibliography

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] S.-I. Amari. Natural gradient learning for over- and under-complete bases in ICA. *Neural Computation*, 11(8):1875–1883, 1999.

[3] S.-I. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

[4] R. M. Balboa, C. W. Tyler, and N. M. Grzywacz. Occlusions contribute to scaling in natural images. *Vision Research*, 41:955–964, 2000.

[5] H. B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

[6] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[7] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[8] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

[9] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[10] G. G. Blasdel. Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12(8):3139–3161, 1992.

[11] R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701, 1999.

[12] G. Buchsbaum and A. Gottschalk. Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London B*, 220:89–113, 1983.

[13] J. Bullier. Neural basis of vision. In H. Pashler, S. Yantis, D. Medin, R. Gallistel, and J. Wixted, editors, *Stevens' Handbook of Experimental Psychology*, volume 1, pages 1–40. John Wiley & Sons, 2002.

[14] J.-F. Cardoso. Source separation using higher order moments. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'89)*, pages 2109–2112, Glasgow, UK, 1989.

[15] J.-F. Cardoso. Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'90)*, pages 2655–2658, Albuquerque, New Mexico, 1990.

[16] J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.

[17] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[18] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

[19] R. Chellappa. *Digital Image Processing.* IEEE Computer Society Press, 1992.

[20] C. H. Chen, L. F. Pau, and P. S. P. Wang, editors. *Handbook of Pattern Recognition & Computer Vision.* World Scientific Publishing, 1993.

[21] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[22] S. S. Chen and R. A. Gopinath. Gaussianization. *Advances in Neural Information Processing Systems*, 13, 2001.

[23] A. Cichocki and S.-I. Amari. *Adaptive Blind Signal and Image Processing.* John Wiley & Sons, 2002.

[24] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[25] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[26] H. Damasio and A. R. Damasio. *Lesion analysis in neuropsychology.* Oxford University Press, 1989.

[27] J. G. Daugman. Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2:1160–1169, 1985.

[28] L. S. Davis, editor. *Foundations of Image Understanding.* Kluwer Academic Publishers, 2001.

[29] G. C. DeAngelis, G. M. Ghose, I. Ohzawa, and R. D. Freeman. Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *Journal of Neuroscience*, 19(10):4046–4064, 1999.

[30] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology*, 69(4):1091–1117, 1993.

[31] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *Journal of Neurophysiology*, 69(4):1118–1135, 1993.

[32] S. C. Douglas. On the convergence behavior of the FastICA algorithm. In *Proc. Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 409–414, Nara, Japan, 2003.

[33] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding*, 91:115–137, 2003.

[34] R. Durbin and G. Mitchison. A dimension reduction framework for understanding cortical maps. *Nature*, 343:644–647, 1990.

[35] J. Erikson and V. Koivunen. Identifiability and separability of linear ICA models revisited. In *Proc. Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003.

[36] J.H. Friedman. Exploratory projection pursuit. *J. of the American Statistical Association*, 82(397):249–266, 1987.

[37] W. S. Geisler, J. S. Perrya, B. J. Superb, and D. P. Galloglya. Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6):711–724, 2001.

[38] M. Girolami and C. Fyfe. Negentropy and kurtosis as projection pursuit indices provide generalised ICA algorithms. Technical report, University of Paisley, Scotland, 1997.

[39] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2nd edition, 2002.

[40] C. G. Gross, C. E. Rocha-Miranda, and D. B. Bender. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neuropsychology*, 35:96–111, 1972.

[41] M. Hämäläinen, R. Hari, R. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1993.

[42] H. H. Harman. *Modern Factor Analysis*. University of Chicago Press, 2nd edition, 1967.

[43] S. Haykin. *Neural Networks - A Comprehensive Foundation*. Prentice Hall, 2nd edition, 1998.

[44] G. E. Hinton, M. Welling, Y. W. Teh, and S. Osindero. A new view of ICA. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation*, San Diego, California, 2001.

[45] J. P. Hornak. The Basics of MRI. Available at `http://www.cis.rit.edu/htbooks/mri/`, 1996–2004.

[46] P. O. Hoyer. Modeling receptive fields with non-negative sparse coding. *Neurocomputing*, 52–54:547–552, 2003.

[47] P. O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.

[48] P. O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.

[49] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 73:218–226, 1962.

[50] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J. of Physiology (London)*, 195:215–243, 1968.

[51] P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[52] J. Hurri and A. Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691, 2003.

[53] J. Hurri and A. Hyvärinen. Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. *Network: Computation in Neural Systems*, 14(3):527–551, 2003.

[54] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[55] A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768, 1999.

[56] A. Hyvärinen, R. Cristescu, and E. Oja. A fast algorithm for estimating overcomplete ICA bases for image windows. In *Proc. Int. Joint Conf. on Neural Networks*, pages 894–899, Washington, D.C., 1999.

[57] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[58] A. Hyvärinen, P. O. Hoyer, and E. Oja. Image denoising by sparse code shrinkage. In S. Haykin and B. Kosko, editors, *Intelligent Signal Processing*. IEEE Press, 2001.

[59] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001. 504 pages.

[60] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[61] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[62] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, 20(7):1237–1252, 2003.

[63] A. Hyvärinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49:151–162, 2002.

[64] M. Inki. Topographic independent component analysis: Theory and applications. Master's thesis, Helsinki University of Technology, 2000.

[65] M. Inki and A. Hyvärinen. Two methods for estimating overcomplete independent component bases. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation*, San Diego, California, 2001.

[66] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[67] E. R. Kandel, J. H. Schwartz, and T. M. Jessel, editors. *Principles of Neural Science*. McGraw-Hill, 4th edition, 2000.

[68] Y. Karklin and M. S. Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14:483–499, 2003.

[69] S. Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. In *Proc. Int. Joint Conf. on Neural Networks (IJCNN'98)*, pages 413–418, Anchorage, Alaska, 1998.

[70] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):56–69, 1982.

[71] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.

[72] B. Kolb and I. Q. Whishaw. *Fundamentals of human neuropsychology*. W. H. Freeman & Co., 1995.

[73] A. Lee and D. Mumford. An occlusion model generating scale-invariant images. In *Proc. IEEE Workshop on Statistical and Computational Theories of Vision*, Fort Collins, CO, 1999. http://www.stat.ucla.edu/ sczhu/Workshops/sctv99/Lee.html.

[74] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[75] T. S. Lee. Image representation using 2D Gabor wavelets. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.

[76] T.-W. Lee, M. Girolami, A.J. Bell, and T.J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *Computers and Mathematics with Applications*, 31(11):1–12, 2000.

[77] T.-W. Lee and M. S. Lewicki. Unsupervised image classification, segmentation, and enhancement using ICA mixture models. *IEEE Transactions on Image Processing*, pages 270 – 279, March 2002.

[78] T.-W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(5), 1999.

[79] M. Lewicki and B. Olshausen. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A: Optics, Image Science, and Vision*, 16(7):1587–1601, 1999.

[80] M. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

[81] R. Manduchi and J. Portilla. Independent component analysis of textures. In *Proc. International Conference on Computer Vision-Volume 2*, pages 1054–1060, 1999.

[82] D. Mumford and B. Gidas. Stochastic models for generic images. *Quarterly of Applied Mathematics*, LIX(1):85–111, 2001.

[83] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[84] E. Oja. Convergence of the symmetrical FastICA algorithm. In *Proc. of 9th International Conference on Neural Information Processing (ICONIP)*, pages 1368–1372, Singapore, 2002.

[85] E. Oja and M. Plumbley. Blind separation of positive sources using non-negative PCA. In *Proc. Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003.

[86] E. Oja and M. Plumbley. Blind separation of positive sources by globally convergent gradient search. *Neural Computation*, 16(9):1811–1825, 2004.

[87] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[88] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[89] B. A. Olshausen, P. Sallee, and M. S. Lewicki. Learning sparse image codes using a wavelet pyramid architecture. In *Advances in Neural Information Processing Systems*, volume 13, pages 887–893. MIT Press, 2001.

[90] S. E. Palmer. *Vision Science – Photons to Phenomenology.* MIT Press, 1999.

[91] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 3rd edition, 1991.

[92] A. Pece. Visual inference with overcomplete linear models. *Journal of Mathematical Imaging and Vision*, 2002.

[93] D. Pollen and S. Ronner. Visual cortical neurons as localized spatial frequency filters. *IEEE Trans. on Systems, Man, and Cybernetics*, 13:907–916, 1983.

[94] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–71, 2000.

[95] C. E. Rasmussen and Z. Ghahramani. Occam's razor. *Advances in Neural Information Processing Systems*, 13, 2000.

[96] D. Regan. *Human perception of objects*. Sinauer Associates, 2000.

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

[98] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997.

[99] A. Said and W. A. Pearlman. An image multiresolution representation for lossless and lossy compression. *IEEE Trans. Image Processing*, 5:1303–1310, September 1996.

[100] M. Schervish. *Theory of Statistics*. Springer, 1995.

[101] B. Schneier. *Applied Cryptography*. John Wiley & Sons, Inc., 1996.

[102] O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001.

[103] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing*, 41:3445–3462, December 1993.

[104] S. M. Sherman and R. W. Guillery. The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society B*, 357(1428):1695–1708, 2002.

[105] M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco. On a common circle: Natural scenes and gestalt rules. *Proc. National Academy of Sciences (USA)*, 98(4):1935–1940, 2001.

[106] E. P. Simoncelli and E. H. Adelson. Noise removal via bayesian wavelet coring. In *Proc. Third IEEE Int. Conf. on Image Processing*, pages 379–382, Lausanne, Switzerland, 1996.

[107] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision*. Chapman & Hall, 1993.

[108] D.G. Stork and H.R. Wilson. Do gabor functions provide appropriate descriptions of visual cortical receptive fields. *J. Opt. Soc. Am. A*, 7(8):1362–1373, August 1990.

[109] N. V. Swindale. The development of topography in the visual cortex: a review of models. *Network*, 7(2):161–247, 1996.

[110] D. R. Tailor, L. H. Finkel, and G. Buchsbaum. Color-opponent receptive fields derived from independent component analysis of natural images. *Vision Research*, 40:2671–2676, 2000.

[111] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820, 1999.

[112] K. Tanaka. Mechanisms of visual object recognition: monkey and human studies. *Current Opinion in Neurobiology*, 7:523–529, 1997.

[113] M. Tessier-Lavigne. Visual processing by the retina. In E. R. Kandel, J. H. Schwartz, and T. M. Jessel, editors, *Principles of Neural Science*, pages 507–522. McGraw-Hill, 2000.

[114] K. Torkkola. Blind separation for audio signals – are we there yet? In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 239–244, Aussois, France, 1999.

[115] D. Y. Ts'o and C. D. Gilbert. The organization of chromatic and spatial interactions in the primate striate cortex. *Journal of Neuroscience*, 8(5):1712–1727, 1988.

[116] H. Valpola. Nonlinear independent component analysis using ensemble learning: Theory. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 251–256, Helsinki, Finland, 2000.

[117] A. van der Schaaf and J. H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, 1996.

[118] J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. Royal Society, Ser. B*, 265:2315–2320, 1998.

[119] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Society, Ser. B*, 265:359–366, 1998.

[120] V. N. Vapnik. *Statistical Learning Theory*. New York: Wiley, 1998.

[121] T. Wachtler, T.-W. Lee, and T. J. Sejnowski. Chromatic structure of natural scenes. *J. Opt. Soc. Am. A*, 18(1):65–77, 2001.

[122] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli. Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In R. Rao, B. Olshausen, and M. Lewicki, editors, *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, 2002.

[123] G. K. Wallace. The JPEG still picture compression standard. *Comm. ACM*, 34:30–44, April 1991.

[124] M. Wertheimer. *Laws of organization in perceptual forms*. Harcourt, Brace & Jovanovitch, London, 1938.

[125] R. W. Williams and P. Rakic. Dispersion of growing axons within the optic nerve of the embryonic monkey. *Proc. Natl. Acad. Sci. USA*, 82:3906–3910, 1985.

[126] S. Wolfram. *A New Kind of Science*. Wolfram Media, Inc., 2002.

[127] David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997.

[128] R. H. Wurtz and E. R. Kandel. Central visual pathways. In E. R. Kandel, J. H. Schwartz, and T. M. Jessel, editors, *Principles of Neural Science*, pages 523–547. McGraw-Hill, 2000.

[129] T. Yu, A. Stoschek, and D. Donoho. Translation- and direction- invariant denoising of 2-D and 3-D images: Experience and algorithms. In *Proceedings of the SPIE, Wavelet Applications in Signal and Image Processing IV*, pages 608–619, 1996.

[130] A. Ziehe and K.-R. Müller. TDSEP—an efficient algorithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skövde, Sweden, 1998.