Towards music perception by redundancy reduction and unsupervised learning in probabilistic models

Samer A. Abdallah

Department of Electronic Engineering,
King's College London.

2002

Towards music perception by redundancy reduction and unsupervised learning in probabilistic models

Towards music perception by redundancy reduction and unsupervised learning in probabilistic models

Samer A. Abdallah

Department of Electronic Engineering,
King's College London.

2002

Ph. D. Thesis:
Towards music perception by redundancy reduction
and unsupervised learning in probabilistic models

This thesis was set in Times Roman and Avant Garde Gothic by the author using LaTeX.

ABSTRACT

The study of music perception lies at the intersection of several disciplines: perceptual psychology and cognitive science, musicology, psychoacoustics, and acoustical signal processing amongst others. Developments in perceptual theory over the last fifty years have emphasised an approach based on Shannon's information theory and its basis in probabilistic systems, and in particular, the idea that perceptual systems in animals develop through a process of unsupervised learning in response to natural sensory stimulation, whereby the emerging computational structures are well adapted to the statistical structure of natural scenes. In turn, these ideas are being applied to problems in music perception.

This thesis is an investigation of the principle of redundancy reduction through unsupervised learning, as applied to representations of sound and music.

In the first part, previous work is reviewed, drawing on literature from some of the fields mentioned above, and an argument presented in support of the idea that perception in general and music perception in particular can indeed be accommodated within a framework of unsupervised learning in probabilistic models.

In the second part, two related methods are applied to two different low-level representations. Firstly, linear redundancy reduction (Independent Component Analysis) is applied to acoustic waveforms of speech and music. Secondly, the related method of sparse coding is applied to a spectral representation of polyphonic music, which proves to be enough both to recognise that the individual notes are the important structural elements, and to recover a rough transcription of the music.

Finally, the concepts of distance and similarity are considered, drawing in ideas about noise, phase invariance, and topological maps. Some ecologically and information theoretically motivated distance measures are suggested, and put in to practice in a novel method, using multidimensional scaling (MDS), for visualising geometrically the dependency structure in a distributed representation.

ACKNOWLEDGEMENTS

Though writing a PhD thesis may sometimes feel like a very solitary occupation, this work would never have been completed without contributions from colleagues, friends and family. First and foremost, I would like to thank my supervisor, Mark Plumbley, for his tireless efforts, his thorough and constructive criticism, always delivered with good humour, his boundless optimism when it comes to assigning deadlines, and his willingness to let me define my own course of research, even when it was perhaps a little too broad to be of any immediate use.

I would also like to thank Adam Stevenson for many hours of abstruse discussions about the 'absolute conic' (an object in projective geometry), which I must admit never fully understood, and never thought would be of any use too me at all, until it emerged that it was lurking in some of the mathematical forms in Chapter 8. I would also like to thank him for informing me what PhD really stands for: Permanent Head Damage.

Thanks go to my family and friends for their support and for putting up with the mood swings, the permanent absent minded preoccupation, and the erratic and generally unsociable behaviour that seem to be associated with the production of a PhD thesis.

Finally, thanks are due to EPSRC for funding my research and my life in London over the last four years.

CONTENTS

LIST OF FIGURES

We thrive in information-thick worlds because of our marvelous and everyday capacities to select, edit, single out, structure, highlight, group, pair, merge, harmonize, synthesize, focus, organize, condense, reduce, boil down, choose, categorize, catalog, classify, list, abstract, scan, look into, idealize, isolate, discriminate, distinguish, screen, pigeonhole, pick over, sort, integrate, blend, inspect, filter, lump, skip, smooth, chunk, average, approximate, cluster, aggregate, outline, summarize, itemize, review, dip into, flip through, browse, glance into, leaf through, skim, refine, enumerate, glean, synopsize, winnow the wheat from the chaff and separate the sheep from the goats.

—Edward R. Tufte, *Envisioning Information* (1990)

PART I

BACKGROUND AND APPROACH

1. INTRODUCTION

Listening to music is one of the more rewarding experiences available to humans. We are capable of extracting much that is meaningful from what is, at one level, nothing more than two acoustic waveforms. From these superficially unstructured signals, we are able to perceive richly structured musical forms.

At another, more practical level, audition is one of the avenues by which we come to a physical awareness of our surroundings and events within it. Again, the processes of auditory perception, as well as visual and other forms of perception, are attuned to structures of all kinds in the world.

In audition, the problem of perception has been cast by Bregman (1990) as one of *auditory scene analysis*, in direct analogy with the idea that vision is concerned with scene analysis. Given some sort of sensory 'image,' perception involves the production of a structural description of the scene, in terms of objects, their states and activities, and their inter-relationships. The difficulty of this task is well communicated by Bregman's analogy for auditory perception, which may be paraphrased as follows: imagine you are on the edge of a lake on which various events are unfolding; there is a motor boat towing a water skier, a sailing dinghy, several swimmers off one of the beaches, children skimming stones, and a dog jumping off the end of a jetty. All of these activities create surface waves of one sort or another. Now, on the beach, imagine two narrow channels dug in the sand, from the edge of the lake and up the beach a little way, with two corks floating, one in each channel, so that they bob up and down with the waves. Given only observations of the two corks, you must answer a series of questions: Can you identify and track the two boats? How many swimmers are there? Can you say when and where the dog hits the water? Can you count the number of skips made by the skimming stones? Although it seems improbably difficult, this is a fairly close analogy for what the auditory system achieves.

This is a description of auditory scene analysis at a very practical level, apparently far removed from musical experience, which would seem to be about much more than the number and locations of the instrumentalists. However, Bregman and others (*e.g.* Scheirer, 1996) suggest that music perception should indeed be considered a form of auditory scene analysis, one in which the 'objects' we are asked to identify are not just the physical objects involved in producing the music, but musical constructions such as chords, melodic and rhythmic phrases, synthetic timbres, and other "chimerical" objects formed by the fusion of disparate physical sources. The implication is that whatever principles govern perceptual organisation in the general case could also apply to music, and when so applied, could explain how and why we hear music the way we do.

1.1 Three problems and several approaches

In trying to construct artificial systems that display musical intelligence, there are at least three different problems that one might aim to solve: (a) to understand music perception in humans and to simulate the computational processes occurring in both the peripheral and central auditory systems, (b) to emulate human performance without necessarily simulating it in detail—a sort of 'black-box' approach—and (c) to implement an artificial musical intelligence capable of solving arbitrary musical problems, such as musicological analysis and transcription, without being limited by human capabilities. The first two have the same end, but try to achieve it by different means. The third has ends that may overlap with, but are not necessarily limited to, the ends of the first two. In trying to solve these problems, researchers have taken several approaches, some of which are listed below:

Auditory modelling aims to follow the processes which are thought to occur in the human auditory system, beginning with simulations of cochlear filtering and continuing with models of hair cell transduction, adaptation, masking, and autocorrelation (Lazzaro and Mead, 1989; Ellis, 1992; Mellinger, 1991).

The engineering approach is more goal oriented, aiming to solve application domain problems using a variety of signal processing and other engineering techniques to represent and analyse sounds. This was described by Leman and Carreras (1996) as the "sonological approach." Examples include the works of Dixon (2000, 2001) and Klapuri et al. (2001).

Perceptual psychology. The Gestalt theory of perception is the basis of much work on auditory scene analysis, with many of the visual grouping rules finding direct analogues in the auditory domain, once some sort of 'auditory image' has been defined (*e.g.* Mellinger, 1991; Ellis, 1994). It has also influenced higher level music theories which operate at the score level, notably that of Lerdahl and Jackendoff (1983). Another contribution from psychology is Gibson's ecological approach (Gibson, 1979) which has informed the work of Leman (1991) and Casey (1998).

Music theory has given researchers a good idea of what sort of structures might be important at higher levels of music perception (*e.g.* Lerdahl and Jackendoff, 1983; Cambouropoulos, 1998).

Connectionist models. A variety of supervised and unsupervised neural network techniques have been applied to musical problems at various levels of representation, aiming to model percepts such as pitch, tonality and rhythm. (See Todd and Loy, 1991 for a good selection of applications.) Notably, it was in the field of neural computation that the concept of unsupervised learning was developed (Barlow, 1989), and its significance (as opposed to supervised learning) has been stressed by Bharucha (1991) and Leman (1991) amongst others.

1.2 Overview of the thesis

The approach taken in this thesis is a combination of those advocated by, *e.g.* , Barlow (1989); Bharucha (1991); Leman (1991), and is based on the idea that the statistical structure of sound and musical sound can drive the development of adaptive processing strategies that would otherwise have to be discovered heuristically by a human designer. The approach is built on three main planks:

Information theory The idea that perception is best understood in the language of information theory is almost as old as information theory itself. Attneave (1954) and Barlow (1959) argued that the processing of sensory data in biological systems is directed towards reducing the redundancy (*i.e.* increasing the efficiency) of internal representations whilst maintaining adequate information content about relevant facets of the external world.

Unsupervised Learning Földiák (1990) states that unsupervised learning "exploits statistical regularities [in data] by using the large amount of unlabelled examples readily available to learn a mapping from raw data to a more meaningful internal representation." Thus, the central premise is that the data itself determines its own fate by virtue of its statistical structure; no external 'teacher' is required, and the role of the engineer is to design a system that is *capable of learning* to perform the required processing, not actually to pin down the specific processes themselves, thus avoiding the 'hand-crafting' to which Cambouropoulos (1998) objected.

Probabilistic Modelling Probability theory is the formal setting for both information theory and unsupervised learning. Adopting explicit probabilistic models for the systems under consideration means that the problems of inference and learning can be precisely defined, and the solutions quantitatively evaluated. It also allows a clean split between the model and any approximations required to implement it. The graphical model formalism (*e.g.* Frey, 1998) provides a rigorous and self-consistent framework for organising probabilistic computations in a distributed system.

The thesis is organised in to two parts. In Part I, some of the methodological issues touched upon in this introduction are discussed. In Chapter 2, an overview of perceptual theory in psychology is given, including the introduction of information theoretic arguments; it also describes some prevalent theories of music perception, and points out how these perceptual theories might fit into a formal probabilistic framework. In Chapter 3, § 3.1 some probabilistic models are described in more detail; in particular, some of the methods that have come out of the neural networks community but can be interpreted probabilistically. In the same chapter, § 3.3 describes some of the time-frequency methods that are commonly used in audio analysis and suggests that they should be judged on the same basis as the adaptive representations described in the section on probabilistic models. Chapter 4 concludes Part I by tying together the many strands of the previous two chapters and setting out the assumptions and methodology to be followed subsequently.

Part II describes some experiments, which are only the first steps in applying the approach described in the first part, yet which yield some interesting results and provide some initial confirmation of the thesis that musical structures should emerge automatically even in highly 'agnostic' unsupervised learning systems. Chapters 5, 6, and 7 describe the application of two methods of redundancy reduction to two audio representations. Chapter 8 describes a way to visualise geometrically the residual redundancy in a representation and then widens the discussion to include concepts of distance and similarity in general. Finally, Chapter 9 summarises the conclusions of Part II and suggests possibilities for further work.

1.3 Original contributions

In this thesis, two existing methods of redundancy reduction based on unsupervised learning are applied to music signals in two forms, yielding some novel results, and a novel method is developed to analyse and visualise any residual redundancy in the resulting representations.

ICA of speech and music waveforms Firstly, independent component analysis (ICA) is applied to raw acoustic data represented as fixed length windowed signals. ICA is perhaps the simplest form of redundancy reduction, being based on a linear, noiseless generative model, and resulting in an optimal *linear* representation of the data. Previously, Bell and Sejnowski (1996) applied ICA in a similar way, but only a short extract of a rather singular form of music (the sound of Tony Bell tapping his teeth) was used as training data. In the present work, two separate and large ensembles of audio data derived from two radio stations, are used, one consisting mainly of speech; the other, of music. The result is two quite different but highly structured linear representations of sound, demonstrating that the sounds themselves are highly structured and that this structure can be discovered in an unsupervised way. Further analysis of the ICA representations using a novel geometric method (see below) reveals that some musically relevant concepts have been incorporated in to them.

Sparse coding of music spectra The second redundancy reduction method applied is sparse coding, which is also derived from a linear causal generative model, but a more general one than that used in ICA, incorporating noise and the ability to develop an *overcomplete* representation. The overall function of the system is no longer linear, requiring a non-trivial inference step involving an iterative optimisation. A novel 'active-set' quasi-Newton optimisation algorithm is developed to address some of the issues specific to sparse coding, yielding faster performance than standard second-order gradient optimisation algorithms under certain conditions.

An analysis of learning in the sparse coder is undertaken, showing how its performance depends on the various details of the model and the data on which it is trained.

The sparse coder is applied to short-term Fourier magnitude-spectra derived from a piece of polyphonic music. The system detects the redundancy due to the harmonic spectra of musical notes, and 'discovers' that the individual notes are the 'independent

causes' of the polyphonic music spectra. The fact that the system is based on an explicit probabilistic causal model means that the inference step, in which the note activities are inferred from the mixed spectrum, is effective enough to drive a transcription and resynthesis of the music.

Geometric visualisation of dependency using multidimensional scaling While the author cannot claim to have invented ICA, multidimensional scaling (MDS) or the use of nonlinear correlations to measure statistical dependence, this is their first application in the particular combination required to visualise geometrically the dependency structure of a distributed representation. The representational units are considered as points in a multidimensional metric space, and a mapping from pair-wise mutual information to metric distance is proposed on heuristic grounds. Since the working hypothesis here is that perceptual process are driven by redundancy reduction, the residual dependencies indicate where further processing is required; thus any method of identifying and localising is likely to be a useful tool in building artificial perceptual systems.

When applied to the ICA representations derived from speech, the wavelet-like basis becomes embedded in a curved two-dimensional manifold, ordered according to time and frequency. Thus, a time-frequency perceptual field emerges quite naturally from the statistical structure of speech. The geometry of the music derived ICA basis is more complex: the residual dependency structure is found to contain evidence of the 12-tone chromatic scale used in Western music, and several types of harmonic relationships. These are harder to visualise in a 2 or 3-dimensional space, but the 3-D arrangement found by MDS reflects some of these relationships.

Other proposals and demonstrations A distinction between two types of similarity in distributed representations is made that has not been clearly made before.

A measure of distance based on ecological ideas and generalised noise model is proposed. The fundamental point here is that a given space is given a *metric* structure from consideration of its structure as a *probability* space.

A link between Wigner Distribution cross-terms and the statistical structure of phase invariant subspaces is demonstrated. The statistical structure of certain quadratic representations of sound (including the Wigner Distribution) is shown to be unsuitable for ICA.

1.4 Notational conventions

Bold type will be used for matrices ($\mathbf{A}, \mathbf{B}$, *etc.*) and vectors ($\mathbf{x}, \mathbf{y}$, *etc.*) when they are to be considered as *arrays*, but not necessarily for vectors in the abstract, that is, as elements of a linear space. The matrix transpose of $\mathbf{A}$ will be written as $\mathbf{A}^T$, and the determinant as $\det \mathbf{A}$. The symbol $\stackrel{\text{def}}{=}$ will signify a definition, the complex-conjugate of x will be denoted by x^*, and $\sqrt{-1}$ will be denoted by i and not j.

Although not directly used in the text, the following definitions may help to clarify the use of random variables and probability density functions. A probability space is a triplet $(\Omega, \mathcal{B}, P)$, where Ω is a set of elementary events, $\mathcal{B}$ is a σ-algebra or σ-field

(closed under countable unions and intersections) containing Ω and subsets of Ω, and $P : \mathcal{B} \mapsto \mathbb{R}$ is a measure that satisfies $P(\Omega) = 1$.

A random variable, then, is a function $X : \Omega \mapsto \mathcal{X}$ such that for each elementary event $\omega \in \Omega$ there corresponds a value $x = X(\omega) \in \mathcal{X}$. Where a clear distinction is required, upper-case symbols (X, Y, *etc.*) will be used to denote random variables and lower-case (x, y, *etc.*) will be used for particular values of them. When the value of a random variable is being considered, *e.g.* $X = 0$ or $X \in \{x_1, x_2\}$, X is actually shorthand for $X(\omega)$. Where the intended meaning is clear from the context, lower-case symbols may be used for both the random variables and their values.

Given a probability space $(\Omega, \mathcal{B}, P)$ and a random variable $X : \Omega \mapsto \mathcal{X}$, a corresponding probability space $(\mathcal{X}, \mathcal{B}_X, P_X)$ is induced, with P_X a probability measure defined so that for $A \in \mathcal{B}_X$, (and hence $A \subseteq \mathcal{X}$)

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\}). \tag{1.1}$$

If, in addition, a volume measure $v : \mathcal{X} \mapsto \mathbb{R}$ is defined on the measurable space $(\mathcal{X}, \mathcal{B}_X)$, then the probability density function $p : \mathcal{X} \mapsto \mathbb{R}$ can be defined as the Radon-Nikodym derivative $\mathrm{d}P_X/\mathrm{d}v$. If $\mathcal{X}$ is finite or countable, then the simplest measure is the cardinality function, also called the counting measure: $v(A) = |A|$, where $A \subseteq \mathcal{X}$ and is therefore countable. In this case, the density function is simply

$$p(x) = P_X(\{x\}) = P(\{\omega \in \Omega : X(\omega) = x\}). \tag{1.2}$$

If X is a continuous random variable with $\mathcal{X} = \mathbb{R}^n$, then $\mathrm{d}v$ can be taken to be the usual volume element $\mathrm{d}v = \prod_{i=1}^{n} \mathrm{d}x_i$, from which the standard definition of the probability density function for continuous variables obtains directly.

Hence, following Amari and Nagaoka (2001), a lower-case p will be used in the text to denote these generalised probability density functions both for continuous random variables, and, on the few occasions where they are required, discrete random variables. When several random variables are being discussed, it will usually be clear which density function is implied from the argument of the function, but if this is not the case, a subscript will be used, *e.g.*

$$p_X(a) = p(x)|_{x=a} = \left.\frac{\mathrm{d}P_X}{\mathrm{d}v}\right|_a. \tag{1.3}$$

An upper-case $P(A)$ will be reserved to denote the probability of the event (*i.e.* set) specified by A.

The notation $\mathrm{E}\,X$ or $\mathrm{E}\,x$ will be used for the expectation of the random variable X, which can be defined in terms of a Lebesgue integral with respect to the measure P_X or the volume measure v discussed previously:

$$\mathrm{E}\,X = \int_{x \in \mathcal{X}} x \,\mathrm{d}P_X = \int_{x \in \mathcal{X}} x p(x) \,\mathrm{d}v. \tag{1.4}$$

Again, this applies both to continuous and discrete random variables. If the expectation is to be taken over a particular distribution, such as a conditional distribution, this will be indicated as a subscript, *e.g.* $\mathrm{E}_{x|y} X$, meaning that the expectation is to be taken over the conditional density $p(x|y)$.

2. PERCEPTION

Introduction

The aims of this chapter are two-fold. The first is to present some relevant background to the problem of music perception. The second, to put it plainly, is to juxtapose in the reader's mind four strands of thought: (a) a synopsis of perceptual theory, (b) that musical experiences are indeed perceptual experiences, which is not necessarily a truism once we begin to characterise perception more precisely, (c) information theory as a tool for thinking about perception, and (d) probabilistic modelling and inference as an analogue for perceptual processes. The formal aspects of probabilistic modelling will be described in more detail in the next chapter, but it underpins many of the ideas described in this one, and also provides the basic setting for information theory.

2.1 What is perception?

According to Fodor (1983, p. 40), "what perception must do is to so represent the world as to make it accessible to thought." Thus, the central problem of perception is to construct such a representation from the collection of signals emanating from sensory transducers. Fodor goes on to say, in more precise terms, that although these signals are best thought of as representing conditions at the "surface" of an organism, the representations produced by perceptual systems are "most naturally interpreted as characterising the arrangements of *things in the world*." The process of going from one to the other is essentially one of *inference*, where "proximal stimulus" provides the evidence or "premises", and the conclusions are "representations of the character and distribution of distal objects." Similarly, Barlow (1990) describes perception as "the computation of a representation that enables us to make reliable and versatile inferences about associations occurring in the world around us."

As a conceptual framework for thinking about perception, this is by no means a recent development. For example, Berkeley, in *A Treatise Concerning the Principles of Human Knowledge* (Berkeley, 1734, § 18) wrote,

> It remains therefore that if we have any knowledge at all of external things, it must be by reason, inferring their existence from what is perceiv'd by sense.

Notwithstanding his use of the word "perceived," the gist of it is the same. Helmholtz's theories about vision and audition hold up remarkably well even today; with regard to the former, he suggested that,

> such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mecha-

> nism, the eyes being used under normal ordinary conditions. (Helmholtz, 1910, §26)

This is an early expression of the idea that perception may be concerned with inferring the worldly *causes* of sensations, the sensations themselves being incidental. Indeed, he added that "we are wont to disregard all those parts of the sensations that are of no importance so far as external objects are concerned." Addressing the commonplace observation that perception seems to be a transparent, effortless activity, Helmholtz argued that this occurs through a process of "unconscious induction."

More or less the opposite view was expressed by Gibson (1966). He claimed that "the senses can obtain information about objects in the world without the intervention of an intellectual process." He also rejected the idea that transduced sensory signals form the sole basis for perception (Gibson, 1979, Ch. 4): "The inputs of the nerves are supposed to be the data on which perceptual processes in the brain operate. But I make a quite different assumption." Ostensibly, such statements are difficult to reconcile with those presented earlier. By "intellectual process", he almost certainly did not mean what we might now call a computational process. What he probably objected to is the need for perceptual *inference*, because, he maintained, under normal conditions, there is enough information available to leave no room for uncertainty. This is an arguable point (discussed further in §2.1.4) but is made more tenable by the idea of *active perception*, which is behind the second of his statements quoted above.

Active perceptual systems are engaged not just in the passive analysis of whatever stimulation happens to be playing over an organism's receptors, but in the active exploration of an "ambient stimulus field". Thus, the visual system is responsible, not just for the analysis of retinal data, but also for the control of systematic eye and head movements designed to extract more information from the available stimulus. It is this information which removes the ambiguity that would otherwise necessitate an inferential system. For example, even slight movements of the head provide strong, unambiguous cues for localisation in 3D for both vision and audition. Biosonar in bats and dolphins is another good example of an intrinsically active perceptual system.

Incidentally, Helmholtz (1910) touched upon similar ideas while discussing his thesis that we perceive *causes*, arguing that, "it is only by voluntarily bringing our organs of sense in various relations to the objects that we learn to be sure as to our judgements of the causes of our sensations." This is a view of active perception as a kind of *experimentation*, by which we distinguish causal relationships from mere coincidences.

Though these ideas will undoubtedly play a role in a more complete understanding of perception, they are somewhat beyond the scope of the present work and will not be pursued.

Perception and sensation It is worth clarifying at this point what these terms will actually mean in the sequel (unless otherwise stated). The colloquial meaning of "sensation" could be defined as a *conscious* awareness of what something 'feels' like, and is probably what Gibson (1966, p. 99) had in mind when he questioned the idea that

sensations are somehow more basic than perceptions. In that context, he took "sensation" to mean a form of conscious awareness, but different from "perception." In particular, he did not use "sensation" to describe the activities of receptors or transducers. For Gibson, sensations are conscious experiences directed inwards rather than outwards, subjective experiences—what something feels like—rather than objective experiences—what it implies about the world out there.

That will not be the usage here. "Sensation" will describe the operation of sense organs, or more properly, receptors such as the inner hair cells in the ear or the rods and cones in the retina, and "perception" will refer both to the "knowledge of external things," and the process by which such knowledge is gained. Sensation therefore provides the raw data for perception, which data will be referred to as "sensory signals" or "raw sense data."

2.1.1 Representation and cognition

The central importance of representation has already been noted. The acquisition of perceptual knowledge can be identified with the formation of representations that fulfil certain goals, or that make plain relevant or interesting aspects of that which is being represented. Hence, we will often talk about "developing representations"—for example, Fodor (1983, p. 29) states that "contemporary cognitive theory takes it for granted that the paradigmatic psychological process is a sequence of transformations of mental representations and that the paradigmatic cognitive system is one which effects such transformations." He goes on to an informal definition of representation: (Fodor, 1983, p. 38) "*Any* mechanism whose states covary with environmental ones can be thought of as registering information about the world; ... the output of such systems can reasonably be thought of as *representations* of the environmental states with which they covary." This notion of covariance can be expressed more precisely in terms of the information theoretic quantity, *mutual information*, which measures how much information one set of variables carries about another, and reaches a maximum when the two are related by an invertible mapping. Thus, a representation could be defined as something which has a relatively 'high' mutual information with the representee. Mutual information and other information theoretic concepts are described further in § 2.2.

Representation* vs. *transmission Fodor's "paradigmatic psychological process," consisting of a sequence of representations each of which is a transformation of the previous one, can be recast as a sequence of 'black box' processing units connected by communications channels (see fig. 2.1). Marr (1982, p. 3) also notes a duality between processing and representation. In doing this, our focus is shifted to the characteristics of the potentially imperfect or noisy channels. Information theory tells us that such channels will have a limited capacity, and that the signals sent down them (*i.e.* the representation) should be tailored to the channel in order to make the best use of the available capacity. This is one of the ideas behind redundancy reduction, originally proposed by Attneave (1954) and further elaborated upon by Barlow (1959), which, depending on the channel characteristics can lead to such processing strategies

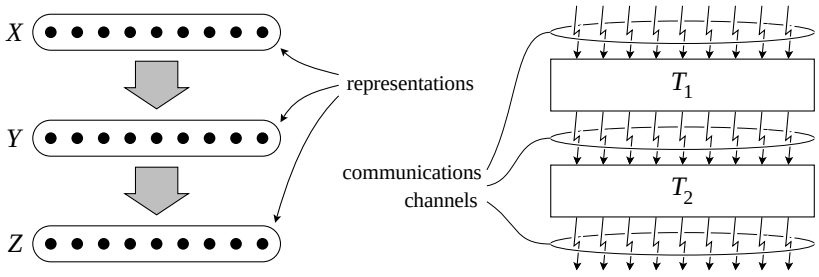


Fig. 2.1: A chain of representations X-Y-Z and an equivalent chain of processing units T_1, T_2 connected by communications channels.

as decorrelation, factorial coding, or topological map formation. All of these can be seen to result from a drive for efficient transmission of information.

Similarity and distance Richardson (1938) proposed that subjective judgements of similarity between stimuli be represented as distances between points in the Euclidean plane. This has since developed into the notion that psychological dissimilarity can be equated with geometric distance in some metric space. As Davidson (1983) notes, this is not necessarily a valid identification, as a distance measure should satisfy a number of constraints to be acceptable as a metric; for example, distances must be symmetric, whereas psychological dissimilarities need not. With this caveat in mind, the geometric visualisation of similarity is still a useful and intuitive aid to understanding.

When it comes to computing distances, questions of representation become paramount. Distance measures such as the Euclidean or Hamming distances (see §8.2) operate on pairs of objects *as represented*—move to a different representation and the distances change. If we wish an artificial cognitive system to assign distances that agree with reported dissimilarity ratings, then we would either have to find a representation such that a simple metric produced the required distances, or we would have to use a more flexible metric capable of achieving the same result with whatever representation was available. In either case, the representation will have a great effect on the details and practicality of the computation.

Shepard's principle of *psychophysical complementarity* (Shepard, 1981) addresses the issue of perceived distance. It says that mental structures should reflect certain transformational symmetries of the real world objects they represent. If an object can be subjected to a group of transformations (say, rotations) without changing its essential identity, this should be mirrored by an isomorphic group of transformations of the internal representation. These would be implemented, not necessarily as physical rotations, or a physically circular structure, but as a set of states and operators that relate in the same way, replicating, for example, the composition of large and small rotations, and the cyclic nature of rotation.

2.1.2 An ecological perspective

From a biological and evolutionary point of view, the machinery of perception is an expensive burden—in humans, the brain is responsible for a large fraction of the body's energy consumption—and must therefore confer considerable advantages to the creature that owns it. This view is more or less taken for granted; for example, quoting Shepard (1999): "The brain has been shaped by natural selection; only those organisms that were able to interpret correctly what goes on in the external world and to behave accordingly have survived to reproduce." Given that there is a limit to the amount of information that can be processed—or equivalently, that there are costs attached to information processing—it is reasonable to suppose that available resources are directed towards the detection and interpretation of events that have some *biological relevance*, such as the presence or absence of food, predators, potential mates and so on; in short, *useful* information. Moreover, there are environmental regularities that have remained constant since life appeared, and are bound to be reflected in biological perceptual systems. This amounts to the *ecological* approach to perception: one that recognises the mutual relationship between the organism and its environment.

Though the details were fleshed-out and argued with some force by Gibson (1966, 1979), the basic idea, like perceptual theorising in general, has a long history; for example, quoting Locke (1706, BII, ch. IX, § 12):

> Perception, I believe, is, in some degree, *in all sorts of animals*; though in some possibly the avenues provided by nature for the reception of sensations are so few, and the perceptions they are received with so obscure and dull, that it comes extremely short of the quickness and variety of sensation which is in other animals; but yet it is sufficient for, and wisely adapted to, the state and conditions of that sort of animals who are thus made…

By Mach's time, evolutionary ideas were current, and he was able to speculate about the biological relevance of the "sensations of tone" which "constitute the means by which we distinguish large and small bodies when sounding, between the tread of large and small animals," and also that "the highest tones presumably are of extreme importance for the determination of the direction from which a sound proceeds." (Mach, 1886, § 13.3.) Of the perception of colour, he thought it "essentially a sensation of favourable or unfavourable chemical conditions of life. In the process of adaptation to these conditions, color-sensation has probably been developed and modified." (Mach, 1886, § 6.2.)

Affordance and attensity Gibson mapped out a more complete ecological theory, stressing that an animal and its environment cannot be considered without one-another. One of the elements of the theory is the perception of *affordances*: "The affordances of the environment are what it *offers* to the animal, what it *provides* or *furnishes*, either for good or ill." (Gibson, 1979, p. 127.) He argued that these affordances are apprehended in a quite direct fashion, that an intrinsic part of the perception of an object is an appreciation of what can be done with it. An apple says "eat me", a predator says "run

away from me", a chair says "sit on me".[1]

As Gibson suggests in his definition, the distinction between *positive* and *negative* affordances (Gibson, 1979, p. 137) is an important part of the perceptual experience and of great relevance to the animal concerned. Shaw et al. (1974) coined the term *attensity* to rate the relevance or usefulness of available information to a particular organism. Things that offer positive or negative affordances would have a high attensity but things that offer neutral affordances have a low attensity, and presumably can be ignored. The relevance or otherwise of information and its possible role in defining what we mean by 'noise' is discussed further in § 2.2.5.

Nature and culture The ecological approach has had several more recent exponents (Marr, 1982; Atick, 1992; Olshausen and Field, 1996; Casey, 1998), and has been adopted specifically in relation to music by Leman (*e.g.* Leman and Carreras, 1996). It may be objected that music is not a 'natural' phenomenon, raising the interesting point that exposure to a musical tradition might constitute an environment of sorts, though a *cultural* one rather than a natural one. To the extent that perceptual systems are shaped by experience rather than by evolution, there should be no difference between natural and cultural influences; the ontogenesis of the auditory system is likely to be as responsive to one as to the other. Windsor (1995) discusses the interplay between natural and cultural factors at some length, and comes to essentially that conclusion.

Another important conclusion to be drawn from all of this is that we should use 'real', or ecologically representative data in our experiments, for example, as training data for learning systems. This data represents the 'environment' to which our would-be perceptual system should be adapted; we would expect it to exhibit different characteristics in different environments.

2.1.3 The objects of perception

Just what is it that we actually perceive? Bregman (1990, p. 9) put it like this: "The goal of scene analysis is the recovery of separate descriptions of each separate thing in the environment. What are these things?" A fairly uncontroversial answer would be that they are *objects*. Bregman goes on to discuss how objects serve as foci for all the perceptual qualities one experiences in response to a scene. This is implicit in the idea of a 'property': it must be a property of something. The object (as a mental construct) plays a syntactic role, binding together properties pertaining to the same physical object. Referring to fig. 2.2, I experience my pudding as {*green, rhubarb*} and {*yellow, custard*}, not just a mish-mash of {*green, yellow, rhubarb, custard*} (school dinners notwithstanding!)

Syntactic structure aside, what is the phenomenal manifestation of 'object-ness?' What aspect of the stimulus triggers the mental organisation? A common observation

[1] In fact, as far as my cat is concerned, my record deck also says "sit on me", but thankfully only when the lid is down. However, humans are just as capable of making fine distinctions in this area—think of what a climber (having obeyed the call of the mountain to 'climb me') does when he reaches the top: he looks with a keen eye for a hollow in the ground or a stone that will make a comfortable chair or picnic table.

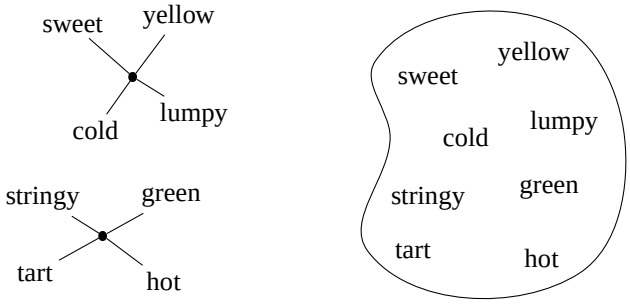


Fig. 2.2: On the left, objects as syntactic constructs, a focus for attached properties or qualities, in this case, of a rhubarb and custart pudding. Contrast this with the unstructured collection of properties on the right.

is that a reliable *association* or *correlation* between a number of features signals the presence of an object, *e.g.* , Berkeley (1734, Part 1, § 1)

> As several of these [sensations] are observ'd to accompany each other, they come to be marked by one name, and so to be reputed as one thing. Thus, for example, a certain colour, taste, smell, figure and consistence having been observ'd to go together, are accounted one distinct thing, signified by the name *apple*.

Mach (1886), in a similar vein, thought that bodies, as we perceive them, are made up of "complexes of sensations" with some relative permanence, and, according to Helmholtz (1910), "experience shows us how to recognise a compound aggregate of sensations as being the sign of a simple object."

The prevailing view is that we perceive the world in terms of objects, their qualities, and their adventures. An object has coherence and continuity, but also variability. The object, as a mental construct, also serves as a scaffold on which to affix properties that seem to be 'hanging around together' in a suspicious way. These mental objects tend to agree with external objects because the latter do indeed tend to leave a trail of physical evidence, like an animal in the woods, that sometimes leaves tracks, spoor, broken twigs, remnants of its last meal and so on; someone who knows what to look for can infer the existence of the animal, and learn a lot about it, without directly seeing it.

Gibson's invariants Gibson (1979, p. 249) took a different position, suggesting *invariance* as a more fundamental construct than objects. Invariants can only be discovered in the context of change, whether in a single thing evolving over time, or conceptually across a set of distinct things: "In the case of the persisting thing, I suggest, the perceptual system simply extracts the invariant from the flowing [stimulus] array; it *resonates* to the invariant structure or is *attuned* to it. In the case of the substantially distinct things, I venture, the perceptual system must *abstract* the invariants." (Gibson, 1966, p. 275.) In short, invariants represent things that remain constant despite change, like the shape, size, and indeed the identity of an object as it moves about.

Gestalt theory The Gestalt psychologists (*e.g.* Köhler, 1947) were concerned with organising parts into wholes. These *gestalten*, these emergent shapes, can have properties, or *gestaltqualitäten* that have meaning only for the whole and not its parts. Köhler gives the example of a musical interval: "fifth-ness" is not a quality possessed by notes, but by the gestalt, the interval. 'Major-ness' or 'minor-ness' of chords is another good example.

Gestalt theory sees perceptual organisation as a grouping process, in which in which parts of the visual image (and they were defined primarily in the context of vision) are allocated to one or other object. These have subsequently been adapted for use in audition, (see, *e.g.* Bregman, 1990; Deutsch, 1999a) largely via the adoption of auditory 'images' which take the form of time-frequency distributions such as spectrograms and the like. Energy in different regions of the time-frequency plane is grouped according to such criteria as onset synchrony, proximity in time or frequency, harmonicity, and the principle of "common fate", which recognises that sounds that undergo, for example, common frequency or amplitude modulations, are likely to have come from the same source. Another component of the theory is the principle of good *prägnanz*, which is invoked when there are competing alternative organisations suggested by the other rules, or when there is missing data. It says that 'good forms' should be preferred, where 'good form' refers to qualities such as simplicity, regularity, symmetry, or continuity. The possible interpretations of good *prägnanz* are discussed further in § 2.1.5.

What are sonic and musical objects? To the extent that perception is concerned with the physical state of the world, auditory objects are just physical objects, just like visual objects. For example, the sound of a clinking glass implies the same physical glass that a visual image would (though perhaps the reliability of the inference may be different in different cases.) A more pertinent question is, what are the auditory manifestations of physical objects? Bregman (1990, p. 9) advocates the idea of *streams*: these are individual sounds, or collections or sequences of sounds that are heard as a unit. A related concept is that of a *source*, that is, a physical object that produces a stream of auditory events. In music, the situation is more complex: there is much less correspondence between what might usefully be called 'musical objects' (such as chords or phrases; see § 2.3) and any physical objects. Bregman calls these musical objects "chimerical," (Bregman, 1990; Scheirer, 1996) seemingly denying them status as first-class objects. Their underlying causes are ultimately not physical but mental constructs in the minds of the composer and the performers, but, I think, no less 'real' or important for that, and as such, the idea that we perceive causes is more general than object based perception.

2.1.4 Mental structure *vs.* stimulus structure

Hochberg (1981) discusses the debate between those who believe that perception is driven by mental structure and those who believe that it is driven by stimulus structure. Hochberg defines stimulus structure as *intra-stimulus constraints*, and mental structure as *intra-response* constraints. Shepard (1981) put it more directly: "Does the world

appear the way it does because the world is the way it is, or because we are the way we are?"

Gestalt theory holds that perceptual organisation is an achievement of the nervous system; *gestalten* do not exist outside the organism. However, Köhler (1947, Ch. 5) did concede that this interpretation may have an objective value: it may "tell us more about the world around us", have "biological value," and "tends to have results which agree with the entities of the physical world."

Mach (1886, § 1.13, p. 29) also had mentalist tendencies: "Bodies do not produce sensations, but complexes of elements make up bodies." However, he did not seem to imply that sensations are amalgamated into bodies in some ad-hoc manner, but rather that there are statistical regularities which are exploited, *i.e.* that mental organisation is a response to ecological stimulus structure.

Gibson, with his theory of the "direct perception" of invariants, and the pick-up of information without an intervening "intellectual process," believed that perception was driven by stimulus structure, not mental structure, even if this stimulus structure is only fully available to an active perceptual system.

Top-down* vs. *bottom-up Though the argument outlined above was conducted primarily in a psychological context, it does have a practical relevance for artificial perceptual systems. Perception dominated by stimulus structure corresponds roughly with bottom-up, or data driven processing, which looks for structures in data without referring to any stored knowledge. Mental structure corresponds with top-down processes, variously known as knowledge driven or, in Psychology, schema driven processes. A schema is a structured representation of high level knowledge, such as constraints, regularities, or archetypal patterns and transitions, the use of which can reduce the ambiguity that may be present in low level data by creating expectations and narrowing the range of possible interpretations in a given context.

Bregman (1990, p. 38), discussing auditory scene analysis and stream segregation, distinguishes "primitive segregation" from "schema-based segregation". He defines a schema as "a [mental] system that is sensitive to some frequently occurring pattern." He also recognises "hypothesis driven" processes, in which the generation of a hypothesis prompts a search for confirming evidence.

The use of high-level knowledge in human perception is not in doubt (*e.g.* , Knill and Richards 1996.) A more pertinent question is, how does high level knowledge get to the top in the first place? In a "knowledge engineering" methodology, it is placed there by a human designer. In unsupervised learning, the acquisition of high level knowledge is itself a data-driven process: it collects over time, 'percolating', as it were, from the bottom up. For example, Leman and Carreras (1996) contrast "*long-term data-driven* perceptual learning with *short-term schema-driven* recognition." This is a sensible distinction to make when there is not enough structure in short spans of data to enable a fully data driven approach, but there is enough structure in data accumulated over longer periods; the schema represents this accumulation of experience.

In conclusion, to quote Shepard's answer to his own question,

> (1) The world appears the way it does because we are the way we are; and (2) we are the way we are because we have evolved in a world that is the way it is.

I would add that not only have we evolved in a world that is the way it is, but each of us has been immersed in it since the day he was born, and therefore has had ample opportunity to build and adapt his mental structures the better to suite whatever environment he happened to have grown up in.

2.1.5 Dealing with uncertainty

Sensory evidence is sometimes uncertain, inconclusive or ambiguous. Usually, we do not notice, filling in the gaps or making assumptions based on prior experience of what is likely; all of this happening below the level of conscious awareness. A good example of this is the phenomenon of perceptual restoration. For example, in speech, phonemes, spoken in context, and masked by noise are often subjectively heard without the listener even noticing that they were missing; this is known as phonemic restoration (Warren, 1970). Similar effects have been reported in music, for example, Deutsch (1999a) describes how perceptual restoration works for notes in a melodic context, in which the missing note can confidently be predicted from the surrounding material.

In Gestalt theory, this sort of situation is handled by the principle of good closure, or *prägnanz*, which says that the mind tends to complete forms with gaps in them. However, as (Bregman, 1990, p. 26) observes, we *do* sometimes see forms with gaps in them: the principle is really for completing *evidence* with gaps in it. Its job is to fill in missing or doubtful data. One may then ask, what does the brain choose to fill-in the gaps with? What is 'good' form?

Probabilistic perception Hochberg (1981) discusses how the Gestalt principle of good *prägnanz* can be interpreted as assuming either the *simplest* (in some sense), or the *most likely* resolution of an ambiguity. Can simplicity be identified with likelihood? Such probabilistic interpretations of perception are quite common in the literature, as the following few examples show. Helmholtz (1910) believed that the most *likely* explanations for sense data were arrived at by a process of "unconscious induction". Mach (1886, § 10.8, p. 213) made the following observation:

> If the visual sense acts in conformity with the habits which it has acquired under the conditions of life of the species and the individual, we may, in the first place, assume that it proceeds according to the principle of probability; that is, those functions which have most frequently excited together before, will afterwards tend to make their appearance together when only one is excited.

Brunswick (as reviewed by Hochberg, 1981) thought that the "Gestalt laws were merely aspects of stimulus organisation, reflecting the probability that any parts of the visual field belonged to the same object." Bregman (1990, p. 24) on the same subject wrote, "It seems likely that the auditory system, evolving as it has in such a world, has developed principles for 'betting' on which parts of a sequence of sensory inputs have arisen from the same source."

Looking ahead to the next section, Attneave (1954) identified good gestalt with a "high degree of internal redundancy." Pomerantz and Kubovy (1981) note that Attneave's economic descriptions (Attneave, 1954) could be interpreted as equating simplicity with economy of coding, though they add, "this interesting approach is unlikely to bear fruit until we know what coding schemes are used in perception. Changes in coding schemes can have an enormous impact on the brevity of codes." In fact, Shannon (1948) showed that coding schemes can indeed be optimised for economy and that this *defines* a measure of simplicity that equates with likelihood.

2.2 Information and redundancy

Not long after the publication of Shannon's *A Mathematical Theory of Communication* (Shannon, 1948), psychologists began to see potential applications in their own field (Miller and Frick, 1949; Miller, 1953; Attneave, 1954). Attneave suggested that information theory is an appropriate way to understand perception, noting that sensory signals carry information redundantly and proposing that perceptual systems are engaged in redundancy reduction. Interestingly enough, Oldfield (1954), without explicitly referring to Shannon's Information Theory, proposed what was in all other respects, a mechanism of adaptive redundancy reduction by successive stages of recoding. Barlow (1959, 1961) was another early advocate of redundancy reduction, arguing that it would benefit other cognitive processes.

The following sections will first examine some of the basic assumptions implicit in the marriage of perception and information theory, and will then describe what can come of such a union.

2.2.1 Information theoretic framework

Central to the application of information theory to perception is the idea that sense data can be treated as a collection of random variables that carry information about the state of the world. Whether or not these are 'random' in some objective sense, there is, from the point of view of the organism at least, a degree of uncertainty about the signals that are about to arrive and the physical events behind them—otherwise, there would be no need for sensation or perception at all. If the random variables representing sense data are denoted by X, with an associated probability function $p(\cdot)$, then the entropy $H(X)$ provides a measure of uncertainty in X:

$$H(X) = \mathrm{E}\log[1/p(X)] = -\mathrm{E}\log p(X), \tag{2.1}$$

where E denotes the expectation operator.

The idea that sense data carries information about the world can be expressed by postulating that the state of the world can also be described by a collection of random variables, S, and that the mutual information between X and S is not zero, or equivalently, they are not statistically independent. The most transparent expression for the

mutual information for our purposes is

$$\begin{aligned} I(X,S) &= H(S) - H(S|X) \\ &= \mathrm{E}\left[\log p(S|X) - \log p(S)\right], \end{aligned} \tag{2.2}$$

which is the *expected drop in uncertainty* about S on learning X.

This sort of treatment can easily be extended to include the putative internal representations computed by cognitive systems. If some internal representation Y is computed from the sense data X, then $I(Y,X)$ and $I(Y,S)$ are further quantities of interest, measuring the information in Y about the stimulus X and the state of the world S respectively. Indeed, Y can be considered a representation of the stimulus X, or the state of the world S, precisely to the extent that $I(Y,X)$ or $I(Y,S)$ are 'high' in some sense.

Gibson on information Gibson (1966) was also very much concerned with information, describing perception as a process of "information pick-up", but felt that the use of Shannon's mathematical framework was inappropriate. He rejected the idea that information is that which brings about a reduction in uncertainty, falling back on a dictionary definition: "that which is got by word or picture," which is unfortunately rather circular considering the subject matter. He also rejected the idea that sense data can be considered as signals in a communications system: "for these signals must be in code and therefore have to be decoded; signals are messages, and messages have to be interpreted." (Gibson, 1979, p. 63.) In fact, he was quite dismissive about the "vast literature nowadays of speculation about the media of communication" which he accused of being "undisciplined and vague." Barlow (1996) mentions some of these concerns, advocating the adoption of an information theoretic approach while at the same time acknowledging that Shannon's original exposition of the theory as a theory of *communication*, with a transmitter, a channel, and a receiver, "is in some ways a poor analogy for the perceptual brain, partly because we must rid ourselves of the idea that there is a homunculus to receive the messages." However, homunculus or no, there is a case to be made that different parts of the brain are in 'communication' with each other, and even if the terminology may seem a little strained when applied to perception, the mathematical forms fit very well.

2.2.2 Structure as redundancy

The idea of stimulus structure was introduced in § 2.1.4. The discussion there focussed on its implications for the division between data driven and knowledge driven processing, but did not address the question of what *structure* is. To do so, we must look at multivariate data.

Consider an n-tuple of real-valued random variables $X = (X_1, \cdots, X_n)$. The individual components can represent simultaneous observations from many receptors, or successive observations from a single receptor, or both. Hochberg's "intra-stimulus constraints" (Hochberg, 1981) imply that the observations are confined to some lower-dimensional subspace or manifold of the full space $\mathbb{R}^n$. In a stochastic system, this hard constraint can be replaced with a soft one: the distribution of X might not strictly

occupy a lower-dimensional manifold of $\mathbb{R}^n$, but it will have some sort of 'shape,' and certain sorts of 'shape' will imply statistical dependencies between the components of X. Informally then, one may characterise structure as any departure from statistical independence, which represents a state, as it were, of 'pure randomness.'

A consequence of this definition is that if X is *un*structured, then it will be impossible to predict one component from the others; that is, the conditional probabilities $p(x_i|\{x_j : j \neq i\})$ will be no different from the marginals $p(x_i)$. Conversely, 'structuredness' implies that it *is* possible to make some of those predictions, or what Attneave (1954) calls "better-than-chance inferences," because some components will carry information about others, and hence there will be *redundancy* in the data. In the case of images, for example, structure implies that some regions of an image can be roughly predicted from others. This is certainly true for symmetric or periodic data, and accords with Barlow's suggestion (Barlow, 1996) that "structure is anything that is regular, repeating, or symmetric."

The principle that redundancy constitutes structure agrees with many examples of what one would intuitively identify as structure. In images, common patterns or features, as well as being distinctive to the eye, will be a source of redundancy because the parts of the feature tend to accompany one-another: the presence of one half is a strong indicator that the other half is to be found adjacently. Objects are a source of redundancy for the same reason. Temporal coherences of many kinds, such as in the overall brightness of visual scenes, in the overall loudness of auditory scenes, and in the set of probable notes in tonal music, are all forms of redundancy, in that the present condition narrows the range of likely future conditions.

Redundancy and Gibson's perceptual systems The concept of redundancy sheds some light on Gibson's ideas about perceptual systems. One of the points he makes (Gibson, 1966) is that perceptual systems should not be defined in terms of sense *organs*, (that is, anatomically) but rather that different, possibly overlapping, sets of receptors work together to form perceptual systems, each dedicated to knowing about a different aspect of the world; they are outward looking systems rather than inward looking ones. In the current context, we might say that perceptual systems are defined by statistical dependencies between receptors. For example, the vestibular and visual organs together constitute a system for the perception of orientation, because they both respond in a correlated fashion to changes of orientation. Similarly, olfactory receptors contribute both to the perception of smells and flavours, depending on whether the signals are correlated with events at the tip of the nose, such as sniffing actions, or with tactile sensations in the mouth, chewing actions, and so on.

2.2.3 Redundancy reduction and factorial coding

Redundancy of the sort described above has two direct implications. One is that the data exhibits some interesting structure that may be worth characterising. The other is that the given representation is not as efficient as it could be. To address both of these points, Attneave (1954) proposed that the job of perceptual systems is to "strip away

some of the redundancy of stimulation," and to construct "economic descriptions" of sense data, through a process of redundancy reduction. Also in support of redundancy reduction, Barlow (1959) argued that the brain could not possibly process and store the vast amounts of information that 3 million sensory nerve fibres are capable of carrying (which he estimated to be of the order of 10^7 bits per second) if it was encoded naively. He also emphasised the need for an *adaptive* encoding: there would be some genetically specified mechanisms of redundancy reduction to deal with ever-present environmental regularities, but there should also be a learning mechanism to deal with those conditions and contingencies peculiar to the individual.

As well as the benefit of producing a more concise encoding of sensory data, the process of identifying and removing 'structured-ness' requires that the structure be, in some sense, 'understood.' It involves an implicit probabilistic model of the data, because in an optimal code, the encoding of a message event A is $\log_2 P(A)$ bits long. Thus, in order to decide how much resource to allocate to representing the message, one must know the probability of its occurrence.

The conclusion of the process would be a completely non-redundant distributed code Y, whose m elements Y_i are statistically independent—a *factorial code*, so called because the joint probability density for independent variables factorises:

$$p_Y(y) = \prod_{i=1}^{m} p_{Y_i}(y_i). \tag{2.3}$$

This is what Harpur (1997) called the "holy grail" of unsupervised learning. It means that all structure has successfully been identified and accounted for. Barlow (1996) observed that is also means that in a dynamic environment, new structure can be identified as "new regularities" or "suspicious coincidences" that "betray the presence of new causal factors in the environment."

Barlow (1990) made an argument for factorial coding based on requirements for versatile and reliable associative learning. In order to identify an association between some stimulus y and another stimulus z, an organism must notice that y accompanies z more often than would be expected by chance if y and z occurred independently. To do this efficiently it must compare $p(y,z)$ with $p(y)p(z)$, and thus requires estimates of the marginal probabilities $p(y)$ and $p(z)$. Barlow described this in the context of a single neuron, with a number of synapses representing the elements of Y. In this case, one can imagine that each synapse is responsible for building up a picture of $p(y_i)$ using locally available information only. If the Y_i are statistically independent, eq. 2.3 applies, and these marginal probabilities can then be multiplied in the cell body to yield the probability of the entire stimulus, $p(y)$.

An important point about Barlow's argument is that it explicitly recognises the desirability of modelling the probability distribution of the stimulus. Factorial coding is useful precisely in that it facilitates a particularly simple computation, but other probabilistic models might do equally well. Although a fuller discussion of probabilistic modelling is deferred to Chapter 3, it is important to bear in mind that building and using these models might be a large part of what perceptual systems do.

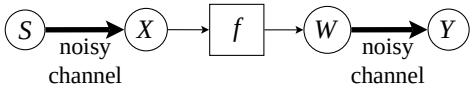


Fig. 2.3: Atick and Redlich's schematic perceptual system. S stands for the state of the world which the system is to represent. X stands for observed sense data, which in general is some noisy transformation of S. The first stage of processing computes $W = f(X)$, which is transmitted along a noisy or constrained channel and received as Y.

Quantifying redundancy Attneave's original expression for redundancy (Attneave, 1959, p. 9) was a measure of the difference between the actual entropy of a representation (computed from its 'true' probability distribution) and a theoretical maximum, $H_{\max}$, computed for that representation under any applicable physical constraints:

$$R_A = \frac{H_{\max} - H(Y)}{H_{\max}}. \tag{2.4}$$

For a neural code, an appropriate constraint might be a fixed average spike rate for each neuron. This is essentially the model that Barlow (1961) adopted in his discussion of redundancy, in which the elements of a code, Y_i are binary with $\mathrm{E}\, Y_i = \alpha$, a constant. In this case, for an N-bit binary code,

$$H_{\max} = N\{\alpha \log \alpha + (1-\alpha)\log(1-\alpha)\},$$

which obtains when the bits are independent. For a real-valued code, the constraint might be a limited range of values or a fixed average power, under which the maximum entropy distributions are uniform or Gaussian respectively. In general, if the constraints apply to each element of the code independently, then the maximum entropy will be obtained with a factorial code.

2.2.4 Information maximisation

Linsker (1988) pointed out that a high-entropy representation is not necessarily an *information* bearing one. Referring back to § 2.2.1, if S denotes the state of some relevant part of the world, then the important quantity is not the entropy of the code, $H(Y)$, but the information it carries, $I(Y,S)$. He proposed that one of the organising principles of perceptual systems is that they should adapt to maximise (under constraint) the mutual information between the internal representations and the environmental states they purport to represent. This is known as the *principle of information maximisation* or "info-max" for short. The constraints generally conspire to restrict the overall representational capacity of the system, encouraging the efficient representation of relevant information and the rejection of noise.

Atick and Redlich (1990) took these considerations into account when they proposed an alternative measure of redundancy based on mutual information rather than entropy. The representation is treated as a constrained communications channel between the perceptual system and higher cognitive systems (as previously illustrated in

fig. 2.1). The constraints act to limit to the capacity of that channel. In the archetypal system illustrated in fig. 2.3, the maximal information transfer $I_{\max}$, is determined by the capacity of the channel between W and Y:

$$I_{\max} = \max_{p(w)} I(Y,W)\Big|_{\mathrm{E}\,Y^2=\text{const.}} \tag{2.5}$$

where the maximisation is over all possible probability distributions of W, with a fixed *power* constraint. In these terms, the redundancy is defined as

$$R_{AR} = \frac{I_{\max} - I(Y,S)}{I_{\max}}. \tag{2.6}$$

Comparing this with Attneave's measure of redundancy in eq. 2.4, both are a fraction of *wasted capacity*, but whereas R_A is sensitive only to the statistical structure of the representation itself, R_{AR} is also sensitive to capacity wasted carrying 'noise', that is, information that is not about interesting things in the world, as encapsulated by S.

There are two approaches to minimising either form of redundancy. One is to maximise the 'content' of the representation, that is, either $H(Y)$ or $I(Y,S)$, the latter equating with Linsker's info-max. In Atick and Redlich's system, this would by done by adjusting the transformation f to make the best use of the W-Y channel . The other is to reduce the 'capacity' $H_{\max}$ or $I_{\max}$, holding the 'content' fixed. This can loosely be thought of as minimising the *cost*, whether measured in numbers of spikes or energy consumption, of coding a given fixed amount of information.

2.2.5 Noise and irrelevant information

Thus far the assumption has been that we know enough about the statistical structure of S to compute $I(S,Y)$; that is, we know the difference between signal and noise. Is this a valid assumption? What are those differences?

Marr's "fundamental hypothesis" (Marr, 1982) was that perceptual systems will discard information that is not relevant to the detection of "stable features," which he defined as those which tend to co-occur. This can be imagined as a sort of voting system amongst receptors: if only one unit reports some stimulus, it is likely to be ignored on the basis that it is probably in error, an aberration peculiar to the internal workings of that unit. But if many units show concerted activity, it is more likely that this is due to events in the world.

Marr's hypothesis amounts to an assumption about the structure, or rather the lack of structure, in noise, namely, that it affects each receptor independently of the others. Thus, it determines which features in the input are going to be important: those which are robust to such noise. It also corresponds exactly with the idea of redundancy: stable features are encoded redundantly, whereas noise is not. Implicit in this is a certain faith in the 'structured-ness' of the world.

Attneave (1954) also noticed that uncorrelated sensory signals tend to be interpreted as noise, observing that even though uncorrelated noise has a high entropy, and thus could potentially carry a lot of information, both visual and auditory white noise have a dull, uniform, and uninformative texture. Redlich (1993) makes a similar observation:

"Desirable input information is often encoded redundantly, so redundancy can be used to distinguish true signal from noise."

These comments have the status of empirical observations, *viz.*, that noise *tends* to appear non-redundantly, to be *structureless*, implying that any structuring is not noise and therefore relevant. This is quite a strong statement. The ecological principles outlined in §2.1.2 suggest that relevance, or *attensity*, is ultimately something which cannot be decided without considering the use to which the information will be put. Atick and Redlich (1990) comment that "neither *noise* nor *signal* is a universal concept, but depend on what is useful visual information to a particular organism in a particular environment." It is well to recognise that this kind of "usefulness" is not something that can be judged in the purely unsupervised framework adopted in this thesis, and to solve real problems that require judicious selection and rejection of information, some element of supervision or reinforcement will eventually be needed.

2.3 Musical structure

The purpose of this section is to give the reader some feeling for what is meant by *musical structure*, bearing in mind the preceding discussions of structure as redundancy, and perception as probabilistic inference. Some familiarity with musical concepts is assumed.

Some types of musical structure are easily identified: for example, the periodicity that characterises pitched sounds; the hierarchy of importance attached to different pitches depending on tonality and harmony; the structuring of time due to metrical regularity and phrasing structure. Other structures, such as those underlying different elaborations of a common theme, are recognisable but harder to characterise rigorously.

West et al. (1987) discuss this "common sense" level of understanding in terms of musical objects:

> With minimal introspection it is possible to identify quite a few *objects* in musical experience—notes, duplets, triplets, phrases, chords, tunes, themes, variations, movements, choruses, songs, violin parts, rests, crescendos, glissandi and so on.

As well as objects, they also recognise *features* of objects, (like pitch, timbre, metre, idiom, loudness, emphasis, tonality, "Wagnerian quality" [*sic*], richness *etc.*) and *functional relationships* between objects, such as harmonic progression, prolongation, anticipation, closure, resolution *etc.*

These observations may be useful, but introspection is not necessarily a reliable guide to building an understanding of psychological processes. Lerdahl and Jackendoff (1983, p. 3) refer to a "musical intuition" guided by "largely unconscious knowledge." There is also a distinction to be made between structure *as experienced*, and structure *as constructed*, that is, compositional structure. This is discussed in §2.3.3, but before moving on, composition, it should be noted, is a psychological process too and therefore not necessarily accessible to consciousness.

2.3.1 Lerdahl and Jackendoff's generative theory

Lerdahl and Jackendoff (1983) took as their goal a description of the "musical intuitions" of an experienced listener. They explicitly limited their scope to Western tonal music, acknowledging that a different theory (at least in the details) would be required to account for the "musical intuitions" of a listener versed in another idiom. Neither is the theory concerned with *how* these intuitions are acquired, though the authors do expect there to be some learning involved.

Inspired by work in linguistic grammars, the theory is expressed as a formal, generative, musical grammar: a set of set rules by which all grammatical "utterances" (that is, well-formed musical passages) can be constructed. These are not intended for the actual composition of music, but for the assignation of structural descriptions of how a given example *would have been* generated by the grammar. The rules fall into two classes—*well-formedness rules* and *preference rules*—and four categories covering different aspects of musical structure: (a) the grouping of events into a hierarchy of phrases, (b) metre, (c) "time-span reduction," a hierarchy of structurally important pitches, and (d) "prolongation reduction," which describes the flow of tension and release in harmony and melody. The well-formedness rules are absolute and indicate which structural descriptions are permissible. The preference rules are flexible and are used to help decide between competing descriptions. Lerdahl and Jackendoff make a point of not giving any quantitative procedure for deciding this competition, citing the difficulty of assigning numerical weights to the rules. One of the benefits of using a *probabilistic* generative model would be a principled way of doing this by choosing the most *likely* structures.

An important aspect of the theory is that the four sets of rules set up four parallel temporal decompositions of the music, each of which is strictly hierarchical. Four inverted trees are built along the same time line, one encoding metre, another phrase boundaries, and so on. The preference rules encourage trees with parallel structures, but the music can dictate otherwise. Indeed, Lerdahl and Jackendoff suggest that pieces giving rise to such commensurate trees elicit a feeling that the music is a little simplistic or boring. It is possible that the tensions set up by conflicting temporal structures are responsible for some of the satisfying richness in music, for example, in syncopation, the structure assigned by a metrical analysis conflicts with the local structure of phenomenal accentuation.

2.3.2 Event hierarchies and tonal hierarchies

Bharucha (1984) describes Lerdahl and Jackendoff's temporal decompositions as *event hierarchies*, contrasting them with *tonal hierarchies*, which, in tonal music, assign a hierarchy of relative importance to different pitches which may occur. Krumhansl (1990, p. 18) also discusses this, citing Meyer (1956, pp. 214–215), who made an assessment of the tonal hierarchy implied by a major key context. For example, in the key of in G major, the most *structurally stable* important tone is the tonic, G, which serves as the reference point for all the other tones. Next come the remaining notes of the major triad, D and B, followed by rest of the pitches in the diatonic scale of G

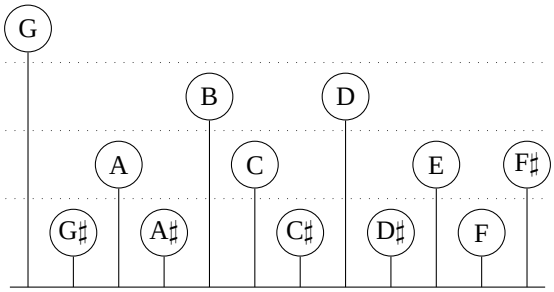


Fig. 2.4: A tonal hierarchy for the key of G major.

major: A, C, E and F♯. Finally, the remaining chromatic notes (G♯, A♯, C♯, D♯, and F) are the least stable (see fig. 2.4). Thus, any note can be assigned a position in the hierarchy on the basis of the the tonality of the context in which it is found. The lower in the hierarchy it is, the less likely it is to 'fit,' or form a comfortable resting point.

Bharucha held that human listeners internalise these relationships in the form of tonal schemata, and, in a series of experiments, Krumhansl (1990) and her co-workers showed that this is indeed the case. She also found that, to a good approximation, the profile of stability ratings assigned to the twelve pitches in the context of a particular key is strongly correlated with the relative probabilities of those pitches occurring in that key. This suggests that the subjects had within them a statistical model of the patterns of pitch use. The obvious first approximation to such a model which would be a table of relative frequencies (in the sense of counts) of the pitches in each key. The next level of complexity would be a table of *transition* probabilities between pairs of notes, that takes into account the sequence of the notes. The evidence is that human listeners do achieve at least this level of sophistication, in that our perception of tonality is affected by a re-ordering of the notes (Deutsch, 1984).

Statistical regularity of the sort implied by tonal structure is precisely what was discussed in §2.2: namely, a form of *redundancy*. The implication is that it may be possible to model pitch perception (including tonal hierarchies) via a process of redundancy reduction. Deutsch (1984) suggested as much when she argued that tonal hierarchies enable pitch structures to be encoded more efficiently. In terms of redundancy, one might say that tonal schemata encode and exploit the redundancy that results from being in a particular key.

2.3.3 Structure and learning in music

The perceptual approach attempts to account for the listener's experience, that is, to describe musical structures as they are perceived. These need not necessarily agree with the structures that the composer or performer intended, perhaps working to some abstruse compositional theory. Leman (1999) points out that music theory tends to describe structures as they *should be* (*i.e.* from a compositional standpoint), but not necessarily how they are *perceived*. Raffman (1993, p. 30) discusses this possibility in terms of alternative grammars, (in the following, "M-grammatical rules" refers to the

psychologically internalised rules that we use to understand music):

> . . . who's to say the correct analysis of the music isn't prescribed by a different set of purely compositional principles? The composer of a tonal work (a fortiori of an atonal one) may follow compositional rules that are not psychologically real and hence no bear no signifi cant relationship to the rules that govern listening. In that case, since an appeal to the compositional rules would likely provide the most coherent and compelling analysis of the music, one one might reasonable insist that *they*, and not the M-grammatical rules, specify the structure of the work.
>
> . . . it is hard to see how experienced listeners could acquire knowledge or understanding of a piece just by hearing it; and *surely* [Raffman's emphasis] tonal music is the sort of thing that can be known in the hearing.

The second paragraph seems questionable in the light of the perceptual theory outlined in the first two sections of this chapter, which points to a situation in which listeners *do* attain high-level knowledge ("M-grammatical rules", schemata) from experience, not necessarily from listening to a piece once, but from repeated exposure to a genre. Raffman does acknowledge this alternative, citing Bharucha's hypothesises (Bharucha, 1984) that familiarity with pieces of music in a given style results in an unconscious abstraction of structural relationships in the music. The resulting schemata then facilitate the subsequent perceptual organisation of music in that style. This suggests that initially, music in a previously unheard style will be interpreted under some pre-existing schema, and will probably not be 'understood' in the sense that the composer intended; it may seem 'difficult,' and may not be fully appreciated. After becoming acquainted with the style, one might infer the new compositional rules and thus come to a better understanding of the work or others cast from the same mould.

2.3.4 The geometric structure of pitch

In this section, we examine the distinction between (non-musical) pitch as a basic auditory percept—which is generally assumed to be a *one-dimensional* correlate of frequency—and *musical* pitch, which seems to be a more complex, *multidimensional* percept. One approach is to suppose that pitches can be represented as points in some, possibly multidimensional, metric space, so that their relationships are expressed spatially, as described in § 2.1.1.

At first blush, it is tempting simply to identify pitch with frequency (or fundamental frequency for harmonic tones) and think no more of it—imagine the tones as points arranged according to frequency along a straight line, perhaps with a logarithmic frequency scale, as illustrated in fig. 2.5(a). The problem with this 1-D representation is that it does not do justice to the generally perceived relationships between pitches with widely separated fundamental frequencies, such as at the Octave (frequency ratio 1:2) or the Fifth (ratio 2:3). Consider the perceived similarity between the three C pitches spread over three octaves in fig. 2.5(a): it is not reflected in the rectilinear configuration, but after a bit of thought (or a leap of insight) one might choose to coil up the logarithmically scaled line into a helix, as in fig. 2.5(b), so that the frequency goes up by a factor of two for each turn around the coil. Similarity at the octave is now

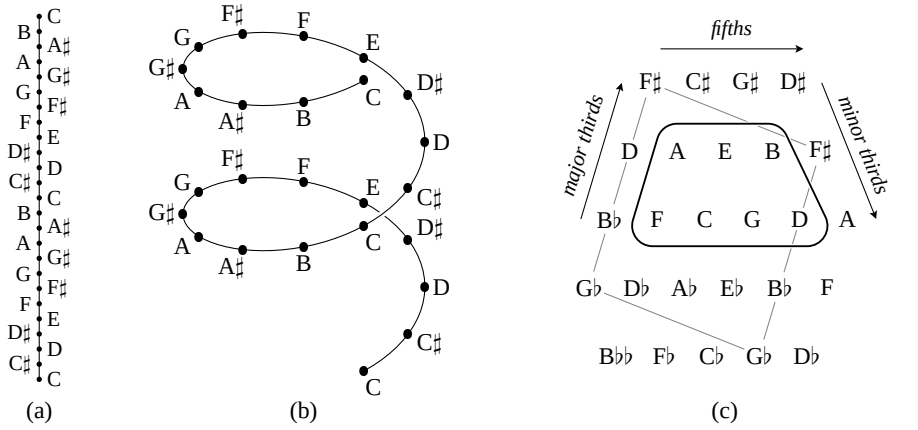


Fig. 2.5: Geometric Pitch Structures: (a) a linear representation of tone height in semitones; (b) a helical representation combines tone height and similarity by octaves (after Shepard, 1982); (c) a 2-D map arranged by fifths and thirds. If enharmonic equivalence is recognised (*e.g.* between F♯ and G♭) then the map becomes periodic in both directions and the grey quadrilateral becomes a repeating cell. The compact region outlined in black contains the diatonic scale of C major.

expressed by a short displacement parallel to the axis of the helix, while similarity in frequency is expressed by small rotations around the axis. Let us recognise that this is quite a clever trick—we have incorporated new information, by jumping from a very simple geometric arrangement to a more complex one in a higher-dimensional space, while retaining the benefits of our geometric intuitions.

In a musical context, experienced listeners are aware of other relationships than the Octave, most notably the Fifth and its inverse the Fourth (frequency ratios, 2:3 and 3:4 respectively) but also major and minor Thirds (4:5 and 5:6) and their inverses, the minor and major Sixths (5:8 and 3:5). How can these relationships be represented geometrically? This sort of exercise has been testing the ingenuity of music theorists and instrument builders since the 19th century. Keislar (1987) reviewed several keyboard designs from the 16th century onwards, which arrange the keys (digitals) in two dimensions in an attempt to bring out their tonal relationships. Ellis, in his appendix to Helmholtz (1885), described a 2-D map of pitches, with one direction moving in Fifths, the other in Thirds. Fig. 2.5(c) is an example of this kind. Schoenberg (1969) drew a similar map to represent key-relatedness (that is, tonal centres, not keys on a keyboard) with separate entries for the major and minor keys. Longuet-Higgins (1987, ch. 7 & 8) proposed an elegant explanation for these 2-D maps in terms of the prime factorisation of the frequency ratio between two notes.

Shepard (1982) explored several geometric representations of pitch, seeking to accommodate melodic (step-wise) relationships in his "melodic map," and harmonic relationships in his "harmonic map." The latter is essentially the same as the ones described by Ellis, Schoenberg, and Longuet-Higgins. However, Shepard went on to embed this 2D map in a 5D space to accommodate simultaneously four types of similarity: proximity in frequency, in octaves, in fifths, and in thirds.

These structures are not just theoretical constructs. The experiments of Krumhansl and Kessler (1982) have shown that musically experienced listeners are able to internalise the distance relationships implied by the various geometries, so that similar structures can be recovered from empirical data. From a different perspective, Leman and Carreras (1996) produced a 2D map of chords using data-driven unsupervised learning techniques rather than human subjects. This is important in the current context because it agrees with the main premise of this report: that musical structure is inherent in musical data, and that any mental structures that we use to process music are there in response to the structure of the data.

Summary

Sensory data is structured in a statistical sense, and the concept of redundancy captures much of this structure. Objects appear as complexes of correlated sensations, and the redundancy induced by this is what betrays the presence of an object and what triggers the perceptual system in its task of organising sensory data.

Music is richly structured and an initial inspection suggests that this structure is indeed captured by the notion of redundancy. Some aspects of music perception are consistent with the hypothesis that the brain builds statistical models of musical phenomena.

Perception itself seems to be a multifaceted process which can be construed as any or all of the following: (a) representation of the environment; (b) gaining information about relevant things in the world; (c) construction of statistical models of sensory data; (d) probabilistic inference of environmental states from sensory data; (e) construction of a causal framework for 'explaining' sensations in terms of hypothetical external objects and events; (f) efficient representation of relevant information through a process of redundancy reduction.

These operations are related in that efficient representation involves a statistical characterisation of the data, and causal models are often the most appropriate for dealing with data generated by causal processes in the world.

The goal of building statistical models of sensory data is one that can in principle be achieved by data-driven unsupervised learning; that is, by using models that adapt to the environment in which they finds themselves. This agrees with the psychological concept of schemata which develop through experience over the life-time of the individual.

3. COMPUTATIONAL MODELS

Introduction

This chapter considers two approaches to auditory modelling and music processing. The first follows directly on from the previous chapter, putting perception in a probabilistic setting and describing in more detail how efficient statistical models can be constructed. Several techniques developed in the neural networks and signal processing communities are listed, and their interpretations as statistical models described.

The second approach is the one that has been more prevalent in music processing until fairly recently, what Cambouropoulos (1998) called the "knowledge engineering" approach, and consists of the application of signal processing and other engineering methods combined with musical fore-knowledge on the part of the designer. The focus will be on linear and quadratic time-frequency representations, and how these relate to some of the adaptive linear representations to be described in § 3.2

3.1 Probability models

Chapter 2 presented two related ways of looking at perception: one is the idea that the brain is involved in building statistical models of sensory data; the other is that it is driven by redundancy reduction and the need to deal efficiently with information in the sense defined by Shannon (1948). Both require a probabilistic characterisation of the data, so either way we are led to the same problem: given some multi-variate random variable X, how do we build up a picture of and do computations with its probability distribution $p(x)$?

For discrete random variables, this could be done by compiling a contingency table consisting of the probabilities of observing each distinct state of X. However, the size of this table grows exponentially with the dimensionality of X: in most cases, storing the table explicitly would be out of the question. Not only would it require an inordinate amount of storage, but filling in the probabilities by counting observations (that is, building a histogram) could be a very lengthy and unreliable process, depending on the sparsity of observations in different parts of the state space. In effect, we would be building a parametric model with as many free parameters as there were distinct states of X. With continuous random variables, the problem is compounded further, since there are an infinite number of states.

Pearl (1988) argues that the key to probabilistic computation in large data spaces is to identify which variables are *relevant* to each other: it allows us to make statements about *conditional independence* between sets of variables, of the form "X is

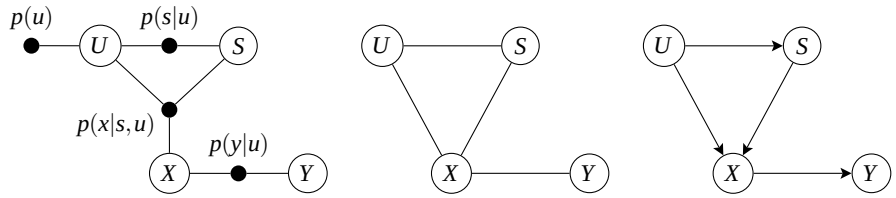


Fig. 3.1: Directed and undirected graphical models. (After Frey, 1998, pp. 10–12.) A *factor graph* (a) explicitly represents the factorisation of the joint probability as $p(y,x,s,u) = p(y|x)p(x|s,u)p(s|u)p(u)$. A undirected graph or *Markov random field* (b) for the same system implies only a factorisation of the form $p(y,x,s,u) = \phi_1(y,x)\phi_2(x,s,u)$. A directed graph or *Bayesian network* (c) uses directed edges to represent conditional probabilities, in this case giving $p(y,x,s,u) = p(y|x)p(x|s,u)p(s|u)p(u)$.

independent of Y given that we know Z." It amounts to making assumptions about the functional form of $p(x,y,z)$, namely, that it factorises, for example, as $\phi_1(x,z)\phi_2(y,z)$. This results in a model with fewer degrees of freedom, thus making training the model from the data through unsupervised learning a more realistic proposition.

A *graphical model* is a particularly convenient and expressive way of codifying assumptions about conditional independence between random variables. The variables are drawn as nodes in a graph, and edges linking them signify certain kinds of dependence and therefore relevance. Graphical models come in several flavours, some of which are illustrated in fig. 3.1. Of these, perhaps Frey's *factor graphs* (Frey, 1998) express the factorisation of the probability function most explicitly, but it is a simple matter to write down the implied factorisation for a directed graphical model, and only slightly less so for an undirected one.

Jordan (1998, Preface) succinctly lists the benefits which flow from this approach: "Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with ... uncertainty and complexity..." They are fundamentally modular systems, and "probability theory provides the glue whereby the parts are combined." Not only is a graphical model a good way to represent a probability model, it also defines an architecture for parallel distributed computation on that model, which in part explains the strong ties between graphical models and neural networks that have developed in recent years: Ghahramani and Roweis (1999) point out that many algorithms developed under the rubric of neural networks (and in other fields) have direct interpretations in terms of probabilistic systems and graphical models; some of these will be examined later in § 3.2.

3.1.1 Latent variables and hidden causes

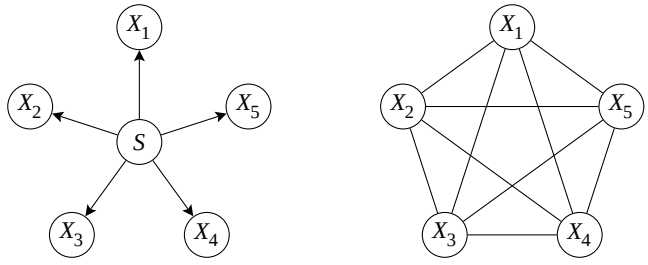
We have seen how a graphical model represents a factorisation of the joint probability distribution of random variables, yet it may not always be possible to find a factorisation sufficient to render the model tractable. If no conditional independencies can be found, the result will be a *complete graph*, that is, one with edges connecting all

pairs of nodes. Faced with such a problem, Pearl (1988, p. 383) argues that the natural human response is to hypothesise some underlying cause for the observed correlations, and that this response is *computationally motivated*. In a graphical model, this means that we *invent* new hidden or latent variables, which though never observed, are able to account for the observed dependencies using a simpler graph than was possible without them. In this way, Bishop (1998) observes, a relatively complex joint distribution over the observed variables can be obtained, by marginalisation, from a simpler distribution over the expanded set of variables. Letting X stand for all the observed variables, S for the latent variables, and $\mathcal{S}$ for range of S, the model defines the joint distribution $p(X,S)$. The observed distribution is then

$$p(x) = \int_{s\in\mathcal{S}} p(x,s)\ \mathrm{d}s. \tag{3.1}$$

Consider the model illustrated in fig. 3.2. If the central node S went unobserved and unrepresented, a complete graph might be required to model the resulting dependency structure. By hypothesising the existence of a sixth unobserved variable, we open up the possibility that a much simpler graph might be capable of modelling the system. This particular graph is directly analogous to Pearl's example (Pearl, 1988, p. 383): if one were to ask five people in the street what time it was, they would give strongly correlated answers. Rather than suspecting some conspiracy between them, it is far more economical to postulate the existence of an underlying 'correct' time, upon which each person independently bases their response.

From a certain point of view, these hidden causes are pure invention; nonetheless, their 'reality' in any physical or metaphysical sense is perhaps of less importance than the computational advantage which they bring, and it is quite conceivable that many of the mental constructs and abstractions we entertain as humans have precisely the same status. Pearl goes as far as to suggest that "these computational advantages ... give rise to the satisfying sensation called in-depth understanding, which people experience when they discover causal models consistent with their observations." This need not be confined to the sort of conscious reasoning Pearl is referring to, but could ap-



(a) (b)

Fig. 3.2: How the introduction of latent variables into a graphical model can simplify the dependency structure. Suppose the (causal) model (a) accurately captures the dependency between a set of observable variables X_i, but the variable S is not observed. An attempt to model the dependencies without including S would result a graph with many more edges, such as the undirected model (b).

ply just as well to unconscious perceptual processes. There are many parallels to be drawn between the graphical models described here and the discussion of perception in Chapter 2. If sensory data constitute the observed variables, then the latent variables represent mental constructs our perceptual systems invent to explain correlations in sensations: precisely the role ascribed to objects in §2.1.3. It is these latent variables which constitute a representation of the external world, which we do not experience directly, yet is the most economical explanation for sensory data. Indeed, in a graphical model, there is no representation at all without latent variables. Add to this their inherent ability to deal consistently with uncertainty, and their potential computational efficiency, and we are lead to the conclusion that latent variable models are a promising computational model of perception.

3.1.2 Inference and learning

Probabilistic inference A typical application of a latent variable model is to infer likely values of the hidden nodes given a set of values for the visible ones. Probability theory dictates how this should be done: before observing X, the marginal density of S is

$$p(s) = \int_{x \in \mathcal{X}} p(x,s)\, \mathrm{d}x, \tag{3.2}$$

where $\mathcal{X}$ is the range of values taken by X. The situation after observing that $X = x$ is described by the *posterior* density

$$p(s|x) = \frac{p(x,s)}{p(x)}, \tag{3.3}$$

where $p(x)$ is defined by eq. 3.1. If the model is *causal* one specified in the form $p(x,s) = p(x|s)p(s)$, then the standard form of Bayes' rule is obtained:

$$p(s|x) = \frac{p(x|s)p(s)}{p(x)}. \tag{3.4}$$

$p(s)$ is the *prior* distribution of S, and is now a given, rather than being computed according to eq. 3.2; it represents what is initially known about S. $p(x|s)$ represents new information about S, gleaned from the data x, and when considered *as a function of* s, is called the *likelihood*, written $l(s|x) \equiv p(x|s)$.

In some applications, the posterior can be used directly, but in those cases where a single 'best' estimate of S is needed, some additional criterion is required. Common choices are to minimise the mean square estimation error by picking the mean of $p(s|x)$, or to pick the single most likely value of S,

$$\hat{s} = \arg\max_s p(s|x), \tag{3.5}$$

otherwise known as *maximum a posteriori* or MAP estimation.

Depending on the structure of the graph, a number of exact and approximate inference algorithms may be applied, some of which take advantage of the graphical representation of the model to reduce the computations required. Examples include *message passing* schemes—variously known as *belief propagation* (Pearl, 1988), the

sum-product algorithm and the *forward backward algorithm* (see Frey, 1998, p. 27)—the *junction tree algorithm* (see Cowell, 1998), and *variational inference* (Jordan et al., 1998). Many of these have the property that the computations can be done locally on the graph and thus implemented efficiently in a distributed fashion.

Learning by induction One of the motivations in using a probabilistic model is that the details of the model can be learned from the data. In a graphical model, there are two aspects to this: one is that the structure of the graph itself may be induced, including the addition of hidden nodes as suggested by data; the other is that any parameters associated with each node may be adjusted. Pearl (1988, Ch. 8) treats the former, while Heckerman (1998) reviews work on both aspects. The discussion here will be confined to parametric optimisation only, and applies to any parametric probability model and not just graphical models.

A parametric model defines a distribution over the observables X given by $p(x|\theta)$, where θ stands for all the parameters. Fitting the model means finding the parameters which best describe the observed data, or, in a Bayesian setting, (see, *e.g.* Heckerman, 1998) finding the posterior distribution over the parameters given the data:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}. \tag{3.6}$$

The prior $p(\theta)$ represents our state of knowledge about the parameters before the observation. If the data consists of a set of T independent observations, (a training set) then the posterior factorises as

$$p(\theta|\{x^{(t)} : 1 \le t \le T\}) = p(\theta)\prod_{t=1}^{T}\frac{p(x^{(t)}|\theta)}{p(x^{(t)})}. \tag{3.7}$$

MAP estimation of θ involves finding the mode of this posterior, but in the case where the training set is large, the data may be assumed to influence the shape of the distribution much more than the prior $p(\theta)$; if the prior is dropped, then a *maximum likelihood* (ML) estimate can be obtained by maximising the log-likelihood:

$$\hat{\theta} = \arg\max_{\theta}\sum_{t=1}^{T}\log p(x^{(t)}|\theta). \tag{3.8}$$

Since we are only interested in the position of the maximum and not its value, we are free to drop the normalisation factors $p(x^{(t)})$ and convert products into sums by taking logarithms. Furthermore, as $T \to \infty$, this sum approximates an expectation:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\log p(x^{(t)}|\theta) = \int_{x\in\mathcal{X}} p(x)\log p(x|\theta)\,\mathrm{d}x \equiv \mathrm{E}\log p(x|\theta), \tag{3.9}$$

where $p(x)$ is the 'true' (*i.e.* observed) distribution of X. Maximisation of this expected log-likelihood is formally equivalent to minimisation of the Kullback-Leibler divergence between the observed distribution and the model distribution, since the two a related by

$$D(p_x\|\,p_{x|\theta}) = \int_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{p(x|\theta)}\,\mathrm{d}x = -\mathrm{E}\log p(x|\theta) - H(X), \tag{3.10}$$

where $H(X)$ is the entropy of the data and is therefore independent of θ. Thus, in the limit of a large training set, MAP estimation, ML estimation, and divergence minimisation are all equivalent.

In the situation where there are latent variables, the procedure is more complicated, since the model distribution $p(x|\theta)$ is now a potentially intractable sum or integral over hidden states, as in eq. 3.1. Heckerman (1998) reviews some of the available methods, including *Monte Carlo* integration, the *expectation–maximisation* (EM) algorithm (Dempster et al., 1977) and the *Gaussian approximation*. Only the last of these will concern us here, as it is the method which Lewicki and Sejnowski (2000) apply to the problem of overcomplete coding—the details are deferred to Chapter 6.

3.1.3 What probabilistic models can do

Chapter 2 presented some of the arguments for a probabilistic approach to perception, and this section has described some of the formal models that could implement this approach. These are some of the capabilities that such a model would inherit in terms of tasks relevant to perception:

Novelty detection The most immediate consequence of building a probability model is that any stimulus can be assigned a numerical probability, which can be used to flag those input patterns that seem unlikely according to the model. This could indicate that something new and interesting is happening, or that the model is inadequate and needs to be re-thought.

Regression If part of the input is missing or thought to be unreliable, the model can be used to supply the most likely completion, in precisely the way that perceptual restoration (see § 2.1.5) is thought to operate.

Creating fantasy data Any probability model can in principle be used to generate samples from the model distribution using a variety of Monte Carlo methods (see MacKay, 1998). The perceptual relevance of this is less certain, but one might speculate that such a process is involved in dreaming.

Inference of latent variables In a latent variable model in which the hidden variables represent the worldly causes of sensory stimulation, the process of inference equates to the maintenance of an internal representation of the external world.

Optimal coding In models where the latent variables are independent—which is one of the goals of causal modelling; see Pearl (1988)—they constitute of themselves an efficient, non-redundant representation. In other cases, the knowledge of the data distribution can be used to guide the construction of an optimal code by determining the code-word length or coding cost for each pattern, according to Shannon's (1948) coding theorem .

Learning The potential of suitably constrained probability models to learn by induction provides a rigorous framework for implementing such aspects of perception as schemata and Barlow's adaptive redundancy reduction, and more specifically, the

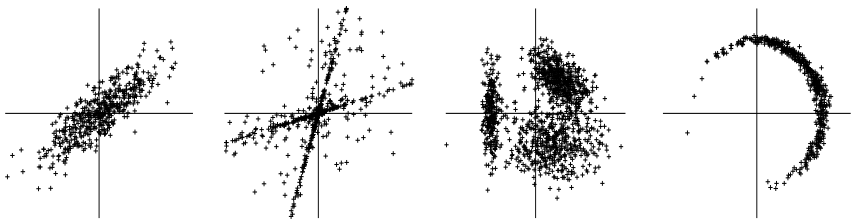


Fig. 3.3: Different data distributions, with structures suited to different probability models: (a) Gaussian data—use PCA, (b) non-Gaussian data—use ICA, (c) clustered data—use Gaussian mixture model, and (d) data on nonlinear manifold—use SOM or GTM. (See text for description of these models.)

acquisition of musical knowledge through listening. The prospect of learning *structure* from data, inventing hidden variables to account for dependencies in the data, is a particularly exciting one as it begins to approach (arguably) what we mean by *intelligence*. For example, the perceived harmonic progression of a piece of music could be represented by hidden variables introduced to account for the redundancy in patterns of note usage.

Principled comparison with other candidate models (Bishop, 1998) notes that the adoption of explicit probability models allows a quantitative comparison between different candidate models of a given data set. Leaving questions of model complexity aside, the best model is the one that assigns the highest likelihood to the data. The probabilistic formulation also allows modular combinations of models to built in a principled manner.

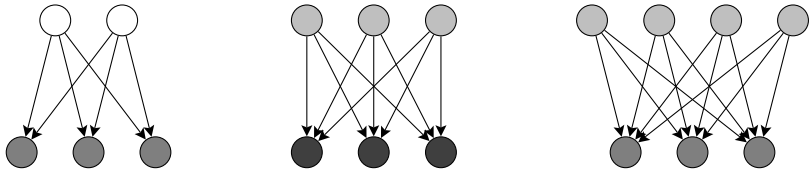
Having looked at probabilistic systems in general, we will now focus on some specific models which will form the background for the methods to be applied in Part II.

3.2 Algorithms for unsupervised learning

This section contains a brief review of some unsupervised learning algorithms and the models implicit within them. Many of them can be fitted into the framework of graphical models described in the previous section, and are suited to different forms of data distribution, some of which are illustrated in fig. 3.3. Ghahramani and Roweis (1999) conduct a more extensive discussion along these lines, showing how many data analysis and signal processing methods can be derived from graphical models.

3.2.1 Principal component analysis and whitening

Principal Component Analysis (PCA; see *e.g.* , Plumbley, 1991) is a well-known technique for dimensionality reduction, whereby high-dimensional data is projected linearly into an M-dimensional subspace whilst maximising the variance of the resulting projection. That subspace is the one spanned by the M eigenvectors of the data covariance matrix with the largest eigenvalues, and is termed the principal subspace.



(a) PCA (b) ICA (c) Sparse coder

Fig. 3.4: Graphical models for (a) probabilistic PCA, (b) ICA, and (c) sparse coding. Hidden nodes are at the top and visible nodes at the bottom. The white nodes are Gaussian, where as light grey nodes are non-Gaussian. The medium gray nodes are conditionally Gaussian and are used to model additive noise in PCA and in sparse coding. The dark grey nodes in the ICA model are deterministic linear functions of the hidden nodes.

In its basic form, PCA is not model based, and might be seen as being a rather ad-hoc procedure; for example, Plumbley (1991) noted that it is not invariant to arbitrary rescaling of the data. However, in certain situations, PCA can be shown to be optimal in a well-defined sense. Linsker (1988) showed that PCA maximises the information transmitted by a linear system on the assumption that the input data is Gaussian with additive spherical Gaussian noise. Plumbley (1991) showed that this is not necessarily true for non-Gaussian data, but that PCA does minimise an upper bound on the information *lost* in the projection into a lower-dimensional subspace.

PCA relies exclusively on second-order statistics (means and covariances) and, as the previous comments suggest, is best suited to Gaussian data, like that in fig. 3.3(a). Tipping and Bishop (1997) formalised this by showing that PCA could indeed be derived from a Gaussian latent variable model which they called "probabilistic PCA." The observed variables x_i are generated by linear combination of uncorrelated unit-variance Gaussian latent variables s_j using

$$x_i = \sum_{j=1}^{M} A_{ij} s_j + e_i, \quad 1 \le i \le N, \tag{3.11}$$

where the A_{ij} are fixed combination weights, and $M < N$ for dimensionality reduction. The e_i are uncorrelated Gaussian noise variables of variance σ_i^2. This results in a Gaussian causal model in which

$$p(s_1,\ldots,s_M) = \prod_{j=1}^{M} \tfrac{1}{\sqrt{2\pi}} \exp -\tfrac{1}{2} s_j^2 \tag{3.12}$$

$$\text{and} \quad p(x_1,\ldots,x_N|s_1,\ldots,s_M) = \prod_{i=1}^{N} \tfrac{1}{\sigma_i\sqrt{2\pi}} \exp -\tfrac{1}{2\sigma_i^2} \left(x_i - \textstyle\sum_{j=1}^{M} A_{ij} s_j\right)^2 \tag{3.13}$$

Thus, the latent variables are marginally independent, and the observed variables are marginally dependent, but conditionally independent given the latent variables. The equivalent graphical model is illustrated in fig. 3.4(a). Maximum likelihood estimates of the parameters A_{ij} and σ_i can be obtained using the EM algorithm, but when the noise is assumed to be isotropic, with the σ_i all equal, the solution can be obtained in closed form. Considering the weights A_{ij} as an $N \times M$ matrix $\mathbf{A}$, Tipping and Bishop

(1997) proved that the data likelihood is maximised when the the M columns of $\mathbf{A}$ span the M-dimensional principal subspace of the N-dimensional data space.

Probabilistic PCA is also related to the technique of *sphering* or *whitening*. Plumbley (1991) discusses how data sphering arises naturally from the principle of information maximisation when *noiseless* Gaussian data must be passed through a noisy, energy-constrained output channel; this is equivalent to redundancy minimisation in the sense of Attneave and Barlow (see eq. 2.4 in §2.2.3.) Both Plumbley (1991) and Atick and Redlich (1990) derived the correction to this when input noise is taken into account; this is equivalent to minimising redundancy as defined in eq. 2.6.

3.2.2 Independent component analysis

Independent component analysis (ICA— Jutten and Herault, 1991) can be thought of as a generalisation of data sphering to non-Gaussian data vectors $\mathbf{x}$, where the aim is to find a linear transformation $\mathbf{s} = \mathbf{W}\mathbf{x}$ which results not just in uncorrelated, but independent output components s_j. Methods based on second-order statistics are invariant to a final rotation of the latent variable space since a spherical Gaussian distribution is spherically symmetric. If the latent variables are non-Gaussian, however, this symmetry breaks down and 'special' directions emerge, as can be seen in fig. 3.3(b).

Restricting themselves to the so-called 'square' problem, in which $\mathbf{s}$ and $\mathbf{x}$ are of the same dimensionality and hence $\mathbf{W}$ is a square matrix, Bell and Sejnowski (1995) introduced an algorithm based on information maximisation. This was later shown (Cardoso, 1997; MacKay, 1996) to be equivalent to maximum likelihood estimation in an explicit probability model, illustrated in fig. 3.4(b). The observed vectors $\mathbf{x} \in \mathbb{R}^N$ are generated according to $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s}$ is a vector of independent source components and $\mathbf{A}$ is a square mixing matrix. The s_j may be thought of as the coordinates of $\mathbf{x}$ relative to a basis formed by the columns of $\mathbf{A}$, in which case the $\mathbf{A}$ may be called a basis matrix. The independence assumption means that $p(\mathbf{s}) = \prod_{j=1}^{N} p(s_j)$, where $p(\cdot)$ is a non-Gaussian probability density assumed *a priori*. Since the mapping $\mathbf{s} \mapsto \mathbf{x}$ is deterministic, linear, and invertible, the probability model for $\mathbf{x}$ is easily derived:

$$p(\mathbf{x}) = \det \mathbf{A}^{-1} p_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x}), \tag{3.14}$$

though writing it in this form obscures the benefit gained from the factorisation of the prior $p_{\mathbf{s}}(\cdot)$. The Bell-Sejnowski algorithm is recovered by parameterising $\mathbf{A}$ as $\mathbf{W}^{-1}$ and estimating $\mathbf{W}$ by gradient ascent on the log-likelihood. We will return to ICA in Chapter 5, where it will be applied to audio data.

3.2.3 Sparse coding

In sparse coding, we aim to represent data using a distributed representation in which only a 'few' elements are 'active' at a time. This is essentially what Barlow (1961) was arguing for when he argued for an economy of *impulses* in a neural representation, rather than economy of *neurons*.

One approach to sparse coding is to assume that a code vector $\mathbf{s} \in \mathbb{R}^M$ represents a data vector $\mathbf{x} \in \mathbb{R}^N$ according to $\mathbf{x} = \mathbf{A}\mathbf{s}$, as in PCA and ICA. The difference is

that the s_j are assumed to be zero with high probability, and we allow $M > N$. The representation is therefore underconstrained, with many vectors **s** representing a given vector **x**. This freedom can be used to minimise the number of non-zero elements of **s**. Further sparsity can be obtained if small reconstruction errors are tolerated, allowing components of **s** with small values to be set to zero. An adaptive sparse coder optimises the matrix **A** to produce the sparsest possible coding, choosing for the columns of **A** the M vectors that are best suited to representing the data. These vectors are sometimes called an overcomplete dictionary, or an overcomplete basis.

Field and Olshausen (1996) implemented such a system by minimising an heuristically constructed cost function. When trained on natural images, the resulting overcomplete basis consisted of a number of localised, oriented, band-pass features, sharing many similarities with the coding found in primary visual cortex. Olshausen (1996) showed that the same algorithm could be derived as an approximation of maximum likelihood learning in a certain latent variable model, illustrated in fig. 3.4(c). The salient features are: (a) there are more hidden nodes than visible ones, (b) the hidden nodes are highly non-Gaussian, with 'sparse' densities, in a sense to be defined in Chapter 6, and (c) the visible nodes have added Gaussian noise, so that the model deals explicitly with noisy data. Inference in this model performs an implicit 'de-noising' of the data, a feature which has been exploited by Hyvärinen (1999).

Lewicki and Sejnowski (2000) used a different and generally more accurate approximation to exact learning. Further details, including the probability model and the derivation of the learning algorithm, will be found in Chapter 6.

3.2.4 Clustering algorithms and mixture densities

More versatile probability models can be built by combining other models. One way of doing this is by constructing *mixtures*, where the probability distribution is modelled as a weighted sum of simpler distributions. Data can be generated from such a model by first choosing one of the source distributions at random, and then drawing a sample from that distribution. This is a system with one *discrete* latent variable U, which encodes the identity of the source distribution that was actually used to generate the data. The resulting probability model is

$$p(\mathbf{x}) = \sum_{u=1}^{M} p(\mathbf{x}|u)p(u), \tag{3.15}$$

where $p(\mathbf{x}|u)$ is the uth source distribution and $p(u)$ its weighting, the probability of choosing that distribution during the generative process. Bishop (1998) provides a general description of the procedures for inference and learning in mixture models.

When the source distributions are relatively disjoint, mixture models can provide a principled way to do clustering analysis and categorisation on data such as that illustrated in fig. 3.3(c). Each cluster or category is modelled as one source distribution, so that inference of the latent variable u given the data **x** naturally results in a probabilistic categorisation of the data.

3.2.5 Topographic maps

The self-organising map (SOM), also known as the topographic feature map or Kohonen map (Kohonen, 1982, 1995), is an unsupervised algorithm for finding a low-dimensional (usually 2-D) nonlinear manifold in high-dimensional data. It maps a low-dimensional discrete lattice smoothly into the data space in such a way that most of the data lies near the mapped lattice points. For example, in fig. 3.3(d), a one-dimensional SOM would fit a string of lattice points along the elliptical arc suggested by the data. Data is then represented by identifying which lattice point lies nearest the data point. The algorithm is designed to encourage neighbouring lattice points to represent to near-by positions in the data space.

Kohonen's algorithm does not have a direct interpretation in terms of a generative model, or even as the optimisation of some objective function. Bishop et al. (1998) addressed this with their *generative topographic mapping* (GTM), which is essentially a mixture model in which each lattice point forms the centre of a spherical Gaussian cluster, but the cluster centres are constrained to lie along a smooth manifold in the data space. In this model, it is not the distinctness or separation of the clusters that is the main concern, but their arrangement, which is designed to preserve distance relationships between points in the lattice after they are mapped into the data space. Luttrell (1994) developed an alternative probabilistic system for topographic mapping based on a noisy vector quantiser.

3.3 Sonological methods

In PCA and in ICA, a statistical characterisation of the data and its representation enables us to think about an optimal linear representation, or a 'best basis.' This section describes some of the standard signal processing methods that attempt to deconstruct the signals linearly in time and frequency. We will see that in many cases, these methods carry an implicit assumption about what this best basis is, raising the possibility that the adaptive methods described above might be capable either of arriving at the same representations in a data driven way, or of finding better representations, at least according to the statistical criteria of redundancy and independence.

3.3.1 Frequency analysis and filter-banks

That the ear is basically a mechanism for doing frequency analysis is an idea that goes back to Ohm and Helmholtz (see 1885, p. 34), the theoretical justification for which was Fourier's theorem on the decomposition of periodic functions into a sum of sinusoids. Subsequently, the majority of auditory theories have been built upon a cochlear model that is functionally a bank of linear band-pass filters, or resonators, though they may differ in details of implementation. For example, Lyon (1984) used a cascade of second-order filters to build the desired frequency responses, whereas Patterson et al. (1988) opted for a set of independently specified "gammatone" filters, which Slaney (1993) implemented using an eighth-order section for each filter in the bank. Regardless of these details, the operation performed is a convolution: the output

of the kth filter is given by

$$y_k(t) = \int_{-\infty}^{\infty} x(t-u)h_k(u)\,\mathrm{d}u, \tag{3.16}$$

where $x(t)$ is the original signal and $h_k(t)$ is the impulse response of the kth filter. This is usually assumed to be causal, with $h_k(t) = 0$ for $t < 0$. The Fourier transform of the the impulse response gives the frequency response of the filter:

$$\tilde{h}_k(\omega) = \int_{-\infty}^{\infty} h_k(t)e^{-i\omega t}\,\mathrm{d}t, \tag{3.17}$$

where a tilde ($\tilde{}$) denotes the Fourier transform of a function. For this to be interpretable as a *frequency* analysis, the filter frequency response should be largely concentrated around a single, nominal centre frequency, ω_k. In this case, it will be convenient to write the impulse response as the real part of the product of a complex exponential and a slowly varying envelope or window function, $H_k(t)$, so that $h_k(t) = \Re H_k(t)e^{i\omega_k t}$. In the frequency domain, the Fourier transform of the window function, $\tilde{H}_k(\omega)$, will be concentrated around zero; multiplication by a complex exponential simply translates it along the frequency axis, giving $\tilde{h}_k(\omega) = \tilde{H}(\omega - \omega_k)$.

Psychoacoustic experiments (*e.g.* Plomp, 1976; Pickles, 1988; Rasch and Plomp, 1999) have investigated the properties of the effective cochlear filters, one of which is that the bandwidths are roughly proportional to the centre frequencies, approximating a constant-Q filter-bank. This bandwidth is called a *critical band*; depending on the experimental procedure used to measure it, (see Moore and Sek, 1995) numerical estimates vary between $\frac{1}{7}$th and $\frac{1}{9}$th of the centre frequency of a particular filter, down to a minimum bandwidth of between 20 and 50 Hz at a centre frequency of around 120 Hz. It is well known (*e.g.* Gabor, 1947) that high frequency resolution (*i.e.* narrow bandwidth) requires a filter with a long impulse response and thus a lower time resolution. Thus, the cochlear filters trade frequency resolution for time resolution at high frequencies, but keep a minimum of time resolution even at very low frequencies.

3.3.2 Time-frequency representations

Vector spaces and basis vectors A filter acts as a linear operator on the signal; in particular, the output at a given time t is, in the terminology of linear algebra, a *linear functional* of the signal; that is, a linear mapping from a vector (the signal) to a scalar. Let us assume that the input signal is a member of a vector space V. A linear functional $h^* : V \mapsto \mathbb{R}$ is then a member of the *dual* vector space V^*. (This and the following results can be found in any textbook of linear algebra, *e.g.* Stoll and Wong 1968.) Now, any linearly independent indexed set of functionals $\{h^*_\alpha \in V^* : \alpha \in \mathcal{A}\}$ (where $\mathcal{A}$ is the set of indices) defines a corresponding linearly independent set of vectors $\{a_\alpha \in V\}$ such that $h^*_\alpha(a_\beta) = \delta_{\alpha\beta}$, where δ is the Kronecker delta. These vectors will form a basis of some subspace S of V, so that any $x \in S$ has a unique coordinate expansion of the form

$$x = \sum_{\alpha\in\mathcal{A}} \xi_\alpha a_\alpha \quad \Longrightarrow \quad h^*_\alpha(x) = \sum_{\beta\in\mathcal{A}} \xi_\beta h^*_\alpha(a_\beta) = \sum_{\beta\in\mathcal{A}} \xi_\beta \delta_{\alpha\beta} = \xi_\alpha,$$

and thus, the functionals give the coordinates of x relative to the basis $\{a_\alpha\}$. In addition, if there is an inner product $(\cdot,\cdot)$ defined on V, then for any functional $h^* \in V^*$, there exists a unique vector $h \in V$ such that $h^*(x) = (h,x)$ for all $x \in V$. This defines an isomorphism between V and V^*, which will be useful when we wish to consider functionals as signals in their own right, with their own time-frequency structure.

Returning to the filter-bank, if the output of each filter is sampled in discrete time, and the filters and sampling times chosen to ensure the linear independence of the corresponding linear functionals, then the system will define a basis of some subspace of the signal space, and the output of the filter-bank over time will be a representation of signals within that subspace.

Tiling of the time-frequency plane The preceding discussion demonstrates that a filter-bank defines a set of linear functionals to be applied to a vector (the signal). If the functionals are linearly independent, the analysis is equivalent to finding the coordinate expansion of a vector relative to a certain basis. A time-frequency representation requires some additional structure.

Firstly, the functionals or basis vectors must be assigned a two dimensional topological structure (separate from that implied by any norm defined on the vector space) with time and frequency defining the coordinate frame. Depending on whether time and frequency are to be discrete or continuous, the functionals may form a lattice, a continuous manifold or a hybrid of both (as in a continuous time filter bank).

Secondly, the notional labelling of each functional with a time and a frequency must be borne out by the operation it actually performs; that is, it must operate on the signal within a time-frequency 'window.' For a filter, this can be interpreted as requiring that the impulse response be localised in both time and frequency, but more generally, the duality between linear functionals and vectors means that we can consider the time-frequency localisation of any functional h^*_α by mapping it into the signal space, using either the mapping $h^*_\alpha \mapsto h_\alpha$ for an inner product space, or $h^*_\alpha \mapsto a_\alpha$ when the functionals define a basis.

Gabor (1947) discussed the idea of time-frequency localisation in relation to hearing by analogy with the position-momentum uncertainty principle of quantum mechanics. Chui (1997, Ch. 2) also describes the construction of time-frequency windows. If $h(t)$ is a signal and $\tilde{h}(\omega)$ its Fourier transform, the mean epoch t_0 and RMS (root-mean-square) duration $2\Delta_t$ are defined as

$$t_0 = \frac{\int t|h(t)|^2 \,\mathrm{d}t}{\int |h(t)|^2 \,\mathrm{d}t}, \qquad \Delta_t = \left\{ \frac{\int (t-t_0)^2|h(t)|^2 \,\mathrm{d}t}{\int |h(t)|^2 \,\mathrm{d}t} \right\}^{1/2} \tag{3.18}$$

Similarly, the mean frequency ω_0 and RMS bandwidth $2\Delta_\omega$ are defined using

$$\omega_0 = \frac{\int \omega|\tilde{h}(\omega)|^2 \,\mathrm{d}\omega}{\int |\tilde{h}(\omega)|^2 \,\mathrm{d}\omega}, \qquad \Delta_\omega = \left\{ \frac{\int (\omega-\omega_0)^2|\tilde{h}(\omega)|^2 \,\mathrm{d}\omega}{\int |\tilde{h}(\omega)|^2 \,\mathrm{d}\omega} \right\}^{1/2} \tag{3.19}$$

In short, these are the first and second moments of the signal's *energy distribution* in the time and frequency domains considered separately. Together, they define the size and position of a window in the time-frequency plane. The uncertainty principle requires

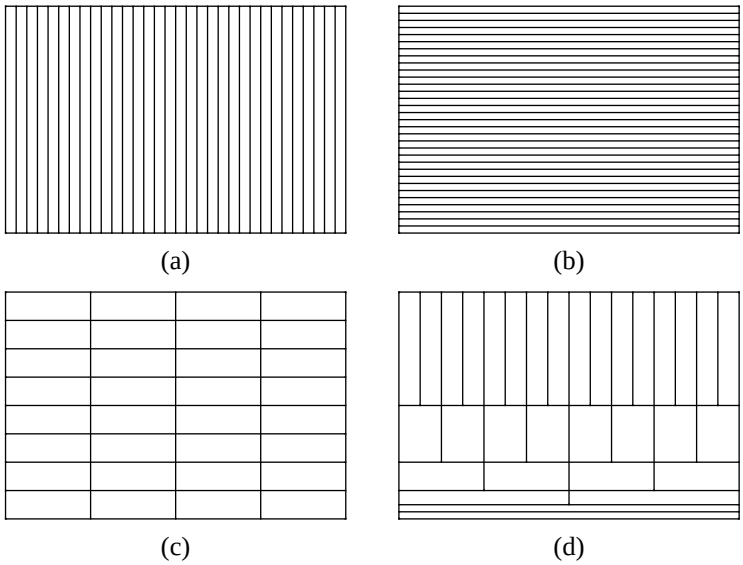


Fig. 3.5: Some alternative tilings of the time-frequency plane (with time horizontal and frequency vertical.) At the two extremes we have (a) the signal in the time domain with the maximum time resolution (1 sample) but no frequency discrimination and (b) the (global) Fourier transform with the maximum frequency resolution but no temporal structure. A short-term Fourier transform (c) uses a rectangular lattice, whereas a wavelet transform (d) uses a frequency dependent tiling, with greater time resolution at high frequencies and greater frequency resolution at low frequencies. Note that every tile in this fi gure has the same area.

that $\Delta_t \Delta_\omega \geq \frac{1}{2}$, a limit which is approached by signals with approximately Gaussian energy profiles in both time and frequency, such as Gaussian windowed sinusoids. Thus, each functional or basis vector of a linear representation can be thought of as occupying a certain region of the time frequency plane; if the representation is suitably designed, they will form a patchwork that covers all relevant regions of the time-frequency plane.

Quadratic, 'energy-like' representations We have just seen how the squared modulus of the signal in either the time or frequency domains yields an 'energy distribution' in time or frequency. These distributions have the convenient property that, by Parseval's identity, the total energy computed from either is the same. Since the point of time-frequency analysis is presumably to localise signal activity in time and frequency simultaneously, it would be useful to be able to consider the squared modulus of a linear time-frequency representation similarly as an energy distribution in time *and* frequency. One desirable property of such a distribution is that the energy should 'add-up' properly: over the whole distribution to give the correct total, but also across frequency or time to give energy densities in time or frequency that agree with those defined by the squared modulus of the signal and its Fourier transform. This question of having the correct marginal distributions will not concern us here, but is discussed by Jeong and Williams (1992) in relation to the Wigner distribution (see below).

The main property of interest here is that the squared modulus of a complex quantity is *phase invariant*. If the linear representation is defined so that each functional returns a complex value whose argument encodes the phase of a locally sinusoidal component of the signal, then the derived quadratic representation will be phase invariant in that sense. Of course, this represents a loss of information, but the evidence from psychoacoustics, beginning with Helmholtz, is that the ear is insensitive to certain kinds of phase information, and hence a phase invariant representation may be a good starting point for a computational model of human hearing.

Some examples In a short term Fourier transform, (STFT) the effective filters consist of windowed complex exponentials, with all frequencies sharing the same window. They all have the same duration and bandwidth, which translates into a uniform rectangular tiling of the T-F plane, illustrated in fig. 3.5(c). The quadratic version of the STFT is the spectrogram or short-term power spectrum.

In a wavelet transform (Chui, 1997) the analysing wavelets (filters) are all dilated and translated versions of a single prototype and the effective bandwidth is directly proportional to the centre frequency. Thus, the tiling rectangles are of different aspect ratios in different parts of the T-F plane as shown in fig. 3.5(d). If the wavelet basis is orthogonal, the basis vectors are the same as the analysing wavelets; otherwise, the basis vectors are dilated, translated versions of a single "synthesising" wavelet; this forms a *bi-orthogonal* wavelet transform. The quadratic version of a wavelet transform is sometimes called a *scalogram* (Flandrin and Rioul, 1990).

3.3.3 The Wigner-Ville distribution and Cohen's class

The Wigner-Ville Distribution (Cohen, 1989) is a quadratic time-frequency representation constructed in such a way that the time-frequency uncertainty principle does not apply: it is not the squared modulus of any filter response or linear function of the signal. The distribution $W(t,\omega)$ is defined as the Fourier transform of an instantaneous auto-correlation function $R_x(t,\tau)$ along the τ direction:

$$R_x(t,\tau) = x(t+\tfrac{1}{2}\tau)x^*(t-\tfrac{1}{2}\tau), \tag{3.20}$$

$$W_x(t,\omega) = \int_{-\infty}^{\infty} R_x(t,\tau)e^{-i\omega\tau}\,\mathrm{d}\tau, \tag{3.21}$$

where the asterisk (*) denotes complex conjugation. Since the Fourier transform is invertible, any bilinear function of the signal and its complex conjugate can be expressed as a *linear* function of the Wigner distribution, and hence it provides a unifying framework for understanding all the quadratic time-frequency representations of real signals. It is invertible, up to complex phase factor, or in the case of real signals, a factor of ± 1. As previously mentioned, it does not force a trade-off between time and frequency resolutions, and has many other desirable properties, including having the correct time and frequency marginals as described above; see Jeong and Williams (1992); Loughlin (1991) for a fuller discussion. One of the less desirable properties is illustrated in fig. 3.6: the so-called 'interference' between components of the signal.

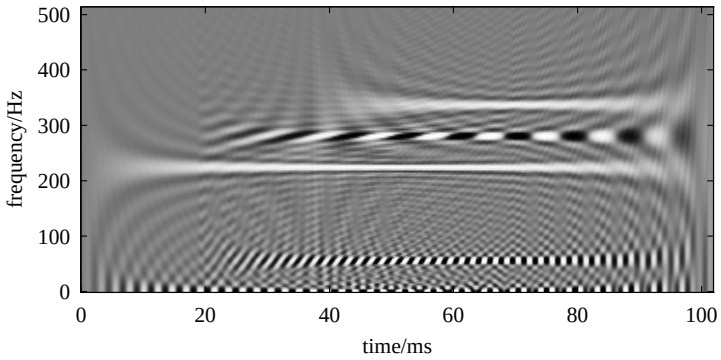


Fig. 3.6: Wigner Distribution of the sum of two sinusoids at approximately 220 Hz and 340 Hz, the higher frequency component starting at around 38 ms. The cross-terms are centred at 280, 60 and 0 Hz. The grey scale goes from negative (black), through zero (grey) to positive (white).

Any spectrogram can be obtained from the Wigner distribution by a 2-dimensional convolution, that is, by *blurring* it with a particular kernel (see fig. 3.7). This has the effect of removing much of the interference and ensuring that the result is strictly non-negative, but at the expense of a loss of resolution, which, unlike the case of the linear representations, represents a loss of information. In fact, a broad class of time-frequency distributions, known as *Cohen's class* (Cohen, 1989), can be obtained by using a general convolution. In a similar way, any scalogram (that is, the squared modulus of a wavelet transform) can be obtained by smoothing with a position dependent kernel, which is tall and narrow at high frequencies, and broad and short at low; this is called *affine smoothing* (Flandrin and Rioul, 1990), because the kernels at different positions are obtained by affine transformation in the time-frequency plane of a single prototype kernel. Scalograms are not in Cohen's class because the smoothing is frequency dependent and hence not a convolution.

3.3.4 Autocorrelation and correlograms

Autocorrelation analysis is a recurring motif in computational models of audition, beginning with Licklider's *Triplex Theory* (Licklider, 1959), and continuing, for example, with Seneff's *Generalised Synchrony Detector* (Seneff, 1984), Lyon's analogue electronic cochlea (Lyon and Mead, 1988) and models based on it (Lazzaro and Mead, 1989), and the *Narrowed Autocorrelation* or NAC (Brown and Puckette, 1989).

The autocorrelation function of a complex signal $x(\cdot)$ is defined as

$$r(\tau) = \int_{-\infty}^{\infty} x(t)x^*(t-\tau)\,\mathrm{d}t = \int_{-\infty}^{\infty} R_x(t,\tau)\,\mathrm{d}t, \tag{3.22}$$

where $R(t,\tau)$ is the instantaneous auto-correlation defined in the previous section. It is the inverse Fourier transform of the power spectral density of the signal. For signals with interesting temporal structure, it is more useful to use the short-term auto-

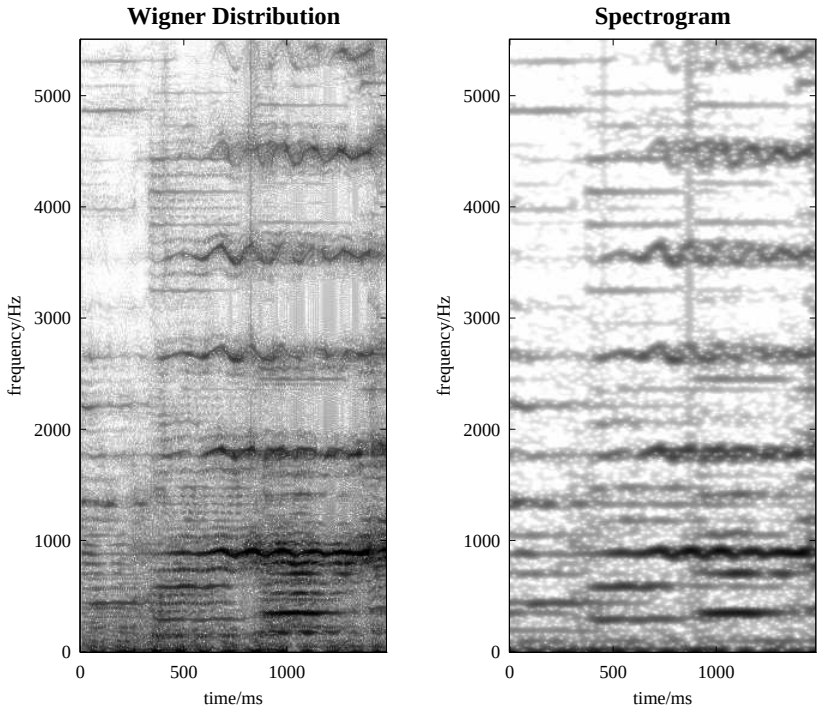


Fig. 3.7: Comparison of two time-frequency distributions: a snippet of Bach's 2nd *Brandenburg Concerto*, featuring harpsichord, flute and violin sounds.

correlation, which is a moving average of the instantaneous auto-correlation function:

$$r(t,\tau) = \int_{-\infty}^{\infty} A(s) R_x(t+s,\tau)\, \mathrm{d}s, \tag{3.23}$$

where $A(\cdot)$ is a window function determining the time over which averaging takes place. The autocorrelation can be used for the detection of periodic signals, because any periodic signal will be strongly correlated with a time-delayed version of itself if the lag is a multiple of the period.

The direct autocorrelation in eq. 3.23 is linearly and invertibly related to the short-term power spectrum, and is thus essentially the same object viewed in two different ways. However, if the short-term autocorrelation is computed for *each* output channel of a filter bank, (or possibly a rectified version of it) then a new object, the auditory *correlogram* (Duda et al., 1990) is obtained. The correlogram is a representation with *three* dimensions: time, frequency and lag. Ellis and Rosenthal (1995) describes why the extra dimension, the lag, is useful in segregating and grouping sounds with common periodic variations.

It should be noted that several researchers (*e.g.* Lyon, 1984; Seneff, 1984) have used autocorrelation analysis based not on the autocorrelation of filter-bank outputs, but on rectified or otherwise nonlinearly transformed versions thereof. This is motivated by what is known of the transduction process in the cochlea, by which the the

filtered signals, represented by motion of the basilar membrane, are converted into neural impulses in auditory nerve; see, for example, Meddis (1983).

3.3.5 The correlogram and the Wigner distribution

In this section, it is shown that a type of correlogram can be obtained from the Wigner distribution by means of a linear transformation equivalent to pre-multiplication followed by smoothing. This correlogram is computed by forming the short-term autocorrelations of each of the outputs of a filter bank. Using complex arithmetic to simplify the algebra, the output of the kth channel can be written as

$$y_k(t) = \int_{-\infty}^{\infty} x(t+s)H_k^*(-s)e^{-i\omega_k s}\,\mathrm{d}s, \tag{3.24}$$

where ω_k is the nominal centre frequency of the filter and $H_k(\cdot)$ is a slowly varying envelope, as described in §3.3.1. The instantaneous autocorrelation of the filter output at time t with lag τ is

$$R_{y_k}(t,\tau) = y_k(t+\tfrac{1}{2}\tau)y_k^*(t-\tfrac{1}{2}\tau). \tag{3.25}$$

After some manipulation and the introduction of new variables, this can be expressed in terms of the Wigner transform $W_x(\cdot,\cdot)$ of the original signal $x(\cdot)$, and the Wigner transform $W_{H_k}(\cdot,\cdot)$ of the envelope $H_k(\cdot)$:

$$R_{y_k}(t,\tau) = \iint W_x(s,\nu)e^{i\tau\nu}W_{H_k}(t-s,\omega_k-\nu)\,\mathrm{d}s\,\mathrm{d}\nu. \tag{3.26}$$

This is a *multiplication* of the original Wigner distribution W_x by a complex exponential in the frequency direction, $e^{i\tau\nu}$, followed by a *convolution* with a kernel W_{H_k} that is itself another Wigner distribution. For a well-behaved envelope $H_k(\cdot)$, the convolution kernel $W_{H_k}(\cdot,\cdot)$ will be localised around the origin. The correlogram itself will be obtained by a further short-term averaging; that is, a further convolution in the time domain, which can be incorporated into the first convolution. Thus, overall, the correlogram at a particular lag τ can be obtained from the Wigner distribution by a multiplication and a smoothing.

How can this result be interpreted? It appears that that correlogram at zero lag ($\tau = 0$) is equivalent to a blurred Wigner distribution, but using a frequency-dependent kernel, as with a scalogram. At non-zero lags, the Wigner distribution is first multiplied by a periodic complex exponential before being blurred. Given that blurring destroys fine structure, this *heterodyning* has the effect of selecting a different set of structures, periodic along the frequency axis on a scale determined by $1/\tau$, to survive the smoothing process. Now, these variations are characteristic of pitched sounds, with the spacing along the frequency axis determined by the periodicity of the sound. Thus, the autocorrelation analysis is able to detect pitch structure even when the individual harmonics cannot be resolved by the filterbank, perhaps because the lower harmonics are missing or masked by noise. This information is made explicit in the correlogram, but the preceding analysis shows that it is also available from the Wigner distribution using only a *linear* transformation, suggesting that it may be possible to design useful phase invariant auditory representations by selecting appropriate linear bases for the Wigner distributions, rather than for the linear time domain version of the signal.

Summary

Graphical models provide a powerful and computationally efficient framework for building and fitting probability models. Latent variable models allow internal representations to be constructed within the graphical model formalism, with well-defined processes of inference and learning. Bayesian networks can be used to model causal relationships, giving quantitative flesh to the idea that perception is concerned with building an internal representation of the world and inferring the physical causes of sensation.

Several of the algorithms discussed in § 3.2 have linear generative models and can be understood as methods that find good linear bases for representing data. Similarly, linear time-frequency representations carry their own implicit linear bases, whereas the quadratic representations, including autocorrelation based representations, can be seen as linear transformations of the Wigner distribution.

Putting these observations together raises some interesting questions about time-frequency representations. When considered as linear bases, on the same footing as the adaptive linear bases in PCA, ICA, or sparse coding, one is lead to ask, are these fixed time-frequency bases optimal with respect to low-redundancy? Perhaps standard methods such as short-term Fourier transforms and wavelet transforms have been found to be so useful precisely because they are easily computable approximations to an optimal linear representation for a wide class of signals. This suggests that, instead of using a given time-frequency representation, we could use ICA or a related method to search for a 'best basis' for a particular class of signals under consideration. This procedure could lead to, for example, alternative tilings of the time-frequency plane, or even representations that cannot simply be interpreted in terms of time and frequency, but nonetheless yield efficient, non-redundant encodings of the data. In the case of quadratic representations derived by linear transformation of the Wigner distribution, one benefit would be that the family of potential bases would include phase invariant representations like the spectrogram.

4. CONCLUDING OBSERVATIONS

The last two chapters presented some of the arguments for a perceptual approach to music processing, and in turn for an approach to perception based on efficient representation in a probabilistic framework. This chapter contains a few observations on what this implies for the implementation of artificial perceptual systems and how that relates to what is known about biological perceptual systems.

4.1 On the construction of artificial perceptual systems

Evolution, design and learning

One of the common threads running through notions of redundancy reduction, efficient representation and transmission of information, and the construction of probabilistic models is that they are all based on the optimisation of an explicitly given objective function. Conceptually, this is an appealing way of framing the problem, as it allows for a clean split between the definition of the objective and the methods used to optimise it. The optimisation procedure may involve various approximations and opportunistic data processing techniques, but as long as the objective function is defined in a rigorous and principled manner, the performance of the system as a whole can still be evaluated objectively, and if the performance is inadequate, it may simply be a question of identifying which approximations need to be improved.

The optimisation is essentially a search over some parameter space; in all but the most trivial systems, an exhaustive search would be out of the question due to the size of the space. For the purposes of the current discussion, we may identify three different ways to manage this search: through design, evolution, or learning.

In natural systems, the scientific argument has traditionally centred on a dichotomy between evolution and learning, since we cannot address the question of a designer without opening up a theological debate. For example, in relation to cognitive processes, Fodor (1983, p. 34) favoured the evolutionary approach: "as the operative notion of mental structure gets richer, it becomes increasingly difficult to imagine identifying the ontogeny of such structures with the registration of environmental regularities… There would seem not to be enough ambient information available to account for the functional architecture that minds are found to have." He did accept that artificial perceptual systems might possibly be constructed, but added that "what does *not* follow is that there is some way of constructing such systems from the information given in *experience*." Others, from Locke (1706) onwards have argued that there are no "innate ideas." In music, Bharucha (1991), Leman (1991) and Cambouropoulos (1998) have all stressed the importance of perceptual learning from empirical data.

Although, at one level, these are unquestionably important concerns, at another, evolution and learning are both just optimisation processes, albeit operating over vastly different time scales and using different mechanisms. In stark contrast, a designed system relies on a designer to perform the search: it is the *designer's* intelligence that is being exercised, often in mysterious ways, relying on leaps of insight, accumulated experience, hunches and so on. Cambouropoulos (1998) made a similar point, contrasting the "hand-crafting" of engineered systems with a process of "empirical induction" in data-driven systems. Though the design approach may ultimately produce usable solutions to engineering problems, it does not really bring us any closer to an understanding of intelligence itself, and arguably, neither does it result in artificially intelligent systems.

Computation routes to optimal representation

Having argued that one distinction between different perceptual architectures derives from how they are developed through some process of optimisation, another arises when we consider the functioning of the end result. A given objective function may be optimised by a certain perceptual representation, yet this same high-level representation may be computable in several different ways, using a different chain of steps or intermediate representations. The most effective computational route will depend on, amongst other things, the nature of the hardware available and any constraints thereby imposed. Nevertheless, if the same objective function is being optimised, we may reasonably expect the final results to be comparable across different implementations.

In particular, when the goal is audition, we need not regard the peripheral sections of the human auditory system as an immutable given, to be modelled as the first step in any artificial auditory system. Nature has had to evolve those particular mechanisms because of the limitations of the hardware available to it, but those restrictions do not necessarily apply to an artificial implementation. For example, simulating the firing of individual neurons is a laborious process that we would like to avoid if possible. Lyon and Mead (1988) note that, "Nerve impulses are by their very nature a horrible medium into which to translate sound. They are noisy and erratic, and can work over only a limited dynamic range of firing rates." Technologically based solutions to the problem of optimal auditory representation may or may not match those found in the human auditory system. The result may not necessarily be a model of *human* hearing, at least not at a low level, but if the computational goals are the same, it may well perform comparably, and if parallels do arise between the biological and artificial processes, this will shed light on why the biological processes are as they are. The process may also generalise readily to situations where there is no biological analogue.

A computational view of the peripheral auditory system

To put the preceding comments into a specific context, let us consider what is known about the functioning of the human auditory system. Can it be understood it as an optimal solution to the problem of efficient representation, or must we accept it as an evolutionary 'accident' without an explanation?

The fundamental constraint is the limited capacity of the individual fibres of the auditory nerve. Nerve fibres can transmit of the order of 100 bits per second (Rieke et al., 1997), whereas an uncompressed CD quality audio signal requires about 700,000 bits per second. Clearly, the information needs to be spread across a number of fibres—there are around 30,000 outgoing fibres in the human auditory nerve (Pickles, 1988). The discussion of §2.2 suggests that the most efficient way to do this for the fibres to be utilised equally and independently, to minimise the redundancy of the auditory nerve representation. The following points are speculative only, but indicate possible interpretations of processes in the inner ear as steps towards this goal.

Cochlear filterbank The acoustic signal is first split into multiple channels in the cochlea, where an effective filterbank produces a number of band-limited signals. The goal of non-redundant representation would require that each channel carry approximately the same quantity of information whilst minimising the statistical independence of the signals. This may provide an explanation for the characteristic shapes of the cochlear filters; indeed, the results presented in Chapter 5 indicate that an optimal linear representation of speech sounds is analogous to a filterbank with wider bandwidths at high frequencies.

Assuming that an effective cochlear filter is of the order of 100 Hz wide or greater (Plomp, 1976), this still represents of the order of thousands of bits per second if coded naively, so further processing to spread the information across many nerve fibres is required.

Nonlinear compression The signal from the cochlear filterbank undergoes rectification and nonlinear compression as part of the neural transduction process in the inner hair cells. This enables the cell to respond to a wider dynamic range of signals than would be possible otherwise.

From an information theoretic point of view, compressive nonlinear transformation emerges as an efficient way to encode a random variable as a bounded value (Nadal and Parga, 1994). The form of the nonlinearity is related to the probability distribution of the data and any assumed input noise. For example, it is easily shown that logarithmic compression is optimal for power-law input distributions of the form $p(x) \propto 1/x$ over some range of x.

Population coding of intensity The inner hair cells have a range of firing thresholds, which means that a group of neighbouring hair cells, experiencing approximately the same disturbance of the basilar membrane, respond differentially to a signal at a given level. This could be interpreted as a population code for local cochlear activity, another case of spreading information across several channels.

Lateral inhibition and adaptation It has already been suggested (Barlow, 1961; Barlow and Földiák, 1989) that lateral inhibition and adaptation are a response to spatial and temporal redundancy in sensory data. The activity of the the inner hair cells may tend to be locally dependent, because of both the structure of natural

sounds and the mechanics of the cochlea. Lateral suppression would then be an effective processing strategy. Considerations of local dependency are taken up again in Chapter 8.

Until more quantitative predictions can be made, these observations have no more than a certain plausibility. Hence, it is worth pursuing the program of developing artificial auditory systems, guided by information theoretic concerns, to see if any of the auditory processes described above emerge spontaneously in an artificial system.

4.2 On the Gestalt approach to perceptual grouping

Much work on computational models of audition (*e.g.* Bregman, 1990; Ellis, 1994; Mellinger, 1991) has been based on Gestalt principles, whereby 'parts' are grouped into 'wholes,' elementary sensations into higher level objects. Bregman (1990) explains the concept of grouping in the visual domain as a "colouring task", in which the aim is conceptually to "colour-in" different parts of the image so that parts belonging to the same object are the same colour. The grouping rules tend to rely on concepts of *locality*, such as proximity and connectedness, and hence presuppose some sort of perceptual 'field' with a topological structure, a 'sensory image' in which objects produce distinct, localised regions of activity.

In the visual domain, the input representation already satisfies these requirements to a large degree and is thus amenable to segmentation and grouping using Gestalt principles. Bregman (1990, p. 6) observes that, "In vision, you can describe the problem of scene analysis in terms of the correct grouping of regions." This is a convenient fact about the retinal image: it has regions. Bregman goes on to ask, "But what about hearing? What are the basic parts that must be grouped to make a sound?" This is a very good question. Those "parts" are not at all apparent in a direct, time-domain representation of the acoustic waveform. Bregman does not dwell on this point, and moves on to a consideration of spectrograms, which, conveniently, are somewhat analogous to images: "Once we see that the sound can be made into a picture, we are tempted to believe that such a picture could be used by a computer to recognise speech sounds." Whereas regions of activity due to different auditory objects may not be distinct in the time domain, (or indeed, the one-dimensional frequency domain of a global Fourier transform) they may become so in the two-dimensional time-frequency plane.

At this point it would be possible simply to accept that the early auditory system produces an auditory 'image' by doing a time-frequency decomposition, and to base any artificial system on data presented in this form. However, as Bregman (1990, p. 8) notes, there are still problems with the the direct application of Gestalt rules in the time-frequency domain. In a spectrogram, different auditory objects can occupy the same region, forming a superposition without one occluding the other in the way that visual objects do; the region must therefore be assigned to *two* objects. The addition of a new rule, grouping by harmonicity, belies that fact that many sounds produce *disjoint* regions of activity: pitched sounds with multiple frequency components are a common occurrence. The adoption of the three-dimensional auditory correlogram by many re-

searchers (Licklider, 1959; Duda et al., 1990; Ellis and Rosenthal, 1995; Mellinger, 1991) may, in part, be a response to the need to obtain simultaneously greater separation between objects and greater locality within objects.

Though these problems are described with specific reference to audition, they could arise in any application of Gestalt rules to a given representation: the representation itself may not be suitable for the application of Gestalt rules. The reason why this has gone relatively unremarked is probably that in vision, which has dominated the research, we are fortunate to be given at the outset a suitable perceptual field. However, even this situation may not be an accident.

We have assumed so far that images are the raw material of visual perception, but the image itself is a product of the visual system: there is no image outside the eye. Visual information about the environment is encoded in the fluctuations of the electromagnetic field at the surface of an organism, but the disturbance at a particular point is a mixture of disturbances caused by all visible objects. The formation of an image by the lens—essentially the computation of a spatial Fourier transform (Hecht, 1987)—goes a long way towards separating the visual objects by segregating light coming from different directions.

Thus, we see that the visual image is not a given, but an intermediate representation, one in which Gestalt principles can successfully be applied. The lens can be seen as an elegant solution to a computational problem. The question then is, what is that computational problem: what principles guide the construction of a perceptual field suitable for Gestalt grouping? Looking ahead briefly to §8.1, it is suggested that the *minimisation*, and the *localisation* of statistical dependency might be the defining characteristics of such a 'Gestalt-ready' perceptual field, but this remains to be seen.

4.3 On previous unsupervised and probabilistic musical systems

This section lists a few examples of music processing systems that demonstrate elements of the approach advocated here, by using probabilistic models or unsupervised learning as a solution to musical problems.

Leman (1991) aimed to show that the perception of tonality in Western music could be learned through exposure to sufficient examples of tonal music. To this end he implemented a Kohonen map which was exposed to musical examples represented in a certain way. Once trained, the map was tested by mapping out which regions which respond to which chords. The result was a well-ordered map of chords, displaying many of the theoretically expected relationships between chords, such as the circle of fifths.

This is an interesting result, but perhaps not as significant as one might expect. The performance of the Kohonen map will be discussed in more detail in §8.1.6, but the proximity relationships it exposes are fully present in the input data if one assumes a Euclidean metric. Thus, the structure of the map is determined by the input representation, which in this case was based on subharmonic templates (see Leman, 1991 for details), carefully chosen to emphasise the kind of harmonic relationships known to ex-

ist in music. However, Leman did repeat the experiment using a representation based on the probabilities of different pitches being part of a given sound. This is a much less ad-hoc definition, incorporating less prior knowledge and fewer expectations about the results, and it is therefore significant that similar results were obtained.

Conklin and Witten (1995) addressed the problem of using an explicit probability model to predict the progression of musical sequences. They used the concept of "instantaneous entropy" to judge how well a particular model predicts a given sequence; this would more conventionally be identified as the *log-likelihood* described in § 3.1.2. Hence, Conklin and Witten were engaged in building maximum likelihood models of music, even though they did not use that term themselves.

As noted in § 3.1, building a single giant contingency table is not a practical approach to constructing a probability model. Conklin and Witten deal with this by defining a "multiple viewpoint" system. Each viewpoint is essentially a representation of some restricted aspect of the musical surface, such that the construction of a contingency table for each viewpoint becomes feasible. The prediction of the entire system is then formed by combining the predictions of each viewpoint, giving more weight to the more confident predictions.

In the language of probabilistic modelling, the system is somewhat like a *mixture of experts* (Jacob et al., 1991), each of which learns its data distribution in the form of a contingency table. However, the rule used to combine evidence is not derived probabilistically. In addition, the representation used for each viewpoint is predefined, and the selection of viewpoints used has a significant effect on the performance of the system. Hence, there is still an element of "hand-crafting" involved.

Cambouropoulos (1998) discussed the modelling of musical structure in terms of the opposition between a "knowledge engineering" methodology and one based on "empirical induction." Referring back to § 2.1.4, Cambouropoulos is addressing the acquisition of high-level knowledge in a system where both top-down and bottom-up processes may occur. In a knowledge engineering methodology, knowledge at the top is explicitly put in place by the designer of the system, and is therefore more a demonstration of *human* rather than artificial intelligence. The empirical induction approach holds that this knowledge can be gained from experience, "by making generalisations on a set of musical phenomena, based on a set of general fundamental principles." The cognitive principles which Cambouropoulos sets out, namely, *economy* and *informativeness*, are strongly reminiscent of the principle of redundancy reduction described in § 2.2. Though not expressed in the language of information theory, they can notionally be equated with the terms $I_{\max}$ and $I(Y,S)$ in eq. 2.6, Atick and Redlich's (1990) expression for redundancy. To these Cambouropoulos adds a *naturalness* principle, which he links to Gibson's (1979) ecological account of perception, discussed previously in § 2.1.2.

Kashino et al. (1998) use a Bayesian network architecture to integrate different knowledge sources in a music transcription system. Similar systems (*e.g.* Godsmark and Brown, 1999) have used a blackboard architecture to achieve this goal, but

a Bayesian network is a more principled way to combine uncertain data. The network used, however, embodies a certain amount of knowledge about the structure of tonal music and so cannot be said to learn it through experience.

Raphael's automatic accompaniment system (Raphael, 2001) is a good example of what can practically be achieved with an explicit graphical model. A Bayesian network describing the structure of a specific piece of music is constructed, specifying how performances may be generated from the structural skeleton by variations in timing and expression. After a number of 'rehearsals,' in which the model parameters are optimised to fit a soloist, the system is able to follow the score in real time while the soloist plays, well enough to play the accompaniment, following the soloist's expressive variations in a complementary manner.

Since the structure of the model closely parallels the score, it is not clear how applicable this system is to general music perception, but, as Raphael notes, there is great potential in such systems if they can be extended to learn structure as well as parameters directly from audio data.

Summary

In this chapter, it was argued that the construction of artificial perceptual systems should be guided by the principles described in Chapter 2, using an optimisation approach that allows for a divergence between the artificial and biological implementations. If perception truly is driven by the goal of efficient representation, the fact that the artificial and biological systems have the same objective should ensure a convergence of the end result, even if the intermediate processes are different. This should allow different implementations to be tailored to different hardware characteristics. It was also argued that this remit should be extended to include even those parts of biological perceptual systems that are genetically specified and are the result of adaptation over evolutionary time scales.

Until the formation of perceptual representations is more fully understood, theories based on Gestalt grouping principles will be incomplete and inapplicable in domains where the appropriate sensory image is not known.

There certainly are precedents for the application of all these ideas in music processing systems, but it is only in recent years that a rigorous framework for constructing complex probabilistic models has arisen in the form of the graphical model formalism. Though it has been applied to many problems in data analysis, it is only just beginning to be applied to music, and hence there are many possibilities for research.

To summarise, the methodology to be followed is to let the statistical structure of sensory (in this case, auditory) data guide the construction of economic representations, whereby redundancy in the input is identified and removed adaptively in a process of unsupervised learning using ecologically representative data. To do this effectively will generally require an explicit probability model, which will therefore form the core of any artificial perceptual system developed in this way.

PART II

EXPERIMENTAL WORK

5. ICA OF AUDIO WAVEFORMS

Introduction

This chapter investigates the application of perhaps the simplest method of redundancy reduction, in which we attempt to construct a factorial code using only a linear transformation of the data: namely, independent component analysis or ICA. The probabilistic model underlying ICA was described in § 3.2.2, and the derivation of a practical algorithm is continued in § 5.1.

Previous work has shown that various flavours of ICA, when applied to natural images, all result in a code made up of elements which respond to broadly similar localised, oriented band-pass features in the images. For example, Field and Olshausen (1996) showed that sparse coding, which can be seen as a generalisation of ICA, of natural images results in a decomposition of the image into a set features localised both spatially and in spatial frequency. Later experiments with ICA (Bell and Sejnowski, 1997; Hyvärinen et al., 1998) have produced comparable results. These features can be interpreted as wavelets or edge detectors, providing an alternative interpretation of the use of wavelet analysis or edge detection in image processing: *viz.*, that these methods are useful precisely because they result in a less redundant representation of images. The features also show some correspondence with receptive fields of simple cells in visual cortex (V1), providing a possible explanation for the development of these cells (Field and Olshausen, 1996).

Moving to auditory applications, Bell and Sejnowski (1996) used ICA on sound, but this was limited to Bell's rendition (on his teeth) of Beethoven's *Für Elise*. Casey (1998) also used ICA, but learning was restricted to one auditory event at a time, not a long exposure to a representative selection of sounds, or what one might call an auditory 'environment.' In this chapter, ICA is applied to such an environment. The data is drawn from two radio stations: one broadcasting mainly speech; the other mainly classical music. Many of the resulting basis vectors are quite wavelet-like, in that they are localised in both time and frequency, and can easily be characterised in terms of their position and spread in the time-frequency plane. Some of them, however, particularly from the set trained on music, do not fit that interpretation very well, and an alternative analysis is required.

5.1 Derivation of an ICA algorithm

Continuing from the ICA model probability model given in eq. 3.14, and parameterising the basis matrix $\mathbf{A}$ as the inverse of a matrix $\mathbf{W}$, the objective function to be

maximised to get a maximum likelihood estimate of $\mathbf{W}$ is the expected log-likelihood,

$$\mathcal{L} = \mathrm{E}_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{A}) = \log\det\mathbf{W} + \mathrm{E}_{\mathbf{x}} \log p_{\mathbf{s}}(\mathbf{W}\mathbf{x}), \tag{5.1}$$

where the expectations are taken over the observed data distribution, and $p_{\mathbf{s}}$ is the density function of the components $\mathbf{s}$. The matrix $\mathbf{W}$ is sometimes called the weight matrix, as it gives the connection weights in a linear network to compute $\mathbf{s} = \mathbf{W}\mathbf{x}$. The gradient of the log-likelihood with respect to $\mathbf{W}$ is

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\mathbf{W}} = (\mathbf{W}^T)^{-1} + \mathrm{E}\left\{[\nabla_{\mathbf{s}} \log p(\mathbf{s})]\mathbf{x}^T\right\}, \tag{5.2}$$

where $\mathbf{s} = \mathbf{W}\mathbf{x}$, and $\nabla_{\mathbf{s}}$ is the vector gradient with respect to $\mathbf{s}$. Since the prior $p(\mathbf{s})$ is assumed to be factorial, with $p(\mathbf{s}) = \prod_{i=1}^{N} p(s_i)$, this gradient is an element-wise function of $\mathbf{s}$, and we may define the vector-valued function $\gamma(\mathbf{s})$ as

$$\gamma(\mathbf{s}) = -\nabla_{\mathbf{s}} \log p(\mathbf{s}), \qquad \text{with} \qquad [\gamma(\mathbf{s})]_i = \gamma(s_i) \stackrel{\text{def}}{=} -\frac{\mathrm{d}}{\mathrm{d}s} \log p(s)\bigg|_{s_i} \tag{5.3}$$

For example, with a Laplacian prior $p(s) = e^{-|s|}$, we simply obtain $\gamma(s) = \operatorname{sgn} s$. A stochastic steepest-ascent gradient algorithm may be constructed by repeatedly updating $\mathbf{W}$ by adding

$$\Delta\mathbf{W} = \eta\left[(\mathbf{W}^T)^{-1} - \langle\gamma(\mathbf{s})\mathbf{x}^T\rangle\right], \tag{5.4}$$

where η is a learning rate parameter, and the angle brackets denote a sample average over a batch of training data, rather than an idealised expectation. This is the algorithm first presented by Bell and Sejnowski (1995), though the derivation is modelled on that of Cardoso (1997). The algorithm requires a matrix inversion and, in common with steepest-ascent algorithms in general, suffers from slow convergence in some cases, because it relies only on local gradient information. Amari et al. (1996) proposed an alteration to the update rule, motivated by considerations of the *natural gradient* (Amari, 1998). The net result is that the weight update is post-multiplied by $\mathbf{W}^T\mathbf{W}$, so that the new increment is

$$\begin{aligned}\Delta\mathbf{W} &= \eta\left[(\mathbf{W}^T)^{-1} - \langle\gamma(\mathbf{s})\mathbf{x}^T\rangle\right]\mathbf{W}^T\mathbf{W} \\ &= \eta\left[\mathbf{I} - \langle\gamma(\mathbf{s})\mathbf{s}^T\rangle\right]\mathbf{W},\end{aligned} \tag{5.5}$$

which has the effect of rescaling the step taken in parameter space so that larger steps are taken in directions where the model is insensitive to variation of the parameters. Both MacKay (1996) and Cardoso and Laheld (1996) also proposed the same modification on different but related grounds which we need not go into here.

5.2 Experiments with speech and music

An ICA model was trained online and in real time on several days' worth of largely unbroken radio output from two stations: BBC Radio 3, broadcasting mainly classical music but with some speech and other music, and BBC Radio 4, which outputs mainly speech. The signal was broken into 512-sample blocks and was passed through

a slow-acting normaliser, which acted both to remove any DC offset and as an automatic gain control with a large time-constant. The aim was *not* to produce individually normalised blocks of samples with zero sum and unit energy, but to compensate for a fixed or slowly varying DC offset associated with an online audio source (in this case a tuner) and for the large dynamic variations over medium to long time-scales that are common in music. This alternation between loud and soft passages was found to destabilise learning when raw, unnormalised input was used, whereas even very slow normalisation using a time constant of the order of 10 s was enough to stabilise the system. The details of normalisation algorithm are not integral to the results presented here and are given in Appendix A.

The normalised 512-sample blocks were presented to the ICA system as the vectors $\mathbf{x}$. Experiments were performed with both a Laplacian prior,

$$p(s) = \exp -|s| \quad \Longrightarrow \quad \gamma(s) = \operatorname{sgn} s \tag{5.6}$$

and a Cauchy prior,

$$p(s) = \frac{1}{\pi(1+s^2)} \quad \Longrightarrow \quad \gamma(s) = \frac{2s}{1+s^2}, \tag{5.7}$$

producing broadly similar results. Indeed, it has been observed (*e.g.* Cardoso, 2000) that ICA is generally quite robust to misspecification of the prior as long as super-Gaussian priors are used for super-Gaussian sources and sub-Gaussian priors for sub-Gaussian sources. Since audio signals tend to be super-Gaussian even in the time domain, and Fourier transforms of audio signals are also highly super-Gaussian, it was safe to assume that the ICA would discover super-Gaussian components.

The analysis produced two sets of 512 basis vectors, one for Radio 3 and one for Radio 4, which are examined in the following sections.

5.3 Preliminary analysis of resulting bases

The speech derived BBC Radio 4 basis will be examined first since the interpretation of the results is clearer than the BBC Radio 3 results.

5.3.1 BBC Radio 4 basis

The basis can be visualised in various ways: in the time domain, in the frequency domain, or in the time-frequency plane. Fig. 5.1 illustrates time domain plots of some of the Radio 4 basis vectors, and their corresponding magnitude spectra.

The basis vectors are, on the whole, quite well localised in both domains, suggesting that it might be useful to characterise them by position and spread in the time and frequency plane. A method for doing this was previously described in § 3.3.2, based on squaring the time and frequency domain representations of a signal to obtain two energy distributions. These were then treated like probability distributions with a mean and a standard deviation to describe the position and width of each, essentially fitting a Gaussian profile to the energy distributions in both time and frequency.

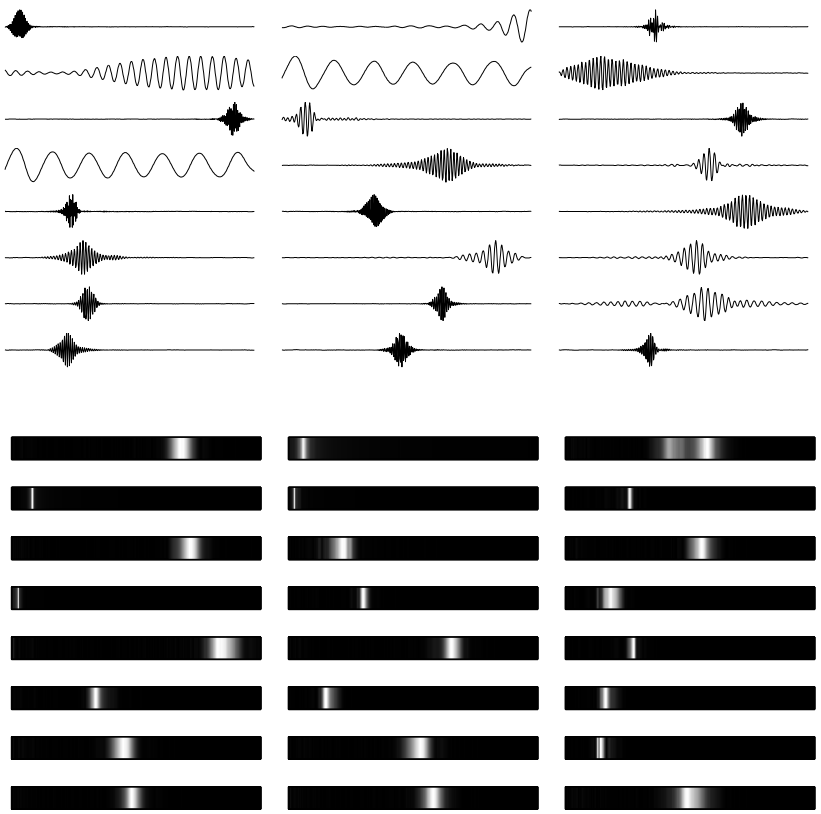


Fig. 5.1: A random sample of the Radio 4 basis vectors trained with a Laplacian prior. The lower plot shows Fourier magnitude spectra of the vectors in the upper plot.

When this procedure was carried out with the ICA basis vectors, it did not seem to adequately capture the time and frequency localisation that was evident in a visual inspection of the basis vectors, because their energy profiles in time and frequency were not sufficiently close to being Gaussian. An alternative characterisation was developed whereby, rather than measuring the mean and standard deviation of each energy distribution, the median and mean absolute value were measured; this produced results that more accurately reflected the shape of each basis vector.

Fig. 5.2 is a combined plot of all 512 basis vectors—each one is represented by an ellipse indicating its position and spread in time and frequency. Fig. 5.3 illustrates more clearly the relationship between centre frequency and bandwidth.

Several observations can be made from the plots: The basis vectors are fairly evenly distributed in time and frequency. The spectral widths are not exactly proportional to the centre frequencies, but there is a general increase in bandwidths at higher frequencies. The very lowest frequencies are not localised in time at all, and cover the full 512-sample width of the input window. In the time domain, these basis vectors consist of sinusoids truncated at the window edges. There is a reversal of the bandwidth trend between 1 and 1.5 kHz. There are also a few anomalous features visible in fig. 5.2 at

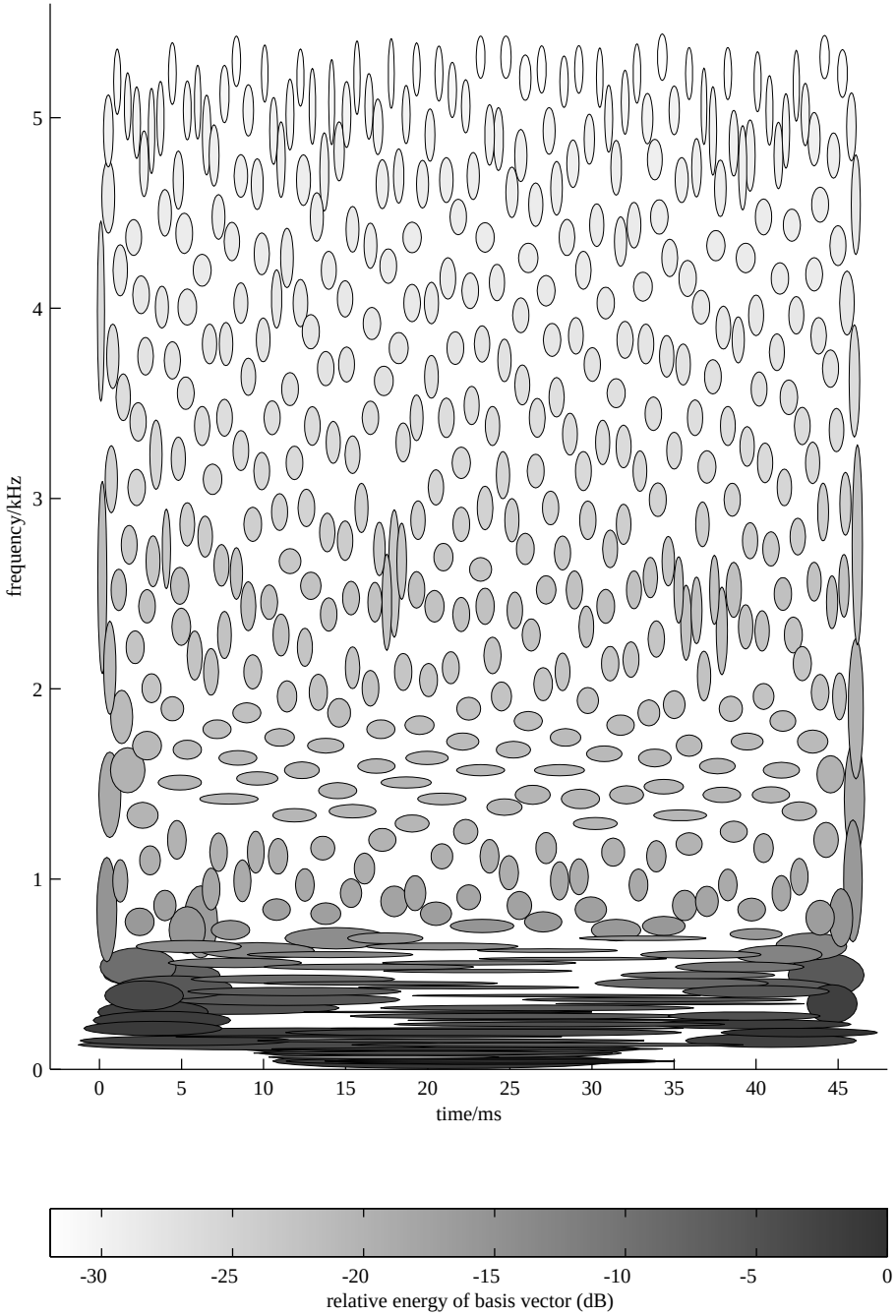


Fig. 5.2: Position and spread in time and frequency of all 512 Radio 4 basis vectors. The grey scale encodes the overall energy (*i.e.* the 2-norm squared) of each basis vector.

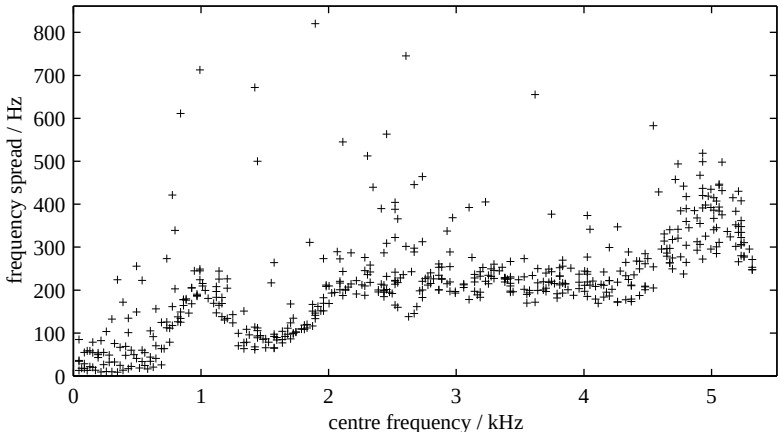


Fig. 5.3: Frequency domain bandwidth v. centre frequency for Radio 4 basis vectors.

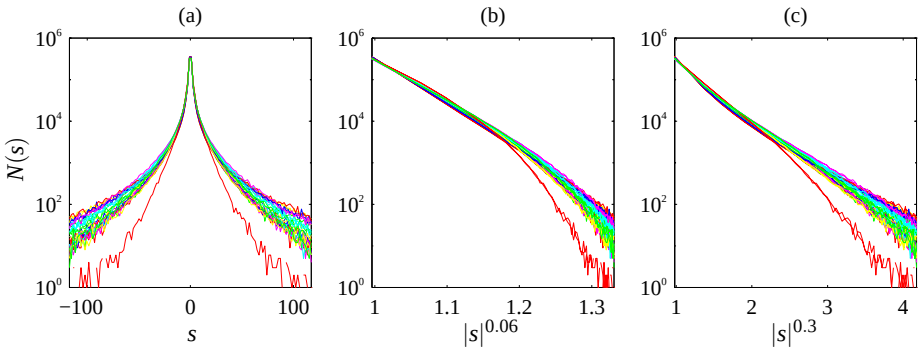


Fig. 5.4: Marginal histograms of some of the Radio 4 output components. Figure (a) is on a log-linear scale, but (b) and (c) plot the histogram bin counts $\log N(s)$ against $|s|^{\alpha}$, where α is a constant. A straight line would indicate that $N(s) \propto \exp -\lambda|s|^{\alpha}$ for some λ, that is, a generalised exponential distribution. These plots show is that a single value of α does not fit the observed histograms over their entire range, but that two different values, approximately 0.06 and 0.3, seem to fit for small and large values of $|s|$ respectively. See fig. 5.8 for an explanation of the anomalous histogram that fits neither model.

around 2.5 kHz near 17 ms and 37 ms.

Overall, the interpretation in terms of wavelets seems appropriate, though not strictly accurate since the 'wavelets' are not all scaled and dilated versions of a single prototype as in a constant-Q wavelet transform: the bandwidths are not directly proportional to the centre frequencies. Neither are the wavelets equivalent to Gabor functions (Gaussian windowed sinusoids).

Fig. 5.4 shows some of the marginal distributions of the resulting components s_i, both confirming their expected super-Gaussianity and showing that neither a Laplacian prior nor a Cauchy prior is an accurate fit.

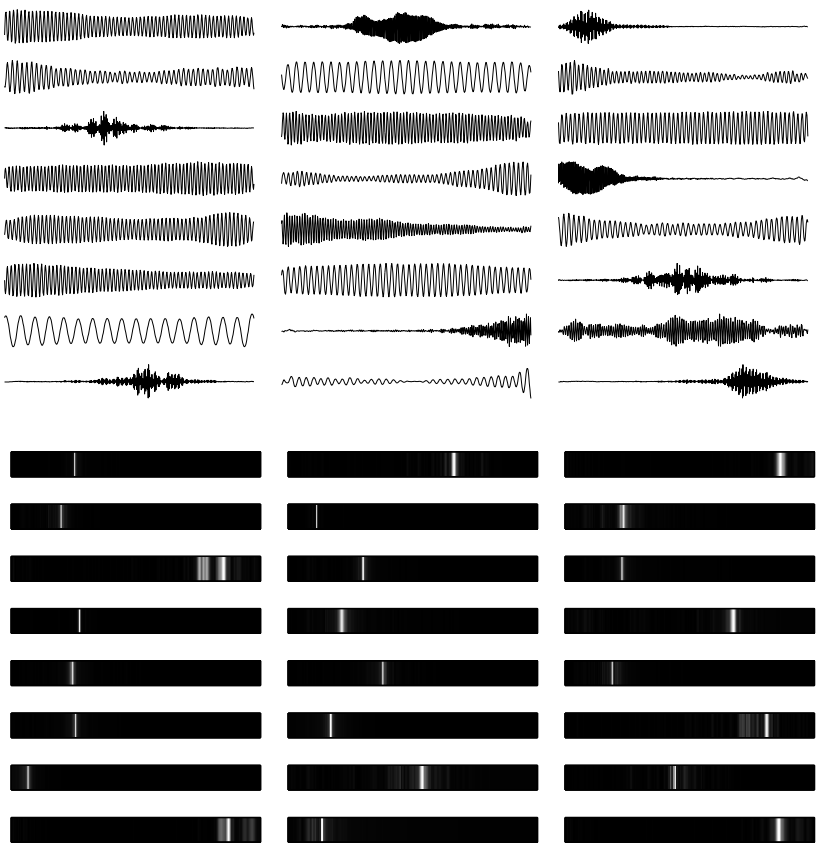


Fig. 5.5: A random sample of the Radio 3 basis vectors trained with a Cauchy prior. (see fi g. 5.1 for explanation.)

5.3.2 BBC Radio 3 basis

When the music derived basis is examined in the same way, the interpretation is not so clear. There are many more full-width sinusoids as in a Fourier basis, but some of the basis vectors appear to be localised neither in time nor frequency, as shown by the large ellipses in fig. 5.6. Closer examination of the least well localised vectors reveals that their energy distribution in frequency is not unimodal and and hence the procedure used to compute position and width does not provide an adequate description of their time-frequency distribution. In the frequency domain, some of the basis vectors have more than one well-localised peak—two of these are illustrated in fig. 5.14. Where these peaks are at multiples of a common fundamental frequency, as it is in those two examples, it is possible that the basis vector represents the upper harmonics of an harmonically rich, lower tone, but it is also possible that the algorithm has not converged properly and has reached a local minimum of the objective function.

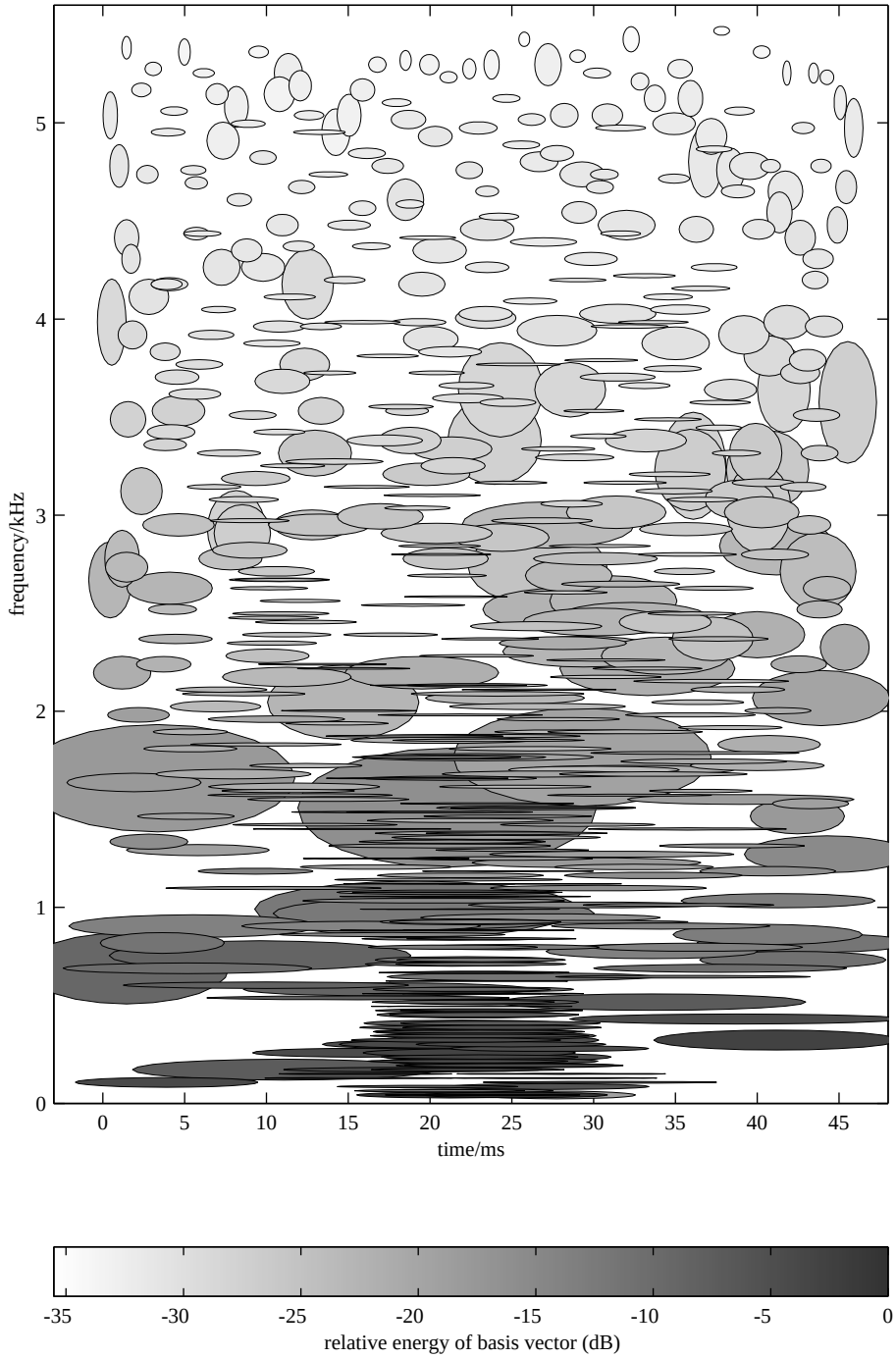


Fig. 5.6: Position and spread in time and frequency of all 512 Radio 3 basis vectors.

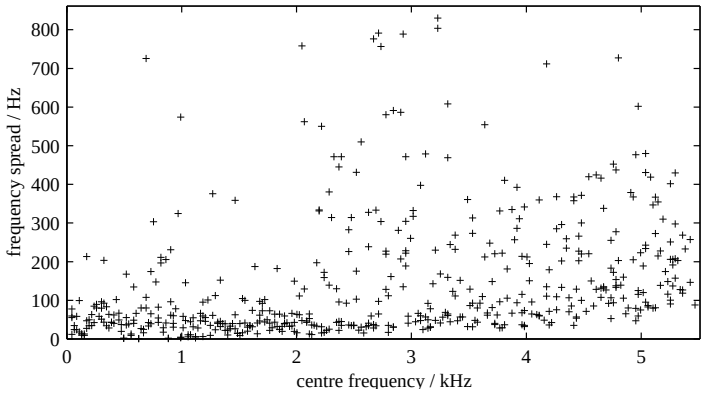


Fig. 5.7: Frequency domain bandwidth v. centre frequency for Radio 3 basis vectors.

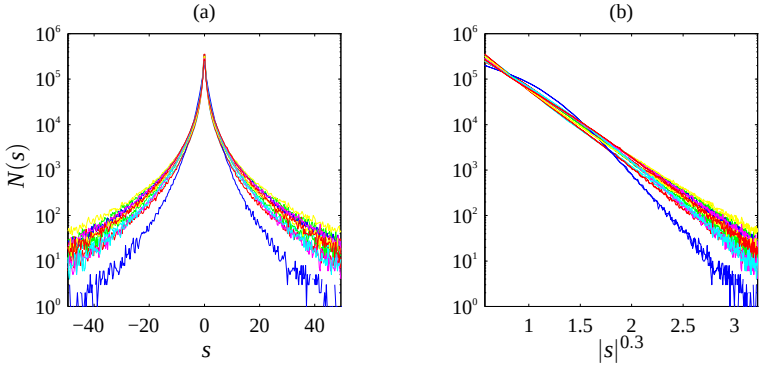


Fig. 5.8: Marginal histograms of Radio 3 output components. The right-hand plot shows that many of the histograms are well approximated by a generalised exponential distribution, $p(s) \propto \exp -|s|^{0.3}$. The anomalous distribution, narrower than the others and with a more rounded peak, is due to one of three components which respond to 50 Hz, which is the power supply frequency in the UK. A similar effect can be seen in fi g. 5.4.

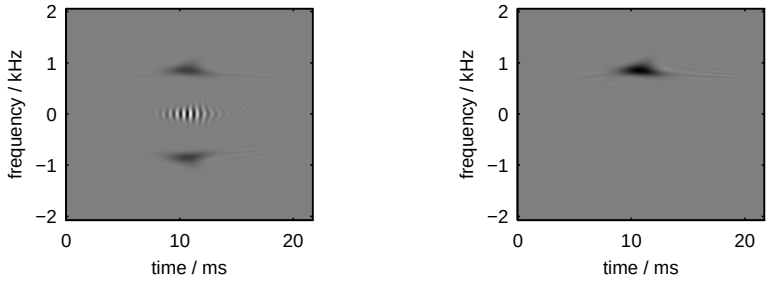


Fig. 5.9: Wigner Distributions of one of the Radio 4 basis vectors in its original form (left) and as an analytic signal (right). An analytic signal is formed by using a Hilbert transform to cancel out the negative frequency components and thus remove the oscillatory cross term in the middle.

5.4 Further analysis using the Wigner distribution

In an effort to gain more insight into what features these basis vectors were encoding, (especially for the Radio 3 set) they were re-examined in the time-frequency plane using the Wigner-Ville distribution, described in §3.3.3 and defined for continuous time signals in eq. 3.21. There are some subtleties involved in defining the discrete time version; in this instance, a "type II quasi-Wigner distribution" was used, which is suited to discrete time, aperiodic signals, and produces a time-frequency distribution which is periodic in frequency; see O'Neill and Williams (1999) for more details.

As described in §3.3.3, the cross-terms are generally a distraction as far as visualisation goes. The usual approach to removing them is some sort of smoothing in the time-frequency plane, which inevitably leads to some loss of resolution (and information). However, in the case of real-valued signals, which produce Wigner distributions that are symmetric in frequency, one set of cross-terms is due to interference between the positive and negative frequency components. These, at least, are easily removed by filtering out the negative frequency components using a Hilbert transform to produce an analytic signal, as recommended by Boashash (1988). This proved to be quite effective for most of the basis vectors (see fig. 5.9.)

A Wigner distribution was computed for each basis vector, then each Wigner distribution was summarised as one or more contours at 70% of its peak value. These were then combined into a single figure containing all 512 basis vectors.

The Radio 4 basis vectors all produced well-localised Wigner Distributions, as illustrated in fig. 5.10. The Radio 3 vectors however, seem to fall in to three groups: (1) many narrow-band features covering most of the width of the window; (2) a number of compact wide-band features towards the top of the spectrum; and (3) a few features with fragmented Wigner distributions. This third group corresponds to those basis vectors which appeared as large ellipses in fig. 5.6.

Looking at their spectra jointly with their Wigner Distributions (see fig. 5.14), it is possible to discern that in many cases, the fragmentation is due to large cross-terms between the components of a multi-harmonic basis vector. Some of these have components whose frequencies are in small integer ratios, which is what one would expect from the spectrum of a low musical note with multiple harmonics.

This still leaves a few basis vectors that defy explanation: some have multiple components which are not in small integer ratios, and some are so irregular that it leads us to suspect that the algorithm has either not converged properly, or has not learned an optimal solution, but fallen into a local minimum.

5.5 Conclusions and further work

This experiment has shown that ICA can learn interesting representations of audio signals. The first obvious conclusion to be drawn is that the two sets of results are *very* different: that statistical structure of musical sounds seems to demand a very different analysis from that required by speech. The speech-derived basis is in many respects like a wavelet basis, with a very clear and regular time-frequency structure. The wavelet

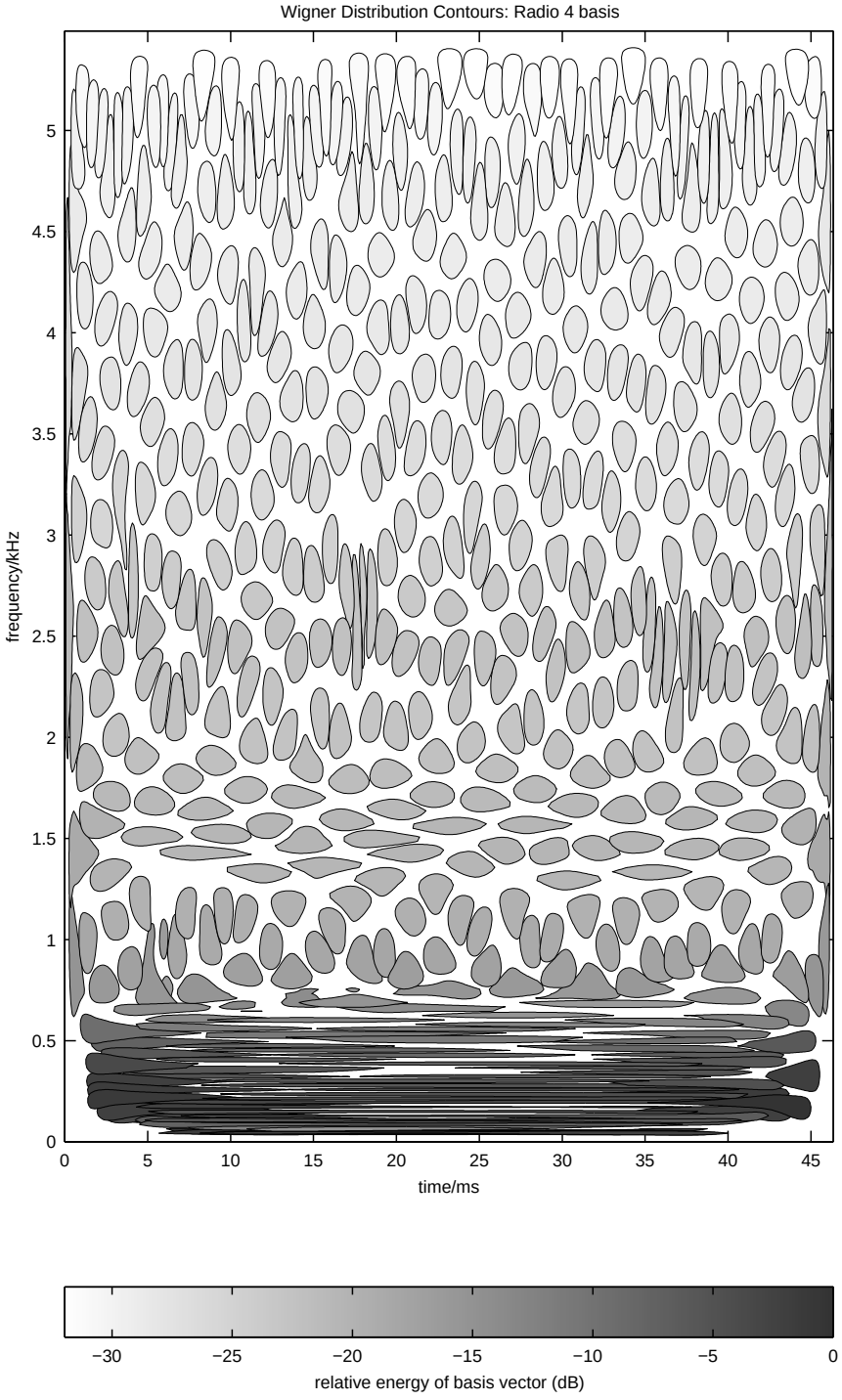


Fig. 5.10: Combined contour plots of all the Radio 4 basis vector Wigner distributions. Each one is represented by a contours at 0.6 times its peak value. The grey scale represents the total energy of each basis vector—not the value of the Wigner distribution.

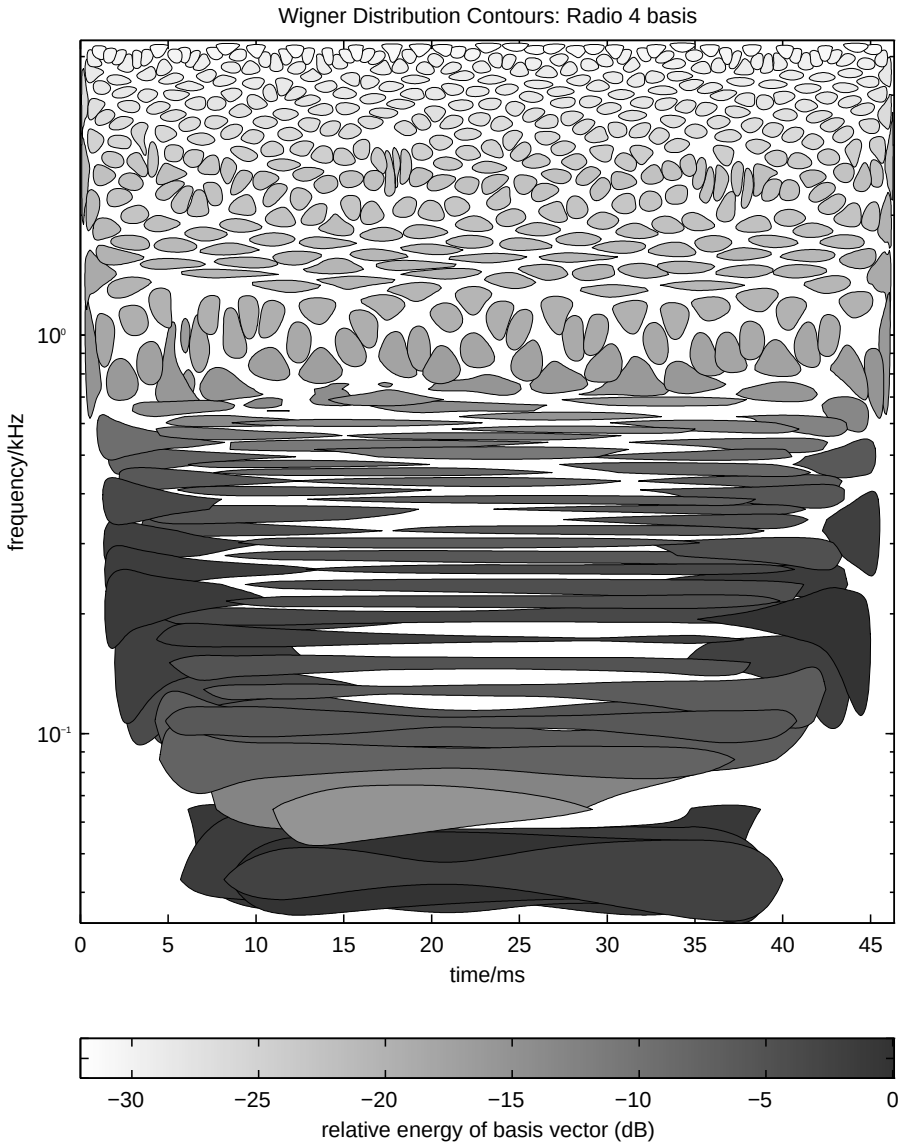


Fig. 5.11: Combined contour plots of all the Radio 4 basis vector Wigner distributions on a logarithmic frequency scale, so that detail at low frequencies is more visible.

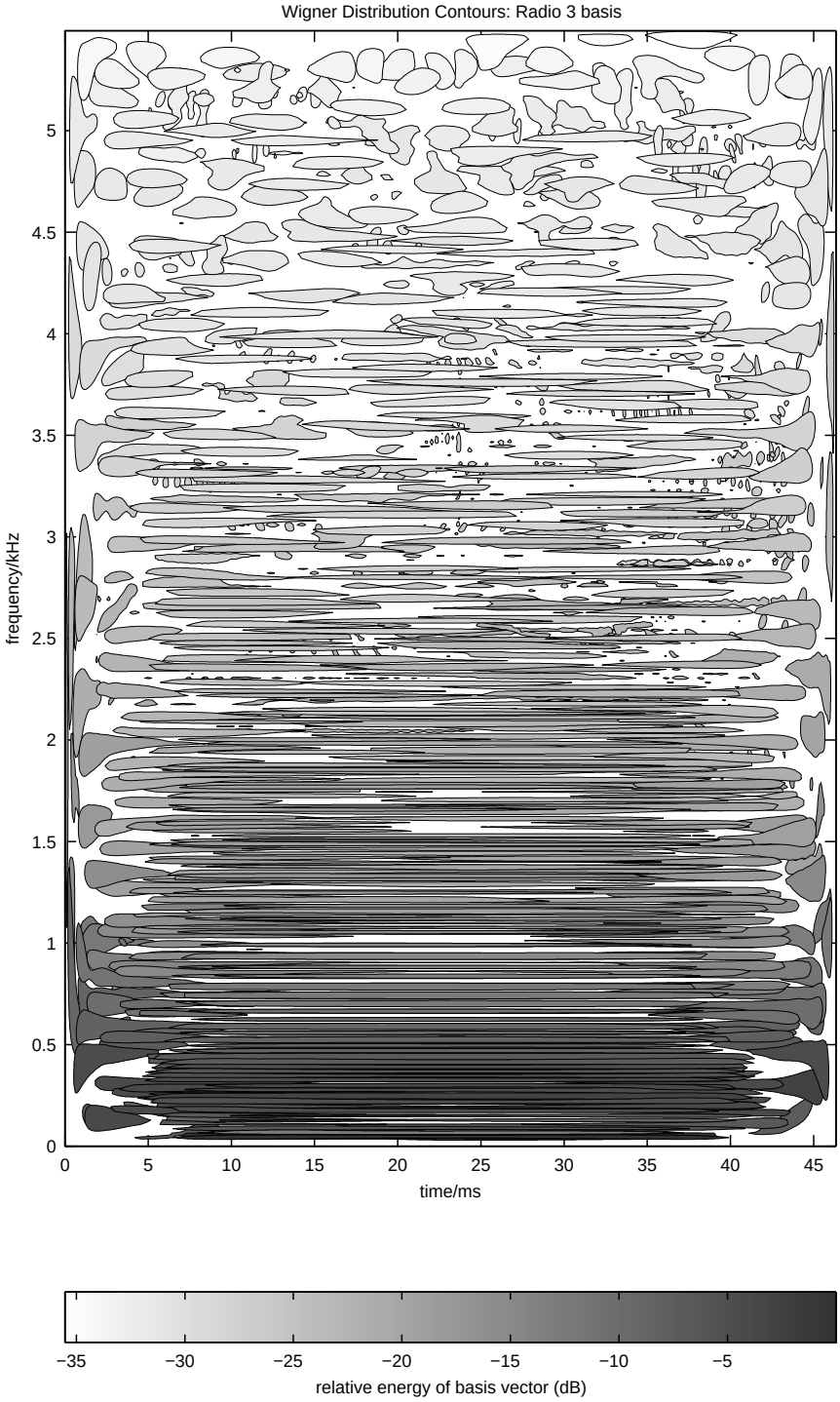


Fig. 5.12: Contour plots of the Radio 3 basis vector Wigner distributions, produced the same way as in fi g. 5.10.

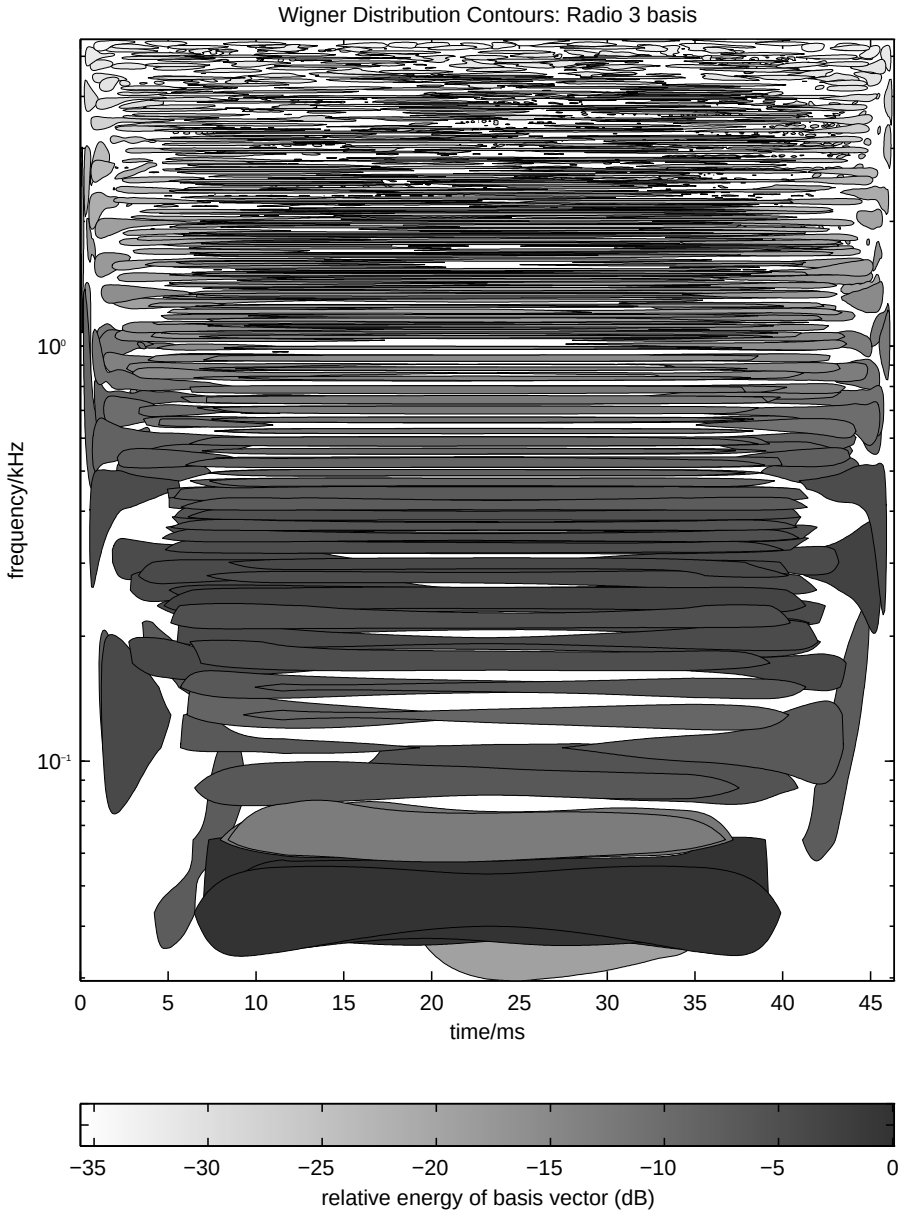


Fig. 5.13: Contour plots of the Radio 3 basis vector Wigner distributions, on a logarithmic frequency scale.

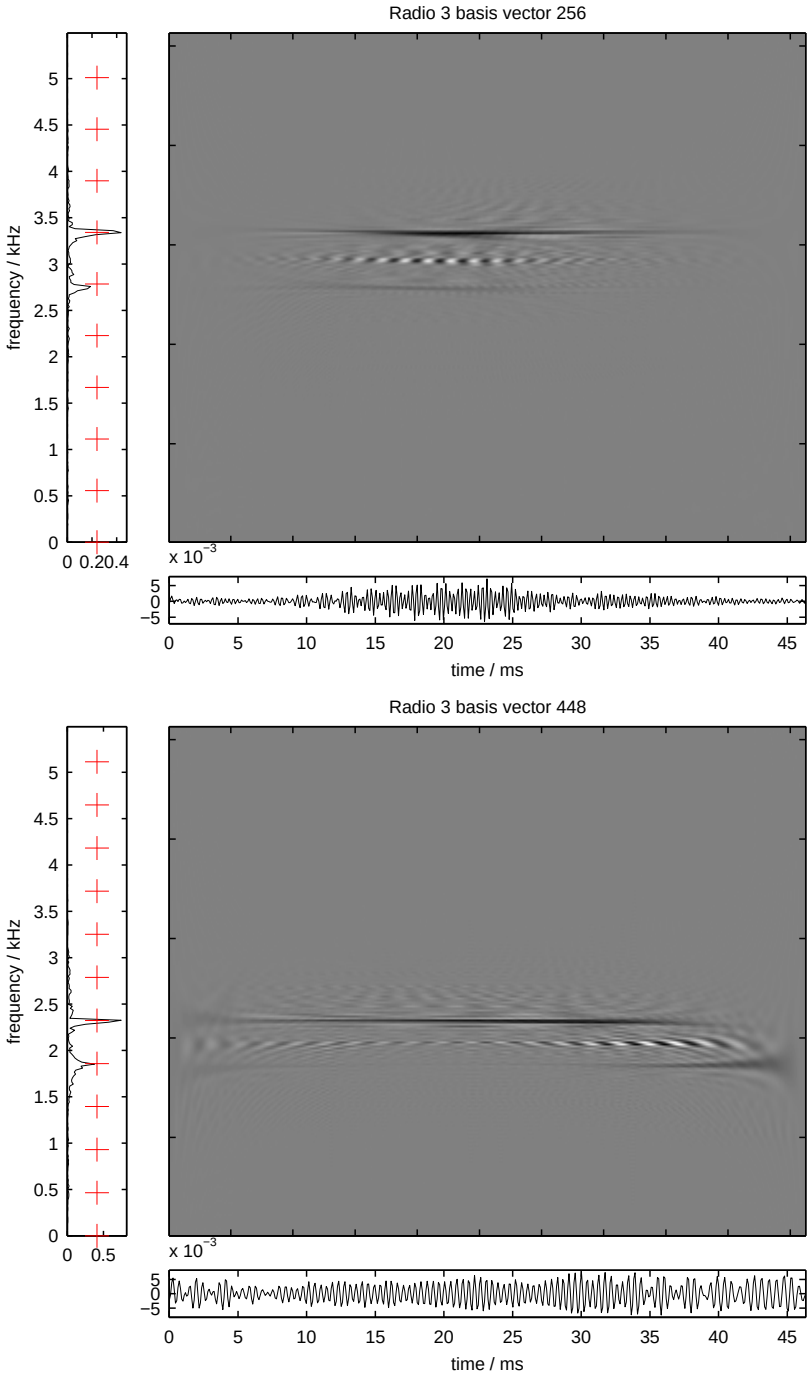


Fig. 5.14: Two of the multi-component Radio 3 basis vectors in time, frequency, and time-frequency. The crosses on the spectral plots are regularly spaced, showing how some of the components are in small integer ratios. Note that in bottom Wigner plot, the two components are consecutive rather than simultaneous, corresponding, in musical terms, to a drop of a Major Third.

bandwidths generally increase with the centre frequency, but there is an clear reversal of this trend between 1 and 1.5 kHz.

ICA experiments with visual scenes have produced results that compare favourably with what is thought to occur in the early stages of the human visual system—the comparison ought to be made between these results and the human auditory system, in particular, the auditory filter-bank (*e.g.* Patterson et al., 1988). This will be the subject of further work, but a cursory investigation suggests that the bandwidths found here are a little narrower than those in the human auditory system. One possible explanation for this is that the continuous speech on which our system was trained may not be the representative auditory environment required to account for the structure of the human auditory system. A closer match might be obtained by training with an environment including more non-speech sounds, such as mechanical noises, animal calls, rustling bushes etc. In this respect, the television may be a better source of training data than radio!

Another interesting avenue of investigation into the speech derived basis is to discover how well adapted (if at all) it is to representing speech. Is the shape of the bandwidth *vs.* centre frequency plot (see fig. 5.3) significant? Does it yield a more efficient coding of speech than a wavelet basis, or other methods?

The music derived results are less conclusive. The basis did include many narrowband sinusoids covering the whole width of the analysis window, suggesting that a Fourier basis is not wholly inappropriate for the analysis of music, at least at a time scale of 50 ms. There were also some wavelet-like features at higher centre frequencies. However there are a significant number of features that are difficult to characterise. Though some consist of harmonically related components, others seem to be very irregular. This may be the result of poor convergence on the part of the ICA algorithm, especially considering the large amount of rather varied training data.

On a more encouraging note, preliminary tests involving listening to the basis vectors has revealed some interesting pitch structure, which is currently under investigation. It appears that there is a periodic variation of bandwidth against frequency with 12 cycles per octave, not visible in fig. 5.7, but which is discernible when the basis vectors are ordered, (according to either centre frequency or bandwidth) concatenated into a long signal, and then listened to.

Finally, an analysis of the marginal distributions of the resulting representation, the components s_i, suggests that a generalised exponential prior, $p(s) \propto \exp -|s|^{\alpha}$ should be used in future experiments, with $0.1 \leq \alpha \leq 0.3$.

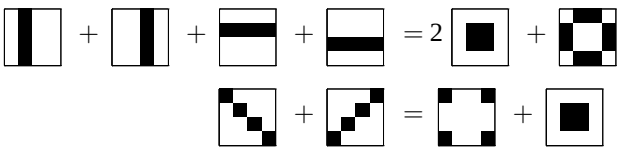
6. SPARSE CODING

Introduction

The aim of sparse coding (Földiák, 1990; Field, 1994) is to find a distributed representation in which "only a fraction of the code elements are actively used to represent a typical pattern" (Harpur, 1997). Information is represented "in terms of a small number of descriptors out of a large set" (Field and Olshausen, 1996). The elements of the representation may be considered to constitute a dictionary of features, of which, only a relatively small number need be combined to describe an observed pattern. This could be taken to an extreme, with exactly one element being activated per pattern, which would be equivalent to a form of quantisation or categorisation; this is not the intention in sparse coding since it implies a form of redundancy. The added requirement that the code be factorial, as well as sparse, discourages this tendency.

As a consequence of having a large dictionary of basic features, it may be possible to reconstruct the input from a number of different combinations of features; that is, the dictionary may be *overcomplete* and the representation *underconstrained*. In a sparse code, this freedom is used to minimise the number of active elements.

Why should it be desirable to have more basic features than apparently necessary? One answer is suggested by the discussion of § 3.1.1: an overcomplete dictionary may be better able to reflect the causal processes behind the observed data. When there are multiple interpretations for a given observation, a sparse coder formulated as a probabilistic causal model can infer the *most likely* explanation for the data in just the same way that perceptual systems are thought to operate, as discussed previously in § 2.1.5. Consider, for example, the patterns in fig. 6.1, and suppose that we observe images formed by linear superposition of a small number of these patterns. Treated as 16-dimensional vectors, these 13 patterns span only a 9-D subspace, due to their linear dependence as the following 'equations' graphically demonstrate:



Hence, the 13 vectors form an overcomplete 'basis' of a 9-D space, and any vector in that space has a non-unique representation in $\mathbb{R}^{13}$. If we now make the assumption that, in the generative process, the coefficient of each basis vector is statistically independent of the others, and has a high probability of being zero, then even if the representation of an observed pattern is non-unique, the most likely decomposition into basis vectors can

Fig. 6.1: Individual patterns of the modified bars data set.

be identified as the one that requires the least number of non-zero coefficients. There is a good chance that this will be the correct 'explanation' for the data, *i.e.* one that agrees with the unobserved generative coefficients. In contrast, a 9-D basis, though able to represent the data without loss of information, would be incapable of identifying the 13 underlying causes. In a sense, the overcomplete representation can 'understand' the data in a way that the 9-D representation cannot.

The following sections develop the theoretical and practical aspects of the framework outlined above, with the aim of applying it (in the next chapter) to spectrograms of polyphonic music. In particular, it was felt that the patterns of harmonics generated by musical notes should be susceptible to this kind of analysis, and that the individual notes might be identified as the basis vectors, in which case, the resulting representation would be a useful step towards obtaining a transcription of the music.

6.1 A generative model

The model to be used has already been introduced informally in the previous section: it is a causal latent variable model in which the observed vectors $\mathbf{x} \in \mathbb{R}^n$ are noisy linear mixtures of m basis vectors, the mixing coefficients s_i being randomly drawn and arranged into a state vector $\mathbf{s} \in \mathbb{R}^m$. Thus, we have

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e}, \tag{6.1}$$

where $\mathbf{A}$ is an $n \times m$ matrix encoding the basis vectors and $\mathbf{e}$ is a Gaussian random vector. An equivalent graphical model is illustrated in fig. 6.2. In practice, the elements of $\mathbf{e}$ will be assumed to be uncorrelated, but for the sake of generality, the analysis will be carried out in terms of an unconstrained covariance matrix $\mathrm{E}\,\mathbf{e}\mathbf{e}^T = \Lambda_{\mathbf{e}}^{-1}$. This gives, putting $\mathbf{e} = \mathbf{x} - \mathbf{A}\mathbf{s}$, the probability density of $\mathbf{x}$ given a known basis matrix $\mathbf{A}$ and source components $\mathbf{s}$:

$$p(\mathbf{x}|\mathbf{A},\mathbf{s}) = \left[\frac{\det \Lambda_{\mathbf{e}}}{(2\pi)^n}\right]^{1/2} \exp -\tfrac{1}{2}\mathbf{e}^T \Lambda_{\mathbf{e}} \mathbf{e}. \tag{6.2}$$

As in ICA, the elements of $\mathbf{s}$ are assumed to be independently drawn from a known continuous density, giving

$$p(\mathbf{s}) = \prod_{i=1}^{m} p(s_i). \tag{6.3}$$

The resulting density model for the observed data is

$$p(\mathbf{x}|\mathbf{A}) = \int_{\mathbb{R}^m} p(\mathbf{x}|\mathbf{A},\mathbf{s})\, p(\mathbf{s})\, \mathrm{d}\mathbf{s}. \tag{6.4}$$

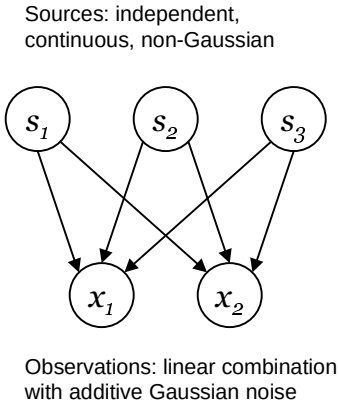


Fig. 6.2: Graphical model of sparse coder with a three-dimensional latent space.

Given such a model and a number of observations of $\mathbf{x}$, we can ask, what is the basis matrix $\mathbf{A}$ most likely to have produced the data? What configuration $\mathbf{s}$ is responsible for each observation? As discussed previously in § 3.1.2, a value of $\mathbf{A}$ most likely to have produced the data can be obtained by maximum likelihood estimation. Then, given a particular observation $\mathbf{x}$ and an estimated basis matrix $\mathbf{A}$, the most probable values for the latent variables $\mathbf{s}$ can be inferred by maximising $p(\mathbf{s}|\hat{\mathbf{A}},\mathbf{x})$. These estimates can be taken as a representation of the data, which by a proper choice of prior $p(s)$ can be encouraged to be a sparse code.

6.1.1 Sparsity in continuous random variables

The introductory discussion described sparse coding in terms of *active* and *inactive* elements, yet, for reasons of computational tractability which will emerge later, the above model is framed in terms of *continuous* random variables with finite density functions. Such variables have an infinitesimal probability of being exactly zero, and thus can hardly be supposed to constitute a sparse code. We must therefore reappraise what we mean by sparsity in this case.

In the literature, (*e.g.* Harpur, 1997; Hyvärinen, 1999) it is usually assumed that inactive elements are zero or *practically zero*, that they are *insignificantly* active. The difficulty lies in defining what 'practically zero' means. The implicit assumption is that those values which are 'close to zero' may be treated as being exactly zero with little or no loss of useful information. In the absence of a model of 'usefulness,' this is not a precise concept.

Several workers (Olshausen, 1996; Lewicki and Sejnowski, 2000; Hyvärinen, 1999) equate sparsity with high kurtosis of a representational element s, defined as

$$\kappa_4(s) = \mathrm{E}\, s^4 - 3(\mathrm{E}\, s^2)^2.$$

It is zero for Gaussians, negative for so-called sub-Gaussian distributions, which have a broader central peak and lighter tails, and positive for super-Gaussian distributions,

which have in comparison a more concentrated peak at zero and heavier tails, for example, the double sided exponential or Laplacian distribution, $p(s) \propto e^{-|s|}$.

Harpur (1997) states that "high kurtosis is a good indicator of high sparseness and low entropy *in the unimodal case*. Its usefulness in more general cases is less clear." The author concurs with this assessment, and would add that such measures based on high order cumulants can be statistically unsatisfactory. Estimates of kurtosis from a finite sample size become less and less reliable as the tails of the distribution become heavier. In fact, for distributions with tails that decay more slowly than s^{-5}, the kurtosis is not even defined, because the integral for E s^4 does not converge.

It may be concluded that kurtosis is not a good way to characterise sparsity, and that it is better to return to the notion of active and inactive elements. Bearing this in mind, the following qualitative definition of sparsity is proposed: namely, that the distribution is *strongly and tightly peaked at zero*, by which is meant:

1. the peak contains much of the 'mass' of the distribution;

2. the peak is narrow in relation to the overall width of the distribution, as characterised by *some appropriate measure*, the implication being that any value smaller than the width of the peak may be set to zero without significant loss of information.

Note that an 'appropriate' width measure need not be variance, which may be poorly defined for some distributions—it could, for example, be a mean absolute value, a maximum value, or a quantile. Given that a distribution satisfies these criteria, then a useful measure of sparsity is simply the probability mass of the peak itself, that is, the probability that a given element is 'inactive.'

Before moving on, it is worth adding that the model described here does not require that the prior $p(s)$ be sparse in the sense described above. In fact, the whole discussion can be framed in terms of *low-entropy* coding (Harpur, 1997). However, since the entropy of a continuous random variable is not invariant to scaling, any assessment of 'low-entropy' must also be made with reference to the notional 'width' of its probability density function.

6.1.2 Relationship with ICA

The model described above is closely related to the standard ICA model with a super-Gaussian prior. It is a natural extension of ICA to the non-square case, where there are more independent components than there are observed mixtures. The inclusion of noise in the model is closely related to the use of overcomplete bases, since the additive noise is equivalent to an extra set of n Gaussian 'independent components.' That the ICA model is not uniquely identifiable with more than one Gaussian source (Comon, 1994), means in this case that the Gaussian sources cannot be unmixed amongst themselves, but this is not an issue since they are treated as noise to be discarded.

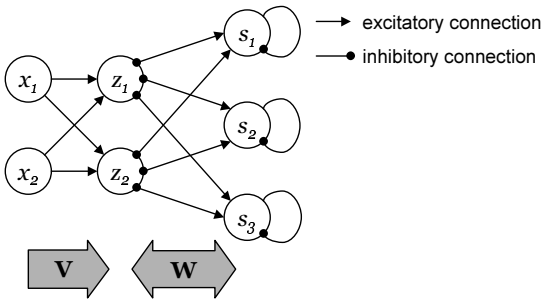


Fig. 6.3: A neural implementation of the activation dynamics.

6.1.3 Inference and the activation dynamics

If **A** is known and **x** is observed, then **s** can be estimated from the posterior density,

$$p(\mathbf{s}|\mathbf{A},\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{A},\mathbf{s})p(\mathbf{s})}{p(\mathbf{x}|\mathbf{A})}. \tag{6.5}$$

A maximum *a posteriori* estimate can then be found at the posterior mode:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} p(\mathbf{s}|\mathbf{A},\mathbf{x}). \tag{6.6}$$

If the posterior is sufficiently smooth, this can be found by gradient ascent on the logarithm of this function, in which case the conditions for zeroing the gradient are

$$\frac{\partial \log p(\mathbf{x}|\mathbf{A},\mathbf{s})}{\partial s_i} + \frac{\partial \log p(\mathbf{s})}{\partial s_i} = 0 \qquad \forall\, 1 \le i \le m. \tag{6.7}$$

It will be useful to define the vector-valued function $\gamma(\mathbf{s})$ as the gradient of the (negative) log-prior: $\gamma(\mathbf{s}) \stackrel{\text{def}}{=} -\nabla_{\mathbf{s}} \log p(\mathbf{s})$. Since $p(\mathbf{s})$ is factorial, its gradient will be an element-wise function defined in terms of the prior density $p(s)$:

$$[\gamma(\mathbf{s})]_i = \gamma(s_i), \qquad \gamma(s) \stackrel{\text{def}}{=} -\frac{\mathrm{d}}{\mathrm{d}s} \log p(s). \tag{6.8}$$

Using this and the expression for $p(\mathbf{x}|\mathbf{A},\mathbf{s})$ from eq. 6.2, the zero-gradient condition can be written as

$$\mathbf{A}^T \Lambda_{\mathbf{e}} (\mathbf{x} - \mathbf{A}\mathbf{s}) - \gamma(\mathbf{s}) = 0. \tag{6.9}$$

Expressed in terms of a dynamic steepest-ascent gradient optimisation, a local maximum of the posterior can be found as a stable fixed point of

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} = \mathbf{A}^T \Lambda_{\mathbf{e}} \mathbf{e} - \gamma(\mathbf{s}), \tag{6.10}$$

though it cannot be guaranteed to be the global maximum unless the posterior is unimodal. These activation dynamics can be interpreted as an *error correcting* mechanism, in which the reconstruction error $\mathbf{e} = \mathbf{x} - \mathbf{A}\mathbf{s}$ induces corrections to **s**, but where each element of **s** is subject to nonlinear decay determined by the function $\gamma(\cdot)$.

It is interesting to observe that the computation can be implemented locally in a three-layer neural network, illustrated in fig. 6.3. The units of first layer are clamped to the input $\mathbf{x}$. The second layer receives feed-forward input from the first layer and feed-back from the third, with activities given by

$$z_k = \sum_{j=1}^{n} V_{kj} x_j - \sum_{i=1}^{m} W_{ki} s_i,$$

where the matrices V_{kj} and W_{ki} represent the connection strengths between the layers. The units of the third layer, in addition to receiving feed-forward input from the second, each have nonlinear self-inhibitory connections, implementing the effect of the term $\gamma(\mathbf{s})$ in eq. 6.10, and are governed by the *dynamic* equation,

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = \sum_{k=1}^{n} z_k W_{ki} - \gamma(s_i).$$

If the matrices $\mathbf{V}$ and $\mathbf{W}$ are chosen so that $\mathbf{W} = \mathbf{VA}$ and $\mathbf{V}^T\mathbf{V} = \Lambda_\mathbf{e}$, (*e.g.* , $\mathbf{V}$ could be the symmetric square-root of $\Lambda_\mathbf{e}$) then it is easily verified that the correct dynamics are produced at the output layer.

The significance of the connection weights $\mathbf{V}$ from the first to the second layer is that they serve to whiten or decorrelate not the data, as is common in ICA algorithms (*e.g.* Karhunen, 1996), but the *noise*. That is, regardless of the noise covariance structure on the input, it appears as decorrelated noise of unit variance at the middle layer. If the noise is decorrelated to begin with, so that $\Lambda_\mathbf{e}$ is diagonal, then so much the better: $\mathbf{V}$ will also be diagonal.

6.1.4 Learning

To estimate the basis matrix $\mathbf{A}$, the following objective function will be maximised:

$$\mathcal{L} = \mathrm{E}_\mathbf{x} \log p(\mathbf{x}|\mathbf{A}), \tag{6.11}$$

where $\mathrm{E}_\mathbf{x}$ denotes the expectation over the observed distribution of $\mathbf{x}$. This is formally equivalent to minimising the Kullback-Leibler divergence between the observed data distribution and that implied by the generative model, given in eq. 6.4. We can gain an insight into how the learning rule is constructed by casting the derivation into the language of statistical physics: defining the energy as

$$\mathcal{E}(\mathbf{s}) = -\log p(\mathbf{x}|\mathbf{A},\mathbf{s}) - \log p(\mathbf{s}), \tag{6.12}$$

the posterior distribution over $\mathbf{s}$ can be written as

$$p(\mathbf{s}|\mathbf{A},\mathbf{x}) = \frac{e^{-\mathcal{E}(\mathbf{s})}}{\mathcal{Z}}, \tag{6.13}$$

where the partition function $\mathcal{Z}$ is defined as

$$\mathcal{Z} = \int_{\mathbb{R}^m} e^{-\mathcal{E}(\mathbf{s})}\, \mathrm{d}\mathbf{s} = p(\mathbf{x}|\mathbf{A}). \tag{6.14}$$

Thus the maximum of the posterior coincides with the minimum of the energy: $\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \mathcal{E}(\mathbf{s})$. To maximise the objective function by gradient ascent, we will need derivatives of $\mathcal{L}$ with respect to the elements of the basis matrix: letting θ stand for some scalar parameter we wish to optimise (*i.e.* an element of $\mathbf{A}$), we find that

$$\frac{\partial \mathcal{L}}{\partial \theta} = \mathrm{E}_{\mathbf{x}} \frac{\partial \log \mathcal{Z}}{\partial \theta}. \tag{6.15}$$

Using eq. 6.14, the derivative of the partition function can be written as

$$\begin{aligned} \frac{\partial \log \mathcal{Z}}{\partial \theta} &= \frac{1}{\mathcal{Z}} \int \frac{\partial}{\partial \theta} e^{-\mathcal{E}(\mathbf{s})} \, \mathrm{d}\mathbf{s} \\ &= -\int \frac{\partial \mathcal{E}(\mathbf{s})}{\partial \theta} \frac{e^{-\mathcal{E}(\mathbf{s})}}{\mathcal{Z}} \, \mathrm{d}\mathbf{s} \\ &= -\int \frac{\partial \mathcal{E}(\mathbf{s})}{\partial \theta} \, p(\mathbf{s}|\mathbf{A},\mathbf{x}) \, \mathrm{d}\mathbf{s}, \end{aligned}$$

which has the form of an expectation over the posterior distribution. Using equations 6.2 and 6.12 and substituting A_{ij} for θ yields

$$\frac{\partial \mathcal{E}(\mathbf{s})}{\partial A_{ij}} = -\left[\Lambda_{\mathbf{e}} (\mathbf{x} - \mathbf{A}\mathbf{s}) \mathbf{s}^T\right]_{ij}. \tag{6.16}$$

Finally, to obtain an online stochastic gradient algorithm, we replace the expectation over $p(\mathbf{x})$ in eq. 6.15 with a sample from the data distribution. If $\mathbf{A}$ is the current estimate of the basis matrix, then the update rule is

$$\begin{aligned} \Delta \mathbf{A} &= \eta \Lambda_{\mathbf{e}} \int (\mathbf{x} - \mathbf{A}\mathbf{s}) \mathbf{s}^T \, p(\mathbf{s}|\mathbf{A},\mathbf{x}) \, \mathrm{d}\mathbf{s} \\ &= \eta \Lambda_{\mathbf{e}} \, \mathrm{E}_{\mathbf{s}|\mathbf{x},\mathbf{A}} \, \mathbf{e}\mathbf{s}^T, \end{aligned} \tag{6.17}$$

where η is a learning rate parameter. The core of this update rule is the Hebbian term $\mathbf{e}\mathbf{s}^T$, but averaged over the posterior density.

6.1.5 Approximations to exact learning

This is an appropriate point to describe the sparse coding algorithms of Field and Olshausen (1996), Harpur (1997), and Lewicki and Sejnowski (2000), and to relate them to eq. 6.17. The integration over the posterior in eq. 6.17 becomes exponentially intractable as the dimensionality of $\mathbf{s}$ grows; an approximation is needed to make the computation manageable.

Delta approximation Both Field and Olshausen's (1996) sparse coder and Harpur's (1997) Recurrent Error Correction (REC) network can be derived by approximating the posterior $p(\mathbf{s}|\mathbf{A},\mathbf{x})$ as an m-dimensional Dirac delta distribution positioned at the posterior mode. Setting $p(\mathbf{s}|\mathbf{A},\mathbf{x}) = \delta(\mathbf{s} - \hat{\mathbf{s}})$ in eq. 6.17 means that $\mathrm{E}_{\mathbf{s}|\mathbf{x},\mathbf{A}} \, \mathbf{e}\mathbf{s}^T = \hat{\mathbf{e}}\hat{\mathbf{s}}^T$, where $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{A}\hat{\mathbf{s}}$, leading to an update of the form

$$\Delta \mathbf{A}_{\mathrm{REC}} = \eta \Lambda_{\mathbf{e}} \hat{\mathbf{e}} \hat{\mathbf{s}}^T. \tag{6.18}$$

Left to itself, this would result in the basis matrix growing without limit and the estimated components of $\hat{\mathbf{s}}$ tending steadily to zero. An explicit normalisation step is required to keep this from happening.

Gaussian approximation Lewicki and Sejnowski (2000) instead use a multivariate Gaussian approximation to the posterior around its maximum, setting

$$p(\mathbf{s}|\mathbf{A},\mathbf{x}) \approx \left[\frac{\det \mathbf{H}}{(2\pi)^M}\right]^{1/2} \exp\left[-\tfrac{1}{2}(\mathbf{s}-\hat{\mathbf{s}})^T\mathbf{H}(\mathbf{s}-\hat{\mathbf{s}})\right]. \tag{6.19}$$

By construction, the mean of the Gaussian is $\hat{\mathbf{s}}$, and its covariance matrix is $\mathbf{H}^{-1}$, where $\mathbf{H}$ is the Hessian of the (exact) log-posterior evaluated at $\hat{\mathbf{s}}$:

$$\mathbf{H} = -\nabla\nabla^T \log p(\mathbf{s}|\mathbf{A},\mathbf{x})\big|_{\hat{\mathbf{s}}} = \mathbf{A}^T\Lambda_{\mathbf{e}}\mathbf{A} - \nabla\nabla^T \log p(\hat{\mathbf{s}}), \tag{6.20}$$

where $\nabla\nabla^T$ is the second-order differential operator, $[\nabla\nabla^T]_{ij} = \partial^2/\partial s_i \partial s_j$. Lewicki and Sejnowski proceed by using this Gaussian approximation to compute the integral for $p(\mathbf{x}|\mathbf{A})$ in eq. 6.4, which they then differentiate to obtain their update rule. The method here has been to differentiate the exact cost function and *then* to approximate, which seems to be a much simpler procedure. The expectation in eq. 6.17 is trivial to compute if the posterior is assumed to be Gaussian:

$$\mathrm{E}_{\mathbf{s}|\mathbf{A},\mathbf{x}}\,\mathbf{e}\mathbf{s}^T = \mathrm{E}_{\mathbf{s}|\mathbf{A},\mathbf{x}}(\mathbf{x}\mathbf{s}^T - \mathbf{A}\mathbf{s}\mathbf{s}^T) = \mathbf{x}\hat{\mathbf{s}}^T - \mathbf{A}(\hat{\mathbf{s}}\hat{\mathbf{s}}^T + \mathbf{H}^{-1}). \tag{6.21}$$

This yields a weight update of the form

$$\Delta\mathbf{A}_{\text{Gauss}} = \eta\Lambda_{\mathbf{e}}(\hat{\mathbf{e}}\hat{\mathbf{s}}^T - \mathbf{A}\mathbf{H}^{-1}), \tag{6.22}$$

where $\hat{\mathbf{e}} = \mathbf{x} - \mathbf{A}\hat{\mathbf{s}}$. Compared with the delta approximation, this expression has gained a decay term $\mathbf{A}\mathbf{H}^{-1}$ which has the effect of solving the problem of unlimited weight growth, essentially by taking into account the width of the posterior around its peak. It means that the basis vectors are automatically normalised so that the marginal distributions of the state variable match the prior.

Lewicki and Sejnowski actually employ a different update rule which gives faster convergence and avoids computing any matrix inverses:

$$\Delta\mathbf{A}_{\text{LS}} = \eta\mathbf{A}[\gamma(\hat{\mathbf{s}})\hat{\mathbf{s}}^T - \mathbf{I}]. \tag{6.23}$$

This can be obtained from eq. 6.22 in the following way. As with the ICA algorithm in §5.1, considerations of the natural gradient (Amari et al., 1996) and the relative gradient (Cardoso and Laheld, 1996) motivate the multiplication of the matrix update by $\mathbf{A}\mathbf{A}^T$. This, coupled with the observation that, according to eq. 6.9, $\mathbf{A}^T\Lambda_{\mathbf{e}}\hat{\mathbf{e}} = \gamma(\hat{\mathbf{s}})$ at a local maximum of the posterior, gives

$$\begin{aligned}\mathbf{A}\mathbf{A}^T\Delta\mathbf{A}_{\text{Gauss}} &= \eta\mathbf{A}\mathbf{A}^T\Lambda_{\mathbf{e}}(\hat{\mathbf{e}}\hat{\mathbf{s}}^T - \mathbf{A}\mathbf{H}^{-1}) \\ &= \eta\mathbf{A}[\gamma(\hat{\mathbf{s}})\hat{\mathbf{s}}^T - \mathbf{A}^T\Lambda_{\mathbf{e}}\mathbf{A}\mathbf{H}^{-1}].\end{aligned} \tag{6.24}$$

If we then neglect the Hessian of the prior, $\nabla\nabla^T \log p(\hat{\mathbf{s}})$, in eq. 6.20, and approximate $\mathbf{H} \approx \mathbf{A}^T \Lambda_\mathbf{e} \mathbf{A}$, then $\mathbf{A}^T \Lambda_\mathbf{e} \mathbf{A} \mathbf{H}^{-1} \approx \mathbf{I}$. Substituted in to eq. 6.24, this yields the Lewicki-Sejnowski learning rule in eq. 6.23, which, as Lewicki and Sejnowski point out, is identical to the natural gradient update rule in the standard noiseless, square ICA problem, except that it relies on a more complex optimisation of $\hat{\mathbf{s}}$ in order to be valid for noisy data and a non-square basis matrix. In this case however, because of the additive noise in the model, it yields only an approximation to the true natural gradient, which can lead to erroneous convergence in some cases, as we will see in § 6.4 and § 6.5.

6.2 The form of the prior

Bearing in mind the intended application of this system (in the next chapter), and observing that the marginal distributions of spectral bands tend to be quite strongly super-Gaussian, we might suspect that any sparse components discovered will be yet more super-Gaussian. This constitutes a form of prior information, and one would like to incorporate this into any sparse coding algorithm in the form of a strongly super-Gaussian prior density $p(s)$. It will therefore be useful to investigate how the system copes with such a prior. A number of observations can be made:

- A Laplacian is the sparsest prior that still guarantees a unimodal posterior. That this is so can be seen by considering the energy function in eq. 6.12:

  $$\mathcal{E}(\mathbf{s}) = -\log p(\mathbf{x}|\mathbf{A},\mathbf{s}) - \log p(\mathbf{s}).$$

  With Gaussian noise, the first term is a positive-definite quadratic form and hence globally convex. This means it will have have exactly one local minimum. If the Hessian of $\log p(\mathbf{s})$ is positive semi-definite everywhere, a unique minimum of the energy can be assured. Since $\log p(\mathbf{s}) = \sum_i \log p(s_i)$, this requires $-\log p(s)$ to have a non-negative second derivative everywhere. Now, because the Laplacian log-prior consists of two linear segments, it is on the verge of non-convexity. If the prior were to become any 'peakier', a multimodal posterior would become a possibility. As soon as the posterior becomes multimodal, not only does the global maximum become harder to find, it becomes less representative of the distribution as a whole, and less useful for approximating the expectation in eq. 6.17. In essence, the various approximations described in Section 6.1.4 break down.

- A gradient discontinuity at zero (as in the Laplacian) is desirable, because it gives a code with many *exact* zeros rather than just very small values, because of the shrinkage effect shown in fig. 6.4(c) and described in § 6.2.1. This is closer to our intuitive concept of sparsity and is more attractive than the alternative of thresholding small values, because it does not involve an arbitrary choice of threshold. If a component is forced to zero because of the shrinkage effect, the 'error-correcting' dynamics in eq. 6.10 allows other non-zero components to compensate.

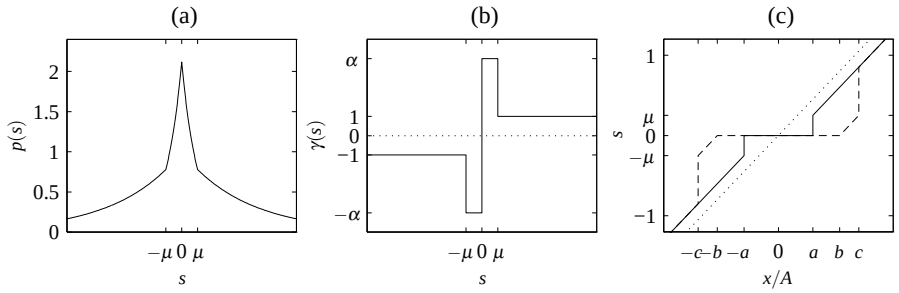


Fig. 6.4: Approximation to 'sparsified' Laplacian, constructed piecewise from exponential segments. (a) Prior, $p(s)$, (b) $\gamma(s) = -(\mathrm{d}/\mathrm{d}s)\log p(s)$, and (c) a multi-valued shrinkage 'function,' which gives the stable solutions of the activation dynamics in a one-dimensional system (*i.e.* , x, A, and s are scalars). The solid and dashed lines indicate local minima of the posterior $p(s|A,x)$. The labelled points on the x/A axis are $a = \sigma^2/A^2 + \mu$, $b = \alpha\sigma^2/A^2$, and $c = \alpha\sigma^2/A^2 + \mu$. The line $\hat{s} = x/A$ is shown dotted.

- One way of looking at a sparse random variable is to treat it as a mixture of zeros and continuous values drawn from some density such as a Gaussian or Laplacian. The resultant (improper) probability density of such a variable will have a Dirac delta distribution at zero. The posterior will potentially be multimodal and will have infinite peaks wherever any of the s_j are zero. Needless to say, this will pose problems for any procedure based on maximum *a posteriori* estimation, which will always yield $\hat{\mathbf{s}} = \mathbf{0}$. This is a situation in which MAP estimation is not appropriate; the correct estimation of such mixed discrete/continuous variable requires special treatment in which the continuous part is marginalised out.

6.2.1 An approximate sparsified Laplacian prior

In order to allow the use of the relatively simple algorithm described earlier, which requires a continuous prior, an approximation to a 'sparsified' Laplacian density was constructed, consisting of two Laplacian pieces. The central part forms a narrow peak, as illustrated in fig. 6.4(a). The parameters $\mu \geq 0$ and $\alpha \geq 1$ control the width and relative mass of the central peak:

$$p(s) = \begin{cases} Ce^{-|s|} & : \ |s| \geq \mu, \\ CKe^{-\alpha|s|} & : \ |s| < \mu, \end{cases} \tag{6.25}$$

where C is a normalisation constant, and $K = e^{\mu(\alpha-1)}$ to ensure continuity. The width parameter μ is intended to be small, in keeping with the working definition of sparsity outlined in § 6.1.1, in which case the distribution's sparsity can be measured as

$$P(|s| < \mu) = \frac{e^{\mu\alpha} - 1}{e^{\mu\alpha} + \alpha - 1}, \tag{6.26}$$

allowing us to calculate the parameters required for a given level of sparsity. The associated function $\gamma(s)$, illustrated in fig. 6.4(b), is given by

$$\gamma(s) = \begin{cases} \operatorname{sgn} s & : |s| \geq \mu, \\ \alpha \operatorname{sgn} s & : |s| < \mu, \end{cases} \tag{6.27}$$

This prior results in a thresholding behaviour when computing $\hat{s}$. Following Hyvärinen (1999), we consider the one-dimensional case ($N = M = 1$), in which case, the stationarity condition in eq. 6.9 gives, at the maximum of the posterior,

$$\tfrac{1}{\sigma^2} A(x - A\hat{s}) = \gamma(\hat{s}), \tag{6.28}$$

which can be written as

$$x/A = g(\hat{s}) = \hat{s} + \tfrac{\sigma^2}{A^2}\gamma(\hat{s}). \tag{6.29}$$

If the function $g(\cdot)$ was invertible, $\hat{s}$ would be obtainable as a function of x: $\hat{s} = g^{-1}(x/A)$. In this case, $g(\cdot)$ is not invertible, but a multi-valued 'function' can be constructed to return all the values of $\hat{s}$ which satisfy eq. 6.29, *i.e.* , all the local maxima of the posterior—see fig. 6.4(c). Where the posterior is bimodal, two local maxima should be compared to obtain the MAP estimate, but in practice, a gradient based optimiser like the one described in the next section may find one or other depending on its initial conditions.

6.3 A modified gradient optimiser

To find the maximum of the posterior—or equivalently, the minimum of the energy function $\mathcal{E}(\mathbf{s})$—an optimisation algorithm is required. Gradient based methods work best with functions that are approximately quadratic near the optimum, but we wish to be able to deal with priors that have a gradient discontinuity at zero, inducing corresponding discontinuities in the energy at coordinate zeros. If the minimum of the energy function occurs at one of these creases, a gradient based optimisation procedure will have problems converging.

To address this, a modified quasi-Newton optimiser was implemented which explicitly recognises the fact that there may be gradient discontinuities at component zeros of the vector being optimised. In a fashion similar to the quadratic programming method of Endres and Földiák (1999), the algorithm maintains a set of active dimensions, and could thus be called an *active set quasi-Newton* optimiser. In fact, the same modification can be applied to any iterative gradient-based optimiser that involves a step length computation; for example, a conjugate gradient version was also implemented.

Both the quasi-Newton optimiser and the conjugate gradient optimiser were based algorithms found in (Press et al., 1992) and the function FMINU in the MATLAB optimisation toolbox. Details such as step length heuristics, termination conditions and Hessian approximations can be found in those sources and will not be stated here.

The optimisation is an iterative procedure, involving a current point $\mathbf{s} \in \mathbb{R}^m$ and two sets I_0 and I_1 which contain the indices of the inactive and active elements of $\mathbf{s}$

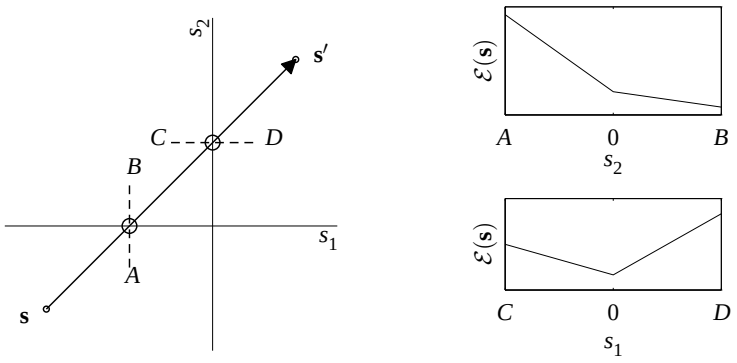


Fig. 6.5: A Two dimensional illustration of the operation of the modified active set optimiser. Given the proposed step from $\mathbf{s}$ to $\mathbf{s}'$, and the local behaviour of the gradient along the segments *AB* and *CD*, the modified optimiser would truncate the step at the *second* zero crossing, and inactivate the *first* coordinate, s_1.

respectively:

$$I_0 \cap I_1 = \emptyset, \qquad I_0 \cup I_1 = \{1, \ldots, m\}.$$

Inactive coordinates are set to zero and do not take part in the current optimisation step, though they may subsequently be activated: $i \in I_0 \implies s_i = 0$. To determine whether or not the ith coordinate should be active, a boolean indicator function $Q(\mathbf{s}, i)$ is defined, which, assuming that $s_i = 0$, depends on the signs of the gradient of the cost function on either side of the point $s_i = 0$. These gradients are given by

$$\partial_i^+ \mathcal{E}(\mathbf{s}) \stackrel{\text{def}}{=} \lim_{s_i \downarrow 0^+} \frac{\partial \mathcal{E}(\mathbf{s})}{\partial s_i}, \qquad \partial_i^- \mathcal{E}(\mathbf{s}) \stackrel{\text{def}}{=} \lim_{s_i \uparrow 0^-} \frac{\partial \mathcal{E}(\mathbf{s})}{\partial s_i}, \tag{6.30}$$

where the limits are taken tending *down* to a value just above zero, and *up* to a value just below zero respectively. The indicator function is

$$Q(\mathbf{s}, i) = \begin{cases} 0 & : [\operatorname{sgn} \partial_i^+ \mathcal{E}(\mathbf{s})][\operatorname{sgn} \partial_i^- \mathcal{E}(\mathbf{s})] \leq 0 \\ 1 & : \text{otherwise}, \end{cases} \tag{6.31}$$

in which $\operatorname{sgn} 0 \stackrel{\text{def}}{=} 0$. If $Q(\mathbf{s}, i) = 0$, then the point $\mathbf{s}$ represents a local minimum in the direction of the ith dimension, and so that dimension should be deactivated. (Note that this definition is correct only because we know that $\gamma(s)$ has a *positive* step at zero, and hence cannot cause a local *maximum*.)

The algorithm

Initialise by inactivating all coordinates, setting

$$I_0 = \{1, \ldots, m\}, \qquad I_1 = \emptyset, \qquad \mathbf{s} = \mathbf{0}.$$

Then, for each iteration of the main loop:

1. Compute proposed new point $\mathbf{s}'$ and step Δ according to the base quasi-Newton or conjugate gradient algorithm, so that $\mathbf{s}' = \mathbf{s} + \Delta$.

2. See if the proposed step would result in any of the components of $\mathbf{s}$ changing sign, by finding the set off all such zero-crossings:

$$Z = \{i \in I_1 | \operatorname{sgn} s_i \neq \operatorname{sgn} s_i'\}$$

3. See if any of the zero-crossings satisfy the inactivation criterion. First, define $\lambda(i)$ as the step size that will take us to the zero crossing in the ith coordinate:

$$\lambda(i) = -s_i/\Delta_i,$$

so that $[\mathbf{s} + \lambda(i)\Delta]_i = 0$. Then find the set Z_0 of zero crossings that satisfy the inactivation criterion:

$$Z_0 = \{i \in Z | Q(\mathbf{s} + \lambda(i)\Delta, i) = 0\}.$$

4. If there are any such zero crossings, choose one (*e.g.* , the first, but I have no rigorous basis for choosing one over the others) and truncate the step there, otherwise, take the proposed step unmodified. ($\leftarrow$ denotes assignment.)

$$\lambda^* = \begin{cases} \min_{i \in Z_0} \lambda(i) & : Z_0 \neq \emptyset \\ 1 & : Z_0 = \emptyset \end{cases}$$
$$\mathbf{s} \leftarrow \mathbf{s} + \lambda^* \Delta.$$

5. Update the active and inactive sets to reflect the new current point $\mathbf{s}$. To do this, two sets are defined:

$$I_- = \{i \in I_1 : Q(\mathbf{s}, i) = 0 \wedge s_i = 0\}$$
$$I_+ = \{i \in I_0 : Q(\mathbf{s}, i) = 1\}.$$

I_- is the set of currently active coordinates that should be deactivated. I_+ is the set of currently inactive coordinates eligible for reactivation. Pick just *one* of these. It is not clear how best to choose the one to be released, as the optimiser may end up in one of several local minima depending on the choice, but the current implementation attempts to rate their 'eligibility', reassigning I_+ to contain only the coordinate with the highest rating:

$$I_+ \leftarrow \{\arg\max_{i \in I_+} \tfrac{1}{2} \left| z_+(\mathbf{s}, i) + z_-(\mathbf{s}, i) \right|\}.$$

All that remains is to transfer the various elements between the active and inactive sets:

$$I_0 \leftarrow (I_0 \backslash I_+) \cup I_-$$
$$I_1 \leftarrow (I_1 \backslash I_-) \cup I_+,$$

where the operator $\backslash$ denotes set difference.

6. Perform any book-keeping required by the quasi-Newton algorithm—in particular update the Hessian approximation using gradient information at the current point. The important point about this is that we update only the sub-matrix defined by the currently active coordinates.

The main loop is subject to the same termination conditions as the unmodified optimiser, except that these apply to the currently active coordinates only. In addition, the main loop terminates if there are no active coordinates.

By inactivating directions in which the local gradient is discontinuous, this algorithm keeps the Hessian approximation from becoming ill-conditioned. Under certain conditions, the algorithm results in significant speed improvements over the equivalent unconstrained quasi-Newton optimiser (see the results of the next section) essentially by exploiting the sparsity we are trying to achieve. Indeed, in use, as the basis converges, the code becomes sparser and the performance improves.

The quasi-Newton version of this algorithm performed better than the equivalent active set conjugate gradient version. These two were, as far as possible, evenly matched, using the same termination conditions and the same line search procedure. It seems that the expense of maintaining a large inverse Hessian approximation is mitigated by the fact that only a small number of coordinates is active at any one time.

It should be noted that this optimisation procedure is quite closely related to *matching pursuits* (Mallat and Zang, 1993), which is an approximate solution to the problem of sparse decomposition in an overcomplete dictionary. Matching pursuits is equivalent to a greedy coordinate descent on the quadratic part of the cost function used here (the error term) and disregarding the sparsity cost. At each iteration, a line minimisation is performed along the coordinate that will yield the greatest reduction in reconstruction error. This tends to limit the number of coordinates which become active, but since their is no explicit sparsity cost corresponding to a prior on the components, the algorithm does not perform the kind of optimal inference that can be done with an explicit latent variable model.

6.4 The bars problem

The algorithm was tested on the patterns described in the introduction and illustrated in fig. 6.1, a variation of the bars data set (Földiák, 1990) designed to be more overcomplete than the original bars data set. The mixing coefficients s_i were drawn from a 'sparsified' Laplacian distribution, that is, a random mixture of Laplacians and zeros in the proportion $z : (1-z)$. Uncorrelated Gaussian noise of variance σ_* was added to the observations $\mathbf{x}$. The parameters z and σ_* were varied between experiments.

Both a Laplacian prior and the sparsified Laplacian prior from § 6.2 were tested using the modified quasi-Newton optimiser described in § 6.3. The sparsified Laplacian prior was parameterised, as described in eq. 6.25, in terms of μ and α, various combinations of which were tested. The results were compared with those obtained using a

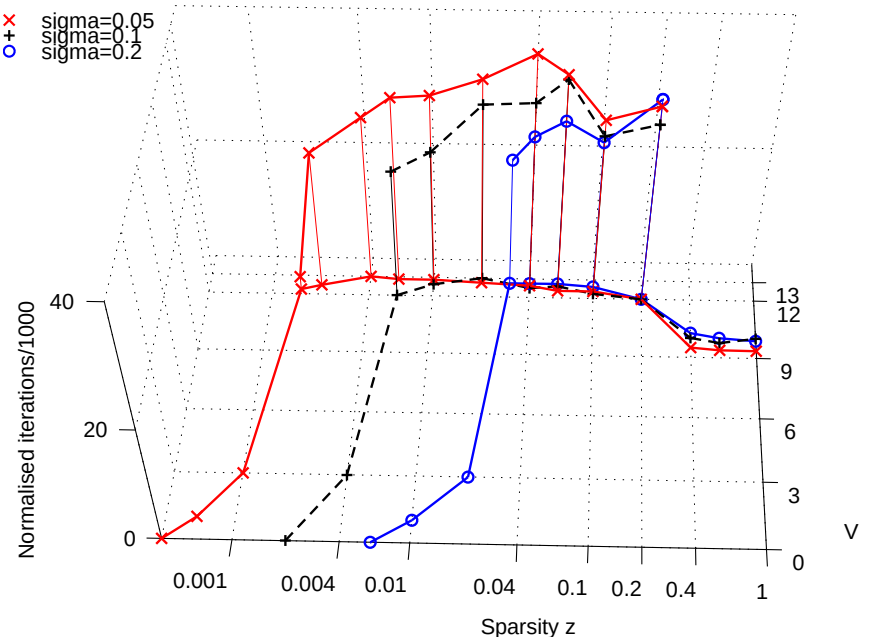


Fig. 6.6: Average numbers (V) of non-zero basis vectors learned using modifi ed optimiser with Laplacian prior, for several sparsities, z, and three values of σ. The vertical axis shows the average number of iterations, taken to converge to the correct 13 basis vectors. The values shown have been 'normalised' by multiplying by $\eta/0.016$, where η is the learning rate that was used for that trial. This removes the trivial linear dependence we would expect between learning rate and iterations to convergence, revealing an underlying tendency for the system to take longer to converge for higher values of z, until $z = 0.1$. With $z > 0.2$, the full basis is not found, so the iterations to convergence are not plotted.

family of smooth sparse priors:

$$p(s) \propto \operatorname{sech}^{1/\beta} \beta s, \tag{6.32}$$

$$\gamma(s) = \tanh \beta s. \tag{6.33}$$

As the 'sharpness' parameter $\beta \to \infty$, this tends to a Laplacian distribution, but for finite β, the distribution and $\gamma(s)$ both remain smooth, so that a standard quasi-Newton optimiser is sufficient to find the maximum of the posterior.

The model noise covariance $\Lambda_{\mathbf{e}}$ was set to $\sigma^{-2}\mathbf{I}$, varying σ between experiments. Learning was performed using batches of 96 input patterns chosen randomly each time (*i.e.* learning was not performed repeatedly on the same batch.) The square (16 by 16) basis matrix A was initialised to $0.1\mathbf{I}$, rather than $\mathbf{I}$, at the start of each run, in order to avoid an initial period during which the weights simply decayed to roughly the right magnitudes before finding the right directions. The learning rate η was of the order of 0.01, but it was found that smaller learning rates were required for high sparsities.

A number of observations can be made about the results (see figures 6.6–6.8):

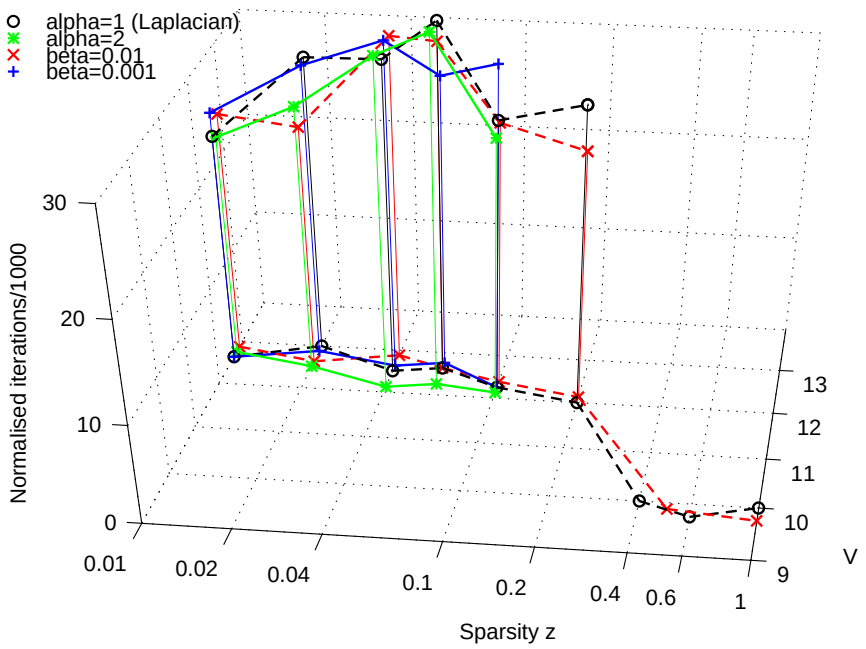


Fig. 6.7: Comparison of learning behaviour using standard and modified optimisers for two values of β (see eq. 6.32) and two values of α (see eq. 6.25). Setting $\alpha = 1$ corresponds to an unmodified Laplacian prior.

Effect of low sparsity Though there were 13 distinct patterns, the basis did not always converge to 13 vectors. In the case of non-sparsified Laplacian input ($z = 1$) it nearly always converged to 9 or 10 basis vectors, which is interesting because the intrinsic dimensionality of the data set is 9. It seems that a certain level of sparsity is required for the algorithm to find an overcomplete basis.

Effect of high sparsity Conversely, too high a level of sparsity also resulted in too few basis vectors being found. The scaling of these vectors is roughly proportional to z, the proportion of Laplacians in the source mixture density, but as the sparsity is increased, at a certain point, all the basis vectors collapse to zero. The point at which this happens depends on the noise parameter σ—the higher the value of σ, the higher the critical value of z.

Performance of modified optimiser The performance advantage of the modified optimiser with zero-crossing detection becomes apparent only when significant numbers of units can be set to exactly zero. The conditions under which this happens involve a trade-off between the sparsity of the input, the actual noise level σ_*, the assumed noise level σ, and the form of the prior.

The standard quasi-Newton optimiser using the smooth prior of eq. 6.32 is quite fast for large β, but slows down as β decreases and the prior becomes more sharply peaked. (These results are not included here.) Holding the other parameters fixed,

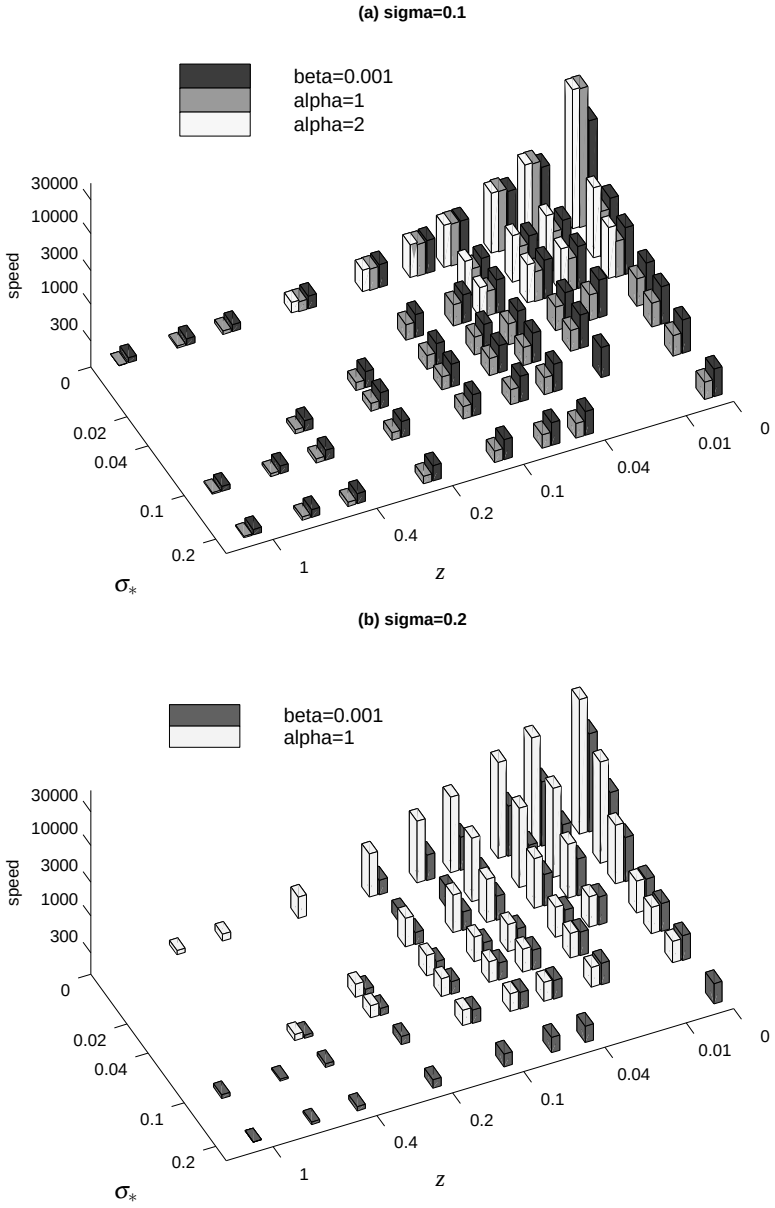


Fig. 6.8: Performance of standard and modified optimisers under various conditions. These results were obtained by running the optimiser with the basis matrix **A** equal to the actual basis used to generate the data. The speeds are quoted in optimisations per second. The upper chart (a) shows how the modified optimiser is faster for high sparsities and low input noise levels, while lower chart (b) illustrates that this effect is more pronounced at high values of σ, the expected noise parameter. In this particular (rather small) problem, the modified optimiser outperforms the standard one over only a restricted range of parameters and is of questionable usefulness. Further experiments with larger problems have shown it in a more favourable light.

there will be a value of β_c such that if $\beta < \beta_c$, the modified optimiser will outperform the standard one.

Effect of modified prior The sparsified prior has a more aggressive de-noising effect on the output, encouraging more of the outputs to remain at exactly zero, reflecting more accurately the actual source distribution. It also results in much faster performance from the modified optimiser.

However, enforcing this greater degree of sparsity too aggressively seems to disrupt learning, resulting in too few basis vectors being learned. This is probably because the shape of the prior admits of a multimodal posterior, with strong local maxima at coordinate zeros. Units that should be activated fail to be so, which stops them from participating in the learning process.

Overall, the algorithm is capable of discovering a sparse code to represent the input that accurately reflects the number of independent basic patterns. The patterns themselves may be linearly dependent and hence form an overcomplete basis within the subspace spanned by them. The success of this is, however, dependent on the sparsity of the input being between certain limits—too low, and an overcomplete basis may not be found; too high, and the basis vectors decay to zero.

6.5 Analysis of learning in one dimension

In this section we examine the dynamics and stability of the Lewicki-Sejnowski learning rule (eq. 6.23) when applied to a one-dimensional 'sparse coder', if we can still call it that. An analysis of even this apparently trivial system does shed some light on the behaviour of the full, multidimensional system.

The focus will initially be restricted to a Laplacian prior. The MAP estimate $\hat{s}$ can then be written in closed form, as a function of the input x. We will first assume noiseless Laplacian data, then we will generalise to the noisy case, and also to the modified prior introduced in §6.2. We will see that that this prior does not have the desired effect at all.

The one-dimensional learning task amounts to fitting the prior distribution to the data distribution by adjusting a single scale parameter A. The continuous time version of the Lewicki-Sejnowski update rule is

$$\frac{\mathrm{d}A}{\mathrm{d}t} = A(\mathrm{E}_x[\gamma(\hat{s})\hat{s}] - 1), \tag{6.34}$$

where the expectation is over the observed distribution of x. The quantity $\gamma(\hat{s})\hat{s}$ will be of importance throughout the derivation, so we will define $\Gamma(\hat{s}) \stackrel{\text{def}}{=} \gamma(\hat{s})\hat{s}$, which, when the dependence of $\hat{s}$ on x is taken into account, can be considered to be a function of x and written $\Gamma(x)$. The expectation of this quantity will then be defined as

$$\mathcal{G} \stackrel{\text{def}}{=} \mathrm{E}_x\Gamma(x) = \int_{-\infty}^{\infty} \Gamma(x)p(x)\,\mathrm{d}x. \tag{6.35}$$

The dynamic equation can then be written as $\mathrm{d}A/\mathrm{d}t = A(\mathcal{G}-1)$. Thus, $\mathcal{G}$ can be thought of as a 'learning signal,' which drives A one way if $\mathcal{G} > 1$ and the other if $\mathcal{G} < 1$. Fixed

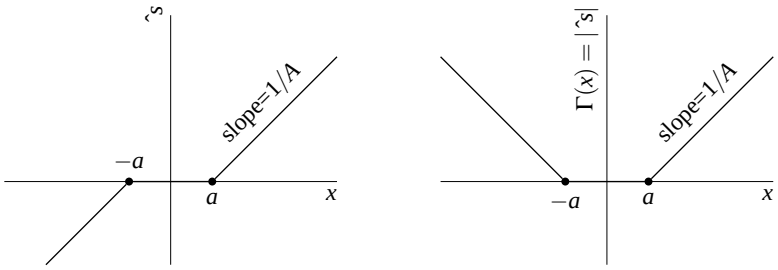


Fig. 6.9: Graph (a) is the 'shrinkage function,' which gives $\hat{s}$ as a function of x. Graph (b) is the function $\Gamma(x)$. The threshold of activation is $a = \sigma^2/A$.

points of the dynamics are to be found at $\mathcal{G} = 1$ but also at $A = 0$. It will be assumed, without loss of generality, that $A \geq 0$ since the priors to be considered are all symmetric about zero.

6.5.1 Using a Laplacian prior

With a Laplacian prior, we have $p(s) = \frac{1}{2}e^{-|s|}$, and hence $\gamma(s) = \mathrm{d}|s|/\mathrm{d}s = \mathrm{sgn}\, s$, (see eq. 6.8). In this case, $\hat{s}$ is a piece-wise linear function of x:

$$A\hat{s} = \begin{cases} x - \sigma^2/A & : x > \sigma^2/A, \\ x + \sigma^2/A & : x < -\sigma^2/A, \\ 0 & : \text{otherwise}, \end{cases} \tag{6.36}$$

where σ is the model noise parameter. This gives, in terms of x,

$$\Gamma(x) = \hat{s}\,\mathrm{sgn}\,\hat{s} = \max\left\{\frac{|x| - \sigma^2/A}{A}, 0\right\}. \tag{6.37}$$

A graph of this function is illustrated in fig. 6.9.

Noiseless Laplacian input

Let us now assume that the data used to drive the learning actually is Laplacian, with a scale parameter A_* distinct from A which is the *model* scale parameter. The data distribution is therefore

$$p(x) = (1/2A_*)e^{-|x|/A_*}. \tag{6.38}$$

Substituting this and eq. 6.37 into eq. 6.35 yields

$$\mathcal{G} = (A_*/A)e^{-\sigma^2/AA_*}. \tag{6.39}$$

At this point it will be useful to introduce a dimensionless parameterisation,

$$\theta = A_*/A, \qquad \lambda = \sigma^2/A_*^2, \tag{6.40}$$

in which terms the result can be stated very simply:

$$\mathcal{G} = \theta e^{-\lambda\theta}. \tag{6.41}$$

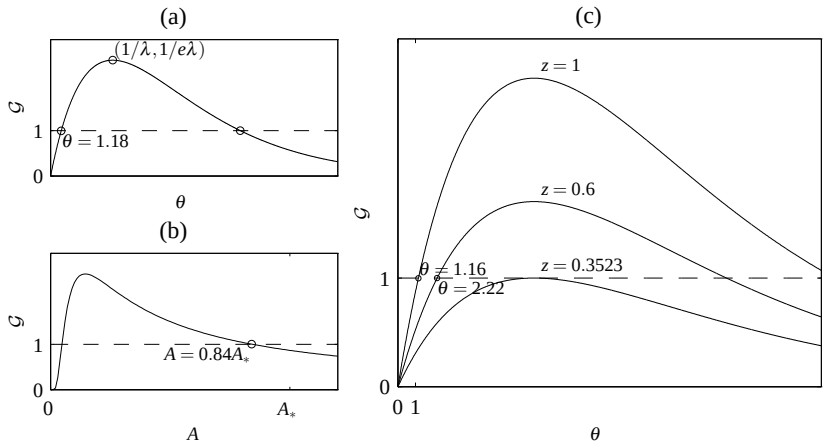


Fig. 6.10: Learning dynamics with a Laplacian prior. (a) $\mathcal{G}$ as a function of $\theta = A/A_*$. Since $dA/dt \propto A(\mathcal{G}-1)$, fixed points exist where $\mathcal{G} = 1$, stable where the slope is positive. (b) The same behaviour in terms of A rather than of θ. The stable fixed point is now on the right. (c) Three curves illustrating what happens when a *sparsified* Laplacian input is used. Increasing the sparsity (by reducing z) results in the loss of the two fixed points. The third curve shows the critical value of z at which this happens. (In these examples, $A_* = 1$ and $\sigma = 0.36$.)

In this equation, θ is the dynamic variable, varying inversely with A. From eq. 6.34, it is clear that fixed points are to be found wherever $\mathcal{G} = 1$. Two such points can be seen in fig. 6.10(a), one of which is stable. For small values of λ (that is, for $\sigma \ll A_*$) the stable solution will be near $\theta = 1$, which is the 'correct' solution, since $\theta \approx 1 \implies A \approx A_*$. If θ is initially beyond the unstable fixed point, it will continue to grow without limit. This means that A will tend to zero, which, thanks to the additional factor of A in the dynamic equation (6.34), is also a stable fixed point of the dynamics. As λ increases, the two fixed points move closer together until a critical point is reached at which they both disappear. The details of this critical point will be derived in the next section for the more general case of 'sparsified' Laplacian input.

Sparsification of the input

A 'sparsified' Laplacian variable can be produced by taking samples from a Laplacian distribution and setting certain fraction of them, say $(1-z)$, to zero. The resulting probability density is

$$p(x) = (z/2A_*)e^{-|x|/A_*} + (1-z)\delta(x), \tag{6.42}$$

where $\delta(x)$ denotes the Dirac delta distribution. This results in a simple modification to the learning behaviour:

$$\mathcal{G} = z\theta e^{-\lambda\theta}. \tag{6.43}$$

The extra factor of z implies that for small λ, the stable solution will no longer be near $\theta = 1$, but near $\theta = 1/z$, so that $A \approx zA_*$: the solution is too small by a factor of z. It

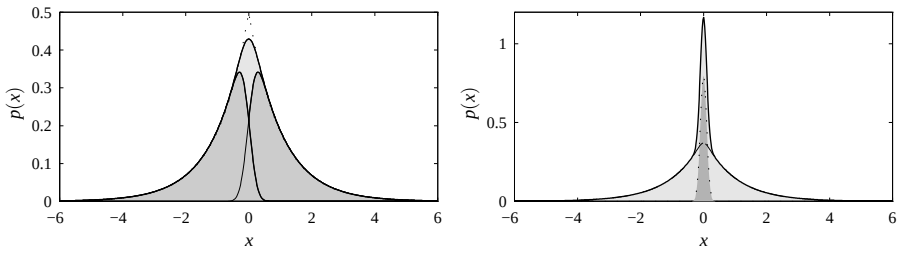


Fig. 6.11: (Left) An example of how a "noisy Laplacian" distribution may be built up from two "noisy exponentials." The dotted line shows the original Laplacian. (Right) A noisy sparsified Laplacian distribution with $z = 0.8$. (The sparsity was set low to keep the central peak from visually overwhelming the plot.)

also means that, for a given λ, it is possible to destabilise that solution by decreasing z to the point where the maximal value of $\mathcal{G}$ is 1 (see fig. 6.10). If the critical point is characterised by the parameters $(\lambda_c, \theta_c, z_c)$, then

$$\lambda_c = z_c/e, \qquad \theta_c = e/z_c, \tag{6.44}$$

or in terms of the original parameters,

$$\sigma_c = A_* \sqrt{z_c/e}, \qquad A_c = A_* z_c/e. \tag{6.45}$$

This fixes the maximum (model) noise level σ for a given level of sparsity, or alternatively, the maximum sparsity for a given noise level. Beyond this range, the only fixed point is at $A = 0$.

Adding input noise

Next, we consider the case where the data really is noisy. The following definitions will help to make subsequent expressions more concise:

$$\mathrm{gss}(t) \stackrel{\mathrm{def}}{=} \tfrac{1}{\sqrt{2\pi}} e^{-t^2/2}, \qquad \mathrm{erf}(t) \stackrel{\mathrm{def}}{=} \int_t^\infty \mathrm{gss}(u)\ \mathrm{d}u. \tag{6.46}$$

Taking the original Laplacian input and adding Gaussian noise of variance σ_*^2 results in a 'noisy Laplacian' probability density given by the convolution of the two original densities. The convolution evaluates to

$$p(x) = \frac{1}{2A_*}\left[\xi\left(\frac{x}{A_*}, \frac{\sigma_*}{A_*}\right) + \xi\left(-\frac{x}{A_*}, \frac{\sigma_*}{A_*}\right)\right], \tag{6.47}$$

where $\xi(u, v)$ is essentially a 'noisy exponential' distribution given by

$$\xi(u, v) = e^{v^2/2} e^{-u} \mathrm{erf}(v - u/v). \tag{6.48}$$

If, instead, noise is added to a sparsified Laplacian specified in eq. 6.42, there will be an extra Gaussian term due to the delta distribution at zero:

$$p(x) = \frac{z}{2A_*}\left[\xi\left(\frac{x}{A_*}, \frac{\sigma_*}{A_*}\right) + \xi\left(-\frac{x}{A_*}, \frac{\sigma_*}{A_*}\right)\right] + \frac{1-z}{\sigma_*}\mathrm{gss}\left(\frac{x}{\sigma_*}\right). \tag{6.49}$$

This eventually yields

$$\begin{aligned}\mathcal{G} = \frac{A_*}{A}\Bigg\{ z\left[\xi\left(\frac{\sigma^2}{AA_*},\frac{\sigma_*}{A_*}\right)+\xi\left(-\frac{\sigma^2}{AA_*},\frac{\sigma_*}{A_*}\right)\right] \\ +2\frac{\sigma_*}{A_*}\,\mathrm{gss}\left(\frac{\sigma^2}{\sigma_* A}\right)-2\frac{\sigma^2}{AA_*}\,\mathrm{erf}\left(\frac{\sigma^2}{\sigma_* A_*}\right)\Bigg\}\end{aligned} \tag{6.50}$$

which can be written in terms of the dimensionless parameters

$$\theta = A_*/A, \qquad \lambda = \sigma^2/A_*^2, \qquad \nu = \sigma_*/A_*, \tag{6.51}$$

as

$$\mathcal{G} = \theta\left\{ z[\xi(\lambda\theta,\nu)+\xi(-\lambda\theta,\nu)] + 2\nu\left[\mathrm{gss}\left(\tfrac{\lambda\theta}{\nu}\right)-\tfrac{\lambda\theta}{\nu}\,\mathrm{erf}\left(\tfrac{\lambda\theta}{\nu}\right)\right]\right\}. \tag{6.52}$$

This expression has several properties:

- The overall form is that of a function of $\lambda\theta$ scaled by $1/\lambda$ (*i.e.* $\lambda^{-1}F(\lambda\theta)$), which means that, as λ is varied, the graph of the function scales in both directions, maintaining its shape and proportions.
- The first term inside the large braces has exactly the form of a noisy Laplacian. This is to be compared with the factor of $e^{-\lambda\theta}$ in eq. 6.43.
- The second term in the large braces is similarly constructed such that (for fixed λ) varying ν results in simultaneous 'horizontal' and 'vertical' scaling in proportion to ν (which measures the actual 'noisiness' of the data.)

One way to visualise the dynamics implied by eq. 6.52 is to graph $\lambda\mathcal{G}$ as a function of $\lambda\theta$ for different values of ν and z. A few examples are shown in fig. 6.12. In each graph, the fixed points of the dynamics for different values of λ (the model noise level as a dimensionless parameter) can be found by intersecting with a horizontal line at λ.

Several conclusions can be drawn from these graphs, which should generalise fairly directly to the multidimensional case:

- Increasing the sparsity of the input makes the basis vectors smaller in proportion to z. (This is exactly what it seen in experiments.) It may also cause the non-zero solutions to disappear, which explains why in experiments, the weights decay to zero when the sparsity is increased beyond a certain point.
- Increasing the model noise parameter σ makes it more likely that the input signal falls under the expected noise level: the outputs will remain, for the most part, clamped to zero and the basis vectors will again fall to zero.
- If the actual noise is greater than expected ($\sigma_* > \sigma$), the scale parameter A may converge to a value that reflects the input noise variance rather than the data. In the multidimensional case, it seems likely that the resulting basis vectors will fit the noise distribution rather than the data distribution, and hence the results will not be useful.

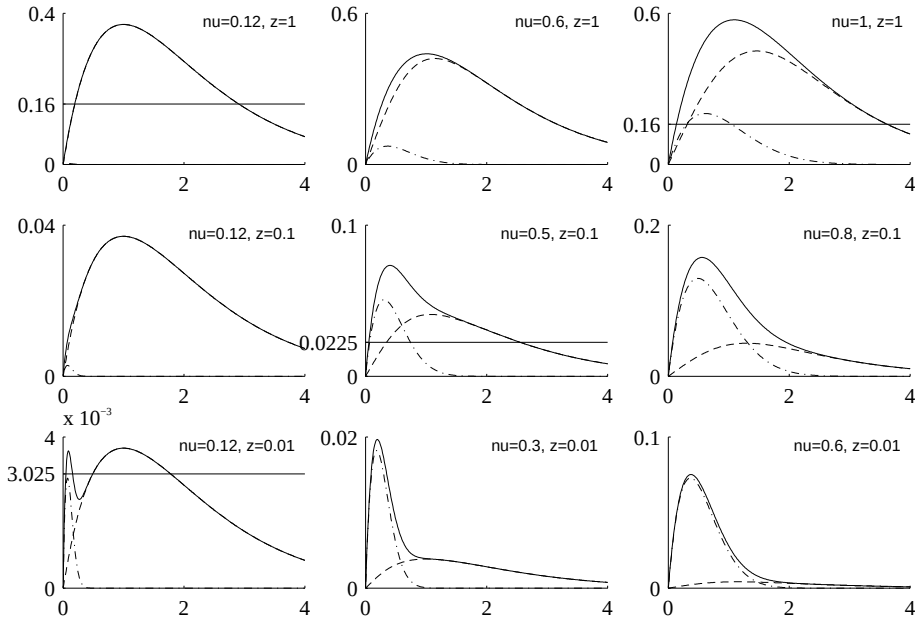


Fig. 6.12: Plots of $\lambda\mathcal{G}$ against $\lambda\theta$ for different values of ν and z. The dashed lines represent the first major term in eq. 6.52, which scales vertically with z. The dash-dot lines represent the second term (due to noise) which scales horizontally with ν and vertically with ν^2. The solid line is the sum and gives the overall behaviour. On some of the plots, the horizontal lines give the fixed points for a certain value of λ. It is possible that in certain situations, (bottom left) there can be *two* stable non-zero fixed points for A.

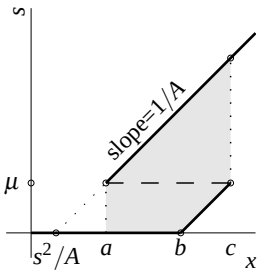
6.5.2 Using the modified prior

As we have seen, one of the effects predicted by the analysis above is that increasing sparsity tends to result in shorter basis vectors, essentially because of a mismatch between the Laplacian prior and the actual source distributions. This effect is amply demonstrated in the results in fig. 7.2 (in the next chapter): the various lengths of the basis vectors reflect the different probabilities of note occurrences.

The modified prior introduced in §6.2 was intended to address this problem, by making a piece-wise exponential approximation to a sparsified Laplacian density, so it may be enlightening to analyse the dynamics of learning using this prior. The gradient of the log-prior is of the form

$$\gamma(s) = \begin{cases} \operatorname{sgn} s & : |s| \geq \mu, \\ \alpha \operatorname{sgn} s & : |s| < \mu. \end{cases} \tag{6.53}$$

with $\alpha > 1, \mu > 0$. The yields a new shrinkage function which is still piece-wise linear, but with more segments (see fig. 6.13). A more serious problem is that the activation dynamics have more than one fixed point, meaning that the shrinkage 'function' is multi-valued. We must decide which of the two branches to use. Referring to fig. 6.13, a point q could be chosen, somewhere between a and c, to jump between the lower



$$a = \frac{\sigma^2}{A} + \mu A,$$
$$b = \alpha \frac{\sigma^2}{A},$$
$$c = \alpha \frac{\sigma^2}{A} + \mu A.$$

Fig. 6.13: Shrinkage function for modified prior. The thick solid lines (two disconnected branches) indicate stable solutions of the activation dynamics, whereas the dashed line indicates unstable fixed points. The grey area indicates where there are two stable solutions available for a particular value of x. (Note that both $b > a$ and $b < a$ are possible.

and upper branches. Letting $\Gamma_-(x)$, (valid for $0 \le x \le c$) denote the lower branch, and $\Gamma_+(x)$, (valid for $x \ge a$) the upper, then

$$\Gamma(x) = \begin{cases} \Gamma_-(x) & \text{for } |x| \le q, \\ \Gamma_+(x) & \text{for } |x| > q, \end{cases} \qquad a \le q \le c. \tag{6.54}$$

However, the issue of choosing a particular point q can be avoided with the following argument: the problem with the un-sparsified Laplacian version was that $\mathcal{G}$ was too small at $A \approx A_*$, the stable fixed point being at $A \approx zA_*$. The modification to the prior is intended to *increase* the value of $\mathcal{G}$ near $A = A_*$. If we can find an upper bound on the new $\mathcal{G}$ and show that it is still insufficient to make up for the sparsity of the input and stabilise a solution $A \approx A_*$, we could conclude that the modified prior does not have the desired effect. Hence, we will integrate over *both* branches of the shrinkage function, this providing an upper bound on the integral regardless of where the discontinuity is placed. Since $p(x)$ is an even function,

$$\begin{aligned} \int_{-\infty}^{\infty} \Gamma(x)p(x)\,\mathrm{d}x &= 2\int_0^{\infty} \Gamma(x)p(x)\,\mathrm{d}x \\ &= 2\int_0^{q} \Gamma_-(x)p(x)dx + 2\int_q^{\infty} \Gamma_+(x)p(x)\;\mathrm{d}x \\ &\le 2\int_b^{c} \Gamma_-(x)p(x)dx + 2\int_a^{\infty} \Gamma_+(x)p(x)\;\mathrm{d}x \end{aligned}$$

regardless of where q lies between a and c.

Now, it would be possible to carry out the integration using a *noisy* sparsified Laplacian input, as done in the previous section, but, in the multidimensional case, any solutions made stable by the addition of noise would most likely reflect the structure of the noise distribution rather than that of the signal. In view of this, the dynamics of the modified system will be derived using a *noiseless* sparsified Laplacian input only. The result is

$$\mathcal{G} \le z\theta e^{-\lambda\theta}\left\{\alpha - (\alpha e^{-\lambda\theta(\alpha-1)} - 1)(1+\mu/\theta)e^{-\mu/\theta}\right\} \tag{6.55}$$

It is easier to see what is going on here if we let $\lambda \to 0$, which corresponds to a low (model) noise regime. (With the Laplacian prior, an increase in λ tended to *reduce* the

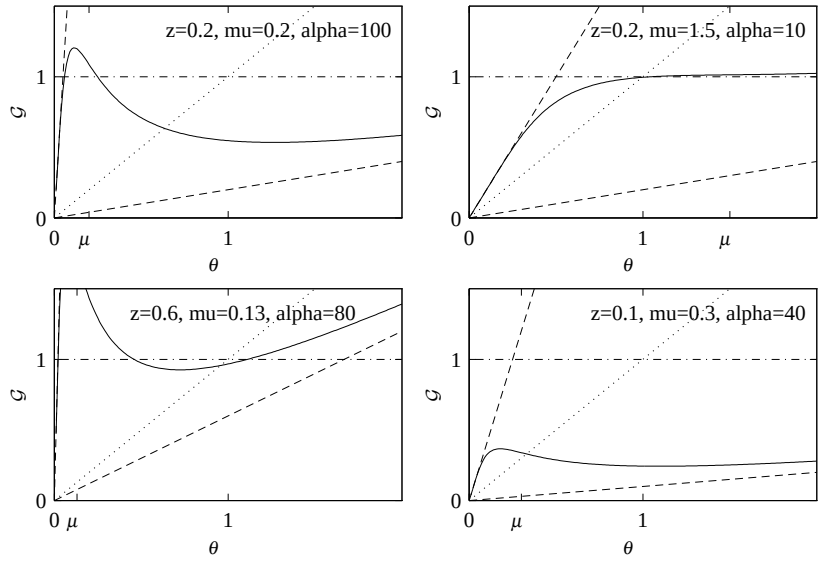


Fig. 6.14: Learning dynamics for the modified prior: various plots (solid lines) of the upper bound on $\mathcal{G}$ described in eq. 6.56. On each plot, the two dashed lines indicate the lines $\mathcal{G} = \alpha z\theta$ and $\mathcal{G} = z\theta$, to which the solid curve tends as $\theta \to 0$ and $\theta \to \infty$ respectively. The dotted line $\mathcal{G} = \theta$ shows where the desired fixed point at $\theta = 1$ lies, at its intersection with $\mathcal{G} = 1$.

value of $\mathcal{G}$ at all points, and thus setting $\lambda = 0$ represents a 'best-case scenario.') We obtain

$$\mathcal{G} \leq z\theta \left\{ \alpha - (\alpha - 1)(1 + \mu/\theta) e^{-\mu/\theta} \right\} \tag{6.56}$$

The factor in braces is approximately α for small θ, tending to 1 for large θ, with a cross-over scale proportional to μ. Hence, the graph of this upper bound (see fig. 6.14) makes a transition from $z\alpha\theta$ to $z\theta$.

The shape of the curve depends on z and α, with z changing only the vertical scaling. The overall size (with fixed proportions) then varies with μ. Experimenting with different combinations of parameters suggests that it is not possible to produce a stable fixed point near $\theta = 1$ unless one of the asymptotes $\mathcal{G} = z\theta$ or $\mathcal{G} = z\alpha\theta$ has a slope of approximately 1, that is, either $z \approx 1$ or $z\alpha \approx 1$. The former case corresponds to non-sparse input and defeats the purpose of introducing the modified prior. The latter requires that μ be of the order of 1 or larger, which means the prior becomes essentially a Laplacian of width z rather than 1. This contradicts the notions of sparsity discussed in § 6.1.1.

Thus, it appears that the modified prior is actually a rather poor approximation to a sparsified Laplacian prior and does not solve the problem of incorrect scaling of the basis and convergence to zero when the data is very sparse. In the true sparsified density, the peak at zero is either infinitely thin or broadened by noise, and cannot provide any information about the scaling of the basis vectors.

6.5.3 Comparison with exact learning

The learning rule we have been using is derived from an approximation to the likelihood $p(x|A)$, and, as we have seen, suffers from a propensity to produce zero length basis vectors under some conditions. To gain a better understanding, the approximation may be compared with the exact likelihood computed by the integral

$$p(x|A) = \int_{-\infty}^{\infty} p(x|A,s)p(s)\,\mathrm{d}s, \tag{6.57}$$

which is tractable in the one-dimensional case.

There are four relevant quantities to investigate: the log-likelihood, $\log p(x|A)$; the objective function $\mathcal{L} = \mathrm{E}\log p(x|A)$; the gradient of the log-likelihood $\partial \log p(x|A)/\partial A$, which provides the stochastic learning signal, and the gradient of the objective function $\partial\mathcal{L}/\partial A = \mathrm{E}\,\partial \log p(x|A)/\partial A$. In each case, the approximation may be compared with the theoretically correct value.

The exact value of $p(x|A)$ is actually the distribution of x produced by the system when treated as a generative model, and is none other than the noisy Laplacian density stated earlier (eq. 6.47). Thus, we can immediately write

$$\log p(x|A) = \log\left[\xi\left(\tfrac{x}{A},\tfrac{\sigma}{A}\right) + \xi\left(-\tfrac{x}{A},\tfrac{\sigma}{A}\right)\right] - \log 2A, \tag{6.58}$$

using the earlier definition of $\xi(u,v)$ (see eq. 6.48).

Lewicki and Sejnowski's (2000) Gaussian approximation to the integral in eq. 6.4 gives, in one dimension,

$$\log p(x|A) \approx \log p(\hat{s}) - \frac{(x-A\hat{s})^2}{2\sigma^2} - \tfrac{1}{2}\log\left|\frac{A^2}{\sigma^2} - \frac{\partial^2 \log p(\hat{s})}{\partial s^2}\right| - \log\sigma. \tag{6.59}$$

Using a Laplacian prior and eq. 6.36 for $\hat{s}$ in terms of x yields

$$-\log p(x|A) \approx \log 2A + \tfrac{1}{2}\log\left|1 + \frac{\sigma^2}{A^2}\frac{\mathrm{d}\gamma(\hat{s})}{\mathrm{d}s}\right| + \begin{cases} \dfrac{x^2}{2\sigma^2}, & |x| < \frac{\sigma^2}{A}, \\ \dfrac{|x| - \frac{\sigma^2}{A}}{A}, & |x| \geq \frac{\sigma^2}{A}. \end{cases} \tag{6.60}$$

At this point a problem emerges: for a sharply peaked prior like a Laplacian, $\gamma(s)$ is discontinuous at $s = 0$, and therefore $\mathrm{d}\gamma(\hat{s})/\mathrm{d}s$ is undefined at zero, highlighting the fact that the Gaussian approximation is a poor one at the sharp peak of the Laplacian density. The learning rule we have been analysing is the result of ignoring the offending term, since for a Laplacian, $\mathrm{d}\gamma(\hat{s})/\mathrm{d}s = 0$ for $s \neq 0$. The remainder of the analysis neglects the term $\mathrm{d}\gamma(\hat{s})/\mathrm{d}s$, but preliminary investigations suggest that, when steps are taken to approximate the effect of the missing term, some of the problems with the approximation are alleviated.

The exact and approximate log-densities are illustrated in fig. 6.15. The approximation consists of two linear pieces and a quadratic piece, whereas the exact version has a shallower curve near $x = 0$.

The next quantity of interest is the gradient of the log-likelihood, $\partial \log p(x|A)/\partial A$; it is this that provides the 'learning signal', which when averaged over x, drives the

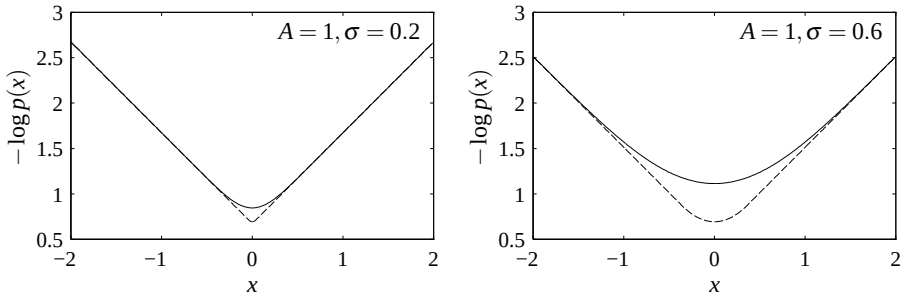


Fig. 6.15: Noisy Laplacian log density $-\log p(x)$ for $A = 1$ and two noise levels, $\sigma = 0.2$ and $\sigma = 0.6$. Solid line: exact version (eq. 6.58), dashed line: approximation (eq. 6.60).

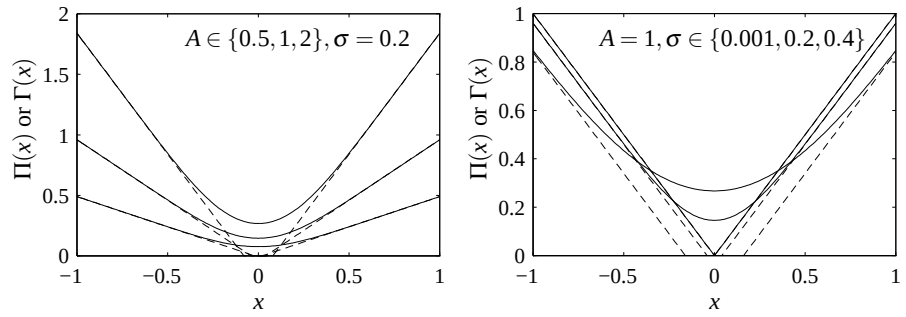


Fig. 6.16: Comparison of the quantity $\Gamma(x)$ (dashed lines) used in the approximate learning rule with the equivalent quantity in the exact derivation, $\Pi(x/A, \sigma/A)$ (solid lines). The left-hand plot shows the variation with A for fixed σ. The right-hand plot shows the variation with σ for fixed A.

overall learning behaviour. Exact differentiation of eq. 6.58 gives:

$$\frac{\partial \log p(x|A)}{\partial A} = \frac{\Pi(\frac{x}{A}, \frac{\sigma}{A}) - 1}{A}, \tag{6.61}$$

where

$$\Pi(u, v) = \frac{u\,[\xi(u, v) - \xi(-u, v)] + 2v\,\mathrm{gss}(u/v)}{\xi(u, v) + \xi(-u, v)} - v^2. \tag{6.62}$$

This should be compared with the approximate version of this obtained by differentiation of eq. 6.60 (ignoring the term in $\mathrm{d}\gamma(\hat{s})/\mathrm{d}s$:

$$\frac{\partial \log p(x|A)}{\partial A} = \frac{|\hat{s}| - 1}{A} = \frac{\Gamma(x) - 1}{A} \tag{6.63}$$

using the earlier definition of $\Gamma(x)$. Thus we may tentatively identify $\Gamma(x)$ as an approximation to $\Pi(\frac{x}{A}, \frac{\sigma}{A})$; the two are compared over a range of parameter values in fig. 6.16. It is clear that the fit is poor when σ is large.

The final step is to compare the exact and approximate versions of the objective function $\mathcal{L}$ and its derivative with respect to A, which are obtained by averaging the

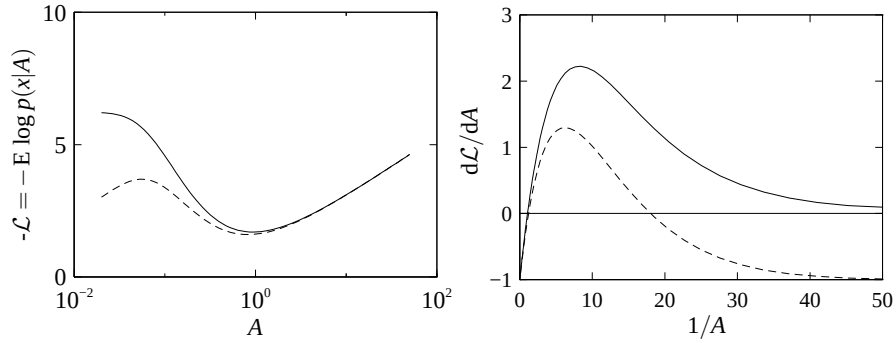


Fig. 6.17: On the left: exact (solid line) and approximate (dashed line) objective functions for $A_* = 1$ and $\sigma = 0.4$. Both version have optima near $A = A_*$, but the approximation also has an unbounded minimum at $A = 0$. On the right, the derivative $\mathrm{d}/\mathrm{d}\mathcal{L}A$ plotted against $1/A$.

log-likelihood $\log p(x|A)$ and its derivative over the observed distribution of x, which will be assumed to be a noiseless Laplacian of width A_* (see eq. 6.38).

The true objective function can be computed by numerical integration, whereas the approximate version can be computed in closed form, yielding

$$-\mathcal{L} \approx \frac{A_*^2}{\sigma^2}\left[1 - \exp -\frac{\sigma^2}{AA_*}\right] + \log 2A. \tag{6.64}$$

This and the numerical integration of the exact objective function are illustrated in fig. 6.17, along with their derivatives with respect to A. The true cost function has a single optimum near $A = A_*$, as one would expect, but the approximation has another at $A = 0$ because of the $\log 2A$ term in eq. 6.64.

Summary and Conclusions

In this chapter, the theoretical aspects of sparse coding were considered. A causal generative model was adopted as the basis for the system, where a sparse code was defined as a *factorial code* in which the marginal distributions of the components are 'sparse,' that is 'strongly and tightly peaked' around zero. A consideration of inference in the model and the energy landscape of the posterior density $p(\mathbf{s}|\mathbf{x},\mathbf{A})$ lead to the conclusion that the Laplacian prior forms a natural watershed between sparse and non-sparse distributions, though, of course, Gaussianity and non-Gaussianity still have an important role to play in ICA related methods, of which sparse coding is an example.

An attempt was made to construct a sparse prior by modifying the Laplacian distribution to add a tight peak at zero. When combined with a modification to the gradient optimisation procedure used to perform MAP inference of the sparse components $\mathbf{s}$, this resulted in some performance gains, but at the expense of poor learning in some circumstances. A more detailed analysis suggested that the modified prior did not have the desired effect and did not successfully model a 'sparsified' Laplacian, that is a

Laplacian density plus a delta distribution at zero, Though the piece-wise Laplacian density may appear visually to be a reasonable approximation, it is actually a poor one for this application.

The analysis (restricted to a simple one-dimensional version of the system) did however explain some of the behaviour of the learning algorithm in situations where the actual data distribution does not match the prior.

The modified optimiser itself was shown to be a useful tool for optimising cost functions with gradient discontinuities at coordinate zero.

One of the more interesting conclusions to be drawn from the experiment with the bars problem is that a certain amount of sparse structure is required in the input if the system is to learn a highly overcomplete basis. If this is not the case, the system learns only enough basis vectors to span the input space.

7. SPARSE CODING OF MUSIC SPECTROGRAMS

Introduction

In this chapter, the methods developed in the previous one will be applied to spectrograms of musical sounds. In a spectrogram, such as the one illustrated in fig. 7.3, each note appears as a regular pattern of harmonics. This structure represents a great deal of redundancy, which a sparse coder may be able to detect and encode.

7.1 Statistical Structure of Spectra

Previous investigations in to the statistical structure of natural sounds have shown that, under several commonly used representations, they have extremely non-Gaussian probability distributions. Attias and Schreiner (1997) analysed speech, music, animal vocalisations, and environmental sounds, using linear band-pass filters of varying widths, and found that, under a wide range of conditions, histograms of the instantaneous amplitude of the filtered audio signals were far from Gaussian. Iordanov and Penev (1999), working from television broadcasts, measured second-order statistics of vectors of waveform samples, from which the principal components of each auditory ensemble were computed. They found that the coefficients of the principal components, that is, the coordinates of the vectors relative to a basis formed by the principal components, had extremely tightly-peaked, heavy-tailed marginal distributions. Moreover, the principal components were essentially sinusoidal, so the coefficients represented a sort of spectral analysis.

Heavy-tailed non-Gaussian distributions can also be observed in the marginal histograms of individual spectral bands in magnitude or power spectrogram (see fig. 7.1): the distributions are strongly peaked toward zero, with tails that decay much more slowly than a Gaussian, which would appear on a semi-logarithmic scale as a parabola curving downwards. In fact, in many cases, they decay more slowly than a Laplacian, which would appear as a straight line of negative gradient.

In fig. 7.1(d), derived from a signal normalised in the same way as the signals used in Chapter 5 (seen Appendix A for details) some further structure is apparent. The distributions appear to fall into two classes, one of which is more sparsely distributed than the other. A preliminary analysis indicates that, as we go up in frequency, the distributions alternate between the two classes at 12 cycles per octave, suggesting that the phenomenon is due the frequencies being in or out of tune with the 12 semitones-

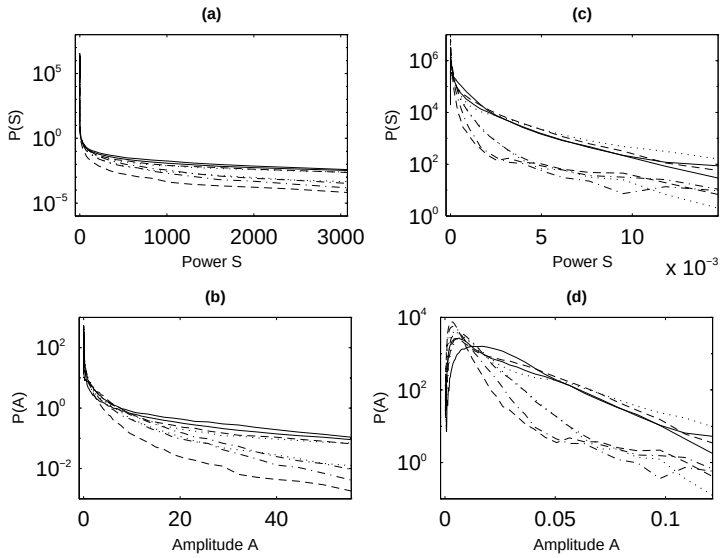


Fig. 7.1: Marginal densities of spectral bands. Each line type indicates the distribution (over time) of activity at a particular frequency. (a) and (b) were taken directly from power (S) and magnitude (A) spectrograms respectively (with $A = \sqrt{S}$), whereas in (c) and (d), the audio signal was fed through a long time-constant automatic gain control, to remove large (time) scale fluctuations in loudness. Even with these variations removed, the densities can still be considered to be quite sparse. This particular signal was take from an ordinary audio CD (the Penguin Cafe Orchestra's *Signs of Life*), but qualitatively similar results were obtained from a number of CDs representing a range of musical styles.

per-octave chromatic scale.

The Use of Spectra as Input Data

The prime motivation for using spectra as input data was that the phase invariance of the spectrogram would enable the system to detect correlations between the different frequency components of a harmonic tone even if the phase relationships were not consistent. The results of Chapter 5 showed that ICA using linear input is not capable of detecting these dependencies; very few multi-component basis vectors were found in the music derived basis. Indeed, the results of Chapter 8 will show that there are dependencies between harmonic components, but these are of a form that cannot be detected by ICA (or sparse coding) in a linear input representation.

These are some of the factors to consider when using spectrograms:

Noise statistics The sparse coder generative model assumes the presence of additive Gaussian noise. However, Gaussian noise in the signal does not appear as Gaussian noise in a power spectrogram. In absence of any other signal, it actually appears as random noise with a (single sided) *exponential* distribution, of the form $p(u) = \lambda e^{-\lambda u}, u \geq 0$. (See Appendix B for details.) If, however, the *magnitude*

spectrogram is used, by taking the square root of the power spectrogram, that noise appears with a distribution of the form $p(v) = 2\lambda v e^{-\lambda v^2}$, where $v = \sqrt{u}$. As well as noise in the original signal, the spectrogram also contains noise-like activity due to spectral leakage. Anderson (1997) describes how this is related to an interaction between windowed harmonic components, and is not phase invariant: thus, phase 'information' appears as spectral 'noise.'

Linearity or otherwise of mixing The sparse coders described the previous chapter assumed a linear generative model, which means that it will theoretically be capable of disentangling the independent causes of the input *if* these causes combine additively. Power spectrograms *do* add linearly if the sources are phase-incoherent, but we do not expect this to be the case for pitched musical instruments that are in tune with each other. If harmonics from two instruments overlap (which they will if the notes are consonant) then we may get constructive, destructive or beating interference, which effects are linear in the time domain (and in the linear Fourier transform) but nonlinear in the spectrogram.

Phase invariance The concept of phase invariance was mentioned in § 3.3.2 and will be discussed further in § 8.3. Some degree of phase invariance seems to be a desirable property for an artificial auditory system and is useful for this particular application, as indicated above; the approximate phase invariance of the spectrogram is therefore an attractive feature.

7.2 Results

The training data was generated from a MIDI encoded piece of music, which was synthesised using an ordinary PC sound card. The synthesis was carried all the way to an analogue signal, which was then resampled using the sound card. Bach's Partita in A minor for keyboard, BWV827, was chosen because it consists mainly of two or three relatively independent lines with few block chords. The hope was that the 'independent components' would turn out to be the notes themselves. A number of pieces in a similar style were also available for testing.

Magnitude spectra were used rather than power spectra in order that the noise be more approximately Gaussian and to reduce the super-Gaussianity of the marginals of spectral bands, as illustrated in fig. 7.1. The modified quasi-Newton optimiser described in the previous chapter was used with the modified Laplacian prior.

7.2.1 Basis vectors

Starting with 96 basis vectors, the basis converged to 55 non-zero vectors, of which 49 appeared to be harmonic note spectra matching the notes in the piece. In order to verify this, the spectral profiles were used to filter white noise. Though the resulting sounds were not realistic reproductions of the harpsichord sound, many pitches were discernible.

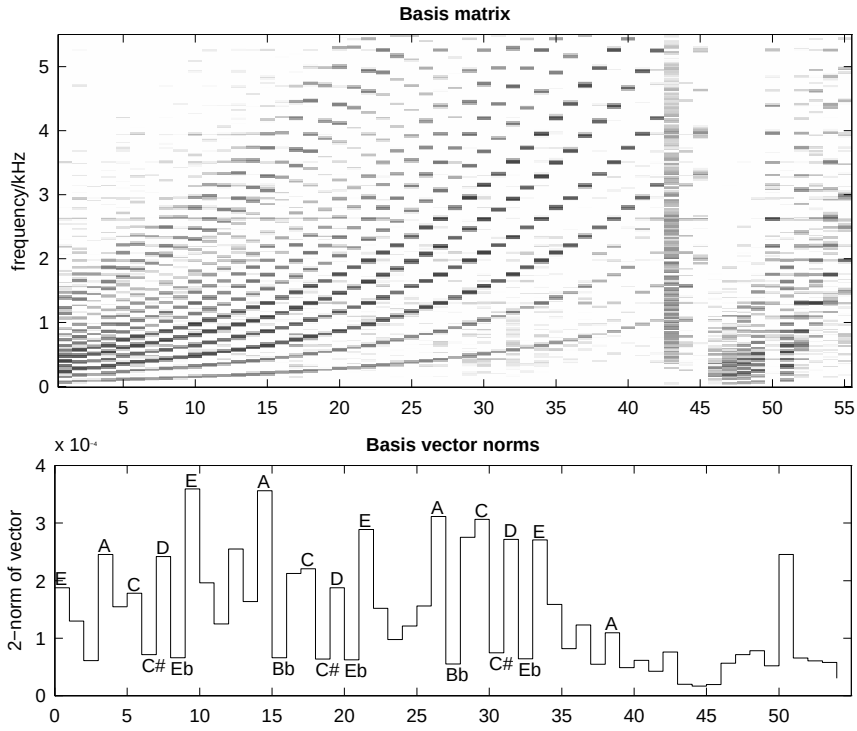


Fig. 7.2: The 55 non-zero basis vectors obtained from sparse coding synthetic harpsichord music. The lower plot show the lengths (2-norms) of the basis vectors, showing that the basis vectors corresponding to the most structurally important pitches in the key (which is A minor) had the largest norms.

On the basis of the reconstructed sounds, the basis was manually reordered by pitch. The result is illustrated in fig. 7.2. The more common notes are represented by longer basis vectors, in agreement with the findings of the previous chapter. The key of the piece (A minor) is plainly visible in the pattern of basis vector norms in fig. 7.2.

7.2.2 Conversion to MIDI and Resynthesis

Once the system was trained, the reordered outputs provided a quite faithful 'piano-roll' transcription of the music, some extracts of which are illustrated in fig. 7.3 and fig. 7.4. Bearing in mind the limited extent of the experiment, it was felt that a rigorous validation of the transcription was not appropriate at this point, and an aural appraisal of the results would be sufficient. To this end, a simple threshold-based MIDI note trigger was added to the system, which, thanks to the modified prior and optimiser, was able to operate in real time. After some trial and error adjustments of the prior parameters μ and α, this resulted in a passable rendition of the original piece, perhaps by a rhythmically-challenged piano student!

Whilst encouraging, this did point out the sensitivity of the procedure to the parameters of the prior. A large α tended to produce a 'cleaner' output, but with some

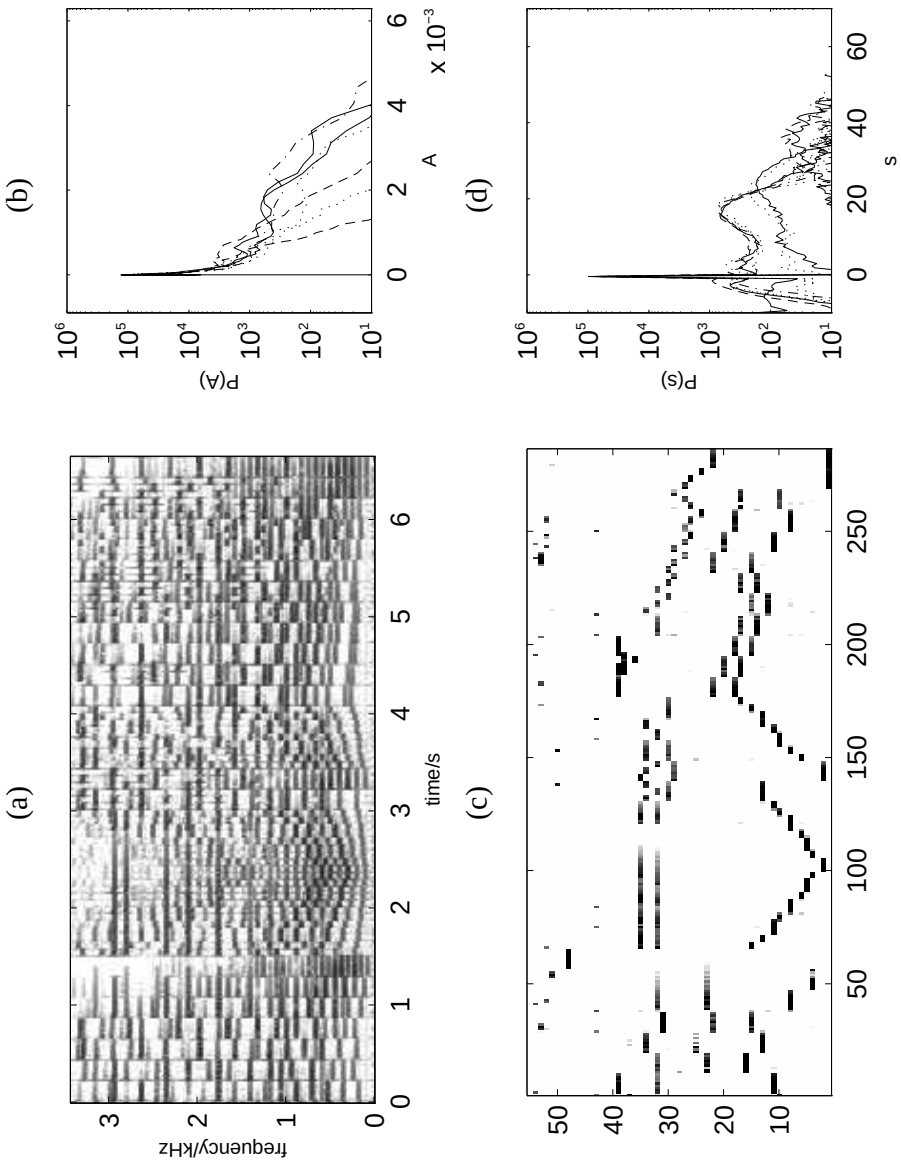


Fig. 7.3: Top: (a) Input spectrogram and (b) histograms of activity at a selection of frequencies. Bottom: (c) output trace and (d) histograms of output components. The output is visibly 'sparser' than the input. The output histograms also show some interesting structure: their bimodality reflects the on/off nature of the notes present in the music.

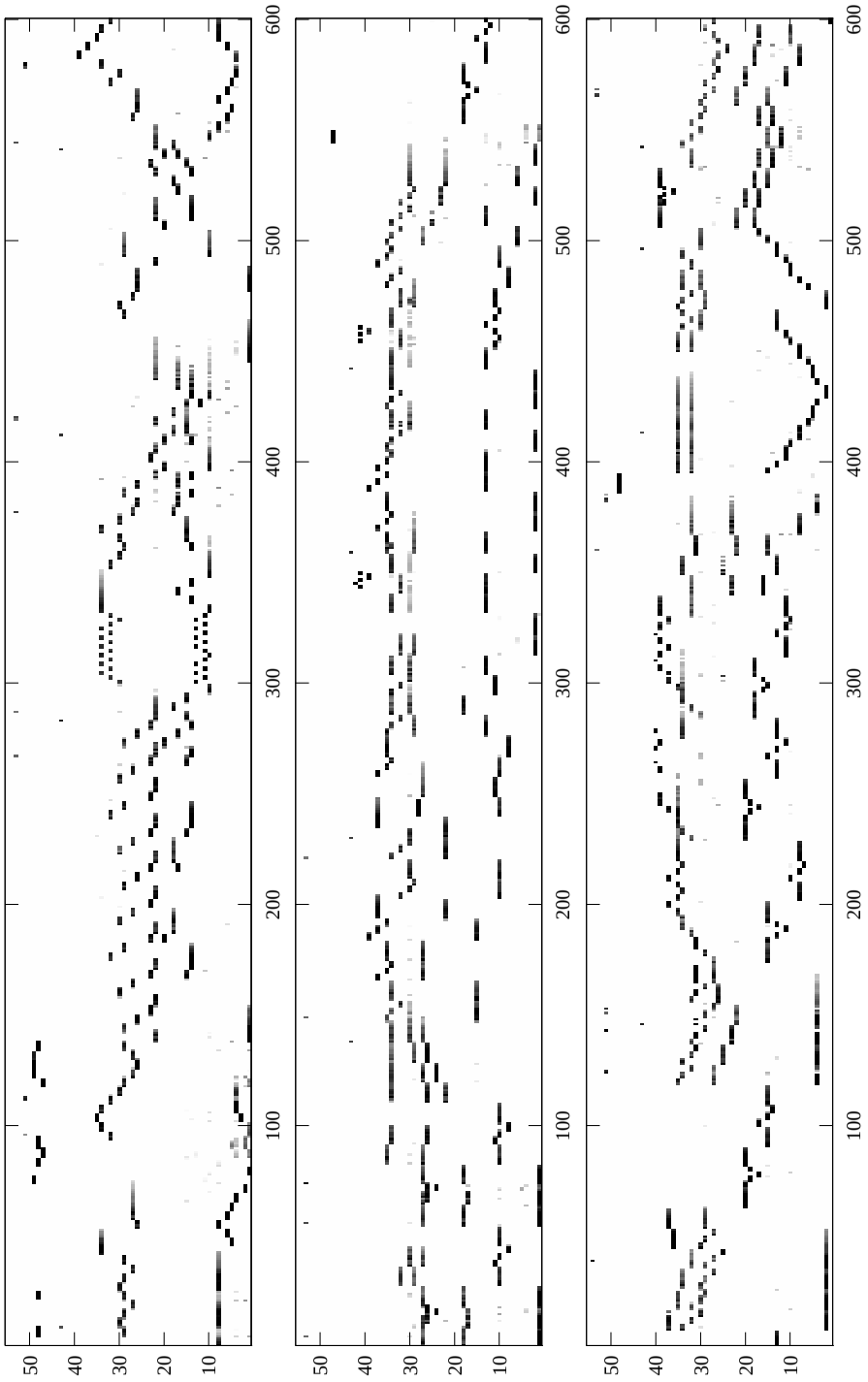


Fig. 7.4: A longer extract of 'piano-roll' output.

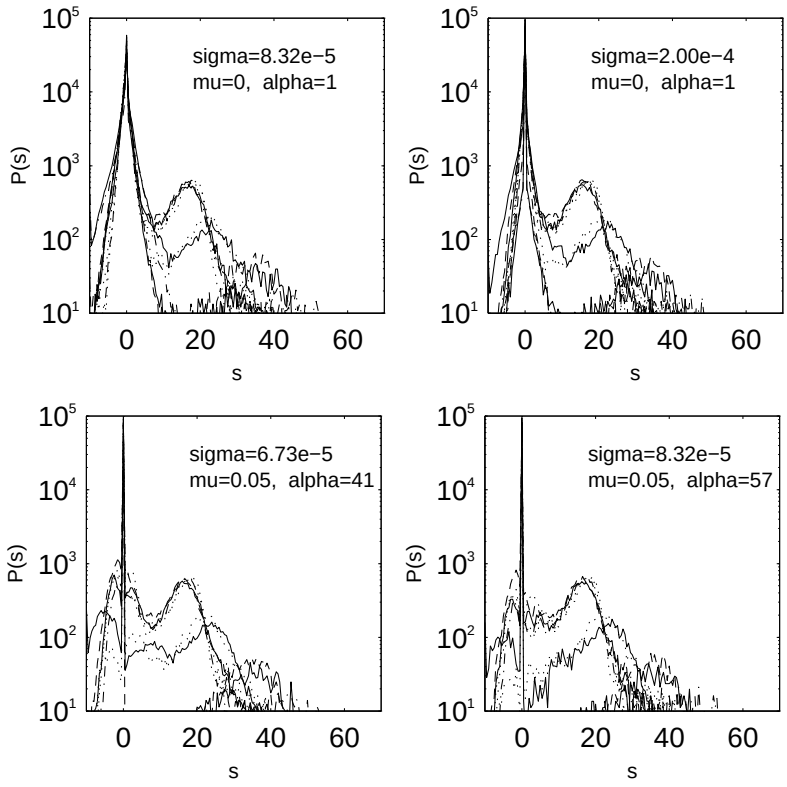


Fig. 7.5: Output marginals for various parameter settings. Each line shows the distribution of activity of a single output unit. See text for discussion.

missing notes. A small α recovered the missing notes, but also allowed many small activations to triggering notes not present in the music. Imposing a threshold of activity for triggering notes improved matters, but seemed a little inelegant, especially after the discussion of § 6.2 about the desire for exact zeros to signal inactivity.

7.2.3 Output component distributions

The marginal histograms of the output components (see fig. 7.3(d) and fig. 7.5) revealed a definite tendency to bimodality, suggesting (accurately in this case) an underlying binary process, namely, the presence or absence of a note. In this case, the use of a fixed threshold for note activation in the MIDI resynthesis is probably not optimal since ideally the threshold would be positioned in the valley between the two modes. An unsupervised procedure to position this threshold would therefore be a useful development of the system.

The distributions in fig. 7.5 also demonstrate the effects of the noise parameter σ and the sparsified Laplacian parameters μ and α (defined in eq. 6.25). The peak at zero tends to be more concentrated both when σ is large and when α is large, though the effect is visibly different in the two cases.

7.2.4 Learning about temporal structure

One of the system's weaknesses thus far is a certain lack of temporal coherence in the output, since the assessment of which notes should be active is made on a moment-by-moment basis. Noise-like activity in the spectrogram causes long notes occasionally to be broken up, producing superfluous onsets, and the introduction of false notes that are not present in the music. These errors are easy for a human observer to spot since we are able to see the whole spectrogram, not just a one-pixel wide strip.

One approach to dealing with this is to expand this time window so that that the sparse coder learns time-frequency basis vectors rather than just spectral profiles. More generally, notes of different instruments have distinctive shapes in the time-frequency plane, including vibrato and frequency glides. For example, the onsets of some wind instruments are often played flat and then brought up to the correct pitch. An algorithm trained on two-dimensional spectrogram patches may thus be able to learn these shapes, and be more capable of discrimination between different instruments.

Preliminary results were obtained with spectrogram patches four pixels wide; these are illustrated in fig. 7.6.

The main drawback with this approach that the input vectors are much larger. For example, if the instantaneous spectra have 256 components, and we choose to train the coder on strips that are 32 points wide, then the input patterns will have 8192 elements. Since the optimisation procedure used to estimate the components is approximately $O(N^3)$ in the dimensionality of the system, this represents a considerable computational burden. (The computation of each step is $O(N^2)$, and it was observed that $O(N)$ steps were required to produce adequate convergence.)

7.3 Conclusions

The main conclusion of this work is that there *is* enough structure in music (or at least certain kinds of music) for a sparse coder to learn about and detect notes in an unsupervised way, even when the music is polyphonic. There is no need to bring any prior musical knowledge to the problem, such as the fact that musical notes have approximately harmonic spectra. This knowledge is acquired during learning. In addition, the MAP activation framework using a sparse prior does a reasonably good job of decomposing an input spectrum into a sum of notes.

The sparse coding model we used is essentially an extension of ICA, since it attempts to construct a sparse *factorial* code, so one might conclude that "notes are the independent components of music." Clearly, in a wider sense, the notes of a piece of music are *not* independent, but, to a first approximation, this assumption proves to be fruitful.

These are some of the more immediate ways in which the system could be improved, and which could form the basis of further work:

Adaptive thresholding of outputs An obvious development arising from the observation of bimodality in the output component distributions is that an adaptive thresh-

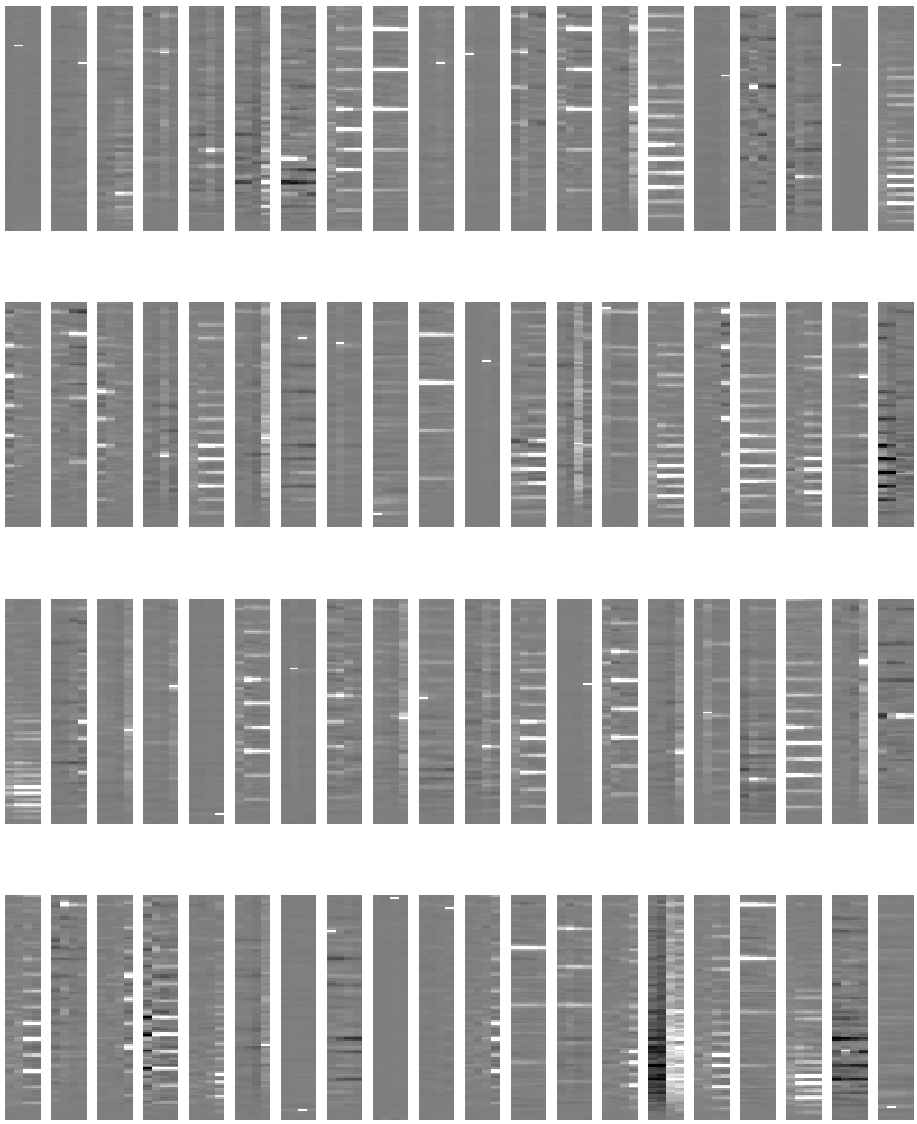


Fig. 7.6: Basis trained on 2-dimensional strips of spectrogram. Each strip is four pixels wide. Many of the basis vectors are clearly attuned to note onsets, and one (bottom row, sixth from right) seems to represent wide-band onset activity. (In these images, white is positive and black is negative.)

old of note activation should be implemented. It may be possible to use a binary mixture model (see § 3.2.4) to achieve this.

A better noise model The sparse coder implemented here assumes additive spherical Gaussian noise on the input ($\Lambda_{\mathbf{e}} = \sigma^{-2}\mathbf{I}$), which does not appear to be very accurate for audio spectrograms. Even if the original signal is noiseless, the spectrogram shows a background of noise-like activity, due to *spectral leakage* (Anderson, 1997), an artifact of the windowed Fourier transform applied to aperiodic signals. The activity at a particular frequency is related to the energy in the signal in the region of that frequency. Since there tends to be more energy at the low end, the 'noise' is more pronounced at low frequencies, and slopes off towards the high. This means that if the noise level σ is set high enough to prevent the system being distracted by high level noise at low frequencies, it fails to pick up significant low level details at high frequencies, where the actual noise level is lower. One option would be manually to set some noise distribution *a priori*. Another is to estimate the noise covariance matrix from the data.

Separable time-frequency basis vectors It was noted that extending the system to use larger input vectors extending over several spectrogram time slots would generally be prohibitively expensive. The situation is fundamentally different from that encountered in image coding, where the assumption of position invariance means that small, square windows can be drawn from anywhere in the image to provide training data. Sounds are generally not invariant to translation in frequency, and the lower part of the spectrogram does not 'look' like the upper part.

One possible solution would be to restrict the basis to *separable* 2D patterns, in which each time-frequency basis vector factorises as a product of a spectral profile and a temporal profile. These patterns could be learned separately: first, the spectral basis vectors would be learned as in the current system; then the temporal basis could be learned by training a sparse coder in the time signal of each output component, essentially producing a set of 'typical' note activation profiles. This might be expected to go a long way towards improving the temporal coherence of the notes produced for certain instrument types, though it will probably be unable to deal with non-separable time-frequency features such as vibrato.

More generally, this system in its current form is not proposed as a practical automatic transcription system. The experiments here have been confined to a very restricted type of synthetic instrument sounds, which have an artificially regular time-frequency structure. Real instruments show considerably more spectral variability which the current system would have trouble coping with. What this work does demonstrate, I hope, is the potential of an unsupervised approach, driven by considerations of redundancy reduction and efficient coding, to solving problems in music processing.

8. SIMILARITY

Introduction

Consider the experiments described in Chapters 5 and 7: in both cases, it was helpful to arrange or order the linear bases developed by ICA or sparse coding. In the sparse coder, it made sense to arrange the basis vectors according to pitch, whereas in the ICA system trained on speech (§5.3.1) it was clear that a two-dimensional arrangement in time and frequency was called for. Similarly, some well-known linear bases come supplied with an intrinsic ordering of their basis vectors. For example, given a Fourier basis, one would conventionally arrange the basis vectors in one-dimension in order of frequency. A wavelet basis, on the other hand, sits naturally in a two-dimensional space, with time along one axis and frequency along the other. The questions we ask in this chapter are, what is it about these bases and the data they represent that makes a given arrangement appropriate and meaningful? By what principle, if any, can these arrangements be discovered in a framework of unsupervised learning?

Hyvärinen et al. (2001) addressed these questions, and developed the method of *topographic ICA* (TICA). This fits an ICA basis into a given predetermined arrangement (such as a two-dimensional sheet or a closed loop) so that neighbouring basis vectors are related in a way to be described in §8.1. However, it does not *discover* the appropriate topology. The work in this chapter follows a slightly different approach which focuses not on predefined neighbourhood relationships, but on *distances* between representational units. The result is a geometric picture of the representation, from which the appropriate topology may, in principle, be deduced.

Two types of similarity Before continuing, it is important to make a clear distinction between the *two* types of similarity that will be discussed in this chapter. The first is similarity between individual elements in a distributed representation. The second is similarity between states of a representation. The topological relationships represented in TICA are of the first kind, whereas the similarity judgements of which we are consciously aware, those between observed patterns of stimulation such as auditory or visual scenes, are presumably of the second type, in that a particular scene is represented by a pattern of activity in the appropriate mental representation.

Unfortunately, and rather confusingly, the two may coincide for a certain class of representational map. In a ‘code-book’ quantiser, the representational units, only one of which is active at a time, correspond directly to observed patterns. It then becomes possible to assign distances between representational units to reflect directly

the distances between the patterns they represent.[1] In the general case, this will not be possible since the each unit encodes patterns only in cooperation with other units.

For the sake of brevity, from here on, similarity between representational units will be called R-similarity, and similarity between patterns will be called P-similarity. They can both be defined formally as follows. Let s stand for a complete sensory scene drawn from a set $\mathcal{S}$. P-similarity is then an expression of the topological and metrical structure of the set $\mathcal{S}$. In contrast, in order to define R-similarity we need to consider a particular distributed representation of elements of $\mathcal{S}$. Assume that to each $s \in \mathcal{S}$ there corresponds an indexed tuple $\xi \equiv (\xi_\alpha, \xi_\beta, \cdots)$, where the indices α, β *etc.* are elements of an indexing set $\mathcal{A}$. (In a linear representation, these indices can be the basis vectors themselves, and the ξ_α will be the coordinates with respect to that basis.) For each $\alpha \in \mathcal{A}$, let ξ_α be an element of Ξ_α, which may be any coordinate space, such as the real numbers, the non-negative reals, the binary set $\{0,1\}$ *etc.* This additional structure introduces two new aspects to consider. One is the topological and metrical structure that may be induced on the product space $\prod_{\alpha\in\mathcal{A}} \Xi_\alpha$ by any assumed metrics on the Ξ_α, the individual coordinate spaces.[2] The other is the possibility of topological or metrical structure on the indexing set $\mathcal{A}$; that is, α and β, both in $\mathcal{A}$, may be more or less similar. This is precisely what R-similarity is concerned with, and defined in this way, is clearly distinct from P-similarity.

Overview §8.1 deals with R-similarity: how it may be defined, and experimental procedures for visualising it. §8.2 is more discursive in nature and describes several possibilities for defining P-similarity in terms of the statistical structure of the set $\mathcal{S}$ in the definitions above. Finally, several of these strands are brought together in a discussion of the concept of phase invariance in §8.3. The general drift of this last section is towards possibilities for further processing after ICA or sparse coding.

8.1 Topographic and geometric representations

Topographic ICA was introduced by Hyvärinen et al. (2001), who began by observing that in ICA, the resulting components and the associated basis vectors have no particular ordering or similarity relationships defined. It would be possible to use ad-hoc criteria such as non-Gaussianity or contribution to data variance to order them, but, they assert, the results are usually not very enlightening. They suggested that the lack of ordering in standard ICA is related to the independence assumption. In practice, with real data, the resulting components rarely achieve perfect independence, with certain pairs of components showing a marked dependence though not of the kind that can be removed by linear transformation. (See *e.g.* Simoncelli, 1999; Schwartz and Simoncelli, 2001, and fig. 8.2.) Thus, Hyvärinen et al. proposed that this residual de-

[1] If we do have 'grandmother' cells, then one's maternal grandmother cell is probably not too far from one's paternal grandmother cell, or indeed one's 'grandfather' cells!

[2] This may be different from the metrical structure induced from the metrical structure of $\mathcal{S}$ by the mapping $s \mapsto \xi$. Indeed, one of the factors influencing the design of the representation may be a requirement to make these two alternative metrical structures agree.

pendency be used to decide which units of the representation, and hence which basis vectors, belong together.

There are sound reasons for bringing together dependent units. One, as we have seen in previous chapters, is as an aid to visualisation and interpretation. More importantly, however, it may be an aid to further processing. If we take the view that redundancy reduction is indeed the goal of perception, then any remaining dependencies after a stage of processing are precisely where more work is required. Localising residual dependency localises any further computations, reducing demands on connectivity in a distributed system. Hyvärinen et al. (2001) described this in terms of *wire length costs* (see also Mitchison, 1995) which could be manifested, for example, as greater energy requirements on long connections, or greater noise.

8.1.1 From mutual information to geometry

Hyvärinen et al. (2001) used the concept of local dependence to guide the construction of a generative model in which similarity between representational units is encoded as a predefined system of neighbourhoods. This in turn defines a topology on the set of units, such as a 2-D sheet or a 1-D closed loop. There is no metric structure as there are no explicit distances involved.

The method presented here takes a different approach: quantitative estimates of residual dependency are used directly to define distances and hence a metric on the representational units. The distance between a pair of units will be small if their mutual dependency is high, zero if they are deterministically related, and infinite if they are independent. This metric is then used to find a geometric arrangement of the units such that the assigned distances are accurately represented in a Euclidean space. There is no prior choice of topology: the dimensionality of the Euclidean embedding space needs to be chosen, but the method used to find the geometric arrangement (multi-dimensional scaling, or MDS) does not prevent the units from taking up positions in a lower-dimensional manifold. Thus, the representation can find its own intrinsic dimensionality. In contrast, if a topographic ICA were to be performed on a data set that required a two-dimensional basis, such the the speech data of § 5.3.1, but a one-dimensional topology was specified, the system would presumably find a good way to arrange the units on a line, but there would be residual dependencies not explained by the one-dimensional topology, and without further analysis to detect this, there would be no hint that a two-dimensional arrangement might be better.

The steps required to go from a given distributed representation to a geometric picture of it will be described in the remainder of this section. The overall procedure can be summarised as follows:

1. Obtain a distributed representation by any means available. The present work is based on ICA, but could easily be repeated with any linear basis, or indeed *any* method at all that produces a distributed representation in which pair-wise dependencies can be measured.

2. Measure the statistical dependency between each pair of units, by estimating their mutual information or other measure of correlation.

3. Convert the dependency measures into metric distances in such a way that high dependency implies proximity.

4. Use MDS to convert the pair-wise distances into a geometric arrangement, that is, a set of points in some embedding space.

The first step has already been covered in Chapter 5, while the last three follow below.

8.1.2 Measuring residual dependence using nonlinear correlation

Statistical dependence, quantified as the mutual information (MI), is difficult to measure directly without prior knowledge about the joint distribution of the two variables being compared. Since residual dependencies exist precisely because the ICA source model has proven to be incorrect, we cannot rely on the prior joint distribution assumed for the sources, even if the marginals are a good fit. The simplest form of dependence, linear correlation, is unlikely to be useful; residual correlations in ICA tend to be small, and some ICA algorithms force them to be zero. Instead, Hyvärinen et al. (2001) proposed that *energy* or *activity* correlations are the primary form of residual dependence, citing a similar conclusion from Schwartz and Simoncelli (2001), who observed that when natural sounds or images are represented using a wavelet or Gabor basis, the coefficients of similar wavelets tend to be non-zero at the same time even if they are not linearly correlated.

In fact, for two-dimensional distributions symmetric in both axes, the dependency is *entirely captured* by the dependency between the absolute values of the variables: if X and Y are two random variables, and $p(x,y) = p(|x|,|y|)$ in all four quadrants, then $I(|X|,|Y|) = I(X,Y)$; this can be shown by direct integration.

In general, correlation of energies might not be the best way to characterise the MI because it involves measuring second-order statistics of the *squares*—that is, fourth order statistics—of what may already be highly super-Gaussian variables. For finite sample sizes, the resulting correlations will tend to have high asymptotic variances, and thus the estimated dependencies will be very noisy. However, it should be possible to compute correlations between *any* nonlinear functions of activity to estimate the dependence: under certain mild restrictions, this will yield zero for independent variables, whereas a non-zero correlation implies dependent variables.

Jutten and Herault (1991) used this principle in their ICA algorithm, but specified that two different odd functions be used. In view of the observations above about measuring *activity* dependencies because of the symmetry of the distributions involved, even functions will be used here. The question is, which particular nonlinear even functions will give the best estimates of the MI?[3]

Consider the simplest case in which two random variables U and V are in fact jointly Gaussian. The dependence between them is fully characterised by their corre-

[3] Bach and Jordan (2001) describes an elegant, but relatively computationally expensive method by which the optimal nonlinearities may be chosen from a given family in an unsupervised way, but this was published after the present work was completed.

lation coefficient, which is defined as

$$\text{corr}[U,V] \stackrel{\text{def}}{=} \frac{\text{cov}[U,V]}{\sqrt{\text{var}\,U\,\text{var}\,V}}, \tag{8.1}$$

$$\text{where} \qquad \text{cov}[U,V] \stackrel{\text{def}}{=} \text{E}\,UV - (\text{E}\,U)(\text{E}\,V), \tag{8.2}$$

$$\text{and} \qquad \text{var}\,U \stackrel{\text{def}}{=} \text{E}\,U^2 - (\text{E}\,U)^2. \tag{8.3}$$

The mutual information can then be expressed in terms of the correlation coefficient only as

$$I(U,V) = -\tfrac{1}{2}\log\left[1-(\text{corr}[U,V])^2\right], \tag{8.4}$$

which is correct for both positive and negative correlations. If U and V are far from being Gaussian, then the MI estimated with this expression may not be accurate; for example, the correlation may be zero even for strongly dependent variables, as illustrated in the first column of fig. 8.2. However, the MI is invariant to invertible transformations of the variables, so if f and g are invertible functions, then $I(U,V) = I(f(U),g(V))$. Furthermore, if f and g could be chosen so that $f(U)$ and $g(V)$ were jointly Gaussian, then the MI could be computed exactly from this nonlinear correlation. Thus, it may be conjectured that for *any* two functions f and g, this nonlinear correlation provides a lower bound on the mutual information:

$$I(U,V) \geq -\tfrac{1}{2}\log\left\{1-(\text{corr}[f(U),g(V)])^2\right\}. \tag{8.5}$$

Recent work by Bach and Jordan (2001) suggests that this may be true, though a formal proof has not been attempted.

Measuring energy correlations, as advocated by Hyvärinen et al. (2001), is equivalent to putting $f(u) = g(u) = u^2$. A cursory look at the distributions of these energies shows that they are very far from being Gaussian, and hence, energy correlations are possibly not the best way to measure the residual dependencies.

What is required is a compressive nonlinearity so that $f(U)$ and $g(V)$ have lighter tails and thus their correlation can be measured more reliably. One possibility is $f(u) = g(u) = \log u$, which produces joint distributions like those shown in the middle columns of figures 8.1 and 8.2. Another possibility is $f(u) = g(u) = -\log p(u)$, where $p(\cdot)$ is the prior density used in the ICA model. For a Laplacian prior, this yields $f(u) = |u|$, but for a more super-Gaussian prior, such as a Cauchy prior (see eq. 5.7), the nonlinearity is more compressive. It must be noted that in both cases, f and g are even functions and hence not invertible, but it still seems a reasonable choice to make given the earlier comment about distributions with four-quadrant symmetry and our wish to measure *activity* correlations.

Several estimates of mutual information are compared in table 8.1. The two methods used are the nonlinear correlation method presented here, and direct summation from the joint histogram, though it must be emphasised that *both* methods are approximations, since the histogram is only a discrete (and often poor) approximation to the underlying continuous probability density.

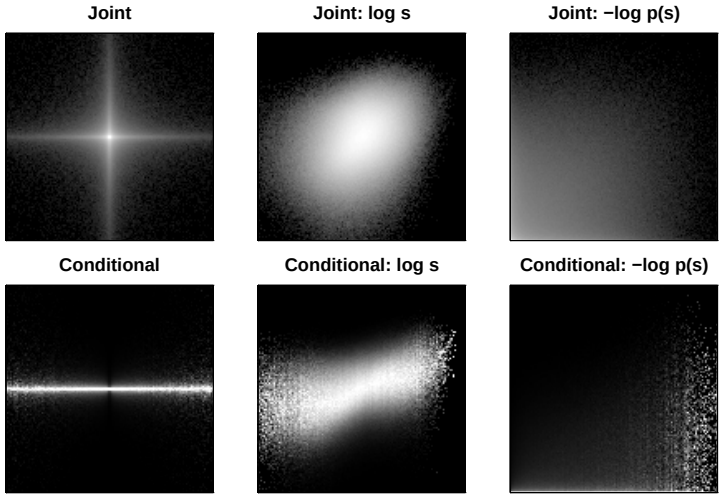


Fig. 8.1: Joint histograms of a pair of ICA outputs s_i and s_j which are nearly independent. The three columns contain histograms (on a logarithmic grey scale) of $[s_i \text{ vs. } s_j]$, $[\log s_i \text{ vs. } \log s_j]$ and $[-\log p(s_i) \text{ vs. } -\log p(s_j)]$, where $p(s)$ is the Cauchy prior. The top row contains joint histograms, whereas in the bottom row, each column of histogram bins has been independently normalised to the same maximum value. (Obtained from ICA experiments with BBC Radio 3, see Chapter 5. The corresponding basis vectors are both sinusoidal with frequencies of 883 Hz and 937 Hz. 82:87.)

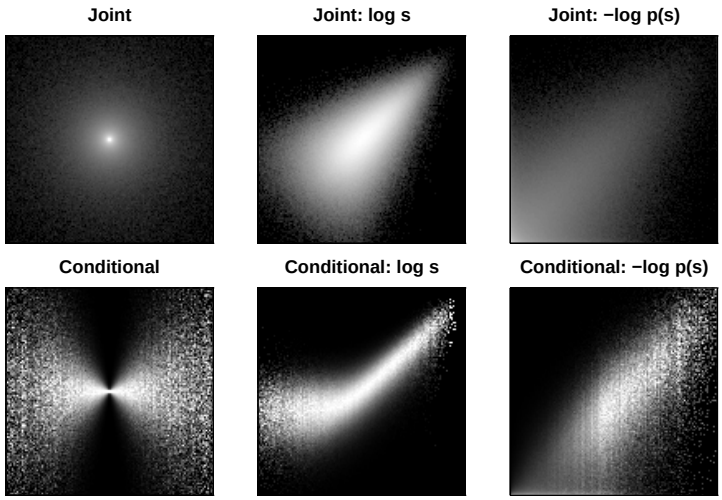


Fig. 8.2: Joint histograms of two strongly dependent components from the BBC Radio 3 ICA results. The corresponding basis vectors are sinusoidal with the same frequency (937 Hz) but in quadrature phase. Note how the untransformed components (first column) show very little correlation, with a circularly symmetric distribution, but after an even, non-linear transformation (other columns), the correlation is high. See fig. 8.1 for further explanation.

	s_{79} *vs.* s_{229} (fig. 8.1)		s_{237} *vs.* s_{229} (fig. 8.2)	
	correlation	**direct**	**correlation**	**direct**
$f(s) = s$	0.0000	0.0457	0.0000	0.3282
$f(s) = \log s$	0.0652	0.0746	0.3228	0.4089
$f(s) = \log(1+s^2)$	0.0419	0.0685	0.4573	0.4022
$f(s) = s^2$	0.006	0.185	0.2663	0.1945

Tab. 8.1: Estimates of MI computed by two methods from the histograms in figures 8.1 and 8.1. Correlation coefficients were computed from the histograms, not from the original samples; eq. 8.4 was then used to estimate the MI in bits. The 'direct' method consists of computing by direct summation the MI of the *discrete* probability distribution implied by the histogram. Note that this is likely to deviate systematically from the MI of the underlying continuous density: this is the source of the differences between the direct estimates. Each row is the result of applying the same methods to the joint histograms of the nonlinearly transformed components, using different nonlinearities. The last row was derived from joint *energy* histograms not illustrated here.

8.1.3 Computing distances

The next step is to find a suitable mapping from mutual information to distance. Let me state at the outset that what follows is a rather heuristic argument in support of a particular form for the mapping. The guiding principle will be that a 'Gaussian-like' noise process, in a sense to be defined later, should result in the familiar Euclidean distance measure. The reasoning behind this will be described in § 8.2.2.

Consider the following thought experiment: a lattice is populated with a number of 'particles' placed at the lattice nodes; these could be photons in an imaging model, for example. This initial configuration constitutes a 'noiseless' image. Then, a noise process is applied in which each particle undergoes a random displacement relative to its initial position, according to a certain probability distribution. This noise model will induce observable statistical dependencies between the numbers of particles at nearby lattice points, even if the original noiseless images are spatially independent. Thus, it should be possible to deduce the spatial configuration of the lattice from statistical data alone. Furthermore, it seems reasonable to design the method to produce a Euclidean metric on the lattice if the distribution of the random displacements is Gaussian.

A number of simplifications are needed to obtain a useful result. Firstly, the noiseless images will be assumed to be spatially independent, so that any measured dependencies can be attributed to the noise process. Secondly, the discrete model outlined above will be replaced with a continuous one, where the lattice spacing tends to zero and the numbers of particles tend to infinity. In this case, the particle distribution is described by a particle density function, and the diffusion caused by the noise process becomes a convolution or filtering of the density function with a Gaussian kernel. Thirdly, in order that the entire system be characterised by second-order statistics, the noiseless images will be assumed to be Gaussian as well as independent. In fact, the

whole problem becomes very similar to *blind deconvolution*, except for the crucial difference that, in blind deconvolution, the geometry of system is known and the convolution kernel deduced, whereas here, the kernel is known and the geometry must be deduced.

After the convolution, the particle densities (or image intensities) at nearby points will be correlated, and their correlation will be directly related to their distance from each other—it will be the given by the autocorrelation function of the filtered image. Since the autocorrelation is the Fourier transform of the power spectral density, which in this case is the product of the initially white spectrum with the Gaussian spectrum of the Gaussian kernel, the result may be written down immediately:

$$\mathrm{corr}[S_\alpha, S_\beta] = \exp -\tfrac{1}{2}[d(\alpha,\beta)]^2, \tag{8.6}$$

where S_α and S_β are random variables representing the particle density at two points α and β, and $d(\alpha,\beta)$ is proportional to the distance between them. Since an overall scaling of the distance is unimportant, there is no need to relate this to the variance of the Gaussian step distribution. Thus, I propose that this simply be inverted to *define* the (squared) distance as

$$d^2(\alpha,\beta) \stackrel{\text{def}}{=} -\log\left\{\mathrm{corr}[S_\alpha, S_\beta]\right\}^2. \tag{8.7}$$

The next step is to generalise the above result to non-Gaussian variables using the mutual information rather that the correlation. The MI for two Gaussians was given in eq. 8.4. Solving this for the correlation coefficient and substituting back into eq. 8.7 gives

$$d^2(\alpha,\beta) \stackrel{\text{def}}{=} -\log\left\{1 - \exp\left[-2I(S_\alpha, S_\beta)\right]\right\}. \tag{8.8}$$

Clearly, this a somewhat speculative definition, but it will serve for the following experiments, and does produce some interesting results. The important point is that the distance is monotonically related to the mutual information and has the correct asymptotic behaviour: $D \to 0$ as $I \to \infty$, and $D \to \infty$ as $I \to 0$.

8.1.4 Geometry by multidimensional scaling

The final step is to generate a set of positions in a Euclidean or other metric space which is consistent with the system of pair-wise distances derived above. Formally, given a metric space E, a metric $d_E : E \times E \mapsto \mathbb{R}$, and a set of representational unit indices $\mathcal{A}$, we seek points $x_\alpha \in E$ such that $d_E(x_\alpha, x_\beta) = d(\alpha,\beta)$ for all $\alpha,\beta \in \mathcal{A}$. This is precisely the problem that multidimensional scaling (MDS) was designed to solve (*e.g.* Cox and Cox, 2001; Davidson, 1983). In practice, it will rarely be possible to make the distances match perfectly, so MDS is formulated as an optimisation problem in which the aim is to minimise a certain *stress* function. In *least-squares* MDS (Torgeson, 1952), the following stress function is minimised:

$$J_1 = \sum_{\{\alpha,\beta\}\subset\mathcal{A}} \Big[d_E(x_\alpha, x_\beta) - d(\alpha,\beta)\Big]^2, \tag{8.9}$$

though there are other possibilities; for example, Sammon (1969) proposed the use of

$$J_2 = \sum_{\{\alpha,\beta\}\subset\mathcal{A}} \frac{\left[d_E(x_\alpha,x_\beta) - d(\alpha,\beta)\right]^2}{d(\alpha,\beta)}. \tag{8.10}$$

The extra factor of $1/d(\alpha,\beta)$ in each term of the sum means that, in comparison to J_1, short distances are given a greater weighting and thus are prioritised in the optimisation: this is saying that it is more important to get the short distances right than the long.

MDS can be formulated as a statistical inference problem (Cox and Cox, 2001, p. 107). In maximum likelihood estimation, quadratic error functions are linked to the assumption of Gaussian noise, and in this case, each stress function can be interpreted as implying a certain error model for the measured target distances $d(\alpha,\beta)$. The first, eq. 8.9 implies additive Gaussian noise of fixed variance. This is not necessarily appropriate for measured distances or distances computed from some statistical method: for example, in some cases one might expect the errors to be larger for long distances. The second stress function in eq. 8.10 goes some way to addressing this, and implies additive Gaussian noise with a *variance* proportional to the actual distance. In fact, one may go further and propose a stress function in which the *standard deviations* of the errors are proportional to the distances:

$$J_3 = \sum_{\{\alpha,\beta\}\subset\mathcal{A}} \left[\frac{d_E(x_\alpha,x_\beta) - d(\alpha,\beta)}{d(\alpha,\beta)}\right]^2, \tag{8.11}$$

All three stress functions were implemented; a steepest-descent algorithm for finding the optimal x_α is given in Appendix C. Fig. 8.3 illustrates the difference between the results obtained with J_1 and J_3 in a two-dimensional MDS. A three-dimensional MDS solution for the same data set produces a curved 2-D manifold embedded in 3-D; the comparison in fig. 8.3 shows that the stress function J_3 is better able to manage the flattening of the curved manifold into a flat 2-D space.

In fact, none of the three stress functions is theoretically correct for the data used here: since the distances used here are derived from measured sample correlation coefficients, we can be more specific about the distribution of the error. For a 2-D Gaussian distribution, the sample correlation coefficient computed from a large sample has a standard deviation of $(1-\rho^2)/\sqrt{n}$, where ρ is the true correlation coefficient and n is the sample size (Mukhopadhyay, 2000). Consulting eq. 8.6, this suggests that the following stress function should be used:

$$J_4 = \sum_{\{\alpha,\beta\}\subset\mathcal{A}} \left[\frac{d_E(x_\alpha,x_\beta) - d(\alpha,\beta)}{1-\exp[-d^2(\alpha,\beta)]}\right]^2. \tag{8.12}$$

Note that the 2-D joint distributions are not in fact Gaussian and so this is still only an approximation to the ideal stress function. It will be interesting to see if the implementation of this stress function yields significantly different results; this will be the subject of further work.

Fig. 8.3: Two of the two-dimensional MDS solutions obtained for the Radio 4 (speech based) representation, using the stress functions J_1 (left) and J_3 (right). The gray scale indicates the nominal centre frequency of each unit as computed in §5.3.1, with low frequencies in black at the bottom of the figures. Further analysis of this configuration makes it clear that it is largely a *time-frequency* representation, with the time-axis horizontal in this figure. The 'tail' extending to the left contains low-frequency units that are not localised in time.

8.1.5 Experimental results with speech and music

The ICA results of Chapter 5 were the starting point for this experiment. An initial inspection of the ICA outputs suggested that many of the components, especially those associated with basis vectors of the same frequency but different phases, were strongly dependent. Figures 8.1 and 8.2 illustrate some measured joint histograms, showing different degrees of dependence but small linear correlations. Related components show the characteristic circular distribution reported by Simoncelli (1999), whereas unrelated components have the more diamond or cross-shaped distribution predicted by the ICA model.

Also illustrated are equivalent joint histograms of nonlinear (rectified) activity distributions. Joint *energy* distributions are not shown, but they are very super-Gaussian as expected and thus their dependency is not very well characterised by a correlation coefficient (see table 8.1). In contrast, using $\log s$ or $\log(1+s^2)$ results in joint distributions which have lighter tails[4] and give better mutual information estimates using the correlation method.

Nonlinear cross-correlation matrices In order to obtain distances between all pairs of components, the full matrix of nonlinear correlation coefficients is required; this was computed as

$$R_{ij} = \mathrm{corr}[f(s_i), f(s_j)], \tag{8.13}$$

where the s_i are ICA output components. The full set of statistics were collected twice, once with $f(s) = |s|$, and once $f(s) = \log(1+s^2)$, which is related to the log-density

[4] Perhaps 'skirts' would be a more appropriate term for such two-dimensional distributions.

Fig. 8.4: Two visualisations of a two-dimensional MDS solution obtained for the Radio 3 (music) representation. On the left, the gray scale indicates the centre frequencies of the units as in fig. 8.3. On the right, however, the gray scale indicates the nominal bandwidths computed in §5.3.2, showing a slight tendency for the wider bandwidth units (lighter) to lie towards the interior of the configuration. Note that this is a flattened version of the frustum of cone which appears in the three-dimensional configuration, and hence the pattern is not very clear.

of the Cauchy distribution. A matrix of pair-wise distances was then computed using

$$D_{ij} = \sqrt{\log R_{ij}^{-2}}, \tag{8.14}$$

as derived in §8.1.3.

Multidimensional scaling results An MDS algorithm was then run on the various distance matrices obtained, using the three different stress functions described in §8.1.4 in embedding spaces of dimension 2 to 8. Lack of dimensions preclude the illustration on paper of most of these results; figures 8.3 and 8.4 show some of the 2-D results and figures 8.5 and 8.6 show two of the 3-D configurations for readers who are able to fuse the stereo pairs.

Interpretation of spatial arrangements The configurations obtained with the BBC Radio 4 speech data show a clear ordering in time and frequency, similar to those illustrated in figures 5.2 and 5.10. This is most apparent when the component activations are displayed in 2 or 3-D in real time: during speech, sibilants are visible as activity in the upper region; plosives are visible as temporally localised strips of activity, and sustained vowel sounds appear as strips of activity localised in frequency. In 3-D, the time-frequency manifold takes on a folded shape, though at present it is not clear whether this is because there are non-local dependencies pulling the whole manifold together, or because the mapping from mutual information to distance has the wrong asymptotic form for large distances.

The results obtained with music (BBC Radio 3) are very different. The configuration is at first quite difficult to interpret, but interactive experimentation has revealed

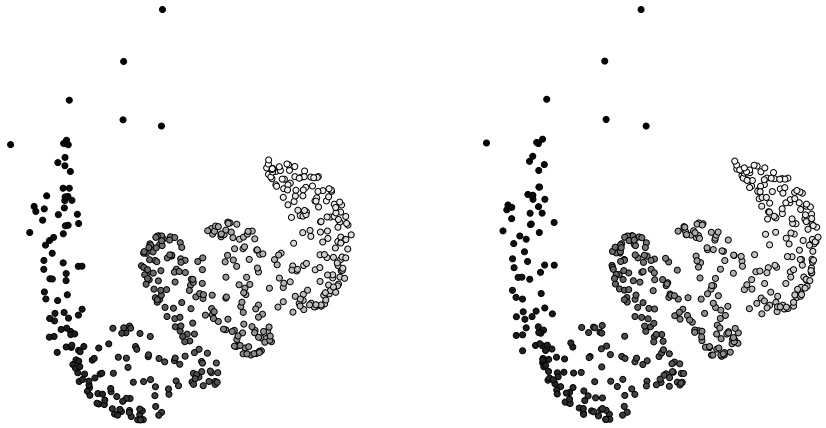


Fig. 8.5: Stereo pair of 3-D MDS results for the Radio 4 basis. The left-eye image is on the left, so you must stare into the distance to fuse the pairs. The grey scale encodes the centre frequency of each unit. In three dimensions, the time-frequency plane visible in fig. 8.3 becomes a curved manifold, viewed edge-on in this figure, so that time axis is perpendicular to the page.

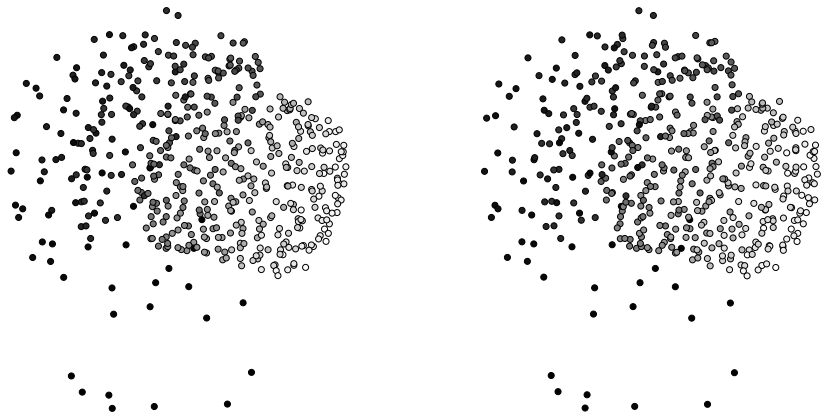


Fig. 8.6: Stereo pair of 3-D MDS results for the Radio 3 basis. The configuration is approximately a frustum of cone, with its axis pointing right, slightly down, and slightly out of the page. High frequencies are at the narrow end.

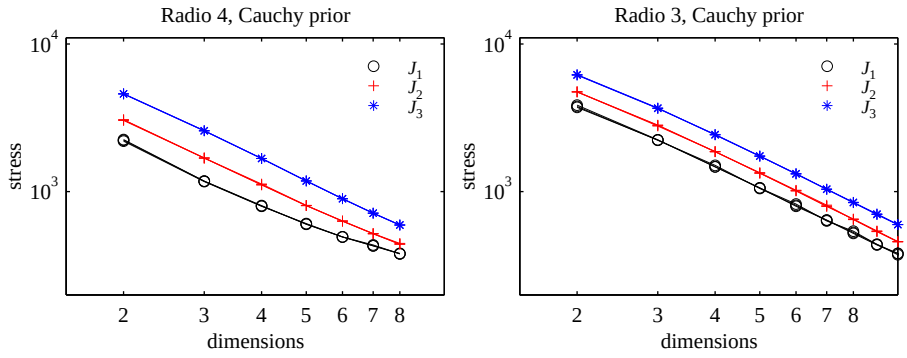


Fig. 8.7: Plots of stress *vs.* dimensionality for Radio 4 and Radio 3 MDS solutions using three different stress functions. The stresses are in arbitrary units and different stress functions cannot be compared directly.

some of its structure. Unfortunately, this is difficult to illustrate statically, but the results may be summarised as follows.

Overall, the geometric arrangement is approximately that of a frustum of cone, with high frequency units towards the narrow end. The units are distributed within the volume of the frustum, not just on the surface. To learn more about the details of the arrangement with respect to pitched sounds, a probe tone was applied and the pattern of activity inspected. The author supplied these probe tones to a real-time implementation of the system by playing his flute into a microphone attached to the computer.

As the pitch of the probe tone was swept upwards, a locus of activity in the representation moved around and up the frustum. Although the arrangement was not perfect, there was a general tendency for units on the surface of the frustum to respond to chromatic notes played in-tune, while the inner units responded to intermediate pitches. In addition, these inner units seemed to be less sharply tuned than the surface units. When viewed along the axis of the frustum, the in-tune chromatic notes formed an approximate circle of Fifths. A idealised schematic of this arrangement is illustrated in fig. 8.8, which should be imagined as the view down the axis of the frustum.

Choosing the dimensionality of the embedding space The MDS stress functions determine the optimal configuration for a given dimensionality of the embedding space, but they do not help us choose that dimensionality. By comparison, this choice is a rather ad-hoc affair. If the distance measurements were noiseless and reflected a real underlying geometric arrangement, then we would expect the stress to go down to zero at the correct dimensionality. In practical applications, however, the data will be noisy, and we will not know if a meaningful embedding space even exists. In the general case, the stress will not go down to zero until we have about as many dimensions as there objects. What we are usually looking for is low dimensional arrangement that captures most of the 'interesting' structure—a requirement loaded with subjective assessments.

The conventional practice is to obtain several configurations for different dimensionalities and plot stress *vs.* dimensionality. If the data set has an 'intrinsic' dimen-

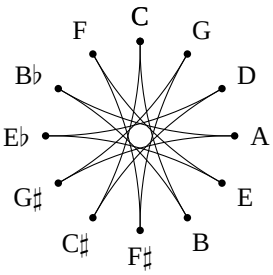


Fig. 8.8: This is an idealised schematic of the pattern of activation generated as a probe tone gradually sweeps up in pitch (see text). The actual arrangement follows this pattern from B♭ round to E but then becomes more diffuse. However, the correlation structure which gives rise to this pattern seems to be consistent across a wide range of frequencies (see fig. 8.9) so it seems likely that the idealised pattern will emerge in MDS solutions in higher dimensions.

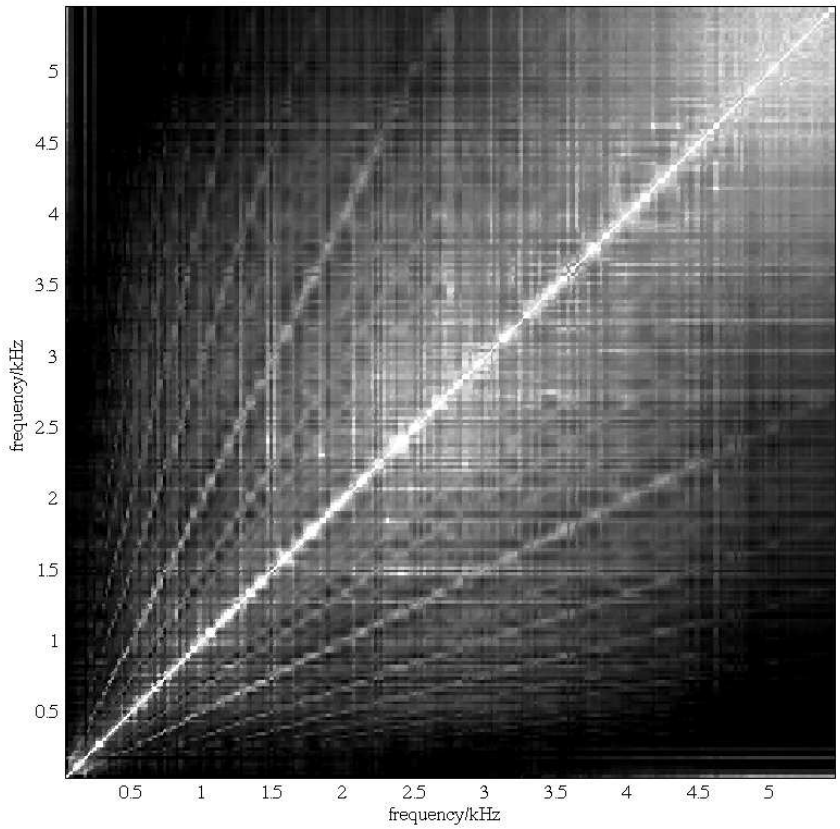


Fig. 8.9: Matrix of non-linear correlation coefficients obtained with the Radio 3 music basis, reordered and scaled by centre frequency to show strong correlations between frequencies in small whole number ratios. There is also a slight 12-cycle per octave 'beading' effect visible.

sionality, it may be apparent in the stress plot as a 'kink', 'elbow' or other discontinuity in the graph. There may be none or several of these, so in the absence of a more principled approach, we must examine each case individually.

The graphs in fig. 8.7 do not have any obvious such points, which makes it difficult to decide which is the appropriate dimensionality. It is possible that an alternative mapping from correlation to distance may have an effect on this.

Visualisation of cross-correlation matrix The emergence of chromatic relationships and a circle of fifths suggests that other harmonic relationships might be encoded in the correlation matrix, but are unable to emerge in a 3-D MDS configuration. It is in fact possible to observe these directly in the correlation matrix if it is reordered appropriately: in fig. 8.9, each element is plotted as a rectangular patch at a position proportional to the centre frequencies of the two basis vectors being correlated. The diagonal lines that are visible at various slopes show that there are indeed correlations, not just at frequency ratios of 1:2 and 2:3 but also at many other harmonic ratios.

8.1.6 Conclusions

The method of nonlinear correlation analysis was certainly able to detect a wealth of residual dependency structure in both the speech and music derived ICA bases developed in Chapter 5. The combination of long and short range distance information was used to find not only topological neighbourhoods for each unit in the representation, as in topographic ICA, but a complete geometrical representation of the dependency structure. One of the clearest conclusions is that, for speech, a two-dimensional time-frequency representation really is appropriate; in the work presented here, a time-frequency manifold emerges from the data in a completely unsupervised way. Equally clearly, this is not the case for music, indicating that a time-frequency picture may be too simplistic a way of looking at the representation of music signals at this level.

It may be thought that Kohonen's self-organising map (see § 3.2.5) achieves a comparable result, but the similarity is only superficial. The Kohonen map finds low-dimensional manifolds in high-dimensional data, but only for data that *already* has a fully specified geometrical structure in a high-dimensional Euclidean space. In contrast, the present method requires only system of pair-wise dependency measurements between entities which need have no pre-existing representation in a metric space. Also, the resulting system uses *distributed* patterns of activity to represent data (since ICA is a multiple case model), rather than the winner-take-all representation of the Kohonen map, which is effectively a single cause model.

In terms of the introductory section of this chapter, Kohonen's map implements a measure of P-similarity, essentially by applying a Euclidean metric to the input space (though it could be argued that it also implements an alternative metric measured within the manifold defined by the map, rather than in the original space) whereas the present geometrical method implements a measure of R-similarity in an abstract space; the MDS embedding space does *not* represent distances between patterns, or input vectors in the case of ICA, but between the resulting components.

One significant drawback of this method as compared with topographic ICA is its computational complexity which, because of the need to measure pairwise correlations, is $O(N^2)$ in the number of representational units. However, since the method is based on the *second-order* statistics of the nonlinearly transformed unit activities, it should be possible to apply a Gaussian latent variable model such as PCA or factor analysis to model what is essentially a large cross-correlation matrix more efficiently.

As an aside, it is interesting to note that both MDS and the method of probe tones are methods which have been used in experimental psychology (Krumhansl, 1990) on data gathered from human subjects, but are applied here to an artificial model, an ICA system, which is in a way, the experimental subject.

8.2 Similarity and distance measures

The previous section was concerned solely with a way of defining R-similarity for a given distributed representation. We will now consider some general points about P-similarity, that is, similarity between complete patterns of activity representing a sensory scene. The focus will be on using *metric* concepts, that is, distances, to express the notion of psychologically experienced or perceived similarity between scenes.

In the theory of metric spaces (*e.g.* Jameson, 1974), distances are subject to a number of formal constraints. Given a set $\mathcal{S}$, a metric or distance measure is a function $d : \mathcal{S} \times \mathcal{S} \mapsto \mathbb{R}$. If, for all $x, y, z \in \mathcal{S}$, the following conditions are satisfied:

$$d(x,x) = 0, \tag{8.15}$$

$$d(x,y) = d(y,x), \tag{8.16}$$

$$d(x,y) + d(y,z) \geq d(x,z), \tag{8.17}$$

$$d(x,y) > 0 \; \forall \; x \neq y, \tag{8.18}$$

then d is a metric on $\mathcal{S}$ and $(\mathcal{S}, d)$ is a metric space. If 8.18 is not satisfied, and $d(x,y) = 0$ for some $x \neq y$, then d is a pseudometric on $\mathcal{S}$.

The situation with subjective dissimilarity is less clear: assuming that it *can* be quantified, one might reasonably impose non-negativity and 8.15 but not necessarily the other conditions. Davidson (1983) discusses this, pointing out that in some contexts, subjective dissimilarity need not be symmetric and need not satisfy the triangle inequality. That possibility will not be pursued here, and we will assume that psychological distances are metric or at least pseudometric.

This raises a few questions. Given a representation consisting of one or more real variables, can the 'usual' metrics (see §8.2.1) yield the psychological distances we wish to reproduce? If not, is there an alternative representation in which the desired distances can be obtained with one of the usual metrics? Or can a more flexible metric be defined on the given representation that can accommodate them? Finally, and most interestingly, are there general principles which explain *why* the psychological distances are what they are? The hypothesis investigated in this section is that these distances are related to the statistical structure of the patterns, that is, they can be derived from a probability measure on S. One desirable characteristic of such a procedure

is that the derived distances can be made representation invariant, rather than defined in terms of a particular representation as with the usual metrics.

8.2.1 The usual metrics

The distance measures in table 8.2 are defined for pairs of objects represented as n-tuples, that is, as elements of a product space such as $\mathbb{R}^n$ or $\{0,1\}^n$. In all cases except the binary Hamming distance, the distance is defined as a sum of coordinate-wise differences, or a simple function thereof. This means that they are well suited to normed vector spaces, that is linear spaces in which vectors can be added and subtracted, and where each vector has a length. In this case, the distance between two vectors can be defined as the length of the vector difference. When viewed as a metric space, this imposes certain restrictions on the distance function. For example, the distance between two points must be invariant to translation, that is, the addition of another vector to both; the distance measure is essentially linear.

Why should this be a problem? If the n-tuple representation of patterns is chosen ad-hoc, and assumed to be a normed linear space, the resulting distance measure may be incapable of replicating the subjective distances we would like it to, no matter how the norm is chosen. A related point is that, for a space to be a vector space, it must define the operations of addition, negation, and multiplication by a scalar. These operations may not have any meaning for sensory patterns—the assumption that a space is a normed linear space is much stronger than the assumption that it is a metric space, and is not necessarily warranted for perceptual representations.

To show that linear spaces are not always appropriate, a simple one-dimensional example will suffice. It is generally accepted that the frequency of tones is perceived on a scale that is close to being logarithmic, that is, the distance between two frequencies f_1 and f_2 is not $|f_1 - f_2|$, but $|\log f_1 - \log f_2|$.

There are various ways of extending the metrics in table 8.2 to make them more flexible. One is to weight the contribution of each coordinate; for example, Cambouropoulos (1998, §8.3) uses a weighted Hamming distance to compute the dissimilarity between two objects:

$$d(x,y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i). \tag{8.19}$$

If each binary coordinate encodes a different property, then the weights adjust the salience of each property in determining distance.

Another possibility in the case of the Euclidean metric is to assume that the coordinate axes are not orthogonal, so that activity in one coordinate might be more or less equivalent to activity in another. This results in the introduction of a symmetric, positive definite metric tensor G_{ij}, with the distance given by

$$d_G(x,y) = \left[\sum_{i,j} (x_i - y_i) G_{ij} (x_j - y_j) \right]^{1/2}. \tag{8.20}$$

Thus, the metric is essentially a quadratic form in $(x-y)$. Putting $G_{ij} = \delta_{ij}$ recovers the Euclidean metric. If the coordinates x_i are relative to a non-orthonormal set of basis

Euclidean	$d_E(x,y) = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2}$	$x,y \in \mathbb{R}^n$
Hamming (between binary tuples)	$d_H(x,y) = \sum_{i=1}^{n}[1 - \delta(x_i,y_i)]$	$x,y \in \{0,1\}^n$
'City block' or 'Manhattan'	$d_{CB}(x,y) = \sum_{i=1}^{n}\lvert x_i - y_i\rvert$	$x,y \in \mathbb{R}^n$
Minkowski of exponent p	$d_p(x,y) = \left[\sum_{i=1}^{n}(x_i - y_i)^p\right]^{1/p}$	$x,y \in \mathbb{R}^n$

Tab. 8.2: Some distance measures.

vectors $\mathbf{a}_i$, so that $\mathbf{x} = \sum_i x_i \mathbf{a}_i$, then $G_{ij} = \mathbf{a}_i \cdot \mathbf{a}_j$: the metric tensor encodes the pair-wise dot products of the basis vectors and hence the angles between the coordinate axes.

What this metric does is to involve a form of similarity between the *coordinate axes*—which is therefore a form of R-similarity—in the computation of P-similarity, demonstrating that the two can be related. If two *units* are R-similar, then two *patterns* that differ only in their distributions of activity over the two units should be P-similar.

An example can be found in the spectral representation of tones: if a sound is represented as the energies in a number of spectral bands, say 10 Hz wide each, and a straightforward Euclidean metric is used, then tones of 200 Hz and 2000 Hz would be no more distant than tones of 200 Hz and 220 Hz. If we wished energy in nearby spectral bands to have similar significance, we might try the following heuristic solution: before computing the Euclidean distance between two spectral vectors $\mathbf{x}$ and $\mathbf{y}$, convolve them both with a blurring kernel:

$$\mathbf{x}^* = \mathbf{U}\mathbf{x}, \qquad \mathbf{y}^* = \mathbf{U}\mathbf{y}, \tag{8.21}$$

where U is a symmetric Toeplitz matrix whose rows are all shifted versions of the convolution kernel. For example, the following matrix (where $0 < b < 1$) will spread activity from each spectral bin to its immediate neighbours:

$$\mathbf{U} = \begin{pmatrix} 1 & b & 0 & \cdots & 0 \\ b & 1 & b & & \vdots \\ 0 & b & 1 & & \\ \vdots & & & \ddots & b \\ 0 & \cdots & & b & 1 \end{pmatrix}. \tag{8.22}$$

In the spectral example, this would result in the 200 Hz and 220 Hz tone vectors being

brought closer together in the Euclidean sense. Computing the distance using

$$\begin{aligned} d^2(\mathbf{x},\mathbf{y}) &= (\mathbf{x}^* - \mathbf{y}^*)^T(\mathbf{x}^* - \mathbf{y}^*) \\ &= (\mathbf{x}-\mathbf{y})^T\mathbf{U}^T\mathbf{U}(\mathbf{x}-\mathbf{y}), \end{aligned} \tag{8.23}$$

shows that the matrix $\mathbf{U}^T\mathbf{U}$ has taken on the role of a metric tensor. If $\mathbf{U}$ has the form indicated above, then then $\mathbf{U}^T\mathbf{U}$ will also be approximately Toeplitz, with each row (except the extremal ones) consisting of a shifted version of

$$(\ \cdots \quad 0 \quad b^2 \quad 2b \quad 1+2b^2 \quad 2b \quad b^2 \quad 0 \quad \cdots\ ).$$

Thus, the procedure is equivalent to assuming that the spectral vectors are measured using a non-orthogonal coordinate system, in which each the coordinate axes for neighbouring spectral bins are not orthogonal, but those for widely separated bins are. We will see shortly that this has an interesting interpretation in terms of noise statistics.

8.2.2 Distance and noise

The previous section examined various increasingly flexible 'numerical recipes' for computing distances in a given representation, with the implicit aim that the metric be capable of reproducing a given set of subjective psychological distances. We now consider what principles might account for those perceived distances themselves. Taking an ecological point of view, it would be useful for an organism if it perceived difference and similarity in relation to the *environmental significance* of different stimuli; that is, subjective dissimilarity should be related to the likely significance of difference. This is intrinsically a subjective notion, because what is important to one organism may not be to another.

Bearing in mind the conclusion of §2.2.5, which equated irrelevant information with noise, the central hypothesis here is that sensory patterns are perceived to be similar to the extent that the difference between them is likely to be due to noise. The simplest illustration of this is the case of Gaussian white noise: in many algorithms the assumption of Gaussian noise results in the use of a mean-square error measure and may be linked with a Euclidean metric. In this particular case, the distance-squared has the form of minus the logarithm of a probability, which can be used to guide any possible generalisations to non-Gaussian or non-additive noise processes; these will be explored in the next section. In its most general form, it should be possible to apply this definition to any kind of noise, originating externally in the world, inside sensory receptors, or inside the cognitive system itself. It is also consistent with previous work by Mitchison (1995), which deals with map formation—that is, maps of P-similarity—driven by a number of objective functions, each of which can be interpreted as minimising the effect of noise in one space on the geometry in another.

Examples To clarify the points made above, a few examples will be given. Firstly, the Hamming distance can be derived from a noise process which inverts each bit of a binary tuple independently with a certain probability, q. The probability that the state

x is transformed into state y is

$$P(x \mapsto y) = \prod_{i=1}^{n} \begin{cases} 1-q & : x_i = y_i \\ q & : x_i \neq y_i \end{cases} \tag{8.24}$$

The negative logarithm of this is

$$\begin{aligned} -\log P(x \mapsto y) &= \sum_{i=1}^{n} \begin{cases} -\log(1-q) & : x_i = y_i \\ -\log q & : x_i \neq y_i \end{cases} \\ &= -n\log(1-q) + \log\tfrac{1-q}{q} \sum_{i=1}^{n} [1 - \delta(x_1, y_i)], \end{aligned} \tag{8.25}$$

which is essentially the Hamming distance offset by $-n\log(1-q)$.

Next, consider an audio signal in additive Gaussian *pink* noise, that is greater at lower frequencies than at high. In the frequency domain, the noise will be uncorrelated but non-uniform, with a variance that drops off towards high frequencies. In this case, a *weighted* Euclidean metric should be used, with a greater weighting at high frequencies, because differences there are less likely to be due to noise.

$$d^2(x,y) = \sum_i \frac{1}{\sigma_i^2}(x_i - y_i)^2, \tag{8.26}$$

where σ_i is the noise variance in the ith element. The reader may have noticed that there is some inconsistency in the identification of the log probability with either a distance (eq. 8.25) or a *squared* distance (eq. 8.26). This is an unresolved issue in the present work, but we may note that it does not affect the relative orderings of the distances so obtained.

Finally, consider the example from the last section in which a metric tensor was used to account for similarity between neighbouring frequency bands. Inspection of eq. 8.20 shows that it can be derived from a Gaussian noise process with covariance equal to the inverse of the metric tensor. Inversion of the metric in the example yields a covariance in which the noise in neighbouring bins is *anti*-correlated, that is, the noise tends to transfer activity from each bin to its neighbours. This is an example of the kind of noise that would lead us to conclude that neighbouring spectral bins should be considered similar. One may also observe that the pre-multiplication by $\mathbf{U}$ in eq. 8.21 has the effect of whitening the noise so that the standard Euclidean metric can be used on the vectors $\mathbf{x}^*$ and $\mathbf{y}^*$.

The foregoing is somewhat reminiscent of the procedure used to compute the Mahalanobis distance, (Mahalanobis, 1936) in which the inverse of a covariance matrix is also used as a metric tensor. The key difference is that the Mahalanobis distance uses the covariance of the *data*, whereas the above method uses the covariance of the *noise*.

8.2.3 Probabilistic distance measures

In this section, several tentative distance measures are proposed. Rather than basing them on a noise model which maps the pattern space on to itself, that is, one which takes

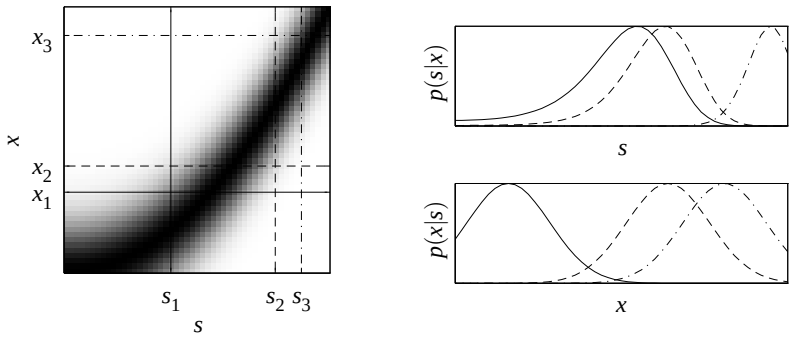


Fig. 8.10: The joint density $p(x,s)$ could be used to induce a metric on the space $\mathcal{X}$ by considering the difference between the conditional densities $p(s|x_1)$ and $p(s|x_2)$ for pairs $x_1, x_2 \in \mathcal{X}$. A metric on $\mathcal{S}$ could be induced similarly. The solid, dashed, and dashed-dotted graphs indicate sections through the joint density at the corresponding lines.

a notionally 'noiseless' pattern and produces a 'noisy' version of it, these measures will be based on a noisy mapping between two distinct spaces. To clarify, let us assume that $\mathcal{X}$ is the space of patterns, and $\mathcal{N}$ is a space of noise variables. The former model would be expressed as a mapping $\mathcal{X} \times \mathcal{N} \mapsto \mathcal{X}$, but the latter would be expressed as a mapping $\mathcal{S} \times \mathcal{N} \mapsto \mathcal{X}$, where $\mathcal{S}$ is a space of unobserved underlying states.

We will assume that every element s of $\mathcal{S}$ is distinct and can carry a different 'meaning.' If we could observe these directly, there would be no need for any concepts of similarity between them at all—each state would be recognisably and reliably distinct from all the others, and any difference would be significant. It is the introduction of noise which creates potential confusion between the elements of $\mathcal{S}$ and of $\mathcal{X}$. We will also assume that the mapping from $\mathcal{S} \times \mathcal{N}$ to $\mathcal{X}$ is not invertible because otherwise, we could simply recover s exactly, and again, there would be no need to consider elements of $\mathcal{X}$ or $\mathcal{S}$ to be more or less similar to each other.

If the noise variables are marginalised, then the system becomes non-deterministic and is summarised by the joint probability distribution of s and x. The distance measures described below are based on the idea that distances in either $\mathcal{S}$ or $\mathcal{X}$ should be chosen to localise the 'spread' of $p(s,x)$. Two observations x_1 and x_2 are to be considered similar to the extent that they could both have been generated from the same hidden state s, but one could also consider two states s_1 and s_2 to be similar to the extent that they could both generate the same observation x. The general principle is illustrated in fig. 8.10.

Coincidence likelihood similarity

First, we will consider a rather simplistic implementation of the above principle: how likely is that two patterns x_1 and x_2 are due to the same underlying state s? Consider two independent realisations of the generative model: (s_1, x_1) and (s_2, x_2). The problem is then to compute the probability that $s_1 = s_2$ given a particular pair (x_1, x_2). This can be done by integrating along the line $s_1 = s_2$ in the two-dimensional density function

$p(s_1, s_2|x_1, x_2)$: if $\Delta s = s_2 - s_1$, then the probability density of Δs evaluated at $\Delta s = 0$ is

$$\begin{aligned} p(\Delta s|x_1, x_2)|_{\Delta s=0} &= \frac{1}{p(x_1)p(x_2)} \iint \delta(s_1 - s_2) p(s_1, x_1) p(s_2, x_2) \, \mathrm{d}s_1 \, \mathrm{d}s_2 \\ &= \int p(s|x_1) p(s|x_2) \, \mathrm{d}s. \end{aligned}$$

Though this expression was derived in rather an ad-hoc manner and will not be pursued any further, it does suggest that we compare $p(s|x_1)$ and $p(s|x_2)$ using a more principled measure of dissimilarity between probability distributions, the obvious choice being the Kullback-Leibler divergence or a related measure.

Distance between probability distributions

The Kullback-Leibler divergence (*e.g.* Cover and Thomas, 1991) between two probability densities p and q on the same space $\mathcal{X}$ is

$$D(p\|q) = \int_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x. \tag{8.27}$$

Strictly speaking, it is not a metric on the space of distributions because it is not symmetric and does not satisfy the triangle inequality. We will overlook the second of these problems for the time being, and consider several possibilities for symmetrising it:

1. Form a weighted combination of the divergences in both directions:

$$\overline{D}(p||q) = \alpha_p D(p\|q) + \alpha_q D(q\|p), \tag{8.28}$$

   where α_p and α_q can either be the same, or chosen according to some scheme that makes sense in the problem at hand. In particular, if p and q are conditional probabilities associated with one of two states, one could use the probabilities of those states to weight the divergences. This is an extension of Kullback's own proposed symmetrisation, which had $\alpha_p = \alpha_q = 1$ (Kullback, 1968).

2. Use the Kullback-Leibler divergence as a *differential* distance measure between densities that are infinitesimally far apart. Given a suitably parameterised family of densities, this can be shown to induce a Riemannian metric on the differentiable manifold formed by the family (see Amari and Nagaoka, 2001). To compute the total distance between two densities in the family, one would integrate the infinitesimal distances along a geodesic in the manifold.

3. The following are suggested in the context of a binary mixture density controlled by a binary random variable U. Suppose that if $U = 0$, then the distribution of x is $p_0(x) \equiv p(x|U = 0)$, but if $u = 1$, then it is $p_1(x) \equiv p(x|U = 1)$. The variable U has a certain distribution $p(u)$, so the joint distribution is

$$p(x, u) = p(x|u)p(u). \tag{8.29}$$

A distance between $p_0(x)$ and $p_1(x)$ can then be obtained by taking one of the following expectations over this joint distribution:

$$D_A = \mathrm{E}_{x,u} \log \frac{p(x|u)}{p(x|\tilde{u})} \qquad D_B = \mathrm{E}_{x,u} \log \frac{p(u|x)}{p(\tilde{u}|x)}, \tag{8.30}$$

where tilde ($\tilde{u}$) denotes binary negation. They can be interpreted as measuring the average log-odds on correct inference of U given an observation x, using either maximum likelihood or maximum a-posteriori inference, respectively. They are essentially a *misclassification likelihood*: they rate how likely are we to confuse two the states. The first (D_A) is actually equivalent to

$$P(U=0)D(p_0 \| p_1) + P(U=1)D(p_1 \| p_0) \tag{8.31}$$

and thus has the same form as the symmetrised K-L divergence in eq. 8.28.

Distance measures in $\mathcal{X}$ and $\mathcal{S}$

Now that a number of possibilities for measuring distance between probability distributions have been suggested, these can be applied to the model system described at the beginning of this section, which is fully described by the joint density $p(s,x)$.

The distance between two observations x_1 and x_2 can be defined as the distance between the two conditional densities $p(s|x_1)$ and $p(s|x_2)$. In effect, this measures the extent to which two observations lead to the same conclusions about s. Similarly, the distance between two states s_1 and s_2 can be equated with the distance between $p(x|s_1)$ and $p(x|s_2)$, which measures the extent to which two states produce the same distribution of observations, once noise has been taken into account.

Distance via mutual information

That observations of the random variable X lead to conclusions about another S suggests that distances in the observation space $\mathcal{X}$ be defined explicitly in terms of a mutual information. Such an expression can be constructed as follows: consider two elements $x_1, x_2 \in \mathcal{X}$. The knowledge that X is *either* x_1 or x_2 results in a certain posterior distribution, $p(s|x \in \{x_1, x_2\})$. Subsequently, on learning the value of X, the posterior reduces to either $p(s|x_1)$ or $p(s|x_2)$, with a concomitant gain of information about S. Intuitively, one might expect that if very little information is gained about S in this final step, then x_1 and x_2 should be considered similar. If this final piece of information is encoded as a binary variable U which is 0 if $X = x_1$ and 1 if $X = x_2$, then the mutual information $I(U,S)$ measures the extent to which learning U supplies information about S. If Q denotes the predicate that $X \in \{x_1, x_2\}$, (which is clearly a precondition for the variable U to have any meaning) then

$$
\begin{aligned}
I(U,S) &= \sum_{u\in\{0,1\}} \int_{\mathcal{S}} p(u,s) \log \frac{p(u,s)}{p(u)p(s|Q)}\, \mathrm{d}s \\
&= p(x_1|Q) \int_{\mathcal{S}} p(s|x_1) \log \frac{p(s|x_1)}{p(s|Q)}\, \mathrm{d}s + p(x_2|Q) \int_{\mathcal{S}} p(s|x_2) \log \frac{p(s|x_2)}{p(s|Q)}\, \mathrm{d}s \\
&= p(x_1|Q) D(p_{s|x_1} \| \, p_{s|Q}) + p(x_2|Q) D(p_{s|x_2} \| \, p_{s|Q}),
\end{aligned}
$$

where

$$
p(s|Q) = \frac{p(s|x_1)p(x_1) + p(s|x_2)p(x_2)}{p(x_1) + p(x_2)}.
$$

In contrast with the symmetrised divergence in eq. 8.28, this is a weighted combination of distances from $p(s|x_1)$ and $p(s|x_2)$ to a *third* distribution, $p(s|Q)$.

8.3 Phase invariance

Hyvärinen et al. (2001) identify a relationship between their topographic ICA and the subspace ICA algorithm proposed earlier by Hyvärinen and Hoyer (1999, 2000), which was designed to perform a certain kind of invariant feature detection, and which can be said to model a certain kind of 'phase invariance.' This provides a link between phase invariance and the R-similarity discussed in § 8.1.

On the other hand, phase invariance in audition can be thought of as a way of ignoring irrelevant information, and is thus related to the P-similarity discussed in § 8.2. Sonological representations which are notionally 'phase invariant,' such as the spectrogram, model this by transforming the audio signal in a certain way and then discarding part of the result. This was one of the motivations in using a spectrogram, rather than a Fourier transform as the raw material for the sparse coder in Chapter 7, and enabled the system to discover harmonic relationships that the linear ICA in Chapter 5 was unable to detect.

In this section, the link between these two concepts of phase invariance will be investigated, and the implications for the use of quadratic representations, of which the spectrogram and the Wigner Distribution are examples, will be discussed.

8.3.1 Phase invariance in audition

Ohm's acoustical law (see, *e.g.* Risset and Wessel, 1999) states that the ear is phase deaf in the context of periodic tones. Specifically, a listener will be unable to perceive any difference between two sounds if their Fourier representations differ only in their patterns of phase relationships, but not in their harmonic amplitudes, even though the sounds may have very different waveforms. However, it is important to note that insensitivity to phase holds only for *periodic* tones. Helmholtz also held that ear is phase deaf under these conditions, and showed that this is indeed true for synthetic vowel sounds (Helmholtz, 1885).

Subsequent experiments (Risset and Wessel, 1999) have shown that the situation is not quite as clear cut: under certain conditions, changes in phase relationships (and

hence the shape of the wave form) are perceptible, however, Risset and Wessel found that the effect is weak, and not significant in a normally reverberant room, because multiple reflections disrupt phase relationships. This observation provides a clue as to why, in ecological terms, the auditory system should discard phase information in this way. If the perceptual machinery is geared towards the segregation of *independent* streams of information, then phase relationships, being dominated by reverberation or other environmental effects, are likely to be separated off. Any information they can yield is likely to be concerned with perception of the acoustic environment rather than the characteristics of the source itself. When it comes to extracting information about the sound source, phase relationships are likely to be of little use, and can be considered as noise according to the ecological definition described in § 8.2.2, at least in the sort of reverberant environment that Risset and Wessel referred to.

Phase invariant representations A phase invariant representation is one that discards 'the phase information'. The problem with this notion is that 'the phase information' is different for different representations. At one extreme, in a (non-windowed) Fourier transform, 'the phase information' includes *all* the timing structure. To give an extreme example, both Gaussian white noise and and the impulsive delta function have the same power spectrum, which is constant at all frequencies. Hence, they could be said to differ only in phase. However, I expect that the reader will agree that, in most contexts, the differences between them would be much more significant than that! In general, *phase* structure is *time* structure, and time structure is important. It is only certain classes of signal (such as periodic or constant signals) that are invariant to time shifts. In a time-frequency representation such as a short-term Fourier transform, or a wavelet transform, the set of signals that differ only in phase will depend on what time-frequency tiling is chosen. For example, Hyvärinen and Hoyer (2000) note that though phase and shift invariance are equivalent in a global Fourier representation, they are different if phase is computed from a local Fourier transform. The conclusion is that phase invariance is only defined *relative* to a particular representation, and thus a sensible approach might be to find the representation in which the resulting form of phase invariance is the most ecologically relevant.

8.3.2 Multidimensional phase spaces and complex cells

Independent subspace analysis or ISA (Hyvärinen and Hoyer, 1999, 2000) attempts to discover a factorial representation by grouping linear components into a number of subspaces. Within a subspace, the components may be dependent, but each subspace should be independent of the others. The energy within each subspace is computed by summing the squares of each component belonging to it, and so these energies should all be independent of each other. The *direction* of the activity in each subspace is discarded; this can be equated with the loss of phase information.

The key feature of a phase invariant subspace in Hyvärinen and Hoyer's model is that, within that subspace, the distribution of data is *spherically symmetric*. For example, in a 2-D subspace consisting of two components x and y, the probability

density would be of the form

$$p(x,y) = h(x^2+y^2). \tag{8.32}$$

Such a distributions *are* factorial: not in Cartesian coordinates but in *polar* coordinates. Thus, the independent components, as it were, of these data are not x and y but r and θ, or indeed, r^2 and θ. Such data can be generated by multiplying an independently drawn radius r with an n-tuple of phase variables distributed uniformly on a hypersphere in n-dimensions. By assuming that phase is unimportant, we imply that the phase variables are noise; in fact, they have a high entropy, uniform distribution, which is, perhaps, consistent with that notion. This leaves $r^2 = x^2+y^2$, the 'energy' within the subspace, as the quantity of interest, and which, it may be noted, is a *quadratic* function of (x,y).

Hyvärinen and Hoyer compare the operation of their ISA system with the so-called complex cells in visual cortex, which seem to behave in an analogous fashion, summing the activities of a number of cells (the simple cells, which behave approximately linearly) in the previous layer.

Furthermore, Hyvärinen et al. (2001) note that, since both ISA and topographic ICA are based on multiplicative models, ISA is a special case of TICA, in which the neighbourhoods form disjoint groups. The complex cell interpretation can still be applied to TICA, but each complex cell computes a local average over overlapping regions of the topographic manifold. This computation of average local energies bears a strong resemblance to the *divisive normalisation* procedures of Schwartz and Simoncelli (2001), though there is an important difference which will be discussed in §8.3.5.

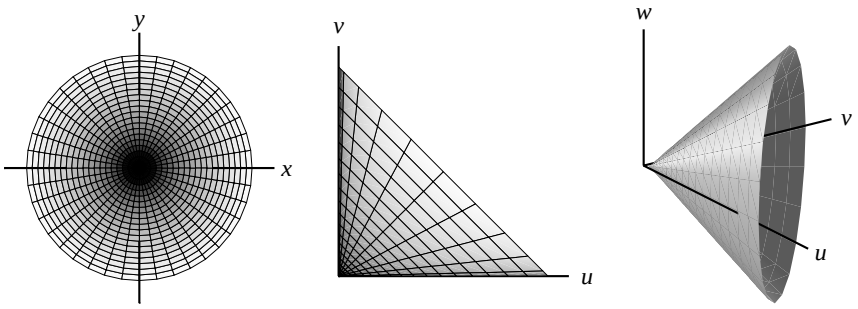
8.3.3 Analysis of a 2-D phase invariant space

We will investigate the simplest possible phase invariant space: the two-dimensional one described above, and illustrated in fig. 8.11(a). This could represent, for example, a sinusoidal signal chosen with random amplitude and phase. The idea is that this 2-D space might be hidden as a linearly transformed subspace in some higher dimensional input, such as a sequence of samples of an audio signal containing many frequencies. In §8.3.4, we will project these findings back in to the more familiar acoustic signal space.

Since we expect that the energy in the subspace will be a quantity of interest, we will consider the general quadratic reparameterisation

$$u = x^2, \quad v = y^2, \quad w = xy\sqrt{2}. \tag{8.33}$$

Thus, any member of the class of quadratic functions of (x,y), of which the energy $r^2 = x^2+y^2$ is one, can be written as a *linear* function of (u,v,w). This method of transforming nonlinearly to a higher-dimensional space where we hope the problem may be simpler to solve is a well-known trick—the new space is known as the *kernel space*, and in this case, we are using a quadratic kernel.



(a) (b) (c)

Fig. 8.11: Visualisations of a phase invariant space as (a) a circularly symmetric distribution in the *xy* plane; (b) the equivalent density in the *uv* plane and (c) the conical manifold on which the distribution lies in *uvw* space. (The gray scales in (a) and (b) are logarithmic.)

Distribution in quadratic kernel space

We wish to examine the probability distribution in the quadratic kernel space *uvw*. The first step is to obtain an expression for the probability density of the data in the *uv* space. The determinant of the Jacobian of the coordinate transformation is

$$|\mathbf{J}| = \begin{vmatrix} \partial u/\partial x & \partial v/\partial x \\ \partial u/\partial y & \partial v/\partial y \end{vmatrix} = \begin{vmatrix} 2x & 0 \\ 0 & 2y \end{vmatrix} = 4xy \tag{8.34}$$

The density function can then be written directly (including an extra factor of 4 to account for the fact that all 4 quadrants of the *xy* plane are mapped on to the first quadrant of the *uv* plane):

$$p(u,v) = 4\frac{p(x,y)}{|\mathbf{J}|} = \frac{h(x^2+y^2)}{xy} = \frac{h(u+v)}{\sqrt{uv}}. \tag{8.35}$$

An example of such a density is illustrated in fig. 8.11(b). From this, the distribution in *uvw* space can be computed. The three coordinates are constrained by $w^2 = 2uv$, which, as is shown in Appendix D, defines the surface of a *cone* with a circular cross-section and an axis which passes through the origin and points along the direction $(1,1,0)$. The cone is illustrated in fig. 8.11(c). The entire distribution in *uvw* space is concentrated on the surface of this cone, and hence, we cannot properly speak of a probability density, but we may consider $p(u,v,w)$ to be a density *in the surface*; that is, a probability mass per unit area of cone, rather than per unit volume.

The mapping from the *xy* disc to the *uvw* cone is not one-to-one: it is quite straightforward to show that a rotation of θ around the origin in the *xy* plane corresponds to a rotation of 2θ around the axis of the cone in *uvw* space, and thus a trip once round the disc in *xy* maps to a path twice round the cone.

We may also show the distribution is uniform around the axis of the cone. Perhaps the simplest way of doing this is to consider how $p(u,v)$ can be obtained from $p(u,v,w)$

by projecting the surface of the cone on to the *uv* plane. The details are given in Appendix D, but the result is that

$$p(u,v,w) = \frac{h(u+v)}{\sqrt{2}(u+v)}. \tag{8.36}$$

That this density is a function of $(u+v)$ only, that is, the distance from the apex of the cone, shows that the density on the surface of the cone is uniform around its axis.

Given this conical distribution and our desire to find phase invariant functions, a sensible linear basis in this *uvw* space would consist of one vector pointing down the axis of the cone and two others in a plane perpendicular to this and to each other, *e.g.* one in *w* direction and other in $(u,-v)$ direction. Will an unsupervised method such as ICA be able to find this basis?

Distribution of linear projection

To answer this question, we will look at the probability density of a general linear function of (u,v,w), which may be thought of as the dot product of a weight vector (a,b,c) with a data vector (u,v,w), giving $s = au+bv+cw$. ICA with a super-Gaussian prior tends to find such projections that are as 'peaky' as possible. In terms of *x* and *y*, we have

$$s = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & c/\sqrt{2} \\ c/\sqrt{2} & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{8.37}$$

Though it would be possible to work directly with this quadratic form, it will simplify matters considerably if we take advantage of the symmetry of the problem. The central matrix in eq. 8.37 is symmetric, and therefore admits of an eigenvalue decomposition, so it will always be possible to rotate to new frame of reference in which it is diagonal and $c = 0$. Since the probability density $p(x,y)$ is circularly symmetric, we can do this without a change of variables. In *uvw* space, this means rotating the vector (a,b,c) around the axis of the cone until it lies within the *uv* plane.

To compute the distribution of *s*, we must integrate the two-dimensional density $p(x,y)$ along a family of curves defined by the quadratic form. Leaving the details to Appendix D, the results are summarised below. Variation of the direction of (a,b) yields a family of distributions. As the direction (a,b) rotates from $(1,1)$—which points down the axis of the cone—towards the edge, the (single-sided) distribution becomes 'peakier.' As it crosses the edge of the cone, it becomes double sided but asymmetric, until finally, it becomes symmetric when $(a,b) = (1,-1)$, perpendicular to the axis. Because the distributions are asymmetric, measuring non-Gaussianity by kurtosis is not very meaningful, but it is quite clear that the distribution is at its 'peakiest' when the direction (a,b) points down the edge of the cone. This suggests that linear methods that search for maximally non-Gaussian projections, of which ICA is an example, will not find the phase invariant projections in quadratic kernel spaces such as this. Preliminary experiments have shown this to be the case: an application of ICA in this 3-D kernel space results in the basis vectors pointing approximately down the edges of the cone as predicted.

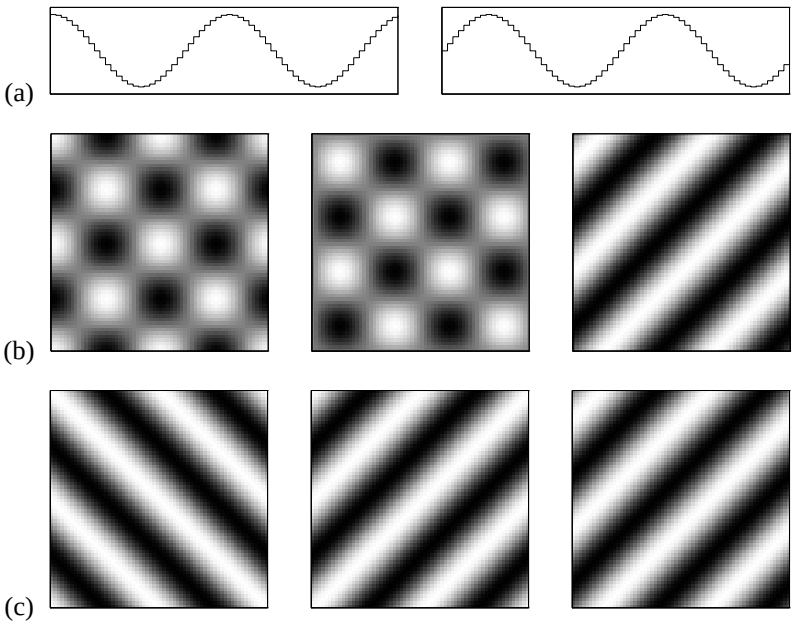


Fig. 8.12: Two alternative bases of the 3-D quadratic kernel space. (a) The two vectors $\mathbf{i}$ and $\mathbf{j}$ that span the phase invariant signal space. (b) The matrices $\mathbf{i}\mathbf{i}^T$, $\mathbf{j}\mathbf{j}^T$ and $\mathbf{i}\mathbf{j}^T + \mathbf{j}\mathbf{i}^T$, which span the quadratic kernel space. (b) The matrices $\mathbf{i}\mathbf{i}^T + \mathbf{j}\mathbf{j}^T$, $\mathbf{i}\mathbf{i}^T - \mathbf{j}\mathbf{j}^T$, and $\mathbf{i}\mathbf{j}^T - \mathbf{j}\mathbf{i}^T$, which form an alternative basis of the quadratic kernel space.

8.3.4 Relevance to Audio Signals and the Wigner Distribution

How do these low-dimensional phase invariant spaces project in to the high-dimensional space of acoustic signals? Let us assume that the 2-D space investigated in the previous section is actually the subspace spanned by two sinusoids of the same frequency but in quadrature phase. Let us also assume for ease of visualisation that the signals are sampled in discrete time and are of finite length, n samples, so we may denote them by $\mathbf{r} \in \mathbb{R}^n$. The 2-D phase invariant space is now a subspace of $\mathbb{R}^n$. The 3-D quadratic kernel space, however, is a subspace of a space of dimensionality $m = \frac{1}{2}n(n+1)$, the space of symmetric matrices $\mathbf{r}\mathbf{r}^T$. Given a vector $\mathbf{r}$, the matrix $\mathbf{r}\mathbf{r}^T$ consists of all possible pair-wise products of elements of $\mathbf{r}$, and so any quadratic function of the vector can be written as a linear function of the matrix. This includes all the quadratic time frequency representations: the spectrogram and the rest of Cohen's class.

The 3-D kernel space was parameterised using the three coordinates u, v and w. In the m dimensional kernel space, these map to three basis 'vectors,' (though they are best visualised as $n \times n$ matrices) which we can specify as follows: Let the two sinusoidal basis vectors be $\mathbf{i}, \mathbf{j} \in \mathbb{R}^n$. These are illustrated in fig. 8.12(a). An input signal $\mathbf{r}$, a sinusoid of unknown amplitude and phase, may then be written as $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$, where x and y are assumed to have a circularly symmetric distribution. This maps to the quadratic kernel space as

$$\mathbf{r}\mathbf{r}^T = x^2\mathbf{i}\mathbf{i}^T + xy(\mathbf{i}\mathbf{j}^T + \mathbf{j}\mathbf{i}^T) + y^2\mathbf{j}\mathbf{j}^T. \tag{8.38}$$

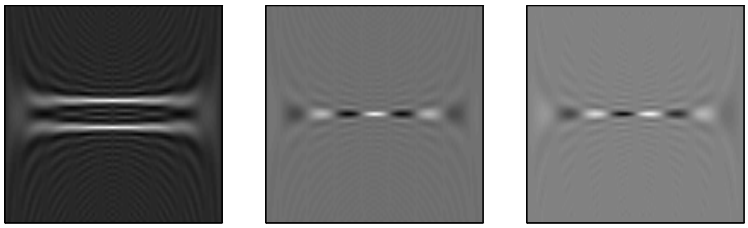


Fig. 8.13: Basis of 3-D quadratic kernel space in the Wigner domain. Note that these are *not* the Wigner distributions of any particular signals, but are obtained by rotating the matrices in fi g. 8.12(c) 45° anticlockwise and Fourier transforming vertically.

We can see by inspection that the u direction is $\mathbf{i}\mathbf{i}^T$, the v direction is $\mathbf{j}\mathbf{j}^T$, and w direction is $(\mathbf{i}\mathbf{j}^T + \mathbf{j}\mathbf{i}^T)/\sqrt{2}$. These are illustrated in fig. 8.12(b). An alternative basis, which yields the phase invariant energy as one of the coordinates, consists of one vector pointing down the axis of the cone and two others perpendicular to it. In the full kernel space, the first becomes $\mathbf{i}\mathbf{i}^T + \mathbf{j}\mathbf{j}^T$, and the other two could be, for example, $\mathbf{i}\mathbf{i}^T - \mathbf{j}\mathbf{j}^T$ and $\mathbf{i}\mathbf{j}^T + \mathbf{j}\mathbf{i}^T$. These are illustrated in fig. 8.12(c).

Why should these be of any interest? It has already been noted that all the quadratic time-frequency representations can be obtained as linear functions in the quadratic kernel space. The Wigner Distribution in particular is invertibly related to the matrix $\mathbf{r}\mathbf{r}^T$, essentially by a one-dimensional Fourier transform along the anti-diagonal of the matrix. The basis vectors found above, being plane waves, all have trivial Fourier transforms, which are illustrated in fig. 8.13. The phase invariant 'energy' direction shows up as the positive and negative frequency components of a sinusoidal signal, while the phase varying components appear as two sets of cross-terms in quadrature phase. Thus, the structure of the cross-terms in the Wigner Distribution is closely related to the structure of phase invariant spaces discussed here.

Note that this conclusion has been drawn from a particular assumption about what constitutes a phase invariant subspace, namely, that spanned by two sinusoids in quadrature phase. This is only an assumption: if an ecological analysis, perhaps using ICA or ISA, reveals that other subspaces exhibit the characteristic spherically symmetric distributions, then different notions of what constitutes a cross-term in the Wigner distribution would result.

Why the Wigner distribution is not suitable for ICA

Many of the time-frequency distributions that are in common use, such as the spectrogram, the wavelet transform energy or scalogram, the Wigner distribution, and the correlogram, can all be computed as linear functions of the quadratic object $\mathbf{r}\mathbf{r}^T$. This suggests that it ought to be possible to design an unsupervised system that finds an optimal representation as a linear function in the quadratic kernel space, hopefully resulting in the discovery of the most appropriate 'phase invariant' representation for a given class of signals, however this is defined.

The results of this section show that signals that contain phase invariant subspaces

have an intricate structure in the quadratic kernel space, and naive applications of linear methods such as PCA or ICA are unlikely to be successful, and some thought will be required to design an appropriate algorithm.

The alternative is to avoid moving into the quadratic kernel space at all, which is precisely what the ISA algorithm is designed to do. The final demonstration of this section is that *any* quadratic function of the data can be computed by an ISA-like model with a layer of linear 'simple cells' and a layer of 'complex cells' that form weighted energy averages. Let the quadratic function be

$$f(\mathbf{x}) = \sum_{i,j} A_{ij} x_i x_j = \mathbf{x}^T \mathbf{A} \mathbf{x}, \tag{8.39}$$

where $\mathbf{A}$ is without loss of generality symmetric. Hence, there exists an eigenvalue decomposition $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$, with $\mathbf{U}$ orthogonal and Λ diagonal. This means that

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{U}\Lambda\mathbf{U}^T \mathbf{x}. \tag{8.40}$$

Letting $\mathbf{y} = \mathbf{U}^T\mathbf{x}$, and assuming that both $\mathbf{x}$ and $\mathbf{y}$ are both n-dimensional, this becomes

$$f(\mathbf{x}) = \mathbf{y}^T \Lambda \mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2, \tag{8.41}$$

showing that any such function can be written as a weighted sum or difference of squares of a suitable set of linear functions of $\mathbf{x}$. This is a sort of generalised 'complex cell', which is allowed inhibitory as well as excitatory connections from the preceding layer of linear 'simple cells.' If many quadratic functions are required, then each 'complex cell' requires at most n 'simple cells' to feed it. Since these may or may not be sharable, the initial linear representation may have to be overcomplete.

8.3.5 Relationship with topographic representation

According to the framework adopted in this chapter, the distinguishing feature of phase invariant subspaces is the spherically symmetric data distribution within them. Independent subspace analysis or ISA (Hyvärinen and Hoyer, 2000) is specifically designed to find these subspaces in high-dimensional data, but the dimensionality of the subspaces must be pre-specified. One way to avoid this requirement is to use the methods of §8.1, since, unless the variables of a spherically symmetric distribution are jointly Gaussian, they will be statistically dependent. The two dependent components in fig. 8.2 show precisely this circular symmetry.

If a linear space does indeed separate into a number of independent subspaces, then the geometric analysis of residual dependency using MDS should reveal a number of distinct clusters, each of which defines a candidate phase invariant subspace. The dimensionality of these subspaces is not predetermined and would emerge from the analysis. The energy in each subspace would be computed by summing the energies of the components in the cluster. Since the clusters should be well separated (because they are assumed to be independent) this can be viewed as a kind of *local averaging* in the MDS embedding space.

This sort of local processing in a topographic representation has been proposed by Schwartz and Simoncelli (2001), who advocate *divisive normalisation*, motivated by the observation that in some distributed representations, a unit's variance can be predicted by the energy of its neighbours. This is a form of redundancy which can be removed by dividing the unit's activity s_i by its predicted variance:

$$\phi_i = \frac{s_i^2}{\sigma_i^2 + \sum_{j \in \mathcal{N}_i} w_{ij} s_j^2}, \tag{8.42}$$

where $\mathcal{N}_i$ is the set of indices of the ith unit's neighbourhood, and the w_{ij} are fixed weights. The predicted variance has the form of a local average with an extra additive term, σ_i^2, to model any residual variance not accounted for by the local energy.

Schwartz and Simoncelli present this in terms of a linear representation with a *predefined topology* (such as a wavelet transform) so that the neighbourhoods can be identified. Both topographic ICA and the method presented here provide data-driven ways of finding a suitable representation in which to do divisive normalisation.

The spherically symmetric distributions of phase invariant subspaces and the local activity correlations of topographic ICA can both be modelled by a *multiplicative* generative process. Consider the two-dimensional space analysed in §8.3.3. If the components x and y are generated according to

$$x = \sigma\phi_x, \qquad y = \sigma\phi_y, \tag{8.43}$$

where σ is a random multiplier, and ϕ_x and ϕ_y are random with a circularly symmetric distribution, then x and y will also have a circularly symmetric distribution. The benefit of this is that ϕ_x and ϕ_y (which could be called 'phase variables') may now be assumed to have a circular *Gaussian* distribution, and hence be *independent* as well circularly symmetric. The fact that they are Gaussian and uncorrelated is perhaps consistent with the idea that phase is a form of multiplicative 'noise.' Given an appropriate distribution of σ, a wide class of circularly symmetric non-Gaussian distributions of the sort illustrated

in fig. 8.2 may be obtained in x and y. This sort of model is called a scale-mixture of Gaussians, which Wainwright and Simoncelli (2000) use to provide a model-based justification of divisive normalisation. Similarly, in the generative model behind topographic ICA, a locally smooth 'energy profile' is multiplied by a set of uncorrelated Gaussian 'phase variables.' The local averages computed by the model 'complex cells' in ISA and topographic ICA are estimates of the scale factors, while the phase variables are discarded.

Note that divisive normalisation appears to be doing the *opposite* of what complex cells do. The complex cells compute local averages, which are assumed to be useful because they begin to capture some invariant structure. Divisive normalisation on the other hand, computes the local averages in order to *remove* them from the input. The result of this normalisation is, in present framework, the phase variables, that is, the 'noise'. Perhaps the real benefit of the divisive process is that it may achieve a *factorial split*, dividing the representation into two sets of variables (the energies and the

phase variables) which, it is hoped, are independent of each other. Either set may have significant non-random structure and thus be an interesting object of further analysis.

Summary

In this chapter, the concepts of similarity and distance were examined from a number of viewpoints. A distinction was made between two types of similarity. The first, P-similarity, pertains to complete sensory scenes and the distributed patterns of activity that might be used to represent them. The second, R-similarity, pertains to the individual units in a distributed representation. In general, the two are distinct, except in the case of certain representations such as 'code-book' vector quantisers or categorisers.

Topographic representation through dependency analysis and MDS It was proposed that statistical dependencies in a distributed representation be used to define not just neighbourhood relationships, as in topographic ICA (Hyvärinen et al., 2001), but a complete *metric*, by defining a mapping from mutual information to distance. This idea was applied to the distributed representations arising from the ICA experiments of Chapter 5. The mutual information between each pair of components was estimated using a method based on measuring correlations between the nonlinearly transformed components. The functions used were chosen to bring out the dependence between pairs of components with symmetric, heavy-tailed joint distributions.

The mutual information estimates were converted into distances using a heuristically derived mapping based on consideration of a certain kind of Gaussian noise process. MDS was then used to to derive low-dimensional spatial representations of the system of distances, resulting in a geometric visualisation of the relationships between the ICA components.

The configurations derived from the speech and music results of Chapter 5 were very different, but both representations had significant residual dependency structure that gave rise to some interesting geometric forms. This should help guide the design of subsequent stages of processing in an artificial perceptual system based on the principle of redundancy reduction as discussed in Chapter 2.

The relationship between distance and noise It was suggested that an ecological measure of P-similarity should be based on the likely significance of difference, and hence that a model of psychological distance could be derived from a consideration of noise. It was shown that several well-known metrics can be derived from an appropriate noise model; for example, a Euclidean metric arises naturally for a system with additive Gaussian noise. As a generalisation of this idea, a few methods for inducing metric structure from probabilistic structure were proposed, mostly based on the Kullback-Leibler divergence as a measure of distance between probability distributions. Such methods have the advantage that they are invariant to a transformation of variables since the Kullback-Leibler divergence is itself invariant to such transformations (Kullback, 1968).

The structure of phase invariant spaces The discussion of phase invariance was prompted by the suggestion (Hyvärinen and Hoyer, 1999) that random variables with a spherically symmetric probability density form a kind of 'phase invariant' subspace. This has implications for the statistical structure of such spaces—in that variables in the subspace will be dependent and thus amenable to analysis with the methods of §8.1—and for the sorts of generative model that could give rise to spherically symmetric data distributions, an important example being a multiplicative generative model. Such a multiplicative model is at the heart of topographic ICA, and also the divisive normalisation models of Schwartz and Simoncelli (2001) and Wainwright and Simoncelli (2000); thus all these systems are closely related.

Finally, the structure of a two-dimensional phase invariant space was analysed. Since the energy in any subspace of any high-dimensional linear space is a quadratic function of the components, the circularly symmetric 2-D distribution was projected into a 3-D quadratic kernel space, in which all such quadratic functions become linear. It was found that this 3-D space has a very particular structure in which there are 'special' directions suitable for forming a 'best basis,' but that these directions are unlikely to be found by methods which maximise super-Gaussianity, such as ICA with a super-Gaussian prior. It was found, however, that when the toy two-dimensional system is projected into a multidimensional space of audio signals, a link between the phase invariant structure and cross-terms in the Wigner distribution emerges, suggesting a new way of looking at these cross-terms, and perhaps a new approach to removing them.

9. CONCLUSIONS AND FURTHER WORK

In this final chapter, the conclusions of the previous four are summarised: the experiments with ICA, sparse coding, and MDS in §9.1, §9.2 and §9.3 respectively, along with any direct consequences in terms of work to be done. The final section suggests some more open ended ideas for further work to do with ways in which probabilistic models can deal with temporal structure.

9.1 Independent Component Analysis of Speech and Music

ICA was applied to two audio ensembles drawn from radio broadcasts: BBC Radio 4, which broadcasts mainly speech, and BBC Radio 3, which broadcasts mainly classical music. The data was represented as blocks of 512 consecutive samples. The result was two linear representations, adapted to speech and music respectively, which in many respects are roughly comparable with Fourier and wavelet representations, but with different time-frequency tilings. Thus, wavelets and sinusoids do indeed seem to be the 'independent components' of speech and music. Further analysis of the residual dependencies using MDS revealed some interesting geometric structure in the bases.

On the one hand, the results suggest that wavelet and Fourier transforms are useful because they are efficiently computable approximations to an optimal linear redundancy reduction for certain types of data. On the other hand, this work demonstrates that these structures are implicit in the statistical structure of sound: they need not be assumed *a priori* and can be discovered through unsupervised learning. It also shows the importance of the ecological approach, showing that the 'environment,' (in this case, an auditory environment) really does have a large impact on the results obtained by approximating an optimal ICA representation.

9.1.1 Speech-Derived Results

The speech basis vectors were essentially temporally-localised sinusoids, except at the lowest frequencies, where the basis vectors spanned the entire 46 ms window visible to the system. The relationship between the basis vector bandwidths and their centre frequencies was not quite the simple proportionality found in a constant-Q wavelet basis—refer back to fig. 5.3 for more details. When plotted as Wigner distribution contours, the basis vectors formed a fairly complete covering tiling of time-frequency plane, with little overlap or unoccupied space.

When the residual dependencies in the representation were estimated using non-linear correlations and visualised spatially using MDS, the elements took up the form of a two-dimensional manifold, arranged almost perfectly according to the time and frequency positions of the basis vectors. Thus, the very notion of a *two* dimensional *time-frequency* representation was found to be inherent in the statistical structure of speech.

9.1.2 Music-Derived Results

The music basis vectors tended to be less temporally localised, including many sinusoids spanning the whole window, similar to those in a Fourier basis, though the frequencies were not distributed as uniformly as in a Fourier basis. In addition, there were some short-duration, high-frequency wavelets, and some basis vectors that could not easily be characterised and were not well localised in time-frequency. The distribution of frequencies and bandwidths showed some evidence of the 12-semitones-per-octave found in Western music, with more sharply tuned vectors alternating with less sharply tuned ones at the rate of 12 per octave, though the effect was quite weak.

The residual dependency structure showed clear evidence of this 'semitone' effect, as well as strong dependencies between vectors with centre frequencies standing in harmonic ratios. When visualised in 3-D, some of this structure was visible. The overall form was that of a frustum of a cone, with high frequencies at the narrow end, sharply tuned units (forming the chromatic scale) towards the surface, poorly tuned 'microtones' towards the axis, and an approximate circle of Fifths around at least part of the surface towards the low frequency end.

9.1.3 Future Developments

Application to more data sets An obvious extension to the work here is to apply ICA to some different ensembles: more percussive music may produce a more wavelet-like basis; musics from other cultures with different scale structures may produce alternative geometric structures. More results of this kind would also begin to give us some idea of how reproducible these bases are; preliminary results suggest that they are quite reproducible.

As more powerful computers become available, it will become practical to use longer blocks of audio data: it will be interesting to see at what point the full-width sinusoids in the music basis become localised wavelets. This will also improve the frequency resolution at the low end.

Comparative analysis of redundancy A quantitative analysis of entropy and redundancy in Fourier and wavelet representations, as well as those derived by ICA, will allow a strict comparison between them: presumably, the ICA representation will have the lowest redundancy, but it is important to confirm this.

Comparison with cochlear filtering The speech basis vectors showed a clear relationship between centre frequency and bandwidth; which should be compared with

physiological and psychoacoustic data about filtering and frequency selectivity in the human auditory system. This may require training an ICA system on a more environmentally representative ensemble than one composed entirely of speech.

Bandwidth structure in speech basis On a related note, it would be interesting to investigate what if any significance there may be in bandwidth structure of the speech basis between 1 and 2 kHz. Does it have anything to do with the structure of formants in speech?

Model driven normalisation The experiments performed here required a gain control stage before the ICA module to maintain stability. The method used was chosen in an ad-hoc manner and was not based on any explicit model of the signal. By considering a multiplicative model of the signal, in which the ICA mixtures are multiplied by a slowly varying scale factor, it may be possible to implement a model based normaliser driven jointly by feedback from the ICA model and a prior model of scale factor variations.

Mixtures of ICA and PCA to model silence It was observed that, during periods of silence, the gain control stage would quickly increase the gain. When the next sound actually arrived, it would be greatly amplified and in some cases would destabilise learning. One way of dealing with this would be to suppose that during periods of silence, gain control should be suspended on the basis it contains no information about the likely loudness of the next sound. The hypothesis is that it is not *silence*, but *quiet sounds* that lead us to expect more quiet sounds, at least over moderately short periods. This is equivalent to a formal mixture model, where 'silence' is modelled separately from non-silence (perhaps as low-level Gaussian noise) so that the non-silence model (including the gain control stage) adapts only when the signal is badly modelled as 'silence.' Roberts and Penny (2001) introduced the mixtures of ICA model, which should be applicable in this instance.

Faster implementation using a sparse weight matrix The ICA bases contained many sinusoids and wavelets, and so it may be possible to use a fast, $O(n \log n)$, Fourier or wavelet transform as a preprocessor. Since both are invertible linear transformations, this should have no effect on the end result, but could allow the implementation of a *sparse* ICA weight matrix, and perhaps an improvement on the current $O(n^2)$ performance.

9.2 Sparse Coding

9.2.1 Sparse Coding Algorithm

A modified 'active set' quasi-Newton optimisation algorithm was developed to deal with the discontinuities in the gradient of the objective function used in maximum *a posteriori* estimation in the sparse coding model. This was found to yield performance improvements under certain conditions.

Experiments with a toy data set showed that a certain amount of sparsity is required to find an overcomplete basis; below this level, only a complete basis is found.

An analysis of learning in a one-dimensional version of the sparse coder yielded some insights into the behaviour of the multidimensional version. The scaling of the basis vectors was shown to be related to their sparsity in the input, due to a mismatch between the assumed prior and the actual distribution of the input. Furthermore, approximations made in the derivation of the learning rule result in the basis vectors converging to zero, rather than the correct value, when the sparsity of the input drops below a certain threshold determined by the assumed noise level.

The analysis also showed that the hybrid 'sparsified' Laplacian prior was largely ineffective in achieving its stated aim, which was to approximate the sources as mixtures of zeros and Laplacian random variables. It did not correct the scaling of the basis vectors, and meant that the posterior distribution of the source given the data could be multimodal. Neither the gradient optimiser, nor the approximations used in the derivation of the learning rule, are equipped to handle this situation.

On the other hand, the hybrid prior did enable quite large gains in performance, though in a somewhat unprincipled way, by forcing many outputs to remain inactive. This would usually prove detrimental to learning, but if the basis had already been learned, the performance gains were substantial. In the musical application (see below) it meant the transcription was 'cleaner,' with fewer extraneous notes, and could perform in real-time.

9.2.2 Application to Music Spectra

The sparse coder was applied to short-term Fourier magnitude spectra derived from some synthesised harpsichord music. The basis vectors, the 'independent components,' were found to be the *note* spectra, and so the system as a whole, after training, was essentially doing an optimal linear decomposition of the input spectra into a set of atomic note spectra, but where the atomic spectra were themselves learned from the music in an unsupervised fashion. Thus, it could form the basis of an automatic transcription system, and was found to be capable of doing audio to midi conversion even with an extremely simple final stage. However, this was with a very restricted data set, and it remains to be seen how it performs with other types of music played on real instruments. The notes produced by the synthetic harpsichord were very consistent, and real instruments produce much more variable sounds.

Regardless, this was not the main point of the experiment: we do not expect to build a polyphonic transcription on a simple linear generative model like this. What it showed was two things. Firstly, there is enough structure in music for musically relevant aspects to emerge through unsupervised learning; in this case, it was the independent existences of notes, each with a certain spectral structure. Secondly, given the limitations of the linear generative model, the system performed quite well, but at a high computational cost. In general, it seems that probabilistic inference in this sort of noisy, multiple-cause graphical model is a powerful technique, but cannot be applied to large amounts of data without some restrictions on the graph topology to allow more

efficient inference, or preprocessing to reduce the number of variables involved.

9.2.3 Future Developments

Improve noise model The current version assumes white noise in the spectral domain. Even if a white noise model was appropriate in the time domain, observation of the Fourier magnitude spectra suggests that the 'noise-like' activity is not distributed uniformly across the spectrum, but tends to be proportional to signal activity—this is a manifestation of *spectral leakage* and fundamentally, is due to phase information 'leaking' into the supposedly phase invariant magnitude spectrum because of the windowing process. This could be modelled as a multiplicative noise process, but since there is generally more activity at the lower end of the spectrum, it would be much simpler, as an initial step, to assume additive, uncorrelated, but non-uniform Gaussian noise, with a smaller variance at high frequencies. This would result in the higher harmonics of musical notes receiving more weight in the sparse coder. Alternatively, the magnitude spectrum could be abandoned as an input representation in favour of a self-organising 'phase invariant' representations of the sort described in § 8.3.

Binary categorisation using mixture models It was noted in Chapter 7 that the marginal distributions of the basis vectors were bimodal, and that this was probably related to the binary nature of note activation. It may be possible to formalise this with an explicit mixture model. Olshausen and Millman (2000) developed a sparse coder using a mixture-of-Gaussians prior, but again, this is at a high computational cost due to the need to integrate over a multimodal posterior.

Fast redundancy reduction as a preprocessing step As mentioned before, the main obstacle to using sparse coding is the computational cost. One way of overcoming this would be to reduce the connectivity of the system considered as a network. It may be possible to do this by using a more efficient method such as ICA, to do an initial phase of redundancy reduction, followed by a topographic organisation to localise the residual dependencies. Then, a sparse coder could be implemented with less than full connectivity. Hoyer and Hyvärinen (2002) have made steps in this direction but their algorithm does not exploit the expected locality of the residual dependencies.

Modelling temporal dependencies The initial experiments with two-dimensional spectrogram patches suggested that note onsets would be an important feature; the use of wider patches should result in more reliable note detection by learning the typical profile of a note in time as well as frequency. The main obstacle to this is the large amount of data involved in representing a patch. Unlike visual images, these spectrogram 'auditory images' are not translation invariant in the frequency direction and hence the patches used as input must span the whole frequency range.

Alternative overcomplete representations The sparse coder used in the present work is not the only system capable of generating overcomplete representations: other,

possibly less computationally burdensome methods have been proposed: Hyvärinen et al. (1999) have developed a method based on *quasi-orthogonality*, and Hinton et al. (2001) suggested a generalisation of Hinton's *product of experts* (PoEs) model which may be applicable.

9.3 Distance and Similarity

9.3.1 Geometric Representation using MDS

The residual dependencies in the ICA representations developed earlier were estimated using correlation of rectified activities. A mapping from correlation coefficient to mutual information to distance was proposed, and the resulting distances used to generate spatial configurations of the representations in 2 and 3 dimensional Euclidean spaces using multidimensional scaling. (Results already summarised in §9.1.)

The nonlinear correlation technique seemed to be quite effective at capturing the dependency when compared with direct estimation of the mutual information using joint histograms, though a more rigorous comparison is required.

9.3.2 Distance and Noise

The relationship between noise and similarity was investigated and a definition of distance in a noisy probabilistic model was proposed on the basis that distance should be related to significance of difference, and significance is related to noise, in that noise is that which is insignificant.

9.3.3 Phase Invariance

The idea that phase invariant spaces are related to spherically symmetric probability distributions was explored, and the two-dimensional case was analysed. The data was found to take on a conical distribution in a quadratic kernel space. One of the conclusions was that searching for phase invariant functions by using ICA in a quadratic kernel space (in particular, the Wigner distribution) would be unsuccessful due to the statistical structure of the data in such a space. However, the analysis did reveal a connection between the structure of the two-dimensional phase invariant subspace and the cross-terms in the Wigner distribution.

9.3.4 Future Developments

Estimation of mutual information by nonlinear correlation A further investigation of the relationship between nonlinear correlation and mutual information is required, and a comparison with other ways of estimating the mutual information. In particular, a proof or otherwise of the conjecture (eq. 8.5) that the MI is bounded by the non-linear correlation is needed, and a more rigorous evaluation of estimation procedure with known distributions should be carried out.

Transformation of target distances Investigation of the effect of a nonlinear transformation of the target distances before the application of MDS. Buja et al. (1998) use this method to combat what they call 'indifferentiation,' which arises when the distances cluster around a non-zero mode and produce an uninformative spherical configuration.

Non-Euclidean spaces Attneave (1950) has argued that 'psychological space' is intrinsically non-Euclidean due to a lack of rotation invariance. He suggested that the 'city block' distance may be more appropriate, and thus may be of use in MDS.

Statistically motivated stress functions The relationship between stress functions and statistical estimation was pointed out in §8.1.4, and an alternative stress function derived from considering the expected sampling distribution of the correlation coefficient. It remains to implement this stress function and compare the results with those obtained earlier.

High dimensional configurations Although MDS was performed on the correlation data in spaces of dimension up to 8, it is difficult to visualise configurations of dimension above 3. Krumhansl (1990) and Shepard (1982) have both suggested that at least 5 dimensions are required to visualise the important pitch relationships, so a careful analysis of the MDS results in high dimensions is required to see if the geometry agrees with Krumhansl's and Shepard's models.

Local processing in MDS embedding space The purpose of the analysis of residual dependency is not primarily visualisation, but to facilitate further processing by detecting and localising those dependencies. There are a number of possibilities that fall under the general heading of 'local processing in MDS space.'

- MDS space filtering and resynthesis would involve multiplying the ICA activities by some continuous function of position in MDS space and then inverting the weight matrix to get back to the input space, in an exact analogue to frequency domain filtering. The end result would, of course, still be a linear transformation, not necessarily equivalent to convolutive filtering. Another possibility is nonlinear filtering in the MDS domain. It remains to be seen whether not this will have any practical applications.
- Schwartz and Simoncelli (2001) advocate *divisive normalisation*, which involves dividing each (linear) unit activity by a locally averaged (rectified) activity: they used energies averaged over time and frequency neighbours in a wavelet basis. In the present context, these averages could be taken over a neighbourhood in MDS space. Dividing by these local averages accomplishes a multiplicative decomposition. Schwartz and Simoncelli imply that normalised activities are the quantities of interest and the averages are nuisance variables, whereas in Hyvarinen's related multiplicative model for topographic ICA, it is the local averages which correspond to the hidden sources, the latent variables of interest, and normalised activities are modelled as a multiplicative Gaussian noise. It seems likely that both are partially correct,

and the important point is that the unnormalised variables are decomposed into two putatively independent sets, thus serving the goal of factorial coding and redundancy reduction. In principle, one would keep *both* sets of variables and look for any remaining structure using another unsupervised method. Hoyer and Hyvärinen (2002) partially implemented this idea by using non-negative ICA after topographic ICA; when trained on natural images, this system learned a form of contour coding.

- Phase invariant subspaces of the sort described in §8.3.2 could be found by looking for clusters of units in MDS space. The advantage of this approach over that used in independent subspace analysis (Hyvärinen and Hoyer, 2000) is that the dimensionality of the subspaces is not predetermined, but judged by comparing distances.

Note that many of these local processes could be implemented without actually generating a spatial configuration, and working directly from the distance matrix.

Phase invariance and the Wigner distribution It may be possible to use knowledge about the structure of the phase invariant subspaces in a signal to develop a new method for removing the cross-terms in the Wigner distribution based on the relationship between the cross-terms and the conical distribution found in §8.3.3.

Use of temporal dependencies The method presented here is based on measuring *instantaneous* correlations, essentially identifying representational units that are active together as 'similar.' This might be adequate in a universe where each moment was independent of the previous, but this is clearly not the case in the real world. A better alternative would be to allow temporal dependencies to influence the characterisation of similarity. For example, in music, the likely *succession* of notes could define a sort of 'horizontal' or 'melodic' relatedness distinct from the 'vertical' or 'harmonic' relatedness that would emerge from an analysis of simultaneously sounded notes. That topographic maps could be defined by the temporal flow of activity was suggested by Licklider (1959), who speculated that the maps in the brain may "favour the location side by side of maxima of activity produced by stimuli that frequently occur in close succession." In the case of auditory maps, he speculates that "in as much as glissandi are probably the most frequent orderly variation of stimulation, an ordinal representation of stimulus frequency arises."

The use of temporal dependencies would naturally give rise to *asymmetric* similarity relationships, and so the inability of spatial models to accommodate asymmetric 'distances' would have to be addressed.

9.3.5 An Ecological Characterisation of Noise

The proposed measures of distance in §8.2.3 depend on an ecological definition of noise as that part of sensory data which has no biological relevance. It may be possible to define this operationally in the following way: for any organism, one may suppose that there are certain intrinsically 'value-laden' signals such as those associated with

hunger or extremes of heat and cold. 'Relevant' signals are those which ultimately correlate with these intrinsically 'good' or 'bad' signals (or the V-set for short); conversely, noise can be defined as any component of sensory data which shows no such statistical dependency. The signals in the V-set are those which an animal ultimately 'cares about,' and noise is that which provides no information about them. The V-set amounts to a set of *reinforcement signals*, but unlike in reinforcement learning (Sutton and Barto, 1998), there is no desired response which brings a reward or punishment—this is still a form of unsupervised learning which passively learns the relationships between neutral sensory data and the V-set, using this to decide which data to keep and what to throw away as noise.

9.4 Time

> "All we composers have to work with are time and sound, and sometimes I'm not so sure about sound."
>
> —Morton Feldman

The methods implemented in this thesis have all been concerned with redundancy over a strictly limited temporal span, and only then through the expedient procedure of windowing the signal into blocks of consecutive samples, converting temporal structure into 'spatial' structure, as it were. One could contemplate dealing with longer-range dependencies in the same way, by just expanding the windows. However, this ignores three crucial points: (1) translation invariance through time: there are no special times and (2) as time-based observers, we cannot have access to data from the future: there is a horizon which continuously moves forward as time passes. This makes dealing with temporal redundancy quite different from dealing with spatial redundancy, as found in, for example, visual images. A visual image is available all at once, whereas auditory data trickles in bit by bit. There is a horizon, a 'now', beyond which no information is available. West et al. (1987) have commented on the "unfoldingness" of time. The third point is that the time dimension is not limited in the way that the spatial dimensions of an image are—temporal dependencies can potentially extend a very long way.

By common agreement, the structuring of time is at the heart of musical experience. For example, West et al. (1987) comment that,

> All information in music is ultimately derived from temporal patterns. This obviously extends beyond the encoding of mere pitch information to temporal organisation over periods of two hours or more. Therefore, it is fundamental to any account of music perception to explain how temporal information is encoded.

To deal with this structure using the methods advocated in this thesis would involve looking at the concepts of redundancy, information, factorial coding and sparsity for long sequences of patterns. Because of the "unfoldingness" of time, *prediction* will probably be an important aspect of the solution (Ellis, 1996).

9.4.1 Multiscale Temporal Structure

The potentially long range of temporal dependencies has implications for the amount of memory required to discover and deal with those dependencies. It is very likely that many natural signals and music contain long-range dependencies, which may require a multi-scale or hierarchical processing approach.

Schmidhuber's *reduced sequence descriptions* (Schmidhuber, 1992b,a) are an elegant and potentially very efficient way of dealing with long range dependencies, working hierarchically but without an explicit tree structure. The system is built from identical blocks whose job it is to predict sequences. Each successive stage is driven by the *failures* in prediction of the previous one. These are presumed to occur relatively infrequently, so later stages operate on progressively longer time-scales and are thus able to see longer range dependencies.

9.4.2 Probability Signals as Meta-data

Rhythmic patterns can be built out of almost any kind of opposition: sound / no sound, soft / loud, bright / dull and so on. The medium by which difference is expressed is (arguably) somewhat less important than the timely delivery of discernible differences. For example, Cambouropoulos (1998) uses a number of variations of the *identity change rule* in his local boundary detection model to determine the accentuation structure of the musical surface. Local discontinuities of any sort signal phrase boundaries or points of metrical significance such as beats or bar lines.

Loudness, pitch, timbre *etc.* are all *qualities of sound*, whereas rhythm is not a quality of any sound: it is an emergent property, a *gestaltqualität*. The 'atoms' of rhythm are pure durations, and rhythms are patterns of durations, but in order to delimit the durations, to outline the patterns, we need timely variations in qualities of sound.

A 'surprise' driven system and a visual analogy Imagine that we have a visual system consisting of a red paint detector and a green paint detector. Now suppose that we want to be able to see red or green painted circles. One option would be to attach a new modules responsible for detecting circles to the red paint detector. This would then be able to see red circles. However, unless there is a common language or protocol that both paint modules conform to, a separate circle detector tailored to the green paint module will be needed to see green circles. Since the two circle detectors are independent, and could, for example, be instances of an adaptive shape detection module, they are not necessarily constrained to agree on what a circle is; the system could not truthfully be said to have an abstract concept of 'circularity' that is independent of concrete (that is, red or green) circles. If a new third colour, or a texture was added to the environment, more independent shape detectors would be required, as the system would be unable to take advantage of the concepts of circularity it had already developed. Ideally, we would like a single module that understood circles independently of how they are rendered. On its own, it would be unable to see circles because, obviously, real circles have to be painted in one or other colour. However, given any set of paint detectors, the module would be able to find circles painted in a variety of colours. As

suggested earlier, what is needed is a common language for communication between any paint detector and a shape detector. That language could be based on probabilities.

Imagine now that the paint detectors are actually probabilistic systems that model the appearance of red or green paint respectively, and expect to see contiguous regions of their own colour. Such a model, when assigning probabilities to regions of an image, will be 'surprised' at the edges of any region of their own paint. Thus, the red paint module will generate a circular pattern of low probability around any red circle, and similarly for the green paint module. The pattern of probabilities for a painted circle will look the same coming from either the red or the green modules, and so the circle detector will have no trouble seeing both and green circles. This architecture extends naturally to any kind of texture, as long as a statistical model of that texture is available, and there is, in principle, no limit to how complex that texture can be—for example, we might be able to detect circles of one kind of tartan against another.

What can we conclude from this whimsical thought experiment? It is that, given a system built from a variety of probabilistic models, all specialised to deal with a certain class of data, the probability signals coming out of the models are a useful form of meta-data. Looking for patterns and shapes in this meta-data allows the abstraction of certain concepts one step removed from their concrete manifestations. Rhythm is an ideal candidate for such an analysis.

Modelling the surprise signal One can envisage a modular system in which each component produces a probability signal, and where separate modules are responsible for learning the structures in these new signals. The metrical structure of music will tend to produce repetitive patterns on several time-scales, and existing beat tracking algorithms (*e.g.* Large and Kolen, 1994) could usefully be employed to model these, but rather than being locked to a particular auditory 'front-end,' they would be able to extract metrical structure from any low level model capable of producing a probability signal. Regular modulations of pitch, timbre, loudness or any other quality would all appear identically to the beat tracker, and so the system could usefully be said to have an abstract concept of rhythm.

Another benefit of this probabilistic approach would be that the components responsible for learning and predicting the surprise signals could feed their predictions back to the low level units, telling them when surprises are likely: in effect, when to 'expect the unexpected.' This could be used to improve the performance of the low level components: when no surprise is expected, they can be made more robust to noise, requiring a lot of evidence to induce a change of state. When a surprise due, more weight can be given to incoming data so that state changes are facilitated at these times. Thus, transitions will tend to be quantised to the metrical grid.

APPENDIX

A. AUDIO SIGNAL NORMALISATION ALGORITHM

This appendix describes the normalisation procedure applied to audio signals before being presented to the ICA system described in Chapter 5. The aim of this process is to reduce the extent of loudness variations over medium to long time scales, not to remove short term dynamics.

Let $\phi(t)$ be a continuous-time audio signal. This is sampled and the samples are arranged into blocks of N so that the jth sample of the mth block is

$$u_j[m] = \phi(T[mL+j]), \qquad 1 \leq j \leq N, \tag{A.1}$$

where T is the sampling period and L is the block hop size. The block mean and standard deviation are computed using:

$$\mu[m] = \frac{1}{N}\sum_{i=1}^{N} u_j[m], \tag{A.2}$$

$$\sigma[m] = \left\{\frac{1}{N}\sum_{i=1}^{N}(u_j[m])^2\right\}^{1/2}. \tag{A.3}$$

Then, low-pass filtered versions of these are updated:

$$\bar{\mu}[m] = \eta_\mu \mu[m] + (1-\eta_\mu)\bar{\mu}[m-1], \tag{A.4}$$

$$\bar{\sigma}[m] = \eta_\sigma \sigma[m] + (1-\eta_\sigma)\bar{\sigma}[m-1], \tag{A.5}$$

where η_μ and η_σ are two adaptation rates. Finally, the jth element of the mth input vector $\mathbf{x}[m]$ is computed as:

$$x_i[m] = (u_i[m] - \bar{\mu}[m])/\bar{\sigma}[m]. \tag{A.6}$$

The adaptation rates η_σ was set to 0.01, which, given a sampling rate of 11.025 kHz and hop size of 512 samples, implies a time-constant of approximately 5 seconds. The adaptation rate η_μ was set to a much smaller value (10^{-4}) on the assumption that the DC offset associated with a particular audio source would be essentially constant.

B. SPECTROGRAMS

B.1 Computational Details

The algorithm used to compute the spectrograms in Chapter 7 is as follows. The audio signal $\phi(t)$ is sampled and windowed into blocks of length $2N$ such that the jth sample of the mth block is

$$u_j[m] = h(j)\phi(T[mL+j]), \qquad 0 \leq j \leq 2N-1, \tag{B.1}$$

where T is the sampling period, L is the hop size by which window is moved at each step, and $h(\cdot)$ is the windowing function. In the sparse coding experiments, a Hamming window was used:

$$h(j) = 0.54 - 0.46\cos\frac{\pi j}{N}. \tag{B.2}$$

The complex DFT (discrete Fourier transform) coefficient v_k for the kth frequency is given by

$$v_k[m] = \frac{1}{\sqrt{2N}} \sum_{j=0}^{2N-1} \left[\exp\frac{i\pi}{N}kj\right] u_j[m], \tag{B.3}$$

where $i = \sqrt{-1}$. The power in the kth frequency is obtained by taking the magnitude squared of the DFT coefficient, but in the experiments, the magnitude was used instead:

$$x_k = |v_k| = \sqrt{v_k^* v_k}, \tag{B.4}$$

where the $*$ operator denotes complex conjugation and the block number r has been dropped for clarity. For real-valued signals, the DFT coefficients form conjugate pairs, with $v_k = v_{2N-k}^*$. This implies that the magnitudes x_k and x_{2N-k} are the same and thus only the coefficients for $0 \leq k \leq N$ need be retained. The kth frequency itself is $k/2NT$.

B.2 Noise Statistics

The following analysis is intended to support the statements of § 7.1 concerning the use of magnitude spectra rather than power spectra as input to the sparse coder described in Chapter 6 and applied to music in Chapter 7.

Consider the DFT of eq. B.3 as a matrix operation on the vector $\mathbf{u} \equiv (u_0, \ldots, u_{2N-1})$ yielding the vector $\mathbf{v} \equiv (v_0, \ldots, v_{2N-1})$, such that $\mathbf{v} = \mathbf{F}\mathbf{u}$. If the window $h(\cdot)$ is rectangular (*i.e.* $h(j) = 1 \;\forall\; j$), then the matrix $\mathbf{F}$, with components $F_{kj} = \exp(-i\pi kj/N)$, is unitary, since $\mathbf{F}^{-1} = \mathbf{F}^\dagger$, where $\mathbf{F}^\dagger$ is the adjoint of $\mathbf{F}$ and is defined as $F_{kj}^\dagger = F_{jk}^*$. Unitarity is the generalisation of orthogonality to complex-valued matrices, and in this case means that Gaussian white noise (also known as spherical Gaussian noise) at the

input **u** appears as spherical Gaussian noise at the output **v**. Let $v_{(r)}$ and $v_{(i)}$ denote the real and imaginary parts of one of the components of the DFT. The energy s at that frequency is then

$$s = v_{(r)}^2 + v_{(i)}^2. \tag{B.5}$$

If the input were to consist purely of spherical Gaussian noise, then these two components would be Gaussian and uncorrelated due to the unitarity of the DFT. Now, for two identically distributed Gaussian random variables X and Y, both with variance $1/\lambda$, it is easily shown that $S = X^2 + Y^2$ is another random variable with a probability density function given by

$$p(s) = \tfrac{1}{2}\lambda e^{-\frac{1}{2}\lambda s}, \quad s \geq 0, \tag{B.6}$$

which is an exponential distribution, indicating that a spherical Gaussian signal appears as exponentially distributed noise in the power spectrogram, as stated in § 7.1. This analysis is not correct if an additional signal is present, but it does suggest that additive Gaussian noise on the signal will induce *super*-Gaussian, or heavy tailed, noise in the power spectrogram. The variable $R = \sqrt{S} = \sqrt{X^2 + Y^2}$, however, is distributed according to

$$p(r) = \lambda r e^{-\frac{1}{2}\lambda r^2}, \quad r \geq 0, \tag{B.7}$$

which, though not Gaussian, does decay approximately as a Gaussian for large r. This suggests that the noise in a magnitude spectrogram will not be heavy-tailed as it was in the power spectrogram.

Note that this analysis does not apply to non-rectangular analysis windows. It is only intended as a rough heuristic argument for using the magnitude spectrum rather than the power spectrum.

C. MULTIDIMENSIONAL SCALING

In this appendix, an algorithm for minimising MDS stress functions (see § 8.1.4) will be given. In the derivation below, a Euclidean metric will be assumed, but other metrics can easily be treated by substituting a suitable alternative to eq. C.2.

Let $\mathcal{A}$ be a set of N objects which we wish to arrange in an M-dimensional Euclidean space (E, d_E) by finding points $x_\alpha \in E$ for each $\alpha \in \mathcal{A}$ such that, for each pair $\alpha, \beta \in \mathcal{A}$, the target distance $d(\alpha, \beta)$ is matched as well as possible by the Euclidean distance $d_E(x_\alpha, x_\beta)$. This is done by minimising a stress function which measures the discrepancy between the two sets of distances. Let us assume the stress function has the form

$$J = \sum_{\{\alpha,\beta\}\subset\mathcal{A}} \tfrac{1}{2}\Big[d_E(x_\alpha, x_\beta) - d(\alpha,\beta)\Big]^2 w_{\alpha\beta}, \tag{C.1}$$

where the sum is taken over each distinct unordered pair $\{\alpha, \beta\}$ and $w_{\alpha\beta}$ is a weighting factor that controls the contribution of each pair to the sum. If $w_{\alpha\beta} = 1 \;\forall\; \alpha, \beta$ then we obtain the stress function J_1 defined in eq. 8.9; if $w_{\alpha\beta} = [d(\alpha,\beta)]^{-1}$, we obtain Sammon's stress function J_2 defined in eq. 8.10; if $w_{\alpha\beta} = [d(\alpha,\beta)]^{-2}$, we obtain the stress function J_3 defined in eq. 8.11. Other functions are permissible, but in this derivation, $w_{\alpha\beta}$ may not depend on the points x_α, as we intend to differentiate the stress with respect to these points.

In order to construct a steepest-descent optimisation algorithm, we will need to compute the derivative of the distance $d_E(x_\alpha, x_\beta)$ with respect to the point x_α. Assuming that the Euclidean space E is also a normed linear space with the usual 2-norm $\|\cdot\|$, we have $d_E(x_\alpha, x_\beta) = \|x_\alpha - x_\beta\|$ and, assuming $x_\alpha \neq x_\beta$,

$$\frac{\partial d_E(x_\alpha, x_\beta)}{\partial x_\alpha} = \frac{x_\alpha - x_\beta}{\|x_\alpha - x_\beta\|}. \tag{C.2}$$

Thus, the derivatives of the stress function are

$$\frac{\partial J}{\partial x_\alpha} = \sum_{\beta\neq\alpha} \frac{x_\alpha - x_\beta}{\|x_\alpha - x_\beta\|}\Big[d_E(x_\alpha, x_\beta) - d(\alpha,\beta)\Big] w_{\alpha\beta}. \tag{C.3}$$

We then iteratively update each point with the mapping

$$x_\alpha \mapsto x_\alpha - \eta \frac{\partial J}{\partial x_\alpha}, \tag{C.4}$$

where η is a positive step length parameter which is gradually reduced as the algorithm converges.

D. CONTOUR INTEGRATION IN PHASE INVARIANT SPACE

This appendix contains mathematical support for the results in §8.3.3 concerning a two-dimensional phase invariant space. Consider the circularly symmetric 2-D probability density

$$p(x,y) = h(x^2+y^2), \tag{D.1}$$

where $h(\cdot)$ is defines the radial profile of the distribution. Next, consider the projection of this 2-D density into the 3-D space *uvw* defined by

$$u = x^2, \quad v = y^2, \quad w = xy\sqrt{2}. \tag{D.2}$$

The resulting distribution is confined to a 2-D manifold within the 3-D space, defined by the constraint $w^2 = 2uv$. This is the equation of a cone, (illustrated in fig. 8.11 in Chapter 8) which can be shown by writing the constraint as a quadratic form:

$$\begin{pmatrix} u & v & w \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = 0. \tag{D.3}$$

This is to be compared with equation of what might be called the 'canonical' cone in 3-D: one with a circular cross-section, a 90° apex at the origin, and its axis along the *w* axis. The canonical cone is defined by $u^2+v^2=w^2$, (which should be clear by inspection) and translates into a diagonal quadratic form:

$$\begin{pmatrix} u & v & w \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = 0, \tag{D.4}$$

where the eigenvalues $(1,1,-1)$ are the distinctive signature of a circular cone with a 90° apex. It is easily verified that the matrix in eq. D.3 can be brought into this form using an eigenvalue decomposition. The eigenvector with the opposite sign to the other two indicates the axis of the cone, and in this case, is $\frac{1}{\sqrt{2}}(1,1,0)$. Thus, the axis lies half way between the *u* and *v* axes, both of which lie in the surface of the cone.

Distribution in three dimensions

It was shown in §8.3.3 by direction transformation that the probability density in the *uv* plane is $p(u,v) = h(u+v)/\sqrt{uv}$ (eq. 8.35). To derive the distribution in 3-D, we will consider how the conical distribution projects down into the *uv* plane. Since the

distribution is concentrated within a 2-D manifold, it does not properly have a density, but we may consider $p(u,v,w)$ to be a *surface* density; that is, a probability mass per unit area. A given small area of cone projects down into a smaller area of the uv plane because of the obliquity of the cone's surface. Instead of deriving the reduction factor geometrically, which would necessitate some rather intricate diagrams, we may proceed algebraically. Differentiating both sides of the constraint $w^2 = 2uv$ gives

$$2w\,\mathrm{d}w = 2u\,\mathrm{d}v + 2v\,\mathrm{d}u, \tag{D.5}$$

which may be written as a dot product,

$$(-v,-u,w)\cdot(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w) = 0. \tag{D.6}$$

This shows that small displacements in the surface of the cone must be perpendicular to the vector $(-v,-u,w)$ and hence that vector must be the surface normal. On projecting into the uv plane, the reduction factor is the cosine of the angle between this surface normal and $(0,0,1)$, which is

$$\frac{(-v,-u,w)\cdot(0,0,1)}{\sqrt{u^2+v^2+w^2}} = \frac{w}{\sqrt{u^2+v^2+2uv}} = \frac{\sqrt{2uv}}{u+v}. \tag{D.7}$$

Taking into account an extra factor of 2 because of the two halves of the cone above and below the uv plane, the final projected density in the uv plane is

$$p(u,v) = 2\cdot\frac{u+v}{\sqrt{2uv}}\cdot p(u,v,w). \tag{D.8}$$

Using the previous expression for $p(u,v)$ (eq. 8.35) and solving for $p(u,v,w)$ gives

$$p(u,v,w) = \frac{\sqrt{2uv}}{2(u+v)}\cdot\frac{h(u+v)}{\sqrt{uv}} = \frac{h(u+v)}{\sqrt{2}(u+v)}, \tag{D.9}$$

as stated in eq. 8.36.

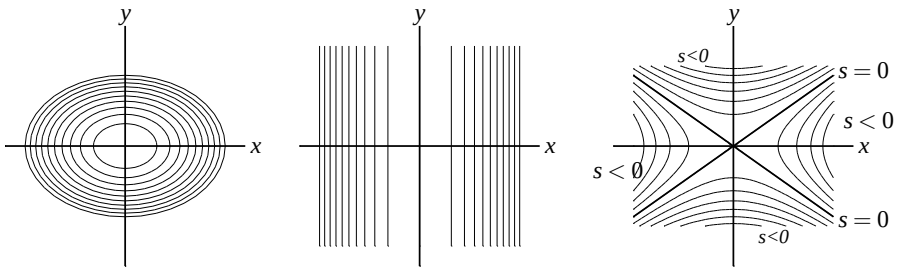
Distribution of projection on to line

Consider a general linear function of the point (u,v,w), $s = au+bv+cw$. This can be written in terms of the original variables x and y as a quadratic form:

$$s = \begin{pmatrix} x & y \end{pmatrix}\begin{pmatrix} a & c/\sqrt{2} \\ c/\sqrt{2} & b \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{r}^T\mathbf{Q}\mathbf{r}, \tag{D.10}$$

where $\mathbf{Q}$ denotes the the central matrix and $\mathbf{r}$ is the 2×1 matrix $(x,y)^T$. Following the discussion of § 8.3.3, we may assume without loss of generality that $c = 0$.

Each value of s defines a curve in the xy plane. Depending on the determinant of $\mathbf{Q}$, this will be an ellipse, two parallel lines, or two branches of an hyperbola (see fig. D.1). To compute the probability density function $p(s)$, we need to consider the area swept out by these curves as s changes.



(a) (b) (c)

Fig. D.1: Constant s contours in the xy plane for the three cases: (a) $a > 0, b > 0$, (b) $a > 0, b = 0$, and (c) $a > 0, b < 0$.

Let $\mathcal{A}(s)$ denote the region of the x,y plane for which $\mathbf{r}^T\mathbf{Q}\mathbf{r} < s$. Reverting briefly to a more formal notation in which S is the random variable and s its values, the probability that $S < s$ is given by integrating the density over $\mathcal{A}(s)$:

$$P(S < s) = \iint_{\mathcal{A}(s)} p(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{D.11}$$

The density function $p(s)$ is the derivative of this integral with respect to its boundary:

$$p(s) = \frac{\mathrm{d}}{\mathrm{d}s}P(S < s) = \frac{\mathrm{d}}{\mathrm{d}s}\iint_{\mathcal{A}(s)} p(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{D.12}$$

The integration can be simplified by a change of variable from (x,y) to (s,θ), where θ will be defined later in a way that is convenient for the particular family of curves defined by the matrix $\mathbf{Q}$. Regardless, if $\mathbf{J}$ is the Jacobian,

$$\mathbf{J} = \begin{pmatrix} \partial x/\partial s & \partial x/\partial\theta \\ \partial y/\partial s & \partial y/\partial\theta \end{pmatrix}, \tag{D.13}$$

then eq. D.11 can be written as

$$P(S < s_0) = \int_0^{s_0} \mathrm{d}s \int \mathrm{d}\theta\, p(s,\theta) \tag{D.14}$$

$$= \int_0^{s_0} \mathrm{d}s \int \mathrm{d}\theta\, |\mathbf{J}| p(x,y), \tag{D.15}$$

in which form, the differentiation with respect to the limit s_0 is trivial. (The appropriate limits of integration for θ will depend on the details of the transformation.) Note that for brevity, the notation $|\cdot|$, rather than $\det\cdot$, is being used for the determinant in this appendix. Substituting in $p(x,y) = h(x^2+y^2)$ and differentiating with respect to the limit s_0 yields

$$p(s) = \int |\mathbf{J}|\, h(x^2+y^2)\,\mathrm{d}\theta. \tag{D.16}$$

This is the general expression to be used for the three classes of curve defined by $|\mathbf{Q}| < 0$, $|\mathbf{Q}| = 0$, and $|\mathbf{Q}| > 0$, though in each case, a different mapping from s,θ to x,y will be used.

Elliptical contours

When $|\mathbf{Q}| > 0$, the curves of constant s are ellipses and the following elliptical-polar coordinate system is convenient:

$$x = \sqrt{(s/a)}\cos\theta, \qquad y = \sqrt{(s/b)}\sin\theta, \tag{D.17}$$

with $\theta \in [0, 2\pi)$. The determinant of the Jacobian evaluates to $|\mathbf{J}| = \frac{1}{2}(ab)^{-1/2}$. In order to evaluate eq. D.16, $x^2 + y^2$ must by expressed in terms of s and θ, which can be done using eq. D.17:

$$x^2 + y^2 = s(a^{-1}\cos^2\theta + b^{-1}\sin^2\theta). \tag{D.18}$$

By making the assignments

$$\alpha = \frac{a+b}{2ab}, \qquad \beta = \frac{a-b}{2ab}, \tag{D.19}$$

this may be written as $x^2 + y^2 = s(\alpha - \beta\cos 2\theta)$. Substituting these pieces into eq. D.16 yields

$$p_+(s) = \frac{1}{2\sqrt{ab}} \int_0^{2\pi} h(s[\alpha - \beta\cos 2\theta])\, \mathrm{d}\theta. \tag{D.20}$$

Note that this integral sweeps out the same area four times; putting $\phi = 2\theta$, changing the limits and multiplying by four gives the final result,

$$p_+(s) = \frac{1}{\sqrt{ab}} \int_0^{\pi} h(s[\alpha - \beta\cos\phi])\, \mathrm{d}\phi. \tag{D.21}$$

Rectilinear contours

When $|\mathbf{Q}| = ab = 0$, the symmetry of the system means that we may assume without loss of generality that $b = 0$. In this case, the lines of constant s will be pairs of straight lines parallel to the y axis, and a convenient coordinate frame is defined by

$$s = ax^2, \qquad \theta = \frac{y}{x\sqrt{a}}. \tag{D.22}$$

This gives $|\mathbf{J}| = 1/(2\sqrt{a})$ and $x^2 + y^2 = s(a^{-1} + \theta^2)$. Symmetry implies that the integration can be done in the first quadrant of the xy plane and the result multiplied by four. The result is

$$p_0(s) = \frac{2}{\sqrt{a}} \int_0^{\infty} h(s[a^{-1} + \theta^2])\, \mathrm{d}\theta \tag{D.23}$$

Hyperbolic contours

Finally, when $|\mathbf{Q}| < 0$, a family of hyperbolae is obtained and a hyperbolic trigonometric transformation is appropriate. However, some care is required since s may now be negative, and depending on the signs of a, b, and s, the hyperbola may be 'horizontal'

or 'vertical' (see fig. D.1). We may assume without loss of generality that $a > 0$ and $b < 0$, in which case, the coordinate transformation can be based on the sign of s:

$$s > 0: \qquad x = \sqrt{|s/a|}\cosh\theta, \qquad y = \sqrt{|s/b|}\sinh\theta, \tag{D.24}$$

$$s < 0: \qquad x = \sqrt{|s/a|}\sinh\theta, \qquad y = \sqrt{|s/b|}\cosh\theta. \tag{D.25}$$

In either case, $|\mathbf{J}| = \frac{1}{2}|ab|^{-1/2}$. The expression for $x^2 + y^2$ is now

$$x^2 + y^2 = \begin{cases} s > 0: & |s|(a^{-1}\cosh^2\theta - b^{-1}\sinh^2\theta) \\ s < 0: & |s|(a^{-1}\sinh^2\theta - b^{-1}\cosh^2\theta) \end{cases} \tag{D.26}$$

which, if we make the same assignments as in the elliptical case:

$$\alpha = \frac{a+b}{2ab}, \qquad \beta = \frac{a-b}{2ab}, \tag{D.27}$$

may be written as

$$x^2 + y^2 = \begin{cases} s > 0: & s(\alpha - \beta\cosh 2\theta), \\ s < 0: & s(\alpha + \beta\cosh 2\theta). \end{cases} \tag{D.28}$$

Taking care over the limits and making the substitution $\phi = 2\theta$ as before, the result is

$$p_-(s) = \frac{1}{\sqrt{|ab|}} \int_0^\infty h(s[\alpha \mp \beta\cosh\phi])\, \mathrm{d}\phi. \tag{D.29}$$

Numerical integration of these density functions over a range of directions (a,b), using the three forms in eq. D.21, eq. D.23, and eq. D.29, yields a family of distributions, all of which are strongly super-Gaussian. When $|\mathbf{Q}| \geq 0$, the distributions are single sided since $s > 0$. When $|\mathbf{Q}| < 0$, the distributions are double sided. The 'peakiest' distribution is obtained when $|\mathbf{Q} = 0|$. The implications of this are discussed in the main text, in §8.3.3.

BIBLIOGRAPHY

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

Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In Touretsky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT Press.

Amari, S. and Nagaoka, H. (2001). *Methods of Information Geometry*. American Mathematical Society / Oxford University Press.

Anderson, E. J. (1997). Limitations of short-time Fourier transforms in polyphonic pitch recognition. Technical report, Department of Computer Science and Engineering, University of Washington.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3(2):213–251.

Atick, J. J. and Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320.

Attias, H. and Schreiner, C. E. (1997). Temporal low-order statistics of natural sounds. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, pages 27–33. MIT Press.

Attneave, F. (1950). Dimensions of similarity. *Journal of American Psychology*, 63:516–556.

Attneave, F. (1954). Some informational aspects of visual perception. *Pyschological Review*, 61(3):183–193.

Attneave, F. (1959). *Applications of Information Theory to Psychology*. Holt, New York.

Bach, F. R. and Jordan, M. I. (2001). Kernel independent component analysis. Technical Report UCB/CSD-01-1166, Division of Computer Science, University of California, Berkely.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *Proceedings of a Symposium on the Mechanisation of Thought Processes*, volume 2, pages 537–559, National Physical Laboratory, Teddington. Her Majesty's Stationery Office, London.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the taskof perception. *Vision Research*, 30:1561–1571.

Barlow, H. B. (1996). Banishing the homonculus. In (Knill and Richards, 1996), page 425.

Barlow, H. B. and Földiák, P. (1989). Adaptation and decorrelation in the cortex. In Durbin, R., Miall, C., and Mitchison, G., editors, *The Computing Neuron*, pages 54–72. Addison-Wesley, Wokingham, UK.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1156.

Bell, A. J. and Sejnowski, T. J. (1996). Learning the higher-order structure of a natural sound. *Network: Computation in Neural Systems*, 7(2):261–266.

Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

Berkeley, G. (1975/1734). *A Treatise Concerning the Principles of Human Knowledge*. Dent and Sons, London. Originally published in 1734.

Bharucha, J. J. (1984). Event heirarchies, tonal heirarchies, and assimilation: A reply to Deutsch and Dowling. *Journal of Experimental Psychology: General*, 113(3):412–425.

Bharucha, J. J. (1991). Pitch, harmony and neural nets: A psychological perspective. In (Todd and Loy, 1991), pages 84–99.

Bishop, C. M. (1998). Latent variable models. In (Jordan, 1998), pages 371–403.

Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234.

Boashash, B. (1988). Note on the use of the Wigner distribution for time-frequency signal analysis. *IEEE Trans. Acoustics, Speech and Signal Processing*, 36(9):1518–1521.

Bregman, A. S. (1990). *Auditory Scene Analysis*. MIT Press, Cambridge, MA.

Brown, J. C. and Puckette, M. S. (1989). Calculation of a narrowed autocorrelation function. *Journal of the Acoustic Society of America*, 85(4):1595–1601.

Buja, A., Swayne, D. F., Littman, M. L., and Dean, N. (1998). XGvis: Inveractive data visualisation with multidimensional scaling. Available at `http://www.research.att.com/areas/stat/xgobi/`. Submitted to *Journal of Computational and Graphical Statistics*.

Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. PhD thesis, University of Edinburgh.

Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114.

Cardoso, J.-F. (2000). Entropic contrasts for source separation. In Haykin, S., editor, *Unsupervised Adaptive Filtering*, volume 1, pages 139–. John Wiley and Son, New York.

Cardoso, J.-F. and Laheld, B. (1996). Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–30.

Casey, M. A. (1998). *Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio*. PhD thesis, MIT Media Laboratory.

Chui, C. K. (1997). *Wavelets: A Mathematical Tool for Signal Analysis*. SIAM, Philadelphia, PA.

Cohen, L. (1989). Time-frequency distributions—A review. *Proc. IEEE*, 77(7):941–981.

Comon, P. (1994). Independent Component Analysis- a new concept? *Signal Processing*, 36:287–314.

Conklin, D. and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley and Sons, New York.

Cowell, R. (1998). An introduction to inference for Bayesian networks. In (Jordan, 1998), pages 9–26.

Cox, T. and Cox, M. A. A. (2001). *Multidimensional Scaling*. Chapman Hall/CRC, London.

Davidson, M. L. (1983). *Multidimensional Scaling*. John Wiley and Sons, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Proceedings of the Royal Statistical Society*, B-39:1–38.

Deutsch, D. (1984). Two issues concerning tonal heirarchies: Comment on Castellano, Bharucha and Krumhansl. *Journal of Experimental Psychology: General*, 113(3):413–416.

Deutsch, D. (1999a). Grouping mechanisms in music. In (Deutsch, 1999b), chapter 9.

Deutsch, D., editor (1999b). *The Psychology of Music*. Academic Press, San Diego, London, second edition.

Dixon, S. (2000). On the computer recognition of solo piano music. In *Proceedings of the Australasian Computer Music Association Conference*, pages 31–37, Brisbane, Australia.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58.

Duda, R., Lyon, R., and Slaney, M. (1990). Correlograms and the separation of sounds. In *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, volume 1, pages 457–61, Pacific Grove, CA.

Ellis, D. P. W. (1992). A perceptual representation of audio. Master's thesis, MIT Electronic Engineering and Computer Science Dept.

Ellis, D. P. W. (1994). A computer model of psychoacoustic grouping rules. In *Proc. 12th Intl. Conf. on Pattern Recognition*, Jerusalem. Available at `http://sound.media.mit.edu/pub/Papers/dpwe-ICPR.ps.gz`.

Ellis, D. P. W. (1996). *Prediction-driven computational auditory scene analysis*. PhD thesis, MIT Electronic Engineering and Computer Science Dept.

Ellis, D. P. W. and Rosenthal, D. F. (1995). Mid-level representations for computational auditory scene analysis. In *Proc. Intl. Joint Conf. on Artificial Intelligence Workshop on Computational Auditory Scene Analysis*, pages 111–117, Montreal.

Endres, D. and Földiák, P. (1999). Quadratic programming for learning sparse codes. In *Intl. Conf. on Artificial Neural Networks*, pages 593–596, Edinburgh.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

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

Flandrin, P. and Rioul, O. (1990). Affine smoothing of the Wigner distribution. In *ICASSP'90*, pages 2455–2458.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64(2):165–170.

Frey, B. J. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press, Cambridge, MA.

Gabor, D. (1947). Acoustical quanta and the theory of hearing. *Nature*, 159:591–594.

Ghahramani, Z. and Roweis, S. (1999). Probabilistic models for unsupervised learning. NIPS Tutorial, available at `http://www.gatsby.ucl.ac.uk/~zoubin/`.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston.

Godsmark, D. and Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366.

Harpur, G. F. (1997). *Low Entropy Coding with Unsupervised Neural Networks*. PhD thesis, Department of Engineering, Cambridge University.

Hecht, E. (1987). *Optics*. Addison-Wesley, Reading, MA, second edition.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In (Jordan, 1998), pages 301–354.

Helmholtz, H. (1954/1885). *On the sensations of tone*. Dover, New York, Second English edition. Republication of the A. J. Ellis translation of 1885.

Helmholtz, H. (1962/1910). *Helmholtz's Treatise on Physiological Optics*. Dover, New York. Originally published in 1910.

Hinton, G. E., Welling, M., Teh, Y. W., and Osindero, S. K. (2001). A new view of ICA. In *3rd Intl. Conf. on Independent Component Analysis and Signal Separation, ICA2001*, pages 746–751, San Diego.

Hochberg, J. (1981). Levels of perceptual organization. In (Kubovy and Pomerantz, 1981), chapter 9.

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

Hyvärinen, A. (1999). Sparse code shrinkage: Denoising of non-Gaussian data by maximum-likelihood estimation. *Neural Computation*, 11(7):1739–1768.

Hyvärinen, A., Cristescu, R., and Oja, E. (1999). A fast algorithm for estimating overcomplete ICA bases for image windows. In *Proc. Int. Joint Conf. on Neural Networks*, volume 2, pages 894 –899, Washington D.C.

Hyvärinen, A. and Hoyer, P. (1999). Emergence of complex cell properties by decomposition of natural images into independent feature subspaces. In *Intl. Conf. on Artificial Neural Networks*, pages 257–262, Edinburgh.

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

Hyvärinen, A., Hoyer, P., and Inki, M. (2001). Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558.

Hyvärinen, A., Oja, E., Hoyer, P., and Hurri, J. (1998). Image feature extraction by sparse coding and Independent Component Analysis. In *Proc. Intl. Conf. on Pattern Recognition (ICPR'98)*, pages 1268–1273, Brisbane.

Iordanov, L. G. and Penev, P. S. (1999). The principal component structure of natural sound. Available from `http://citeseer.nj.nec.com/201279.html`.

Jacob, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Jameson, G. J. O. (1974). *Topology and Normed Spaces*. Chapman and Hall, London.

Jeong, J. and Williams, W. J. (1992). Kernel design for reduced interference distributions. *IEEE Transactions on Signal Processing*, 40(2):402–12.

Jordan, M. I., editor (1998). *Learning in Graphical Models*. MIT Press, Cambridge, MA.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An introduction to variational methods for graphical models. In (Jordan, 1998), pages 105–161.

Jutten, C. and Herault, J. (1991). Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10.

Karhunen, J. (1996). Neural approches to Independent Component Analysis and source separation. In *Proc. 4th European Symposium on Artificial Neural Networks (ESANN'96)*, pages 249–266, Bruges, Belgium.

Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. (1998). Application of the Bayesian probability network to music scene analysis. In Rosenthal, D. F. and Okuno, H., editors, *Computational Auditory Scene Analysis*, pages 115–137. Lawrence Erlbaum, Mahwah, NJ.

Keislar, D. (1987). The history and principles of microtonal keyboards. *Computer Music Journal*, 11(1):18–28.

Klapuri, A., Eronen, A., Seppännen, J., and Virtanen, T. (2001). Automatic transcription of music. In *Symposium on Stochastic Modeling of Music, 14th Meeting of the FWO Research Society on Foundations of Music Research*, Ghent, Belgium. Available at `http://www.cs.tut.fi/sgn/arg/klap/ATOM_article_2001.pdf`.

Knill, D. C. and Richards, W., editors (1996). *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, UK.

Köhler, W. (1947). *Gestalt Psychology*. Liveright, New York.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin.

Krumhansl, C. L. (1990). *The Cognitive Foundations of Musical Pitch*. Oxford University Press, New York.

Krumhansl, C. L. and Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4):334–368.

Kubovy, M. and Pomerantz, J. R., editors (1981). *Perceptual Organization*. Lawrence Erlbaum, Hillsdale, NJ.

Kullback, S. (1997/1968). *Information Theory and Statistics*. Dover, New York. Originally published in 1968.

Large, E. and Kolen, J. (1994). Resonance and the perception of musical meter. *Connection Science*, 6:177–208.

Lazzaro, J. and Mead, C. (1989). Silicon modeling of pitch perception. *Proceedings of the National Academy of Sciences, USA*, 86:9597–9601.

Leman, M. (1991). The ontogenesis of tonal semantics: Results of a computer study. In (Todd and Loy, 1991), pages 100–127.

Leman, M. (1999). Naturalistic approaches to musical semiotics and the study of causal musical signification. In Zannos, I., editor, *Music and Signs – Semiotic and Cognitive Studies in Music*, pages 11–38. ASCO Art and Science.

Leman, M. and Carreras, F. (1996). The self-organization of stable perceptual maps in a realistic musical environment. In Assayag, G., Chemillier, M., and Eloy, C., editors, *Proc. Troisième Journée d'Informatique Musicale, JIM96*, pages 156–169, Caen.

Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, 12:337–365.

Licklider, J. C. R. (1959). Three auditory theories. In Koch, S., editor, *Pyschology: A Study of a Science*, volume 1, pages 41–144. McGraw-Hill, New York.

Linsker, R. (1988). Self-organisation in a perceptual network. *Computer*, 21(3):105–117.

Locke, J. (1961/1706). *An Essay Concerning Human Understanding*. Dent and Sons, London. Fifth edition, originally published in 1706.

Longuet-Higgins, H. C. (1987). *Mental Processes: Studies in Cognitive Science*. MIT Press, Cambridge, MA.

Loughlin, P. J. (1991). New properties to alleviate interference in time-frequency representations. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 3205–3208.

Luttrell, S. P. (1994). A Bayesian analysis of self-organizing maps. *Neural Computation*, 6(5):767–794.

Lyon, R. F. (1984). Computational models of neural auditory processing. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, page 36.1.1, San Diego.

Lyon, R. F. and Mead, C. (1988). An analog electronic cochlea. *IEEE Trans. on Acoustics Speech and Signal Processing*, 36(7):1119–1144.

Mach, E. (1959/1886). *The Analysis of Sensations*. Dover, New York.

MacKay, D. J. C. (1996). Maximum likelihood and covariant algorithms for independent component analysis. Unpublished manuscript available at `http://wol.ra.phy.cam.ac.uk/mackay`.

MacKay, D. J. C. (1998). Introduction to monte carlo methods. In (Jordan, 1998), pages 175–204.

Mahalanobis, P. C. (1936). Mahalanobis distance. *Proceedings National Institute of Science of India*, 49(2):234–256.

Mallat, S. G. and Zang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. on Signal Processing*, 41(12):3397–3415.

Marr, D. (1982). *Vision*. Freeman, San Fransisco.

Meddis, R. (1983). Simulation of auditory neural transduction. *Journal of the Acoustic Society of America*, 83:1056–1063.

Mellinger, D. K. (1991). *Event Formation and Separation in Musical Sound*. PhD thesis, Department of Computer Science, Standford University.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. University of Chicago Press.

Miller, G. A. (1953). What is information measurement? *American Psychologist*, 8:3–11.

Miller, G. A. and Frick, F. C. (1949). Statistical behaviouristics and sequences of responses. *Psychological Review*, 56:311–324.

Mitchison, G. J. (1995). A type of duality between self-organizing maps and minimal wiring. *Neural Computation*, 7(1):25–35.

Moore, B. C. J. and Sek, A. (1995). Auditory filtering and the critical bandwidth at low frequencies. In et al., G. A. M., editor, *Advances in Hearing Research*, pages 425–440. World Scientific, Singapore.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference*. Dekker, New York.

Nadal, J.-P. and Parga, N. (1994). Nonlinear neurons in low-noise limit: A factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5:565–581.

Oldfield, R. C. (1954). Memory mechanisms and the theory of schemata. *British Journal of Psychology*, pages 14–23.

Olshausen, B. A. (1996). Learning linear, sparse, factorial codes. Technical Report AI Memo 1580, Artificial Intelligence Lab, MIT.

Olshausen, B. A. and Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339.

Olshausen, B. A. and Millman, K. J. (2000). Learning sparse codes with a mixture-of-Gaussians prior. In A., S. S., Leen, T. K., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 841–847. MIT Press.

O'Neill, J. C. and Williams, W. J. (1999). Shift covariant time-frequency distributions of discrete signals. *IEEE Transactions on Signal Processing*, 47(1).

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (1988). SVOS final report: The auditory filterbank. Technical Report APU 2341, Applied Psychology Unit, Medical Research Council, UK.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan-Kaufmann, San Mateo, CA.

Pickles, J. O. (1988). *An Introduction to the Physiology of Hearing*. Academic Press, London.

Plomp, R. (1976). *Aspects of Tone Sensation*. Academic Press, London.

Plumbley, M. D. (1991). On information theory and unsupervised neural networks. Technical Report CUED/F-INFENG/TR.78, Department of Engineering, Cambridge University.

Pomerantz, J. R. and Kubovy, M. (1981). Perceptual organization: An overview. In (Kubovy and Pomerantz, 1981), chapter 13.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T., editors (1992). *Numerical recipes in C*. Cambridge University Press, Cambridge, UK.

Raffman, D. (1993). *Language, Music and Mind*. MIT Press, Cambridge, MA.

Raphael, C. (2001). Synthesizing musical accompaniments with Bayesian belief networks. *Journal of New Music Research*, 30(1):59–67.

Rasch, R. and Plomp, R. (1999). The perception of musical tones. In (Deutsch, 1999b), chapter 4.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, 5(2):289–304.

Richardson, M. W. (1938). Multidimensional psychophysics. *Psychological Bulletin*, 35:659–660.

Rieke, F., Warland, D., de Ruyter van Stevenink, R., and Bialek, W. (1997). *Spikes*. MIT Press, Cambridge, MA.

Risset, J.-C. and Wessel, D. L. (1999). Exploration of timbre by analysis and synthesis. In (Deutsch, 1999b).

Roberts, S. J. and Penny, W. D. (2001). Mixtures of independent component analysers. *Lecture Notes in Computer Science*, 2130:527–.

Sammon, J. W. (1969). A nonlinear mapping for data analysis. *IEEE Transactions on Computers*, C-18:401–409.

Scheirer, E. D. (1996). Bregman's chimerae: Music perception as auditory scene analysis. In *Proc. 4th Int. Conf. on Music Perception and Cognition*, pages 317–322, Montreal.

Schmidhuber, J. (1992a). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4(2):234–242.

Schmidhuber, J. (1992b). Learning unambiguous reduced sequence descriptions. In Lippmann, R. P., Moody, J. E., and Hanson, S. J., editors, *Advances in Neural Information Processing Systems*, volume 4, pages 291–298. Morgan-Kauffman.

Schoenberg, A. (1969). *Structural Functions of Harmony*. Norton.

Schwartz, O. and Simoncelli, E. P. (2001). Natural sound statistics and divisive normalisation in the auditory system. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 166–172.

Seneff, S. (1984). Pitch and spectral estimation of speech based on an auditory synchrony model. In *Proc. Intl. Conf. on Acoustics, Speech and Signal Processing*, page 36.2.1.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656.

Shaw, R. E., McIntyre, M., and Mace, W. M. (1974). The rôle of symmetry in event perception. In MacLeod, R. B. and Pick, H., editors, *Perception: Essays in Honour of James J. Gibson*. Cornell University Press, Ithaca, NY.

Shepard, R. N. (1981). Psychological complimentarity. In (Kubovy and Pomerantz, 1981), chapter 10.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review*, 89(4):305–333.

Shepard, R. N. (1999). Cognitive psychology and music. In Cook, P. R., editor, *Music, Cognition, and Computerised Sound*, chapter 3, pages 21–35. MIT Press, Cambridge, MA.

Simoncelli, E. P. (1999). Modeling the joint statistics of images in the wavelet domain. In *Proceedings of the 44th Annual Meeting of the SPIE*, pages 188–195, Denver.

Slaney, M. (1993). An efficient implementation of the Patterson-Holdsworth auditory filterbank. Technical Report #35, Apple Computer Co.

Stoll, R. R. and Wong, E. T. (1968). *Linear Algebra*. Academic Press, New York.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA.

Tipping, M. and Bishop, C. M. (1997). Probabilistic principal component analysis. Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, Birmingham, UK.

Todd, P. M. and Loy, D. G., editors (1991). *Music and Connectionism*. MIT Press, Cambridge, MA.

Torgeson, W. S. (1952). Multidimensional scaling, I: Theory and method. *Psychometrika*, 17:401–419.

Tufte, E. R. (1990). *Envisioning Information*. Graphics Press, Cheshire, CT.

Wainwright, M. J. and Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. In A., S. S., Leen, T. K., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 855–861. MIT Press.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167:392–393.

West, R., Cross, I., and Howell, P. (1987). Modelling music as input output and as process. *Psychology of Music*, 15(1):7–29.

Windsor, W. L. (1995). *A Perceptual Approach to the Description and Analysis of Acousmatic Music*. PhD thesis, Department of Music, City University, London.