AVAILABLE AT
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT•

Neural Networks 16 (2003) 1325–1352

Neural Networks

www.elsevier.com/locate/neunet

2003 Special Issue

Learning and inference in the brain

Karl Friston*

The Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK

Received 16 December 2002; revised 26 June 2003; accepted 26 June 2003

Abstract

This article is about how the brain data mines its sensory inputs. There are several architectural principles of functional brain anatomy that have emerged from careful anatomic and physiologic studies over the past century. These principles are considered in the light of representational learning to see if they could have been predicted a priori on the basis of purely theoretical considerations. We first review the organisation of hierarchical sensory cortices, paying special attention to the distinction between forward and backward connections. We then review various approaches to representational learning as special cases of generative models, starting with supervised learning and ending with learning based upon *empirical Bayes*. The latter predicts many features, such as a hierarchical cortical system, prevalent top-down backward influences and functional asymmetries between forward and backward connections that are seen in the real brain.

The key points made in this article are: (i) hierarchical generative models enable the learning of *empirical* priors and eschew prior assumptions about the causes of sensory input that are inherent in non-hierarchical models. These assumptions are necessary for learning schemes based on information theory and efficient or sparse coding, but are not necessary in a hierarchical context. Critically, the anatomical infrastructure that may implement generative models in the brain is hierarchical. Furthermore, learning based on empirical Bayes can proceed in a biologically plausible way. (ii) The second point is that backward connections are essential if the processes generating inputs cannot be inverted, or the inversion cannot be parameterised. Because these processes involve many-to-one mappings, are non-linear and dynamic in nature, they are generally non-invertible. This enforces an explicit parameterisation of generative models (i.e. backward connections) to afford recognition and suggests that forward architectures, on their own, are not sufficient for perception. (iii) Finally, non-linearities in generative models, mediated by backward connections, require these connections to be modulatory, so that representations in higher cortical levels can interact to predict responses in lower levels. This is important in relation to functional asymmetries in forward and backward connections that have been demonstrated empirically.

Keywords: Inference; Predictive coding; Generative models; Information theory; Bayesian

1. Introduction

This article uses the relationship among computational models of representational learning as a vehicle to illustrate how theoretical approaches to neuronal information processing can help understand the structure of functional brain architectures. We start by reviewing two principles of brain organisation, namely *functional specialisation* and *functional integration* and how they rest upon the anatomy and physiology of hierarchical cortico-cortical connections in the brain. Section 3 deals with the nature and learning of representations from a theoretical or computational perspective. This section reviews *supervised* (e.g. connectionist) approaches, *information theoretic* approaches and those predicated on *predictive coding*. The review examines the heuristics behind the various schemes using the framework of *generative models*. We then introduce learning based on *empirical Bayes* that is enabled by hierarchical generative models. The key focus of this section is on the functional architectures and assumptions implied by each model. Representational learning based on information theory can, in principle, proceed using only forward connections. However, this is only tenable when processes generating sensory inputs are invertible and independent. Invertibility is precluded when the cause of a percept and its context interact. These interactions create a problem of contextual invariance that can only be resolved using internal or generative models. Contextual invariance is necessary for categorisation of sensory input (e.g. category-specific responses) and represents a fundamental problem in perceptual synthesis. Generative or forward models can

* Tel.: +44-207-833-7457; fax: +44-207-813-1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk (K. Friston).

0893-6080/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2003.06.005

solve this problem using predictive coding but only if the distribution of causes is known a priori. Empirical Bayes allows these priors to be learned, along with the model itself, using hierarchies of backward and lateral projections that prevail in the real brain. In short, hierarchical models of representational learning are a natural choice for understanding real functional architectures and, critically, confer a necessary role on backward connections.

2. Functional specialisation and integration

2.1. Background

The brain appears to adhere to two fundamental principles of functional organisation, namely functional integration and functional specialisation, where the integration within and among specialised areas is mediated by effective connectivity. The distinction relates to that between 'localisationism' and '(dis)connectionism' that dominated thinking about cortical function in the 19th century. Since the early anatomic theories of Gall, the identification of a particular brain region with a specific function has become a central theme in neuroscience. However, functional localisation per se was not easy to demonstrate. For example, a meeting that took place on August 4th 1881 addressed the difficulties of attributing function to a cortical area, given the dependence of cerebral activity on underlying connections (Phillips, Zeki, & Barlow, 1984). This meeting was entitled 'Localisation of function in the cortex cerebri'. Goltz, although accepting the results of electrical stimulation in dog and monkey cortex, considered that the excitation method was inconclusive, because the behaviours elicited might have originated in related pathways, or current could have spread to distant centres. In short, the excitation method could not be used to infer functional localisation because localisationism discounted interactions, or functional integration among different brain areas. It was proposed that lesion studies could supplement excitation experiments. Ironically, it was observed on patients with brain lesions some years later (see Absher & Benson, 1993) that led to the concept of 'disconnection syndromes' and the refutation of localisationism as a complete or sufficient explanation of cortical organisation. Since that time, there has been a great endeavour to integrate electrophysiological findings and those inspired by the lesion-deficit model. This convergence has led to an increasingly refined understanding of the segregation–integration axis. Much of this understanding rests on the many-to-one and one-to-many mappings between structure and function. For example, the functional specialisation of motion-selective units in MT has been characterised with remarkable finesse (e.g. Liu & Newsome, 2003), yet lesions to MT alone are not sufficient to produce stable deficits in motion perception. Results like these suggest that several neuronal systems may support the same function (i.e. degenerate many-to-one mappings). Conversely, one cortical system may contribute to many functions: motion cues serve many purposes in primate vision. Consequently, akinetopsia, a defect of movement perception due to cerebral lesions, may comprise a range of motion-related defects. To address this issue Rizzo, Nawrot, and Zihl (1995) explored the perceptual profiles in an akinetopsia subject L.M. who had extensive bilateral lesions of the dorsolateral visual association cortex that spared primary visual cortex. "Surprisingly, L.M. also had trouble perceiving 2-D shapes defined by non-motion signals including 'on' and 'off' transients, dynamic and static binocular disparity, and static texture cues." This sort of finding speaks to a one-to-many structure–function relationship.

Functional localisation implies that a function can be localised in a cortical area, whereas specialisation suggests that a cortical area is specialised for some aspects of perceptual or motor processing, where this *specialisation* can be anatomically *segregated* within the cortex. The cortical infrastructure supporting a single function may then involve many specialised areas whose union is mediated by the functional integration among them. Functional specialisation and integration are not exclusive; they are complementary. Functional specialisation is only meaningful in the context of functional integration and vice versa.

2.2. Functional specialisation and segregation

The functional role, played by any component (e.g. cortical area, sub-area, neuronal population or neuron) of the brain, is defined largely by its connections. Certain patterns of cortical projections are so common that they could amount to rules of cortical connectivity. "These rules revolve around one, apparently, overriding strategy that the cerebral cortex uses—that of functional segregation" (Zeki, 1990). Functional segregation demands that cells with common functional properties be grouped together. There are many examples of this grouping (e.g. laminar selectivity, ocular dominance bands and orientation domains in V1). This architectural constraint necessitates both convergence and divergence of cortical connections. Extrinsic connections, between cortical regions, are not continuous but occur in patches or clusters. This patchiness has, in some instances, a clear relationship to functional segregation. For example, the secondary visual area V2 has a distinctive cytochrome oxidase architecture, consisting of thick stripes, thin stripes and inter-stripes. When recordings are made in V2, directionally selective (but not wavelength or colour selective) cells are found exclusively in the thick stripes. Retrograde (i.e. backward) labelling of cells in V5 is limited to these thick stripes. All the available physiological evidence suggests that V5 is a functionally homogeneous area that is specialised for visual motion. Evidence of this nature supports the idea that patchy connectivity is the anatomical infrastructure that underpins functional segregation and specialisation.

The notion that connections underpin segregation scales from the level of stripes in V2 to processing streams that encompass many cortical areas. For example, current models partition the primate visual system into dorsal (magno) and ventral (parvo, konio) streams. Evidence for this derives from the pattern of projections between V1 and V2. Recently this evidence has been re-evaluated (Sincich & Horton, 2002) who demonstrate that "V1 output arises from just two sources: patch columns and inter-patch columns. Patch columns project to thin stripes and inter-patch columns project to pale and thick stripes. Projection of inter-patches to common V2 stripe types (pale and thick) merges parvo and magno inputs, making it likely that these functional channels are distributed strongly to both dorsal and ventral streams."

2.3. The anatomy and physiology of cortico-cortical connections

If specialisation rests upon connectivity, then important organisational principles should be embodied in the neuroanatomy and physiology of extrinsic connections. Extrinsic connections couple different cortical areas, whereas intrinsic connections are confined to the cortical sheet. There are certain features of cortico-cortical connections that provide strong clues about their functional role. In brief, there appears to be a hierarchical organisation that rests upon the distinction between *forward* and *backward* connections. The designation of a connection as forward or backward depends primarily on its cortical layers of origin and termination. Some characteristics of cortico-cortical connections are presented below and are summarised in Table 1. The list is not exhaustive, nor properly qualified, but serves to introduce some important principles that have emerged from empirical studies of visual cortex.

Hierarchical organisation. The organisation of the visual cortices can be considered as a hierarchy of cortical levels with reciprocal extrinsic cortico-cortical connections among the constituent cortical areas (Felleman & Van Essen, 1991). Forward connections run from lower to higher areas and backward connections from higher to lower. Lateral connections connect regions within a hierarchical level. The notion of a hierarchy depends upon a distinction between extrinsic forward and backward connections.

Reciprocal connections. Although reciprocal, forward and backward connections show both a microstructural and functional asymmetry. The terminations of both show laminar specificity. Forwards connections (from a low to a high level) have sparse axonal bifurcations and are topographically organised, originating in supragranular layers and terminating largely in layer IV. Backward connections, on the other hand, show abundant axonal bifurcation and a more diffuse topography, although they can be patchy (Angelucci et al., 2002b). Their origins are bilaminar/infragranular and they terminate predominantly in supragranular layers (Rockland & Pandya, 1979; Salin & Bullier, 1995). Extrinsic connections show an orderly convergence and divergence of connections from one cortical level to the next. At a macroscopic level, one point in a given cortical area will connect to a region 5–8 mm in diameter in another. An important distinction between forward and backward connections is that backward connections are more divergent. For example, the divergence region of a point in V5 (i.e. the region receiving backward afferents from V5) may include thick and interstripes in V2 whereas its convergence region (i.e. the region providing forward afferents to V5) is limited to the thick stripes (Zeki & Shipp, 1988). Furthermore, backward connections are more abundant than forward connections and transcend more levels. For example the ratio of forward efferent connections to backward afferents in the lateral geniculate is about 1:10/20. Another important distinction is that backward connections will traverse a number of hierarchical levels whereas forward connections are more restricted. For example, there are backward connections from TE and TEO to V1 but no monosynaptic connections from V1 to TE or TEO (Salin & Bullier, 1995).

Table 1
Some key characteristics of extrinsic cortico-cortical connections in the brain

Hierarchical organisation
The organisation of the visual cortices can be considered as a hierarchy (Felleman & Van Essen, 1991)
The notion of a hierarchy depends upon a distinction between forward and backward extrinsic connections
This distinction rests upon different laminar specificity (Rockland & Pandya, 1979; Salin & Bullier, 1995)
Backward connections are more numerous and transcend more levels
Backward connections are more diffuse than forward connections (Zeki & Shipp, 1988)

Forwards connections	Backwards connections
Sparse axonal bifurcations	Abundant axonal bifurcation
Topographically organised	Diffuse topography
Originate in supragranular layers	Originate in bilaminar/infragranular layers
Terminate largely in layer IV	Terminate predominantly in supragranular layers
Post-synaptic effects through fast AMPA (1.3–2.4 ms decay) and GABA_A (6 ms decay) receptors	Modulatory afferents activate slow (50 ms decay) voltage-sensitive NMDA receptors

Functionally asymmetric forward and backward connections. Functionally, reversible inactivation (e.g. Girard & Bullier, 1989; Sandell & Schiller, 1982) and neuroimaging (e.g. Büchel & Friston, 1997) studies suggest that forward connections are driving, always eliciting a response, whereas backward connections can also be modulatory. In this context, modulatory means backward connections modulate responsiveness to other inputs. At the single cell level "inputs from drivers can be differentiated from those of modulators. The driver can be identified as the transmitter of receptive field properties; the modulator can be identified as altering the probability of certain aspects of that transmission" (Sherman & Guillery, 1998).

The notion that forward connections are concerned with the promulgation and segregation of sensory information is consistent with: (i) their sparse axonal bifurcation, (ii) patchy axonal terminations, and (iii) topographic projections. In contradistinction, backward connections are generally considered to have a role in mediating contextual effects and in the co-ordination of processing channels. This is consistent with: (i) their frequent bifurcation, (ii) diffuse axonal terminations, and (iii) more divergent topography (Crick & Koch, 1998; Salin & Bullier, 1995). Forward connections meditate their post-synaptic effects through fast AMPA (1.3–2.4 ms decay) and GABA_A (6 ms decay) receptors. Modulatory effects can be mediated by NMDA receptors. NMDA receptors are voltage-sensitive, showing non-linear and slow dynamics (~50 ms decay). They are found predominantly in supragranular layers where backward connections terminate (Salin & Bullier, 1995). These slow time-constants again point to a role in mediating contextual effects that are more enduring than phasic sensory-evoked responses. The clearest evidence, for the modulatory role of backward connections (that is mediated by 'slow' glutamate receptors) comes from corticogeniculate connections. In the cat lateral geniculate nucleus, cortical feedback is partly mediated by type 1 metabotropic glutamate receptors, which are located exclusively on distal segments of the relay-cell dendrites. Rivadulla, Martinez, Varela, and Cudeiro (2002) have shown that these backward afferents enhance the excitatory centre of the thalamic receptive field. "Therefore, cortex, by closing this corticofugal loop, is able to increase the gain of its thalamic input within a focal spatial window, selecting key features of the incoming signal" (Rivadulla et al., 2002).

The asymmetry between forward and backward connections maps nicely onto the distinction between driving and modulatory effects proposed by Sherman and Guillery (1998). (i) Cross-correlograms from driving inputs have sharper peaks than modulatory inputs, (ii) there are likely to be few drivers but many modulators for any cell, and (iii) drivers act through (fast) ionotropic receptors, whereas modulators also activate metabotropic receptors with a slow and prolonged post-synaptic effect.

In relation to the status of hierarchical cortical organisation, it should be noted that the hierarchical ordering of areas is a matter of debate and may be indeterminate. On the basis of computational neuroanatomic studies, Hilgetag, O'Neill, and Young (2000) conclude laminar hierarchical constraints that are presently available in the anatomical literature are "insufficient to constrain a unique ordering" for any of the sensory systems analysed. However, basic hierarchical principles were clearly evident. Indeed, the authors note "All the cortical systems we studied displayed a significant degree of hierarchical organisation" with the visual and somato-motor systems showing an organisation that was "surprisingly strictly hierarchical".

In what follows we will consider hierarchies as entities in their own right. However, there are probably many hierarchical brain systems that are interconnected. The schematic in Fig. 1 shows how several hierarchical structures could be organised in the brain. This schematic is inspired by Mesulam's (1998) notion of sensory-fugal processing over "a core synaptic hierarchy, which includes the primary sensory, upstream unimodal, downstream unimodal, heteromodal, paralimbic and limbic zones of the cerebral cortex" (see Mesulam, 1998 for more details).

There are many mechanisms that are responsible for establishing connections in the brain. Connectivity results from interplay between genetic, epigenetic and activity- or experience-dependent mechanisms. In utero, epigenetic mechanisms predominate, such as the interaction between the topography of the developing cortical sheet, cell migration, gene expression and the mediating role of gene–gene interactions and gene products such as cell adhesion molecules (CAMs). Following birth, connections are progressively refined and re-modelled with a greater emphasis on activity- and use-dependent plasticity. These changes endure into adulthood with ongoing reorganisation and experience-dependent plasticity that subserves behavioural adaptation and learning. In brief, there are two basic determinants of connectivity. (i) *Structural plasticity*, reflecting the interactions between the molecular biology of gene expression, cell migration and neurogenesis in the developing brain. (ii) *Synaptic plasticity*, activity-dependent modelling of the pattern and strength of synaptic connections. This plasticity involves changes in the form, expression and function of synapses that endure throughout life.

It is interesting to note that forward and backward connections evidence a structural plasticity that is, neurodevelopmentally, quite distinct. The laminar organisation of cortico-cortical projection neurons (reflected in the percentage of supragranular projecting neurons—SLN%) characterises cortical pathways as forward or backward. The developmental reduction of SLN% is a widespread phenomenon in the neocortex and is a distinctive feature of backward pathways (Batardiere et al., 2002). Recent studies by Batardiere et al. (2002) suggest that forward and backward connections "exhibit different developmental processes and patterns of connections linking cortical

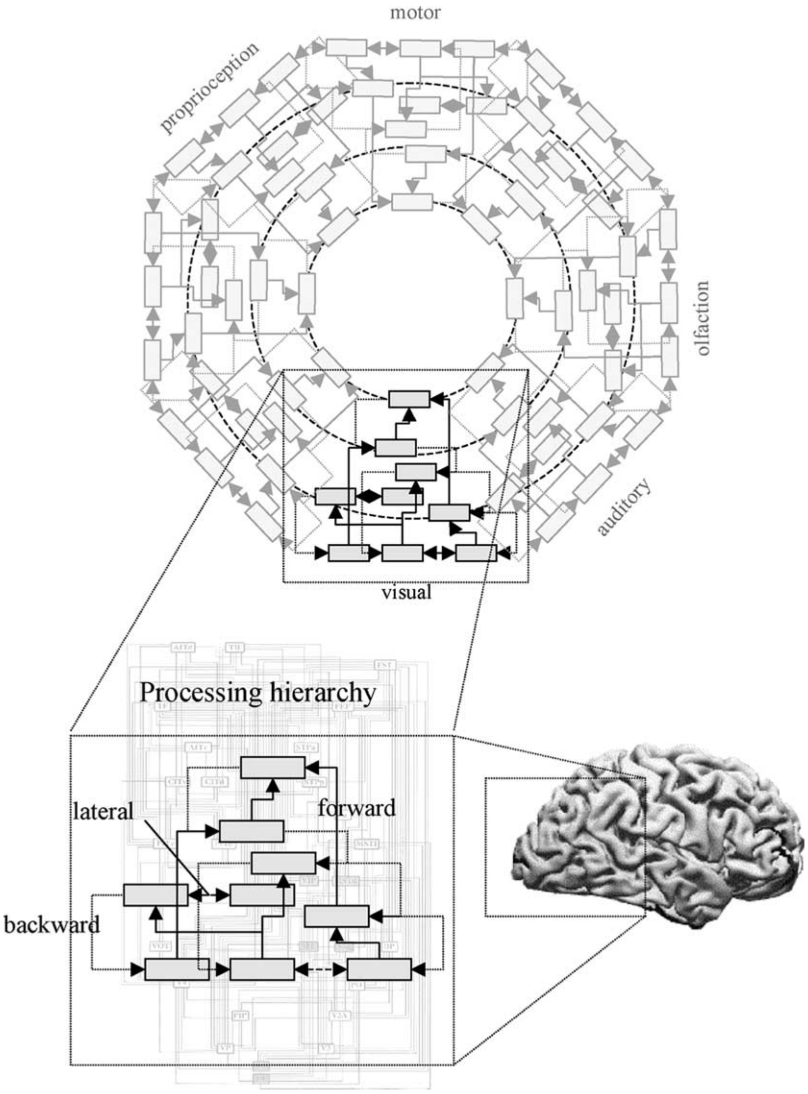

Fig. 1. Schematic illustrating hierarchical structures in the brain and the distinction between forward, backward and lateral connections.

areas and their hierarchical relations are established prenatally, independently of regressive phenomena".

In the post-developmental period, synaptic plasticity is an important functional attribute of connections in the brain and is thought to subserve perceptual and procedural learning and memory. This is a large and fascinating field that ranges from molecules to maps (see for example Buonomano & Merzenich, 1998; Martin, Grimwood, & Morris, 2000). Changing the strength of connections between neurons is widely assumed to be the mechanism by which memory traces are encoded and stored in the central nervous system. In its most general form, the synaptic plasticity and memory hypothesis states that. "Activity-dependent synaptic plasticity is induced at appropriate synapses during memory formation and is both necessary and sufficient for the information storage underlying the type of memory mediated by the brain area in which that plasticity is observed" (see Martin et al., 2000 for an evaluation of this hypothesis). A key aspect of this plasticity is that it is generally associative.

Associative plasticity. Synaptic plasticity may be transient (e.g. short-term potentiation STP or depression STD) or enduring (e.g. long-term potentiation LTP or LTD) with many different time constants. In contrast to short-term plasticity, long-term changes rely on protein synthesis, synaptic re-modelling and infrastructural changes in cell processes (e.g. terminal arbours or dendritic spines) that are mediated by calcium-dependent mechanisms. An important aspect of NMDA receptors, in the induction of LTP, is that they confer associativity on changes in connection strength. This is because their voltage-sensitivity only allows calcium ions to enter the cell when there is conjoint pre-synaptic release of glutamate and sufficient post-synaptic depolarisation (i.e. the temporal association of pre- and post-synaptic events). Calcium entry renders the post-synaptic specialisation eligible for future potentiation by promoting the formation of synaptic 'tags' (e.g. Frey & Morris, 1997) and other calcium-dependant intracellular mechanisms.

In summary, the anatomy and physiology of cortico-cortical connections suggest that forward connections are driving and commit cells to a pre-specified response given the appropriate pattern of inputs. Backward connections, on the other hand, are less topographic and are in a position to modulate the responses of lower areas to driving inputs from either higher or lower areas (see Table 1). For example, in the visual cortex Angelucci et al. (2002b) used a combination of anatomical and physiological recording methods to determine the spatial scale and retinotopic logic of intra-areal V1 horizontal connections and inter-areal feedback connections to V1. 'Contrary to common beliefs, these (monosynaptic horizontal) connections cannot fully account for the dimensions of the surround field (of macaque V1 neurons). The spatial scale of feedback circuits from extrastriate cortex to V1 is, instead, commensurate with the full spatial range of centre–surround interactions. Thus these connections could represent an anatomical substrate for contextual modulation and global-to-local integration of visual signals.'

Finally, brain connections are not static but are changing at the synaptic level all the time. In many instances this plasticity is associative. Backwards connections are abundant in the brain and are in a position to exert powerful effects on evoked responses, in lower levels, that define the specialisation of any area or neuronal population. Modulatory effects imply the post-synaptic response evoked by pre-synaptic input is modulated by, or interacts with, another. By definition this interaction must depend on non-linear synaptic or dendritic mechanisms. In Section 3 we describe a theoretical perspective, provided by generative models, that highlights the functional importance of hierarchies, backward connections, with non-linear coupling and associative plasticity.

3. Representational learning

This section compares and contrasts the heuristics behind three prevalent computational approaches to representational learning, *supervised learning*, and two forms of self-supervised learning based on *information theory* and *predictive coding*. This section concludes with the introduction of learning based on empirical Bayes. These approaches are considered within the framework of *generative models*. This section follows Dayan and Abbot (2001, pp. 359–397) to which the reader is referred for more detailed background. A more heuristic discussion of these issues can be found in Friston (2002a,b). The more mathematical sections are divided into a conceptual overview and a computational subsection for the interested reader.

We start with an overview of representations in which the distinctions among various approaches can be seen clearly. An important focus of this section is the interaction among causes of sensory input. These interactions create a problem of contextual invariance. In brief, it will be shown that this problem points to the adoption of generative models where interactions among causes of a percept are modelled explicitly in backward connections. After establishing a framework that covers representational learning in a general way, specific examples are reviewed. Each example is presented as a generalisation of the preceding example, by successively relaxing assumptions or constraints under which learning proceeds. We start with supervised learning in which both the prior distribution of the underlying causes of sensory inputs and the processes generating them are known. We then consider self-supervised approaches in which the generative processes are learned but under the assumption of independent causes. Predictive coding is presented as an example that relaxes the independence assumption but still depends upon known priors. Finally, we consider empirical Bayes in which both the generating process and priors are learned. At this point neuronal implementation is considered in sufficient depth to make predictions about the anatomical and functional architectures that would be needed to implement empirical Bayes in the brain. We conclude by relating theoretical predictions with the neurobiological principles listed at the end of the previous section.

3.1. The nature of inputs, causes and representations

Here a representation is taken to be a neuronal event that represents some 'cause' in the sensorium. Causes are simply the states of processes generating sensory data or input. It is not easy to ascribe meaning to these states without appealing to the way that we categorise things, perceptually or conceptually. High-level conceptual causes may be categorical in nature, such as the identity of a face in the visual field or the semantic category a perceived object belongs to. In a hierarchical setting, high-level causes may induce priors on lower-level causes that are more parametric in nature. For example, the perceptual cause 'moving quickly' may show a one-to-many relationship with representations of different velocities in V5 (MT) units. Causes have relationships to each other (e.g. 'is part of') that often have a hierarchical structure. This hierarchical ontology is attended by ambiguous many-to-one and one-to-many mappings (e.g. a table has legs but so do horses; a wristwatch is a watch irrespective of the orientation of its hands). This ambiguity can render the problem of inferring causes from sensory information under-determined or ill posed.

Even though causes may be difficult to describe, they are easy to define operationally. Causes are quantities or states that are necessary to specify the products of a process generating sensory information. To keep things simple, let us frame the problem of representing causes in terms of a deterministic non-linear generative function

$$u = G(v, \theta) \tag{1}$$

where v is a vector (i.e. a list) of underlying causes in the environment (e.g. the velocity of a particular object, direction of radiant light, etc.), and u represents some sensory inputs. $G(v, \theta)$ is a function that generates inputs from the causes. θ are the parameters of generative model. Unlike the causes, they are fixed quantities that have to be learned. We shall see later that the parameters correspond to connection strengths in the brain's model of how inputs are caused. Non-linearities in Eq. (1) represent interactions among the causes. Second-order interactions are formally identical to interaction terms in conventional statistical models of observed data. These can often be viewed as contextual effects, where the expression of a particular cause depends on the context established by another. For example, the extraction of motion from the visual field depends upon there being sufficient luminance or wavelength contrast to define the surface moving. Another ubiquitous example, from early visual processing, is the occlusion of one object by another. In the absence of interactions we would see a linear superposition of both objects but the visual input, caused by the non-linear mixing of these two causes, renders one occluded by the other. At a more cognitive level the cause associated with the word 'HAMMER' will depend on the semantic context (that determines whether the word is a verb or a noun). These contextual effects are profound and must be discounted before the representations of the underlying causes can be considered veridical.

The problem the brain has to contend with is to find a function of the input that recognises or represents the underlying causes. To do this, the brain must effectively undo the interactions to disclose contextually invariant causes. In other words, the brain must perform some form of non-linear unmixing of causes and context without knowing either. The key point here is that this non-linear mixing may not be invertible and that the estimation of causes from input may be fundamentally ill posed. For example, no amount of unmixing can discern the parts of an object that are occluded by another. The mapping $u = v^2$ provides a trivial example of this non-invertibility. Knowing u does not uniquely determine v, which could be negative or positive. The corresponding indeterminacy, in probabilistic learning, rests on the combinatorial explosion of ways in which stochastic generative models can generate input patterns (Dayan, Hinton, & Neal, 1995). The combinatorial explosion represents another example of the uninvertible 'many-to-one' relationship between causes and inputs.

In probabilistic learning, one allows for stochastic (i.e. random) components in the generation of inputs and recognising a particular cause becomes probabilistic. Here the issue of deterministic invertibility is replaced by the existence of an inverse conditional probability (i.e. recognition) density that can be parameterised. Although not a mathematical fundament, parameterisation is critical for the brain because it has to encode the parameters of these densities with biophysical attributes of its nervous tissue. In what follows, we consider the implications of this problem. In brief, we will show that one needs separate (approximate) recognition and generative models that induces the need for both forward and backward influences. Separate recognition and generative models resolve the problem caused by generating processes that are difficult to invert and speak to a possible role for backward connections in the brain.

3.2. Generative models and representational learning

3.2.1. Conceptual overview

In this subsection we introduce the basic framework within which one can understand learning and inference. This framework rests upon generative and recognition models, which are simply functions that map causes to sensory input or vice versa. The objective of learning is to build internal models that can explain observed inputs in terms of some inferred causes. Making inferences about causes (e.g. the most likely cause, or how certain we are that the cause falls within some interval) depends on some representation of the relative probabilities of values the causes can take. This entails representing the probability distribution or density of the causes. The key density is the conditional or posterior density that summarises the likelihood of any cause given the input. This section establishes the nature of this density and how it relates to the underlying models.

Generative models afford a generic formulation of representational leaning in a supervised or self-supervised context. There are many forms of generative models that range from conventional statistical models (e.g. factor and cluster analysis) and those motivated by Bayesian inference and learning (e.g. Dayan et al., 1995; Hinton, Dayan, Frey, & Neal, 1995). The goal of generative models is "to learn representations that are economical to describe but allow the input to be reconstructed accurately" (Hinton et al., 1995). Representational learning is framed in terms of estimating probability densities of the causes. This is referred to as posterior density analysis in the estimation literature and posterior mode analysis if the inference is restricted to estimating the most likely cause. The mode of a distribution is the location of its maximum. Although density learning is formulated at a level of abstraction that eschews many issues of neuronal implementation (e.g. the dynamics of real-time learning), it provides a unifying framework that connects the various schemes considered below. It is important to appreciate the distinction between simply estimating the most likely cause (i.e. mode) and the broader problem of inference. Inference entails estimating the conditional density not just its mode.

3.2.2. Inference vs. learning

Eq. (1) relates the unknown causes v and some unknown parameters θ, to observed inputs u. The objective is to make *inferences* about the causes and *learn* the parameters. Inference may be simply estimating the most likely causes and is based on the products of learning. A useful way of thinking about the distinction between inference and learning is in terms of how one accounts for the patterns or distribution of inputs encountered. Fig. 2 shows a very simple example with a univariate cause and a bivariate observation. Observations are denoted by dots in the right hand panel and cluster around a curvilinear line. A parsimonious way of generating dots like these would be to move up and down the line and add a small amount of observation error. The position on the line corresponds to the state of the single cause and the probability of selecting a particular position to the probability density of the causes on the right. Inference means ascertaining the probability of each potential cause given an observation. *Estimation* refers to estimating the most likely cause, denoted in Fig. 2 by $\hat{v}$. This estimate is the closest point on the line to the observation that a priori has a reasonable probability of being selected. This simple example introduces the notion of representing observations in terms of points that lie on a low dimensional manifold in observation space, in this case a line. The dimensions of this manifold are the causes. The shape and position of the manifold depends on the parameters θ. These have to be known or learned before inference about any particular observation can proceed. This learning requires multiple observations so that the manifold can be placed to transect the highest density of observations. In short, representational learning can be construed as learning a low dimensional manifold onto which data can be projected with minimum loss of information. This manifold is an essential component of generative models.

The goal of learning is to acquire a recognition model for inference that is effectively the inverse of a generative model. The generative model creates data from causes and the inverse model recognises causes from data. Learning a generative model corresponds to making the density of inputs, implied by a generative model $p(u;\theta)$, as close as possible to those observed $p(u)$. The generative model is specified in terms of a prior distribution over the causes $p(v;\theta)$ and the *generative* distribution or likelihood of the inputs given the causes $p(u|v;\theta)$. Together, these define the marginal distribution that has to be matched to the input distribution

$$p(u;\theta) = \int p(u|v;\theta)p(v;\theta)\mathrm{d}v \tag{2}$$

See Fig. 2. Once the parameters of the generative model have been learned, through this matching, the posterior density of the causes, given the inputs is given by the recognition model, which is defined in terms of the *recognition* distribution

$$p(v|u;\theta) = \frac{p(u|v;\theta)p(v;\theta)}{p(u;\theta)} \tag{3}$$

However, as considered above, the generative model may not be easily inverted and it may not be possible to parameterise the recognition distribution. This is crucial because the endpoint of learning is the acquisition of a useful recognition model that can be applied to sensory inputs. One solution is to posit an approximate recognition distribution $q(v;u,\phi)$ that is consistent with the generative model and that can be learned at the same time. The approximate recognition distribution has some parameters ϕ, for example, the strength of forward connections. The first question addressed in this section is whether forward connections are sufficient for representational leaning.

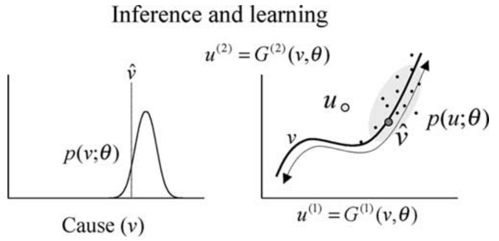

Fig. 2. Schematic of a simple model with a univariate cause and a bivariate observation. Observations are denoted by dots in the right hand panel and cluster around a curvilinear line. A parsimonious way of generating dots like these would be move up and down the line and add a small amount of random error. The position on the line corresponds to the state of the single cause and the probability of selecting a particular position the probability density of the causes on the right.

3.3. Density estimation and EM

3.3.1. Conceptual overview

In this subsection we introduce a general scheme for representational learning using expectation maximisation. Section 3.2 established that the objective is to estimate the parameters of an approximate recognition density $q(v;u,\phi)$ for some generative model. This objective can be split into two steps. First, ensure that the recognition density is consistent with the generative model, noting one is the inverse of the other. Second, adjust the parameters of the generative model to fully account for the data. These two steps correspond to the expectation and maximisation steps, respectively. We will now look more closely at what these steps entail and specify operationally what they are trying to attain in terms of an objective function. An objective function is a function of the parameters and specifies how 'good' they are. As we will see below, the objective function embodies both the internal consistency of the recognition and generative models and the likelihood of the data given the generative model.

A key point made in this subsection is that if the generative model can be inverted easily, then there is no need for the expectation step because internal consistency between the generative and recognition models is assured. From a neurobiological perceptive this means only one set of (recognition) parameters is required. Conversely, if the inversion is difficult, both recognition and generative parameters are needed. We will see later this means backward connection become necessary.

3.3.2. The objective function

In density learning, representational learning has two components that are framed in terms of expectation maximisation (EM, Dempster, Laird, & Rubin, 1977). Iterations of an **E**-step ensure the recognition approximates the inverse of the generative model and the **M**-step ensures that the generative model can predict the observed inputs. Probabilistic recognition proceeds by using $q(v; u, \phi)$ to determine the probability that v caused the observed sensory inputs. EM provides a useful procedure for density estimation that helps relate many different models within a framework that has direct connections with statistical mechanics. Both steps of the EM algorithm involve maximising a function of the densities that corresponds to the negative free energy in physics

$$
\begin{aligned}
F &= \langle l(u) \rangle_u \\
l &= \int q(v; u, \phi) \ln \frac{p(v, u; \theta)}{q(v; u, \phi)} \mathrm{d}v \\
&= \langle \ln p(v, u; \theta) \rangle_q - \langle \ln q(v; u, \phi) \rangle_q \\
&= \ln p(u; \theta) - \mathrm{KL}\{q(v; u, \phi), p(v|u; \theta)\}
\end{aligned}
\tag{4}
$$

This objective function comprises two terms. The first is the expected log likelihood of the inputs under the generative model. The second term is the Kullback–Leibler (KL) divergence[1] between the approximating and true recognition densities. Critically, the KL term is always positive, rendering F a lower bound on the expected log likelihood of the inputs. Maximising F encompasses two components of representational learning: (i) it increases the likelihood of the inputs produced by the generative model and (ii) minimises the discrepancy between the approximate recognition model and that implied by the generative model. The **E**-step increases F with respect to the recognition parameters ϕ, ensuring a veridical approximation to the recognition distribution implied by the generative parameters θ. The **M**-step changes θ, enabling the generative model to reproduce the inputs

$$
\mathbf{E} \quad \phi = \max_{\phi} F \qquad \mathbf{M} \quad \theta = \max_{\theta} F \tag{5}
$$

There are a number of ways of motivating the free energy formulation in Eq. (4). A useful one, in this context, rests upon the problem posed by non-invertible models. This problem is finessed by assuming it is sufficient to match the joint probability of inputs and causes under the generative model $p(u, v; \theta) = p(u|v; \theta)p(v; \theta)$ with that implied by recognising the causes of inputs encountered $p(u, v; \phi) = q(v; u, \phi)p(u)$. Both these distributions are well defined even when $p(v|u; \theta)$ is not easily parameterised. This matching minimises the divergence

$$
\begin{aligned}
&\mathrm{KL}\{p(v, u; \phi), p(v, u; \theta)\} \\
&= \int q(v; u, \phi)p(u) \ln \frac{q(v; u, \phi)p(u)}{p(v, u; \theta)} \mathrm{d}v\, \mathrm{d}u \\
&= -F - H(u)
\end{aligned}
\tag{6}
$$

This is equivalent to maximising F because the entropy of the inputs $H(u)$ is fixed. This perspective is used in Fig. 3 to illustrate the **E** and **M** steps schematically. The **E**-step adjusts the recognition parameters to match the two joint distributions, while the **M**-step does exactly the same thing but by changing the generative parameters. The dependency of the generative parameters, on the input distribution, is

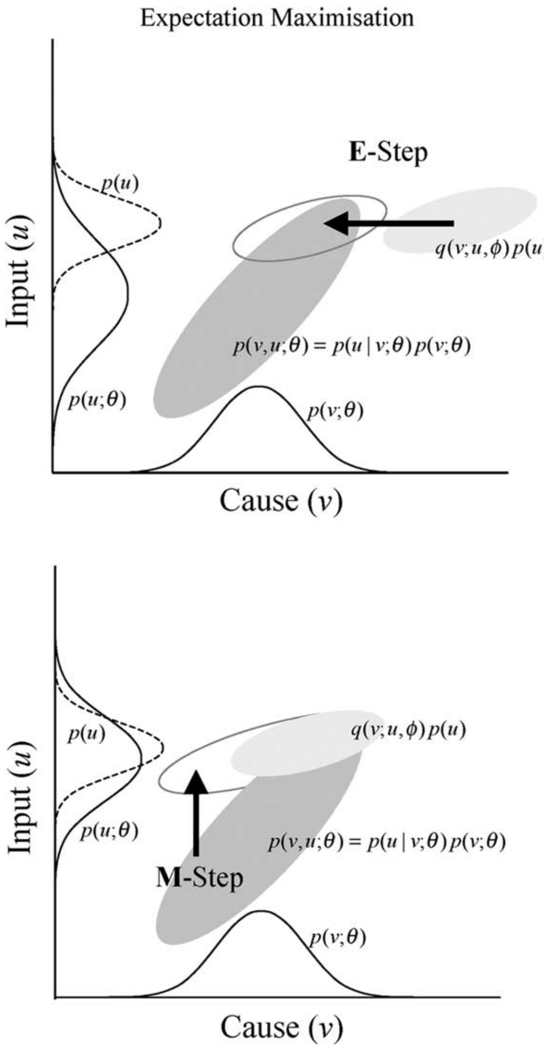

Fig. 3. Schematic illustrating the two components of EM. In the **E**-step the joint distribution of causes and inputs under the recognition model changes to approximate that under the generative model. This refines the recognition model. In the **M**-step the joint distribution under the generative model changes to approximate that under the recognition model. This reduces the difference between the distribution of inputs implied by the generative model and that observed.

[1] A measure of the distance or difference between two probability densities.

mediated vicariously in the **M**-step through the recognition. In the setting of invertibility, where $q(v; u, \phi) = p(v|u; \theta)$ the divergence in Eq. (6) reduces to $\mathrm{KL}\{p(u), p(u; \theta)\}$. As above, the **M**-step then finds parameters that allow the model to simply match the observed input distribution (i.e. maximise the expected likelihood).

3.3.3. Invertibility

This formulation of representational leaning is critical for the thesis of this section because it suggests that backward and lateral connections, parameterising a generative model, are essential when the model is not invertible. If the generative model is invertible then the KL term in Eq. (4) can be discounted by setting $q(v; u, \phi) = p(v|u; \theta)$ with Eq. (3) and learning reduces to the **M**-step (i.e. maximising the expected likelihood).

$$F = \langle \ln p(u; \theta) \rangle_u \tag{7}$$

See Fig. 4 (upper panel). In principle, this could be done using a feedforward architecture corresponding to the inverse of the generative model. However, when processes generating inputs are non-invertible (in terms of the parameterisation of the recognition density) a generative model and approximate recognition model are required that are updated in **M**- and **E**-steps, respectively. In short, non-invertibility enforces an explicit parameterisation of the generative model in representational learning. In the brain this parameterisation may be embodied in backward and lateral connections.

3.3.4. Deterministic recognition

Another special case arises when the recognition is deterministic. The recognition becomes deterministic when $q(v; u, \phi)$ is a Dirac δ-function over its mode $v(u, \phi)$ (i.e. reduces to a point). In this instance, posterior density analysis reduces to a posterior mode analysis at which point inference and estimation coincide. They are equivalent in the sense that inferring the posterior distribution of causes is the same as estimating the most likely cause given the inputs (the maximum a posteriori or MAP estimator). Here the integral in Eq. (4) disappears, leaving the joint probability of the inputs and their cause to be maximised

$$F = \langle \ln p(v(u), u; \theta) \rangle_u = \langle \ln p(u|v(u); \theta) + \ln p(v(u); \theta) \rangle_u \tag{8}$$

Notice, again, that this objective function does not require $p(v|u; \theta)$ and eschews the inversion in Eq. (3). An illustration of the **E**-step for deterministic recognition is shown in Fig. 4 (lower panel). In this article the distinction between deterministic and stochastic relates to inference and refers to form of the recognition density. It should be noted that learning could also employ a deterministic or stochastic ascent on F. We will deal largely with deterministic learning schemes.

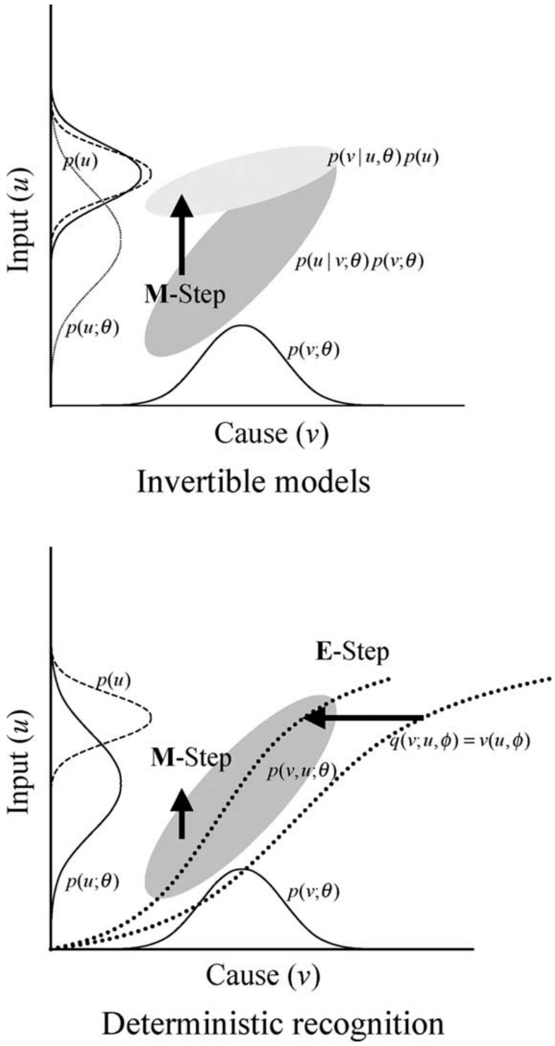

Fig. 4. As for Fig. 2 but for two special cases of learning in which the generative model is invertible (upper panel) and in which the recognition model is deterministic (lower panel). When the model is invertible only the generative parameters need to be learned in the **M**-step. This learning reduces the difference between the distribution of inputs implied by the generative model and that observed. In deterministic recognition the recognition density reduces to a Dirac δ-function (i.e. point) over the point estimator of the causes $v(u, \phi)$.

EM enables exact and approximate maximum likelihood density estimation for a whole variety of generative models that can be specified in terms of prior and generative distributions. Dayan and Abbot (2001) work though a series of didactic examples from cluster analysis to independent component analyses, within this unifying framework. For example, factor analysis corresponds to the generative model

$$p(v; \theta) = N(v : 0, 1) \qquad p(u|v; \theta) = N(u : \theta v, \Sigma) \tag{9}$$

Namely, the underlying causes of inputs are independent normal variates that are mixed linearly and added to Gaussian noise to form inputs. $N(u : \theta v, \Sigma)$ means a normal distribution over u with a mean of θv and variance–covariance Σ. In the limiting case of $\Sigma \to 0$ the ensuing model become deterministic and conforms to PCA. By simply assuming non-Gaussian priors one can specify generative models for sparse coding of the sort proposed by Olshausen and Field (1996)

$$p(v;\theta) = \prod p(v_i;\theta) \qquad p(u|v;\theta) = N(u : \theta v, \Sigma) \tag{10}$$

where $p(v_i;\theta)$ are chosen to be suitably sparse (i.e. heavy-tailed) with a cumulative density function that corresponds to the squashing function in Section 3.5.5. The deterministic equivalent of sparse coding is ICA that obtains when $\Sigma \to 0$. The relationships among different models are rendered apparent under the perspective of generative models. In what follows we consider a series of models entailing assumptions about the generation of sensory inputs that are relaxed one by one. At each point we consider whether they could be implemented plausibly in the brain.

3.4. Supervised learning

3.4.1. Conceptual overview

To start we will review briefly supervised learning, using connectionist models in cognitive science as a paradigm example. Supervised learning deals with the simplest problem in which the parameters of the generative model are known, allowing one to generate simulated sensory inputs from causes with a known prior distribution. Although supervised learning schemes have an established utility in helping understand some aspects of functional architectures in the brain they are not candidates for models of representational learning. This is because their supervised aspect means the generative model is already known. From the point of view of expectation maximisation, only the first step is required to find the parameters of the recognition density. In this subsection we place supervised learning in the framework described above and touch upon some of their useful applications in cognitive neuroscience.

3.4.2. Implementation

In supervised schemes the generative model is pre-specified and only the recognition parameters need to be learned. The generative model is known in the sense that any cause determines the input, either deterministically or stochastically. In this case only the **E**-step is required in which the parameters ϕ that specify $q(v;u,\phi)$ change to maximise F. The only term in Eq. (4) that depends on ϕ is the divergence term, such that learning reduces to minimising the expected difference between the approximate recognition density and that required by the generative model

$$\mathbf{E} \qquad \phi = \max_{\phi} F = \min_{\phi} \langle \mathrm{KL}(q(v;u,\phi), p(v|u;\theta)) \rangle_u \tag{11}$$

This can proceed probabilistically (e.g. contrastive Hebbian learning in stochastic networks, Dayan & Abbott, 2001, p. 322) or deterministically. In a deterministic setting, the connection strengths ϕ (usually connecting multiple layers of nodes) are changed, typically using the delta rule, such that the distance between the modes of the approximate and desired recognition distributions are minimised over all inputs. This distance is typically measured by the average sum of squared difference between the recognised and true causes (see Fig. 5, upper panel). Supervised learning, of this sort, is equivalent to non-linear function approximation, a perspective that can be adopted on all supervised learning of deterministic mappings with neural nets. Note that any scheme, based on supervised learning, requires the processes generating inputs to be known a priori and as such cannot be used by the brain. However, supervised learning has a substantial role in understanding some aspects of functional anatomy.

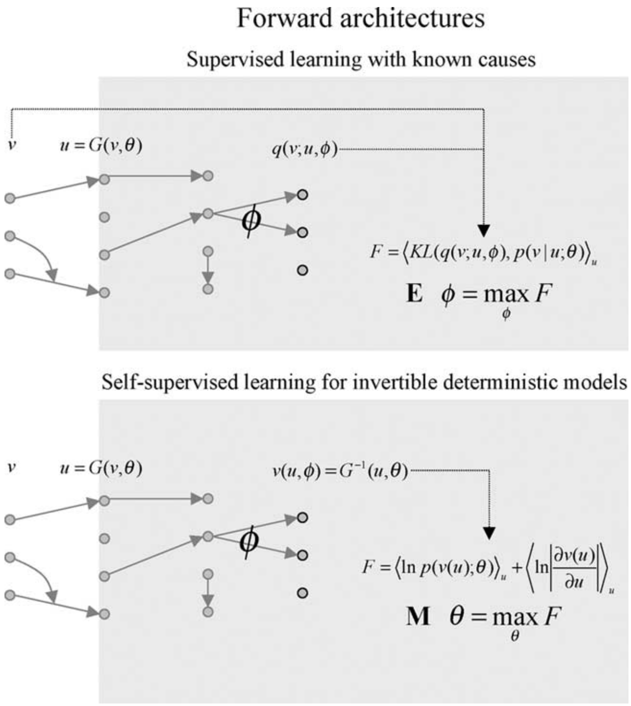
Forward architectures
Supervised learning with known causes
v
$u = G(v,\theta)$
$q(v;u,\phi)$
ϕ
$F = \langle KL(q(v;u,\phi), p(v|u;\theta)) \rangle_u$
$\mathbf{E} \quad \phi = \max_{\phi} F$
Self-supervised learning for invertible deterministic models
v
$u = G(v,\theta)$
$v(u,\phi) = G^{-1}(u,\theta)$
ϕ
$F = \langle \ln p(v(u);\theta) \rangle_u + \left\langle \ln \left| \frac{\partial v(u)}{\partial u} \right| \right\rangle_u$
$\mathbf{M} \quad \theta = \max_{\theta} F$

Fig. 5. Schematic illustrating architectures with forward connections that are sufficient when the generative model is known (supervised learning) or can be inverted (infomax). The circles represent nodes in a network and the arrows represent a few of the connections. See the main text for an explanation of the equations and designation of the variables each set of nodes represents. The light grey boxes encompass connections and nodes within the model. Connection strengths are determined by the free parameters of the model ϕ (forward connections).

3.4.3. Category specificity and connectionism

Supervised learning in the context of connectionism is an approach that has proved very useful in relating putative cognitive architectures to neuronal ones and, in particular, modelling the impact of brain lesions on cognitive performance. Semantic memory impairments can result from a variety of pathophysiological insults, including Alzheimer disease, encephalitis and cerebrovascular accidents (e.g. Nebes, 1989; Warrington & Shallice, 1984). The concept of category specificity stems from the work of Warrington and colleagues (Warrington & McCarthy, 1983; Warrington & Shallice, 1984) and is based on the observation that patients with focal brain lesions have difficulties in recognising or naming specific categories of objects. Patients can exhibit double dissociations in terms of their residual semantic capacity. For example, some patients can name artefacts but have difficulty with animals, whereas others can name animals with more competence than artefacts. These findings have engendered a large number of studies, all pointing to impairments in perceptual synthesis, phonological or lexico-semantic analysis that is specific for certain categories of stimuli. There are several theories that have been posited to account for category specificity. Connectionist models have been used to adjudicate among some of them.

Connectionist (e.g. parallel distributed processing or PDP) techniques use model neuronal architectures that can be lesioned to emulate neuropsychological deficits. This involves modelling semantic networks using connected units or nodes and suitable learning algorithms to determine a set of connection strengths (Rumelhart & McClelland, 1986). Semantic memory impairments are then simulated by lesioning the model to establish the nature of the interaction between neuropathology and cognitive deficit (e.g. Hinton & Shallice, 1991; Plaut & Shallice, 1993). A compelling example of this sort of approach is the connectionist model of Farah and McClelland (1991): patterns of category-specific deficits led Warrington and McCarthy (1987) to suggest that an animate/inanimate distinction could be understood in terms of a differential dependence on functional and structural (perceptual) features for recognition. For example, tools have associated motor acts whereas animals do not, or tools are easier to discriminate based upon their structural descriptions than four-legged animals. Farah and McClelland (1991) incorporated this difference in terms of the proportion of the two types of semantic featural representations encoding a particular object, with perceptual features dominating for animate objects and both represented equally for artifacts. Damage to visual features led to impairment for natural kinds and conversely damage to functional features impaired the output for artefacts. Critically the model exhibited category-specific deficits in the absence of any category-specific organisation. The implication here is that an anatomical segregation of structural and functional representations is sufficient to produce category-specific deficits following focal brain damage. This example serves to illustrate how the connectionist paradigm can be used to relate neuronal and cognitive domains. In this example, connectionist models were able to posit a plausible anatomical infrastructure wherein the specificity of deficits, induced by lesions, is mediated by differential dependence on either the functional or structural attributes of an object and not by any (less plausible) category-specific anatomical organisation per se.

In summary, connectionist models specify distributed profiles of activity over (semantic) primitives that are induced by (conceptual) causes and try to find connection parameters that emulate the inverse of these known mappings. They have been used to understand how the performance (storage and generalisation) of a network responds to simulated damage, after learning is complete. However, connectionism has a limited role in understanding representational learning per se because the brain is not given a generative model, it must be learned. Next we will look at self-supervised approaches that do not require the generative distribution to be known a priori.

3.5. Information theory and efficient coding

3.5.1. Conceptual overview

In the previous section we had assumed both the generative $p(u|v;\theta)$ and prior $p(v;\theta)$ distributions were known. In this section we consider self-supervised schemes that can be regarded as a generalisation of supervised learning and that do not require the parameters of the generative distribution. We first consider deterministic models and then turn to stochastic models. This section focuses on the links between the EM schemes and information maximisation procedures. These rest on the prior assumption that the causes are independent such that $p(v;\theta) = \prod p(v_i;\theta)$ where v_i represents the ith cause. This equation simply states the prior probability of several causes is the product of each considered alone. This factorisation means the causes do not interact, or depend on each other, in terms of their expression.

It transpires that the independence assumption renders the objective function equivalent to a measure of how efficiently inputs are encoded. This is reflected in the average information (i.e. entropy) expressed by the inferred causes. This is important because it connects density learning with a large body of work that uses information theory to understand how the brain might encode its sensory inputs in an efficient fashion with minimal loss of information. In short, we will see that the principle of maximum information transfer (i.e. infomax, Linsker, 1990) is exactly the same as expectation maximisation under the prior assumption of independent causes.

3.5.2. Implementation

For invertible deterministic models the KL term in Eq. (4) can be discounted and learning reduces to maximising

$$\begin{aligned} F &= \langle \ln p(u;\theta)\rangle_u = \langle \ln p(v(u);\theta)|\frac{\partial v(u)}{\partial u}|\rangle_u \\ &= \langle \ln p(v(u);\theta)\rangle_u + \langle \ln|\frac{\partial v(u)}{\partial u}|\rangle_u \\ &= \sum \langle \ln p(v_i(u);\theta)\rangle_u + H(v(u);\phi) - H(u) \end{aligned} \tag{12}$$

The first term of the last line serves to constrain the marginal distribution of the estimated causes while the second requires their entropy to be maximised under these constraints. This is the essence of *infomax* procedures (Linsker, 1990) and can be understood as maximising the mutual information between the estimated causes and inputs as discussed below.

3.5.3. Infomax

There have been many compelling developments in theoretical neurobiology that have used information theory (e.g. Barlow, 1961; Foldiak, 1990; Linsker, 1990; Oja, 1989; Optican & Richmond, 1987; Tononi, Sporns, & Edelman, 1994; Tovee, Rolls, Treves, & Bellis, 1993). Many appeal to the principle of maximum information transfer (e.g. Atick & Redlich, 1990; Bell & Sejnowski, 1995; Linsker, 1990). This principle has proven extremely powerful in predicting some of the basic receptive field properties of cells involved in early visual processing (e.g. Atick & Redlich, 1990; Olshausen & Field, 1996). This principle represents a formal statement of the common sense notion that neuronal dynamics in sensory systems should reflect, efficiently, what is going on in the environment (Barlow, 1961). In the present context, the principle of maximum information transfer (infomax; Linsker, 1990) suggests that a model's parameters should maximise the mutual information between the sensory input u and the evoked responses or outputs $v(u, \phi)$. This maximisation is usually considered in the light of some sensible constraints, e.g. the presence of noise in sensory input (Atick & Redlich, 1990) or dimension reduction (Oja, 1989) given the smaller number of divergent outputs from a cortical area than convergent inputs (Friston et al., 1992).

The mutual information between inputs and responses is given by

$$I(u, v) = H(u) + H(v) - H(u, v) = H(v) - H(v|u) \tag{13}$$

where $H(v|u)$ is the conditional entropy or uncertainty in the response, given the input. For deterministic recognition there is no such uncertainty and this term can be discounted (see Bell & Sejnowski, 1995). More generally

$$\frac{\partial}{\partial \phi} I(u, v; \phi) = \frac{\partial}{\partial \phi} H(v; \phi) \tag{14}$$

It follows that maximising the mutual information is the same as maximising the entropy of the responses. The infomax principle (maximum information transfer) is closely related to the idea of efficient coding. Generally speaking, redundancy minimisation and efficient coding are all variations on the same theme and can be considered as the infomax principle operating under some appropriate constraints or bounds. The key thing that distinguishes among the various information theoretic schemes is the nature of the constraints under which entropy is maximised. One useful way of looking at constraints is in terms of efficiency.

3.5.4. Efficiency and redundancy

The efficiency of a system can be considered as the complement of redundancy (Barlow, 1961), the less redundant, the more efficient a system will be. Redundancy is reflected in the dependencies *or mutual information among the outputs* (c.f. Gawne & Richmond, 1993)

$$I(v; \phi) = \sum H(v_i; \phi) - H(v; \phi) \tag{15}$$

Here $H(v_i; \phi)$ is the marginal entropy of the ith output. Eq. (15) implies that redundancy is the difference between the joint entropy and the sum of the marginal entropies. Intuitively this expression makes sense if one considers that the variability in activity of any single unit corresponds to its entropy. Therefore, an efficient neuronal system represents its inputs with the minimal excursions from baseline firing rates. Another way of thinking about Eq. (15) is to note that maximising efficiency is equivalent to minimising the mutual information among the outputs. This is the basis of approaches that seek to de-correlate or orthogonalise the outputs. To minimise redundancy one can either minimise the entropy of the output units or maximise their joint entropy, while ensuring the other is bounded in some way. Olshausen and Field (1996) present a nice analysis based on sparse coding. Sparse coding minimises redundancy using single units with low entropy. Sparse coding implies coding by units that fire very sparsely and will, generally, not be firing. Therefore, one can be relatively certain about their (quiescent) state, conferring low entropy on them.

The relationship between Eqs. (12) and (15) rests on the nature of the prior distribution. If we relax constraints on the form of the marginal distributions of $v_i(u)$ such that $p(v(u); \theta) = \prod p(v_i(u))$ then F becomes a functional of the recognition parameters

$$\begin{aligned} F &= \sum \langle \ln p(v_i(u)) \rangle_u + H(v; \phi) - H(u) \\ &= -\sum H(v_i; \phi) + H(v; \phi) - H(u) \\ &= -I(v; \phi) - H(u) \end{aligned} \tag{16}$$

This has exactly the same dependence on the parameters as the objective function employed by infomax in Eq. (15). In this context, the free energy and the information differ only by the entropy of the inputs. In other words minimising the free energy is the same as minimising the redundancy or maximising the efficiency with which the causes of inputs are encoded. This equivalence rests on use of maximum entropy priors of the sort used in sparse coding. Although the infomax and density learning approaches have the same objective, their heuristics are complementary. Infomax is motivated by maximising the mutual information between u and $v(u, \phi)$ under some constraints. The generative model approach takes its heuristics from the assumption that the causes of inputs are independent and possibly non-Gaussian. This results in a prior with maximum entropy $p(v; \theta) = \prod p(v_i; \theta)$. The reason for adopting non-Gaussian priors (e.g. sparse coding and ICA) is that the central limit theorem implies mixtures of causes will have Gaussian distributions and therefore something that is not Gaussian is unlikely to be a mixture.

3.5.5. Invertible models

It is often the case that the constraints on the marginal distributions are absorbed into the parameters of the recognition function. In this instance learning reduces to maximising $H(v;\phi)$. A simple example of this is PCA, which samples the subspace of the inputs that have the highest entropy (e.g. Foldiak, 1990). PCA is conventionally regarded as maximising $H(v;\phi)$, for deterministic recognition $v(u,\phi) = \phi u$ where the sum of squared connections in each row of ϕ is constant. Similarly, ICA finds non-linear functions of the inputs that maximise the joint entropy (Bell & Sejnowski, 1995; Common, 1994). The marginal entropies are constrained by passing the outputs through a sigmoid squashing function $v(u,\phi) = g(\phi u)$ so that the outputs lie in a bounded interval (hypercube). Learning can then proceed by maximising $\langle \ln|\partial v(u)/\mathrm{d}u|\rangle_u$ in Eq. (12). See Eq. (10) for the implicit generative model, in which the outputs are not bounded but forced to have cumulative density functions that conform to the squashing function g.

As noted above, PCA and ICA are based on linear deterministic generative models with independent Gaussian and non-Gaussian priors, respectively. Generalising to non-deterministic models with Gaussian errors gives factor analysis and sparse coding, respectively (see Eqs. (9) and (10)). In principle, all these models are probabilistically invertible. However, it is difficult to parameterise the inverse of the sparse coding model in Eq. (10) and a deterministic approximating recognition density is used (see below).

3.5.6. Non-invertible models

In the context of invertible deterministic generative models, the parameters of the recognition model specify the generative model and only the recognition model, i.e. forward connections constructing $v(u) = v(u,\phi)$ need to be instantiated. If the generative model cannot be inverted, or the inversion cannot be parameterised, the recognition model is not defined and analytic infomax schemes are precluded. In this instance, one has to parameterise both an approximate recognition and generative model as required by EM. This enables the use of non-linear generative models, such as the Helmholtz machine (Dayan et al., 1995) for binary stochastic systems and non-linear PCA for parametric deterministic models (e.g. Dong & McAvoy, 1996; Friston et al., 2000; Karhunen & Joutsensalo, 1994; Kramer, 1991; Taleb & Jutten, 1997). The latter schemes typically employ a 'bottleneck' architecture that forces the inputs through a small number of nodes. The output from these nodes then diverges to produce the predicted inputs. The approximate recognition model is implemented in connections to the bottleneck nodes and the generative model by connections from these nodes to the outputs. Non-linear transformations, from the bottleneck nodes to the output layer, recapitulate the non-linear mixing of the real causes of the inputs. If such a scheme was implemented in the brain, one would envisage the recognition to be implemented in forward connections and learned in the **E**-step. Critically, the generative model would be mediated by backward connections from the bottleneck nodes, so that the output nodes were juxtaposed to the inputs being predicted (see Fig. 6, upper panel). The backward connections are updated in the **M**-step. After learning, the activity of the bottleneck nodes can be treated as estimates of the causes. These representations are obtained by projection of the input onto a low-dimensional curvilinear manifold that is encompassed by the activity of the bottleneck nodes (c.f. Fig. 2).

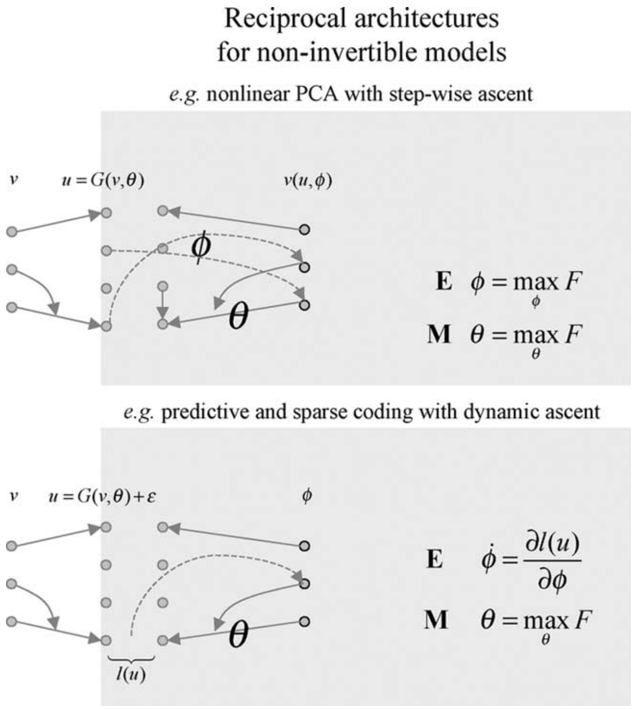
Reciprocal architectures
for non-invertible models
e.g. nonlinear PCA with step-wise ascent
v
$u = G(v,\theta)$
$v(u,\phi)$
ϕ
θ
E $\phi = \max_{\phi} F$
M $\theta = \max_{\theta} F$
e.g. predictive and sparse coding with dynamic ascent
v
$u = G(v,\theta) + \varepsilon$
ϕ
θ
$l(u)$
E $\dot{\phi} = \frac{\partial l(u)}{\partial \phi}$
M $\theta = \max_{\theta} F$

Fig. 6. Schematic illustrating architectures with forward and backward connections that are necessary when the generative model is unknown and cannot be inverted. This figure uses the same format as Fig. 4. Learning involves changes in the free parameters of the model ϕ (that correspond to forward connections in the upper panel and neuronal states in the lower panel) and θ (backward connections). Non-linear effects are implied when one arrow connects with another. The broken arrows represent forward influences to the higher level from the input level.

3.5.7. Summary

In summary, ICA and like-minded approaches try to find some deterministic recognition function of the inputs that maximises information transfer. Compared to supervised schemes, this has the fundamental advantage that the algorithm is unsupervised by virtue of the fact that the causes and generating process are not needed for leaning. This is shown in Fig. 5, which portrays the main difference between supervised and self-supervised learning in terms of whether F is a function of the true causes. An important aspect of the infomax principle is that it goes a long way to explain functional segregation in the cortex. One perspective on functional segregation is that each cortical area is segregating its inputs into relatively independent functional outputs. This is exactly what infomax predicts. See Friston (2000) and references therein, for an example of how infomax can be used to predict the segregation of processing streams from V2 to specialised motion, colour and form areas in extrastriate cortex.

Infomax assumes that the causes are independent. While this may be sensible for simple systems, it is certainly not appropriate for more realistic hierarchical processes that generate sensory inputs (see below). This is because correlations among causes at any level are induced by changes at supraordinate levels. In Section 3.6 we will look at a more general implementation of representational learning that encompasses any specified form for the priors.

3.6. Predictive coding

3.6.1. Conceptual overview

In the previous section, we considered representational learning from the infomax perspective afforded by independence assumptions about the prior distributions. In this section we relax the independence assumptions and relate predictive coding to the free energy formulation of EM. Predictive coding is based on minimising prediction error. This prediction error is the difference between the observed input and that predicted on the basis of the generative model and inferred causes. It is relatively easy to show that maximising the objective function above is equivalent to minimising prediction error, under some constraints. The nature of these constraints emerge naturally from the EM formulation and are considered, usefully, in the light if ill-posed inverse problems and regularisation (e.g. in machine vision).

This section introduces predictive coding as a general implementation of EM that frees itself from the difficulties of parameterising complicated recognition densities by estimating the first few moments (mean or expectation and covariance) for each input. This leads to an **E**-step that can be implemented ‘on-line’ and takes us a step closer to a neurobiological implementation.

3.6.2. Implementation

To do this we consider how probabilistic, non-linear generative models are learned, under Gaussian assumptions about the stochastic components

$$p(v,\theta) = N(v : \mu_p, \Sigma_p)$$
$$p(u|v;\theta) = N(u : G(v,\theta), \Sigma_u) \qquad (17)$$

where θ includes the prior expectation, prior covariance and observation error covariance (μ_p, Σ_p, Σ_u). This can be formulated as a non-linear observation model with two levels

$$u = G(v,\theta) + \varepsilon_u \qquad v = \mu_p + \varepsilon_p \qquad (18)$$

where $\mathrm{Cov}\{\varepsilon_u\} = \Sigma_u$ and $\mathrm{Cov}\{\varepsilon_p\} = \Sigma_p$. This perspective will be useful in the next section when we generalise to hierarchical models. In the previous section most of the recognition models were deterministic which simplified the parameterisation of the recognition density. Here the approximate recognition density is taken to be probabilistic and Gaussian

$$q(v;u;\phi) = N(u : \phi, \Sigma_q) \qquad \Sigma_q = (J^T\Sigma_u^{-1}J + \Sigma_p^{-1})^{-1}$$
$$J = \frac{\partial G(\phi,\theta)}{\partial v} \qquad (19)$$

The covariance of the recognition density is approximated here by the inverse of the negative curvature of the log posterior $\ln p(v|u;\theta)$ evaluated at its mode. Critically, this covariance is a function of the mode or MAP estimator ϕ meaning that it does not have to be learned. From Eq. (4)

$$\begin{aligned} l(u) &= \langle \ln p(v,u;\theta)\rangle_q - \langle \ln q(v;u,\phi)\rangle_q \\ &= \langle \ln p(u|v;\theta)\rangle_q + \langle \ln p(v;\theta)\rangle_q + H(q;\phi) \\ &\approx -\frac{1}{2}\xi_u^T\xi_u - \frac{1}{2}\xi_p^T\xi_p - \frac{1}{2}\ln|\Sigma_u| - \frac{1}{2}\ln|\Sigma_p| + \frac{1}{2}\ln|\Sigma_q| \end{aligned}$$
$$\xi_u = \Sigma_u^{-1/2}(u - G(\phi,\theta)) \qquad \xi_p = \Sigma_p^{-1/2}(\phi - \mu_p) \qquad (20)$$

with equality for linear models. The **E**-step corresponds to maximising ϕ with respect to the expectation $F = \langle l(u)\rangle_u$. A general solution to this is to find $\phi(u)$ that satisfies $\partial l/\partial\phi = 0$ *for each input*. This is the approach adopted by Olshausen and Field (1996) in their implementation of sparse coding. Assuming dynamics are fast, in relation to changes in input, this can be implemented with gradient ascent. This gives, ignoring time constants

$$\mathbf{E}\ \dot{\phi} = \frac{\partial l(u)}{\partial\phi} \qquad \mathbf{M}\ \dot{\theta} = \frac{\partial F}{\partial\theta} \qquad (21)$$

3.6.3. The nature of predictive coding

There is a subtle but key departure from the previous sections implied by this scheme. In previous schemes we had treated ϕ as the parameters of some static recognition function of the inputs. Here, the recognition parameters are input-specific and correspond to the mode of the recognition distribution for each input. From a neuronal perspective this means that $\phi(u)$ are not forward connections (c.f. connectionist and infomax schemes) but are dynamically encoded by the activity of units in the brain (c.f. predictive and sparse coding). This distinction is illustrated schematically in Fig. 6. An important consequence of this general solution to learning the recognition parameters is that **E**-step learning and inference become the same thing. This is because finding the mode of the recognition distribution, for each input, is the same as inferring its most likely cause. In other words, ϕ is both a parameter of the recognition density and an estimate of the cause. This is nice because one does not have to posit distinct neuronal mechanisms for inference and learning. Both are implicit in the **E**-step, which can be implemented using neuronal-like dynamics that conform to some gradient ascent. In the current discussion, we will take this to be the essence of predictive coding, namely any scheme that finds the mode of the recognition density by dynamically minimising prediction error in an input-specific fashion. The advantages of predictive coding are that inference is implicit in the **E**-step and the use of gradient ascent provides a general solution to arbitrarily complicated and non-linear generative models. Note that predictive coding does not imply deterministic recognition; the recognition density may have a covariance defined by Eq. (19), which is a function of Σ_u and Σ_p. These can be encoded by neuronal activity or, as shown later, the strength of (lateral) connections.

Predictive coding, when defined in this way refers to learning and inference schemes that employ a particular process in the **E**-step, based on minimising predictive error. Consequently, predictive coding is not defined by assumptions about the generative or prior densities (c.f. supervised learning and infomax). Indeed predictive coding makes no generic assumptions other than constraints on the manifolds implicit in any generative functions. However, it should be remembered predictive coding is essentially a process for, as opposed to a category of, representational learning.

The **E**-step is trying to find

$$\min_{\phi}(\xi_u^T\xi_u + V(\phi)) \qquad V(\phi) = \xi_p^T\xi_p - \ln|\Sigma_q| \tag{22}$$

This can be regarded as a minimisation of (whitened) prediction error ξ_u subject to some constraints, here denoted by the potential $V(\phi)$ that does not depend on the input. These constraints comprise a term that penalises deviations from prior expectations and a second term whose dependency on ϕ is mediated by non-linearities in the generative model, through the conditional covariance, Eq. (19). For linear models and non-linear models with deterministic recognition, this second component disappears. These constraints can be viewed in terms of regularisation, a perspective that provided the heuristics for early generative models in machine vision.

3.6.4. Predictive coding and the inverse problem

In predictive coding, the dynamics of units are trying to predict the inputs. As with infomax schemes, the representational aspects of any unit emerge spontaneously as the capacity to predict improves with learning. There is no a priori 'labelling' of the units or any supervision in terms of what a correct response should be (c.f. connectionist approaches). The only correct response is one in which the implicit internal model of the causes and their non-linear mixing is sufficient to predict the input with minimal error, under some constraints.

Conceptually, predictive coding and generative models are related to 'analysis-by-synthesis' (Neisser, 1967). This approach to perception, from cognitive psychology, involves adapting an internal model of the world to match sensory input and was suggested by Mumford (1992) as a way of understanding hierarchical neuronal processing. The idea is reminiscent of Mackay's epistemological automata (MacKay, 1956) which perceive by comparing expected and actual sensory input (Rao, 1999). These models emphasise the role of backward connections in mediating the prediction, at lower or input levels, based on the activity of units in higher levels.

Predictive coding schemes can also be regarded as arising from the distinction between forward and inverse models adopted in machine vision (Ballard, Hinton, & Sejnowski, 1983; Kawato, Hayakawa, & Inui, 1993). Forward models generate inputs from causes (c.f. generative models), whereas inverse models approximate the reverse transformation of inputs to causes (c.f. recognition models). This distinction embraces the non-invertibility of generating processes and the ill-posed nature of inverse problems. As with all underdetermined inverse problems the role of constraints becomes central. In the inverse literature a priori constraints usually enter in terms of regularised solutions. For example: "Descriptions of physical properties of visible surfaces, such as their distance and the presence of edges, must be recovered from the primary image data. Computational vision aims to understand how such descriptions can be obtained from inherently ambiguous and noisy data. A recent development in this field sees early vision as a set of ill-posed problems, which can be solved by the use of regularisation methods" (Poggio, Torre, & Koch, 1985). The architectures that emerge from these schemes suggest that "feedforward connections from the lower visual cortical area to the higher visual cortical area provides an approximated inverse model of the imaging process (optics), while the backprojection connection from the higher area to the lower area provides a forward model of the optics" (Kawato et al., 1993). See also Harth, Unnikrishnan, and Pandya (1987). The connection between this perspective on forward influences and the error minimisation can be seen by finessing the gradient decent in the **E**-step using a Newton–Raphson scheme. For deterministic recognition

E

$$\dot{\phi} = -\left(\frac{\partial^2 l(u)}{\partial\phi^2}\right)^{-1}\frac{\partial l(u)}{\partial\phi} \tag{23}$$

$$= (J^T\Sigma_u^{-1}J + \Sigma_p^{-1})^{-1}(J^T\Sigma_u^{-1/2}\xi_u - \Sigma_p^{-1/2}\xi_p)$$

When $\Sigma_u \rightarrow 1$ and $\Sigma_p^{-1} \rightarrow 0$ this reduces to $\dot{\phi} = J^-\xi_u$ where J^- the generalised inverse of $J = \partial G/\partial v$. In other words, forward influences on the recognition parameters correspond to passing the prediction error through the inverse of the generative or forward model (Kawato et al., 1993). The use of Eq. (23) in the posterior mode analysis of non-linear dynamical systems is discussed in Friston (2002c).

3.6.5. Summary

Predictive coding is a strategy that has some compelling (Bayesian) underpinnings and embeds two concurrent processes. (i) The parameters of the generative or forward model change to emulate the real world mixing of causes, using their current estimates (**M**-step) and (ii) these estimates change to best explain the observed inputs, using the current forward model (**E**-step). Both the parameters and the estimates minimise prediction error. To finesse the inverse problem, posed by non-invertible generative models, constraints are required. These resolve the problem of non-invertibility that confounds simple infomax schemes but emphasise the dependency of representational learning on priors. In the final subsection we consider representational learning when the parameters of the prior distribution are unknown and must be learned. This represents the last generalisation of representational learning that eschews knowledge about the parameters of both the generative and prior distributions. This generalisation speaks to hierarchical models and the notion of empirical Bayes.

3.7. Cortical hierarchies and empirical Bayes

3.7.1. Conceptual overview

The problem with predictive coding is that the brain cannot construct priors μ_p and Σ_p de novo. They have to be learned along with the generative parameters. In Bayesian estimation, priors are estimated from data using empirical Bayes. Empirical Bayes harnesses the hierarchical structure of a forward model, treating the estimates at one level as prior expectations for the subordinate level (Efron & Morris, 1973). This provides a natural framework within which to treat cortical hierarchies in the brain, each providing constraints on the level below. This approach models the world as a hierarchy of systems where supraordinate causes induce, and moderate, changes in subordinate causes. For example, the presence of a particular object in the visual field changes the incident light falling on a particular part of the retina. A more intuitive example is provided in Fig. 7. These priors offer contextual guidance towards the most likely cause of the input. Note that predictions at higher levels are subject to the same constraints, only the highest level, if there is one in the brain, is free to be directed solely by bottom-up influences (although there are always implicit priors). If the brain has evolved to recapitulate the causal structure of its environment, in terms of its sensory infrastructures, it is interesting to reflect on the possibility that our visual cortices reflect the hierarchical causal structure of our environment.

In this section we introduce hierarchical models and extend the parameterisation of the ensuing generative model to cover the priors. This means that the constraints required by predictive coding and regularised solutions to inverse problems in the previous section, are now embraced by the learning scheme and are estimated in exactly the same way as the parameters of the generative distribution. These new parameters are refereed to as hyperparameters and are absorbed into he **M**-step to maximise the same objective function. This empirical approach to hyperparameter estimation rests upon a hierarchical structure for the generative model; indeed the term hyperparameter only has meaning in a hierarchical context. Hierarchical models are important because, as we will see, they encompass all the observation models mentioned above. Furthermore, hierarchical models may have a special standing in relation to hierarchical cortical organisation in the brain.

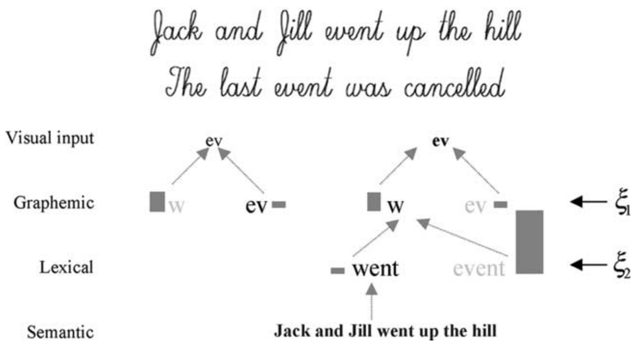
Jack and Jill went up the hill
The last event was cancelled
Visual input
Graphemic
Lexical
Semantic
ev
w
ev
ev
w
ev
went
event
Jack and Jill went up the hill
ξ_1
ξ_2

Fig. 7. Schematic illustrating the role of priors in biasing towards one representation of an input or another. Upper panel: On reading the first sentence 'Jack and Jill went up the hill' we perceive the word 'event' as 'went' despite the fact it is 'event' (as in the second sentence). However, in the absence of any hierarchical inference the best explanation for the pattern of visual stimulation incurred by the text is the grapheme 'ev'. This would correspond to the maximum likelihood estimate and would be the most appropriate in the absence of prior information, from the lexical and semantic context, about which is the most likely grapheme. However, within hierarchical inference the semantics (provided by the sentence) provide top-down predictions about the word, which in turn predicts the graphemes and, finally, the visual input. The posterior estimate is accountable to all these levels. When the semantic prior biases in favour of 'went' and 'w' we tolerate a small error as a lower level of visual analysis to minimise the overall prediction error. Lower panel: (left) The grapheme 'ev' is selected as the most likely cause of visual input; (right) The letter 'w' is selected, as it is (i) a reasonable explanation for the sensory input and (ii) conforms to prior expectations induced by lexico-semantic context. The bars represent prediction error, which is minimised over all levels to attain the most likely cause.

3.7.2. The nature of hierarchical models

Consider any level i in a hierarchy whose causes v_i are induced by corresponding causes in the level above v_{i+1}. The hierarchical form of the implicit generative model is

$$u = G_1(v_2, \theta_1) + \varepsilon_1 \qquad v_2 = G_2(v_3, \theta_2) + \varepsilon_2 \qquad v_3 = \cdots \tag{24}$$

with $u = v_1$ c.f. Eq. (18). Technically, these models fall into the class of conditionally independent hierarchical models when the stochastic terms are independent at each level (Kass & Steffey, 1989). These models are also called *parametric empirical Bayes* (PEB) models because the obvious interpretation of the higher-level densities as priors led to the development of PEB methodology (Efron &

Morris, 1973). Often, in statistics, these hierarchical models comprise just two levels, which is a useful way to specify simple shrinkage priors on the parameters of single-level models. We will assume the stochastic terms are Gaussian with covariance $\Sigma_i = \Sigma(\lambda_i)$. Therefore, θ_i and λ_i parameterise the means and covariances of the likelihood at each level

$$p(v_i|v_{i+1};\theta) = N(v_i : G_i(v_{i+1},\theta_i), \Sigma_i) \tag{25}$$

This likelihood of v_i also plays the role of a prior on v_i that is jointly maximised with the likelihood of the level below $p(v_{i-1}|v_i;\theta)$. This is the key to understanding the utility of hierarchical models. By learning the parameters of the generative distribution of level i one is implicitly learning the parameters of the prior distribution for level $i-1$. This enables this learning of prior densities.

Although λ_i are parameters of the forward model they are sometimes referred to as hyperparameters and in classical statistics correspond to variance components. We will preserve the distinction between θ_i and λ_i because they may correspond to backward and lateral connections strengths, respectively.

The hierarchical nature of these models lends an important context-sensitivity to recognition densities not found in single-level models. This is shown in Fig. 8, which should be compared with Fig. 2. The key point here is that high-level causes v_{i+1} determine the prior expectation of causes v_i in the subordinate level. This can completely change the marginal $p(v_{i-1};\theta)$ and recognition $p(v_i|v_{i-1};\theta)$ distributions upon which inference in based. From the manifold perspective on inference, the part of the manifold $G_{i-1}(v_i;\theta_{i-1})$ highlighted by prior expectations, changes from input to input in a context-dependent way (see Fig. 8). The context established by priors is not determined by preceding events but is immediate and conferred by higher hierarchical levels. For example, in Fig. 7 the semantic context induced by reading one of the sentences has a profound effect on the most likely graphemic cause of the visual input subtended by 'ev'. The dual role of $p(v_i|v_{i+1};\theta)$ as a likelihood or generative density for level i and a prior density for level $i-1$ is recapitulated by a dual role for MAP estimates of v_i. From a bottom-up perspective, these correspond to parameters (modes) of the recognition densities. However, from a top-down perspective they also act as parameters of the generative model by interacting with θ_{i-1} in $G_{i-1}(v_i,\theta_{i-1})$ to give the prior expectation of v_{i-1}.

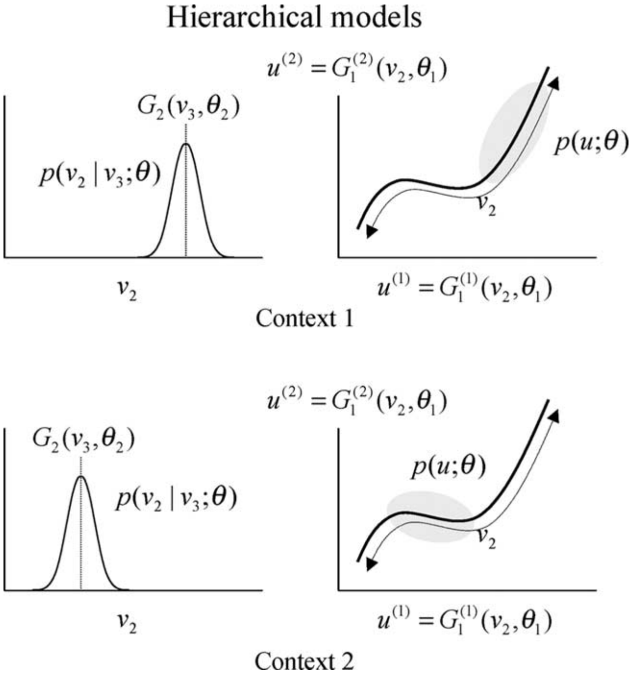

Fig. 8. Hierarchical models embody context-sensitivity not found in single-level models (c.f. Fig. 2). High-level causes v_{i+1} determine the prior expectation of causes v_i in the subordinate level. Changes in v_{i+1} can completely change the marginal $p(v_{i-1};\theta)$ and recognition $p(v_i|v_{i-1};\theta)$ distributions upon which inference in based.

It may seem restrictive to make Gaussian assumptions about the stochastic terms in Eq. (24). However, non-linearities in $G_i(v_{i+1},\theta_i)$ can transform Gaussian distributions into arbitrarily complicated non-Gaussian distributions. A simple example is given in Fig. 9 in which a bimodal marginal density, of a bivariate input, is induced by a univariate Gaussian density at the level above. In short, non-linear hierarchical models, under Gaussian assumptions are equivalent to non-hierarchical models under non-Gaussian assumptions. Perhaps the simplest example of this is the three-level model

$$u = \theta_1 v_2 + \varepsilon_1 \qquad v_2 = G(v_3,\theta_2) + \varepsilon_2 \qquad v_3 = \varepsilon_3 \tag{26}$$

This is formally identical to the (non-Gaussian) sparse coding model in Eq. (10) where $\mathrm{Cov}\{\varepsilon_1\} = \Sigma$, $\mathrm{Cov}\{\varepsilon_2\} = 0$ and $\mathrm{Cov}\{\varepsilon_3\} = 1$. $G(v_3,\theta_2)$ plays the role of a probability integral transform that renders the cumulative distribution of v_2 the same as g in Section 3.5.5. When $\mathrm{Cov}\{\varepsilon_1\} \to 0$ Eq. (24) reduces to the model adopted in ICA.

3.7.3. Implementation

The biological plausibility of the empirical Bayes in the brain can be established fairly simply. To do this, a hierarchical scheme is described in some detail. A more

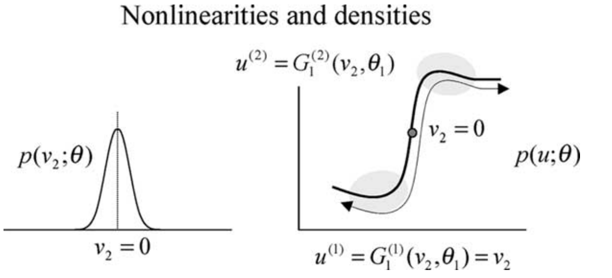
Nonlinearities and densities
$u^{(2)} = G_1^{(2)}(v_2,\theta_1)$
$p(v_2;\theta)$
$v_2 = 0$
$p(u;\theta)$
$v_2 = 0$
$u^{(1)} = G_1^{(1)}(v_2,\theta_1) = v_2$

Fig. 9. Non-linearities in $G_i(v_{i+1},\theta_i)$ can transform Gaussian distributions of v_{i+1} into arbitrarily complicated non-Gaussian densities for $v_i = G_i(v_{i+1},\theta_i) + \varepsilon_i$. In this illustration a bimodal marginal density, of a bivariate input, is induced by a univariate Gaussian density at the level above.

thorough account of this scheme, including simulations of various neurobiological and psychophysical phenomena, will appear in future publications. For the moment, we will address neuronal implementation at a purely theoretical level, using the framework developed above.

For simplicity, we will assume deterministic recognition such that $q(\phi(u); u) = 1$. In this setting, with conditional independence, F comprises a series of log likelihoods

$$\begin{aligned}\ell(u) &= \langle \ln p(u, v; \theta) \rangle_q = \ln p(u, \phi_2, \ldots; \theta)\\ &= \ln p(u|\phi_2; \theta) + \ln(\phi_2|\phi_3; \theta) + \cdots\\ &= -\frac{1}{2}\xi_1^T\xi_1 - \frac{1}{2}\xi_2^T\xi_2 - \cdots - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\ln|\Sigma_2|\\ &\quad - \cdots\end{aligned} \tag{27}$$

$$\begin{aligned}\xi_i &= \phi_i - G_i(\phi_{i+1}, \theta_i) - \lambda_i\xi_i\\ &= (1 + \lambda_i)^{-1}(\phi_i - G_i(\phi_{i+1}, \theta_i))\end{aligned}$$

c.f. Eq. (20). Here $\Sigma_i^{1/2} = 1 + \lambda_i$. In the setting of neuronal models the (whitened) prediction error is encoded by the activities of units denoted by ξ_i. These error units receive a prediction from units in the level above[2] and connections from the principal units ϕ_i being predicted. Horizontal interactions among the error units serve to de-correlate them (c.f. Foldiak, 1990), where the symmetric lateral connection strengths λ_i hyper-parameterise the covariances of the errors Σ_i, which are the prior covariances for level $i - 1$.

The estimators ϕ_i and the connection strength parameters perform a gradient ascent on the compound log probability.

$$\begin{aligned}&\mathbf{E}\quad \dot{\phi}_{i+1} = \frac{\partial l(u)}{\partial \phi_{i+1}} = -\frac{\partial \xi_i^T}{\partial \phi_{i+1}}\xi_i - \frac{\partial \xi_{i+1}^T}{\partial \phi_{i+1}}\xi_{i+1}\\ &\mathbf{M}\quad \dot{\theta}_i = \frac{\partial F}{\partial \theta_i} = -\left\langle \frac{\partial \xi_i^T}{\partial \theta_i}\xi \right\rangle_u\\ &\dot{\lambda}_i = \frac{\partial F}{\partial \lambda_i} = -\left\langle \frac{\partial \xi_i^T}{\partial \lambda_i}\xi \right\rangle_u - (1 + \lambda_i)^{-1}\end{aligned} \tag{28}$$

This is the simplest version of the most general learning algorithm considered so far. It is general in the sense that is does not require the parameters of either the generative or prior distributions. It can model non-invertible, non-linear generation of sensory inputs and encompasses complicated hierarchical processes. Furthermore, each of the learning components has a relatively simple neuronal interpretation (see below)

When G_i models dynamical processes (i.e. is effectively a convolution operator) this gradient ascent is more complicated. In a subsequent paper we will show that, with dynamical models, it is necessary to minimise the prediction error and their temporal derivatives. An alternative is to assume a simple hidden Markov model for the dynamics and use Kalman filtering (c.f. Rao & Ballard, 1998). For the moment, we will assume the inputs change sufficiently slowly for gradient ascent not to be confounded.

3.8. Theoretical implications for neuronal implementation

The scheme implied by Eq. (28) has four clear implications or predictions about the functional architectures required for its implementation. We now review these in relation to cortical organisation in the brain. A schematic summarising these points is provided in Fig. 10. In short, we arrive at exactly the same four points presented at the end of the previous section.

Hierarchical organisation. Hierarchical models enable empirical Bayesian learning of prior densities and provide a plausible model for sensory inputs. Single-level models that do not show any conditional independence (e.g. those used by connectionist and infomax schemes) depend on prior constraints for unique inference and do not call upon a hierarchical cortical organisation. On the other hand, if the causal structure of generative processes is hierarchical, this will be reflected, literally, by the hierarchical architectures trying to minimise prediction error, not just at the level of sensory input but at all levels (notice the deliberate mirror symmetry in Fig. 10). The nice thing about this architecture is that the responses of units at the ith level ϕ_i depend only on the error at the current level and the immediately preceding level. This follows from conditional independence and is important because it permits a biologically plausible implementation, where the connections driving the error minimisation only run forward from one level to the next.

Reciprocal connections. As established at the beginning of this section, the non-invertibility of processes generating sensory data induces a need for both forward and backward connections. In the hierarchical model, the dynamics of principal units ϕ_{i+1} are subject to two, locally available, influences. A likelihood or recognition term mediated by forward afferents from the error units in the level below and an empirical prior conveyed by error units in the same level. Critically, the influences of the error units in both levels are mediated by linear connections with a strength that is exactly the same as the [negative] effective connectivity of the *reciprocal* connections from ϕ_{i+1} to ξ_i and ξ_{i+1}. Functionally, forward and lateral connections are reciprocated, where backward connections generate predictions of lower-level responses. Effective connectivity is simply the change in a neuronal unit (neuron, assembly or cortical area) induced by inputs from another (Friston, 1995). In this case $\partial\xi_i/\partial\phi_{i+1}$ and $\partial\xi_{i+1}/\partial\phi_{i+1}$.

Effective connectivity in the forward direction is the reciprocal (negative transpose) of that in the backward direction $\partial\xi_i/\partial\phi_{i+1} = -\partial G_i/(\phi_{i+1}, \beta_i)_i/\partial v_{i+1}$ that is a function of the generative parameters. Lateral connections,

[2] Clearly, in the brain, backward connections are not inhibitory but, after mediation by inhibitory inter-neurons, their effective influence could be rendered so.

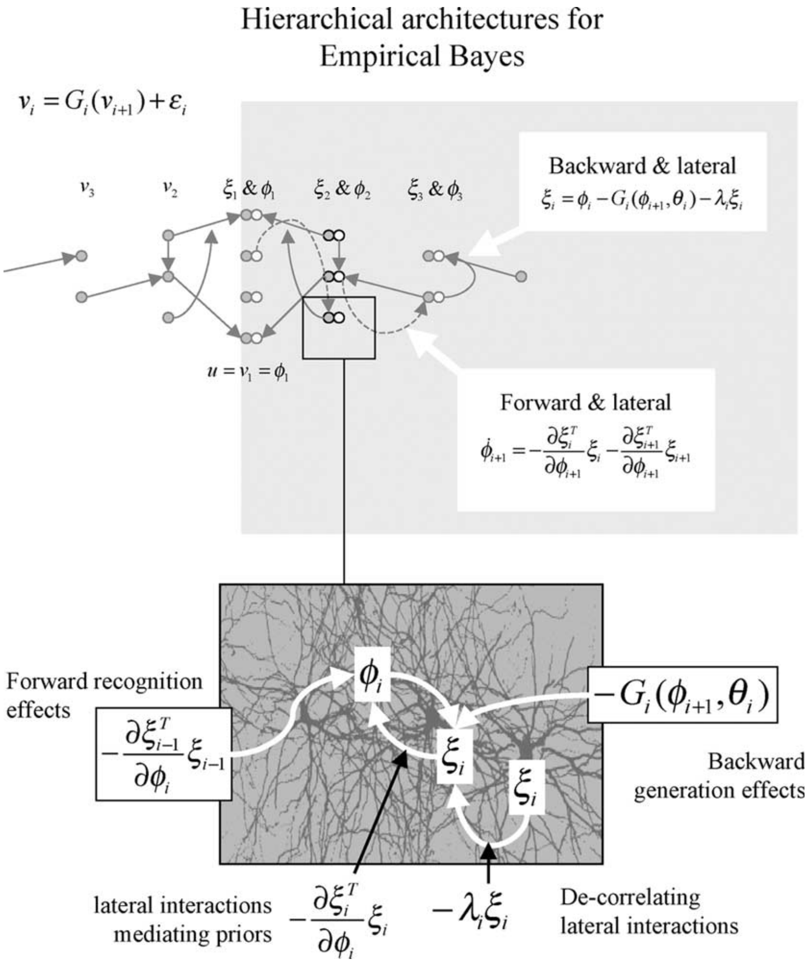
Hierarchical architectures for Empirical Bayes
$v_i = G_i(v_{i+1}) + \varepsilon_i$
v_3
v_2
ξ_1 & ϕ_1
ξ_2 & ϕ_2
ξ_3 & ϕ_3
Backward & lateral
$\xi_i = \phi_i - G_i(\phi_{i+1}, \theta_i) - \lambda_i \xi_i$
$u = v_1 = \phi_1$
Forward & lateral
$\dot{\phi}_{i+1} = -\frac{\partial \xi_i^T}{\partial \phi_{i+1}} \xi_i - \frac{\partial \xi_{i+1}^T}{\partial \phi_{i+1}} \xi_{i+1}$
Forward recognition effects
$-\frac{\partial \xi_{i-1}^T}{\partial \phi_i} \xi_{i-1}$
ϕ_i
ξ_i
ξ_i
$-G_i(\phi_{i+1}, \theta_i)$
Backward generation effects
lateral interactions mediating priors
$-\frac{\partial \xi_i^T}{\partial \phi_i} \xi_i$
$-\lambda_i \xi_i$
De-correlating lateral interactions

Fig. 10. Upper panel: Schematic depicting a hierarchical extension to the predictive coding architecture, using the same format as Fig. 4. Here hierarchical arrangements within the model serve to provide predictions or priors to representations in the level below. The open circles are the error units and the filled circles are the states encoding the conditional expectation of causes in the environment. These change to minimise both the discrepancies between their predicted value and the mismatch incurred by their own prediction of the level below. These two constraints correspond to prior and likelihood terms, respectively (see main text). Lower panel: a more detailed picture of the influences on principal and error units.

within each level, mediate the influence of error units on the principal units and intrinsic connections λ_i among the error units decorrelate them, allowing competition among prior expectations with different precisions (precision is the inverse of variance). In short, lateral, forwards and backward connections are all reciprocal, consistent with anatomical observations.

Functionally asymmetric forward and backward connections. The forward connections are the reciprocal of the backward effective connectivity from the higher level to the lower level, extant at that time. However, the functional attributes of forward and backward influences are different. The influences of units ϕ_{i+1} on error units in the lower level ξ_i instantiate the forward model $\xi_i = \phi_i - G_i(\phi_{i+1}, \theta_i) - \lambda_i \xi_i$. These can be non-linear, where each unit in the higher level may modulate or interact with the influence of others, according to the non-linearities in $G_i(\phi_{i+1}, \theta_i)$. In contradistinction, the influences of units in lower levels do not interact when producing changes at the higher level, because their effects are linearly separable (see Eq. (28)). This is a key observation because the empirical evidence, reviewed in the previous section, suggests that backward connections are in a position to interact (e.g. though NMDA receptors expressed predominantly in the supragranular layers receiving backward connections). Forward connections are not. It should be noted that, although the implied forward connections $-\partial \xi_i / \partial \phi_{i+1}^T$ mediate linearly separable effects of ξ_i on ϕ_{i+1}, these connections might be activity- and time-dependent because of their dependence on ϕ_{i+1}. In summary, non-linearities, in the way sensory inputs are produced, necessitate non-linear interactions in the generative model that are mediated by backward influences but do not require forward connections to be modulatory.

Associative plasticity. Changes in the parameters correspond to plasticity in the sense that the parameters control the strength of backward and lateral connections. The backward connections parameterise the prior expectations of the forward model and the lateral connections hyperparameterise the prior covariances. Together they parameterise the Gaussian densities that constitute the priors (and likelihoods) of the model. The plasticity implied can be seen more clearly with an explicit parameterisation of the connections. For example, let $G_i(v_{i+1}, \theta_i) = \theta_i v_{i+1}$. In this instance

$$\begin{aligned}\dot{\theta}_i &= (1 + \lambda_i)^{-1}\langle \xi_i \phi_{i+1}^T \rangle_u \\ \dot{\lambda}_i &= (1 + \lambda_i)^{-1}(\langle \xi_i \xi_i^T \rangle_u - 1)\end{aligned} \tag{29}$$

This is just Hebbian or associative plasticity where the connection strengths change in proportion to the product of pre- and post-synaptic activity; for example, $\langle \xi_i \phi_{i+1}^T \rangle$. An intuition about Eq. (29) obtains by considering the conditions under which the expected change in parameters is zero (i.e. after learning). For the backward connections this implies there is no component of prediction error that can be explained by estimates at the higher level $\langle \xi_i \phi_{i+1}^T \rangle = 0$. The lateral connections stop changing when the prediction error has been whitened $\langle \xi_i \xi_i^T \rangle = 1$.

Non-diagonal forms for λ_i complicate the biological interpretation because changes at any one connection depend on changes elsewhere. The problem can be finessed slightly by rewriting the equations as

$$\dot{\theta}_i = \langle \xi_i \phi_{i+1}^T \rangle_u - \lambda_i \dot{\theta}_i \qquad \dot{\lambda}_i = \langle \xi_i \xi_i^T \rangle_u - \lambda_i \dot{\lambda}_i - 1 \tag{30}$$

where the decay terms are mediated by integration at the cell body in a fashion similar to that described by Friston, Frith, and Frackowiak (1993). Furthermore the expectations can be approximated by a trace of the associative term. For example, $\tau \dot{T} = \xi_i \phi_{i+1}^T - T$, where the trace T may correspond to the accumulation of post-synaptic tags (e.g. Frey & Morris, 1997) mentioned above. Finally, one should note that changes in lateral and self-connections encoding precision do not have to be mediated by long-term changes in plasticity. They could change on a short time-scale, through classical neuromodulatory effects. However, this would entail a slightly different parameterisation that pooled over error units.

It is evident that the predictions of the theoretical analysis coincide almost exactly with the empirical aspects of functional architectures in visual cortices highlighted by the previous section (hierarchical organisation, reciprocity functional asymmetry and associative plasticity). Although somewhat contrived, it is pleasing that purely theoretical considerations and neurobiological empiricism converge so precisely.

3.9. Summary

In summary, predictive coding lends itself naturally to a hierarchical treatment, which considers the brain as an empirical Bayesian device. The dynamics of the units or populations are driven to minimise error at all levels of the cortical hierarchy and implicitly render themselves posterior modes (i.e. most likely values) of the causes given the data. In contradistinction to supervised learning, hierarchical prediction does not require any desired output. Indeed predictions of intermediate outputs at each level emerge spontaneously. Unlike information theoretic approaches they do not assume independent causes. In contrast to regularised inverse solutions (e.g. in machine vision) they do not depend on a priori constraints. These emerge spontaneously as empirical priors from higher levels. Fig. 11 is a schematic that reviews the issues covered in this section from the point of view of model estimation.

The overall scheme implied by Eq. (28) relates comfortably to the hypothesis (Mumford, 1992), “on the role of the reciprocal, topographic pathways between two cortical areas, one often a ‘higher’ area dealing with more abstract information about the world, the other ‘lower’, dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops”. We have seen that supervised, infomax and regularised models require prior assumptions about the distribution of causes. The final section introduced empirical Bayes to show that these assumptions are not necessary and that priors can be learned in a hierarchical context. Furthermore, we have tried to show that hierarchical prediction can be implemented in brain-like architectures using mechanisms that are biologically plausible.

Clearly, there are many aspects of neuronal information processing that have not been included in the theoretical considerations above. For example, we have not specified the detailed role of local neuronal circuits or the potentially important role of horizontal connections within areas, or lateral connections among areas. Furthermore, a number of the dynamical aspects of neuronal interactions have been ignored (e.g. timing of responses, conduction delays synchrony, etc.). A number of these issues, particularly the dynamic aspects are the subject of current work, in many units, that represents an exciting confluence of estimation theory, statistics, computational neuroscience and neurobiology. See, for example, Pouget, Deneve, and Duhamel (2002) for a theoretical treatment of multimodal representations within a Bayesian framework.

4. Discussion

4.1. Representational leaning

The formulation of representational learning in terms of generative models embodies a number of key distinctions,

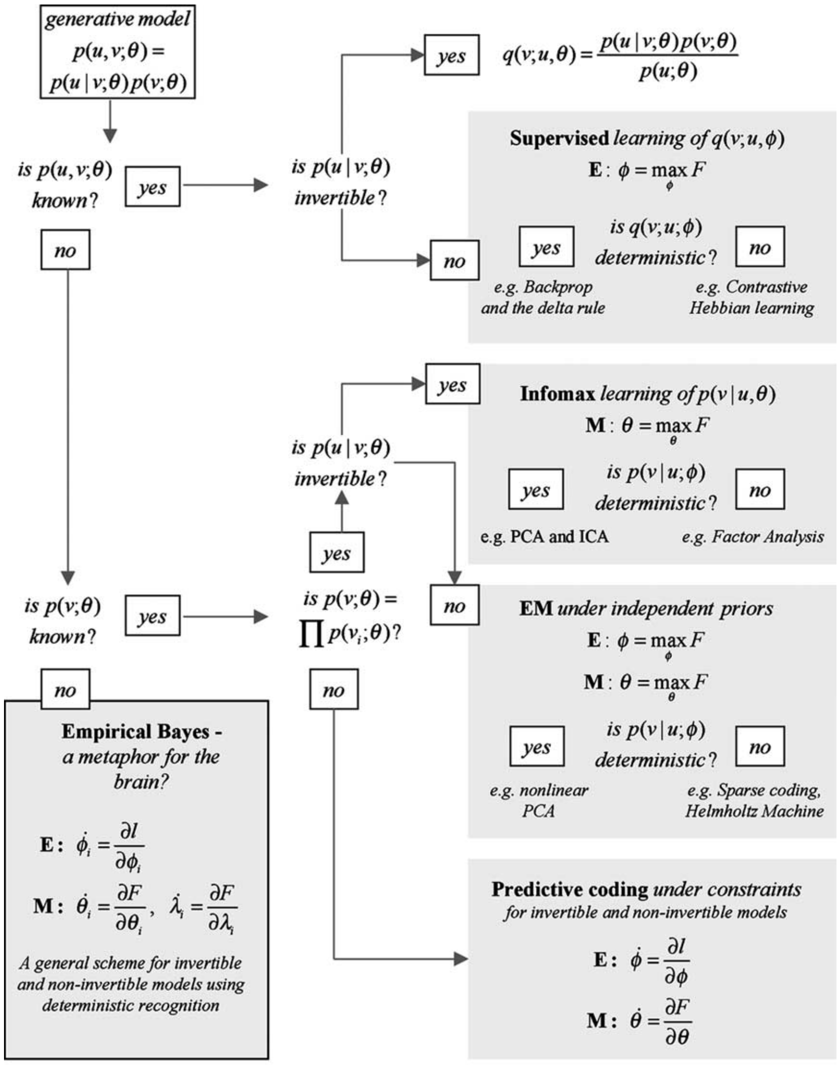
generative model
$p(u,v;\theta) = p(u|v;\theta)p(v;\theta)$
is $p(u,v;\theta)$ known?
yes
no
is $p(u|v;\theta)$ invertible?
yes
$q(v;u,\theta) = \frac{p(u|v;\theta)p(v;\theta)}{p(u;\theta)}$
no
Supervised learning of $q(v;u,\phi)$
$\mathbf{E}: \phi = \max_{\phi} F$
yes
is $q(v;u;\phi)$ deterministic?
no
e.g. Backprop and the delta rule
e.g. Contrastive Hebbian learning
yes
is $p(u|v;\theta)$ invertible?
Infomax learning of $p(v|u,\theta)$
$\mathbf{M}: \theta = \max_{\theta} F$
yes
is $p(v|u;\phi)$ deterministic?
no
e.g. PCA and ICA
e.g. Factor Analysis
yes
is $p(v;\theta)$ known?
yes
is $p(v;\theta) = \prod p(v_i;\theta)$?
no
no
no
EM under independent priors
$\mathbf{E}: \phi = \max_{\phi} F$
$\mathbf{M}: \theta = \max_{\theta} F$
yes
is $p(v|u;\phi)$ deterministic?
no
e.g. nonlinear PCA
e.g. Sparse coding, Helmholtz Machine
Empirical Bayes - a metaphor for the brain?
$\mathbf{E}: \dot{\phi}_i = \frac{\partial l}{\partial \phi_i}$
$\mathbf{M}: \dot{\theta}_i = \frac{\partial F}{\partial \theta_i}, \quad \dot{\lambda}_i = \frac{\partial F}{\partial \lambda_i}$
A general scheme for invertible and non-invertible models using deterministic recognition
Predictive coding under constraints
for invertible and non-invertible models
$\mathbf{E}: \dot{\phi} = \frac{\partial l}{\partial \phi}$
$\mathbf{M}: \dot{\theta} = \frac{\partial F}{\partial \theta}$

Fig. 11. A taxonomy of procedures for representational learning. The various schemes are organised to reflect their dependence on different assumptions about generative and prior distributions that are necessary to learn a recognition distribution. See main text for a full discussion of the schemes and mathematical notion.

(i) the distinction between invertible vs. non-invertible models, (ii) deterministic vs. probabilistic recognition and (iii) static vs. dynamic recognition. If the inverse of the generative model cannot be parameterised an approximate recognition is required. This invokes the need for an explicit parameterisation of both the recognition and generative densities and suggests an important role for forward and backward connections in the brain. Invertible models can, in principle, be implemented using only forward connections because the recognition parameters are completely specified by the generative model. However, non-linear, hierarchical and dynamic aspects of the sensorium render easy inversion highly unlikely. Most of the examples in this section have focussed on deterministic recognition where neuronal dynamics encode the most likely causes of the current sensory input. The distinction between deterministic and probabilistic representation addresses a deeper question about whether neuronal dynamics represent the state of the world or embody probability densities of those states. Classically, patterns of activity have been treated as encoding the value of the stimulus (e.g. the orientation of a contour). Pouget, Dayan, and Zemel (2003) explore the more recent suggestion “that neural computation is akin to a Bayesian inference process, with population activity patterns representing uncertainty about stimuli in the form of probability distributions (e.g., the probability density function over the orientation of a contour).”

Does the brain estimate or infer? From the point of view of hierarchical models, the states of units encode the mode of the posterior density at any given level. This can be considered a point recognition density or an estimate. However, the states of units at any level also induce a prior density in the level below. This is because top-down influences provide a prior expectation in the context of a prior covariance that is encoded by the strength of lateral connections. These covariances render the generative model probabilistic. By summarising densities in terms of their modes, using neuronal activity, the posterior and prior densities can change quickly with sensory inputs. However, this does entail unimodal densities. From the point of view of a statistician, this may be an impoverished representation of the world that compromises inference, especially when the posterior distribution is multimodal. However, it is exactly this approximate nature of recognition that preoccupies psychophysicists and psychologists; the emergence of unitary, deterministic perceptual representations in the brain is commonplace and is of special interest when the causes are ambiguous (e.g. illusions and perceptual transitions induced by binocular rivalry and ambiguous figures).

The arguments in the preceding section clearly favour predictive coding, over supervised or information theoretic frameworks, as a more plausible account of functional brain architectures. However, it should be noted that the differences among them have been deliberately emphasised. For example, predictive coding and the implicit error minimisation results in the maximisation of information transfer. In other words, predictive coding conforms to the principle of maximum information transfer, but in a distinct way. The infomax principle is a principle, whereas predictive coding represents a particular scheme that serves that principle. There are examples of infomax that do not employ predictive coding (e.g. transformations of stimulus energy in early visual processing; Atick & Redlich, 1990) that may be specified genetically or epigenetically. However, predictive coding is likely to play a much more prominent role at higher levels of processing for the reasons detailed in the previous section.

In a similar way predictive coding, especially in its hierarchical formulation, embraces PDP principles found in connectionist schemes. The representation of any cause depends upon internally consistent representations of subordinate and supraordinate causes in lower and higher levels. These representations induce and maintain themselves, across and within all levels of the sensory hierarchy, through dynamic and reentrant interactions (Edelman, 1993). The same PDP phenomena (e.g. lateral interactions leading to competition among representations) can be observed. For example, the lateral connection strengths embody what has been learnt empirically about the prior covariances among causes. A prior that transpires to be very precise (i.e. low variance) will receive correspondingly low strength inhibitory connections from its competing error units (recall $\Sigma_i^{1/2} = 1 + \lambda_i$). It will therefore supervene over other error units and have a greater corrective impact on the estimate causing the prediction error. Conversely, top-down expectations that are less informative will produce errors that are more easily suppressed and have less effect on the representations. In predictive coding, these dynamics are driven explicitly by error minimisation, whereas in connectionist simulations the activity is determined solely by the connection strengths established during training.

In addition to the theoretical bias toward generative models and predictive coding, the clear emphases on backward and recursive dynamics make it a more natural framework for understanding neuronal infrastructures. Figs. 5 and 6 show the fundamental differences among supervised, infomax and predictive schemes. In the supervised and infomax schemes the connections are universally forward. In the predictive coding scheme the forward connections (broken line) drive the prediction so as to minimise $l(u)$ whereas backwards connections (solid lines) use these representations of causes to emulate mixing enacted by the real world. The non-linear aspects of this mixing imply that only backward influences interact in the predictive coding scheme. Section 2 assembled some of the anatomical and physiological evidence suggesting that backward connections are prevalent in the real brain and could support non-linear mixing through their modulatory characteristics. Before turning to electrophysiological evidence for backward connections we consider the implications for classical views of receptive fields and the representational capacity of neuronal units.

4.2. Context, causes and representations

The Bayesian perspective suggests something quite profound for the classical view of receptive fields. If neuronal responses encompass a bottom-up likelihood term and top-down priors, then responses evoked by bottom-up input should change with the context established by prior expectations from higher levels of processing. Consider the example in Fig. 7. Here a unit encoding the visual form of ‘went’ responds when we read the first sentence at the top of this figure. When we read the second sentence ‘The last event was cancelled’ it would not. If we recorded from this unit we might infer that our ‘went’ unit was, in some circumstances, selective for the word ‘event’. This might be difficult to explain without an understanding of hierarchical inference and the semantic context the stimulus was presented in. In short, under a predictive coding scheme, the receptive fields of neurons should be context-sensitive. The remainder of this subsection deals with empirical evidence for these extra-classical receptive field effects.

Generative models suggest that the role of backward connections is to provide contextual guidance to lower levels through a prediction of the lower level’s inputs. When this prediction is incomplete or incompatible with the lower area’s input, an error is generated that engenders changes in the area above until reconciliation. When, and only when, the bottom-up driving inputs are in harmony with top-down prediction, error is suppressed and a consensus between the prediction and the actual input is established. Given this conceptual model, a stimulus-related response or ‘activation’ corresponds to some transient error signal that drives the appropriate change in higher areas until a veridical higher-level representation emerges and the error is ‘cancelled’ by backwards connections. Clearly the prediction error will depend on the context and consequently the backward connections confer context-sensitivity on the functional specificity of the lower area. In short, the activation does not just depend on bottom-up input but on the difference between bottom-up input and top-down predictions.

The prevalence of non-linear or modulatory top-down effects can be inferred from the fact that context interacts with the content of representations. Here context is established simply through the expression of causes other than the one in question. Backward connections from one higher area can be considered as providing contextual modulation of the prediction from another area. Because the effect of context will only be expressed when the thing being predicted is present these contextual afferents should not elicit a response by themselves. Effects of this sort, which change the responsiveness of units but do not elicit a response, are a hallmark of modulatory projections. In summary, hierarchical models offer a scheme that allows for contextual effects; firstly through biasing responses towards their prior expectation and secondly by conferring a context-sensitivity on these priors through the modulatory component of backward projections. Next we consider the nature of real neuronal responses and whether they are consistent with this perspective.

4.3. Neuronal responses and representations

Classical models (e.g. classical receptive fields) assume that evoked responses will be expressed invariably in the same units or neuronal populations irrespective of the context. However, real neuronal responses are not invariant but depend upon the context in which they are evoked. For example, visual cortical units have dynamic receptive fields that can change from moment to moment (c.f. the non-classical receptive field effects modelled in (Rao and Ballard, 1998)). A useful synthesis of data for the macaque visual system that highlights the anatomical and physiological substrates of context-dependent responses can be found in Angelucci, Levitt, and Lund (2002a). A key conclusion of the authors is that "feedback from extrastriate cortex (possibly together with overlap or inter-digitation of coactive lateral connectional fields within V1) can provide a large and stimulus-specific surround modulatory field. The stimulus specificity of the interactions between the centre and surround fields, may be due to the orderly, matching structure and different scales of intra-areal and feedback projection excitatory pathways."

There are numerous examples of context-sensitive neuronal responses. Perhaps the simplest is short-term plasticity. Short-term plasticity refers to changes in connection strength, either potentiation or depression, following pre-synaptic inputs (e.g. Abbot, Varela, Sen, & Nelson, 1997). As noted by Fuhrmann, Segev, Markram, and Tsodyks (2002) "Synaptic transmission in the neocortex is dynamic, such that the magnitude of the post-synaptic response changes with the history of the pre-synaptic activity. Therefore each response carries information about the temporal structure of the preceding pre-synaptic input spike train." In brief, the underlying connection strengths, that define what a unit represents, are a strong function of the immediately preceding neuronal transient (i.e. preceding representation). A second, and possibly richer, example is that of attentional modulation that can change the sensitivity of neurons to different perceptual attributes (e.g. Treue & Maunsell, 1996). It has been shown, both in single unit recordings in primates (Treue & Maunsell, 1996) and human functional fMRI studies (Büchel & Friston, 1997), that attention to specific visual attributes can profoundly alter the receptive fields or event-related responses to the same stimuli.

These sorts of effects are commonplace in the brain and are generally understood in terms of the dynamic modulation of receptive field properties by backward and lateral afferents. There is clear evidence that horizontal connections in visual cortex are modulatory in nature (Hirsch & Gilbert, 1991), speaking to an interaction between the functional segregation implicit in the columnar architecture of V1 and the neuronal dynamics in distal populations. These observations suggest that lateral and backwards interactions may convey contextual information that shapes the responses of any neuron to its inputs (e.g. Kay & Phillips, 1996; Phillips & Singer, 1997) to confer on the brain the ability to make conditional inferences about sensory input. See also McIntosh (2000) who develops the idea from a cognitive neuroscience perspective "that a particular region in isolation may not act as a reliable index for a particular cognitive function. Instead, the *neural context* in which an area is active may define the cognitive function." His argument is predicated on careful characterisations of effective connectivity using neuroimaging.

4.3.1. Examples from neurophysiology

Here we consider the evidence for contextual representations in terms of single cell responses, to visual stimuli, in the temporal cortex of awake behaving monkeys. If the representation of a stimulus depends on establishing representations of subordinate and supraordinate causes at all levels of the visual hierarchy, then information about the high-order attributes of a stimulus must be conferred by top-down influences. Consequently, one might expect to see the emergence of selectivity, for high-level attributes, *after* the initial visually evoked response (although delays vary greatly, it typically takes about 10 ms for spike volleys to propagate from one cortical area to another and about a 100 ms to reach prefrontal areas). This is because the representations at higher levels must emerge before backward afferents can reshape the response profile of neurons in lower areas. This temporal delay, in the emergence of selectivity, is precisely what one sees empirically: Sugase, Yamane, Ueno, and Kawano (1999) recorded neurons in macaque temporal cortex during the presentation of faces and objects. The faces were either human or monkey faces and were categorised in terms of identity (whose face it was) and expression (happy, angry, etc.). "Single neurones conveyed two different scales of facial information in their firing patterns, starting at different latencies. Global information, categorising stimuli as monkey faces, human faces or shapes, was conveyed in the earliest part of the responses. Fine information about identity or expression was conveyed later", starting on average about 50 ms after face-selective responses. These observations demonstrate representations for facial identity or expression that emerge dynamically in a way that might rely on backward connections. These influences imbue neurons with a selectivity that is not intrinsic to the area but depends on interactions across levels of a processing hierarchy.

A similar late emergence of selectivity is seen in motion processing. A critical aspect of visual processing is the integration of local motion signals generated by moving objects. This process is complicated by the fact that local velocity measurements can differ depending on contour orientation and spatial position. Specifically, any local motion detector can measure only the component of motion perpendicular to a contour that extends beyond its field of view (Pack & Born, 2001). This 'aperture problem' is particularly relevant to direction-selective neurons early in the visual pathways, where small receptive fields permit only a limited view of a moving object. Pack and Born (2001) have shown "that neurons in the middle temporal visual area (known as MT or V5) of the macaque brain reveal a dynamic solution to the aperture problem. MT neurons initially respond primarily to the component of motion perpendicular to a contour's orientation, but over a period of approximately 60 ms the responses gradually shift to encode the true stimulus direction, regardless of orientation".

Friston (2002a,b) presented a number of examples from functional neuroimaging that demonstrated the context-sensitivity of evoked brain responses and the use of effective connectivity to establish interactions between bottom-up and top-down influences. Recent neuroimaging studies have addressed predictive coding explicitly, with some compelling results: Murray, Kersten, Olshausen, Schrater, and Woods (2002) used functional MRI to measure responses in V1 and a higher object processing area, the lateral occipital complex, to visual elements that were either grouped into objects or arranged randomly. They "observed significant activity increases in the lateral occipital complex and concurrent reductions of activity in primary visual cortex when elements formed coherent shapes, suggesting that activity in early visual areas is reduced as a result of grouping processes performed in higher areas. These findings are consistent with predictive coding models of vision that postulate that inferences of high-level areas are subtracted from incoming sensory information in lower areas through cortical feedback."

Recent developments in functional mapping, at the cellular level, may disclose more details about the specific contribution of backward and lateral connections. These advances involve the use of extra-cellular electrode recordings, optical imaging and three-dimensional anatomical reconstruction cells in conjunction with the GABA inactivation paradigm and related interventions (see Kisvarday, Crook, Buzas, & Eysel, 2000). The predictions of the empirical Bayesian model reviewed above are clear: Disabling backward and lateral afferents should destroy context sensitivity. Principal units should still be able to 'recognise' stimulus configurations but will simply attain the 'maximum likelihood' estimate of their cause, unconstrained by priors or contextual information. Error units will respond exuberantly, because prediction error cannot be cancelled by constructs from higher levels of synthesis. Learning-related deficits may be expressed as a failure of repetition suppression leading to an overall picture of disinhibition and context-insensitive responses.

4.4. Perception and action

This paper has deliberately restricted its focus to perceptual synthesis in (visual) cortical hierarchies. However, it is useful to note the close links between predictive coding in perception and motor control. These links exist at a number of levels. First, the conjoint use of forward (generative or *predictive*) and inverse (recognition or *controller*) models has been central to theories of motor control and action for many years (see Wolpert & Kawato, 1998). Indeed the use of prediction error to drive changes is probably more established in this context. Here, the prediction error is used to adjust motor executive processes to minimise the discrepancy between the consequences of action and that predicted (by a forward model) given top-down signals. Hitherto, we have portrayed error units as suppressing themselves, in a reentrant fashion, though supraordinate principal units. In motor systems error signals self-suppress, not through neuronally mediated effects, but by eliciting movements that change bottom-up proprioceptive and sensory input. This unifying perspective on perception and action suggest action is both perceived and caused by its perception. The behaviour of co-evolving forward and inverse models, embedded in the real world, is a fascinating area that links perception and action and even encompasses communication. For example, Wolpert, Doya, and Kawato (2003) have examined the extent to which motor commands acting on the body can be equated with communicative signals acting on other people and suggest that "computational solutions for motor control may have been extended to the domain of social interaction."

The behaviour of embodied agents, capable of empirical Bayesian inference, is another area we have not considered. It is interesting to reflect on what might happen if principal units were connected to motor effectors. The activity of principal units is directed by their ability to suppress prediction error. In this paper we have only dealt with neuronally mediated suppression through backward connections. However, it is perfectly possible for this suppression to be mediated though the physical world by changing the sensorium or the way it is sampled. This leads to the interesting conjecture that much innate orientating and tracking behaviour is simply a reflection of the brain's inherent tendency to maintain a predictable sensory input.

5. Conclusion

In conclusion, the representational capacity and inherent function of any neuron, neuronal population or cortical area in the brain is dynamic and context-sensitive. Functional integration, or interactions among brain systems, that employ driving (bottom-up) and backward (top-down) connections, mediate this adaptive and contextual specialisation. We have seen that most models of representational learning require prior assumptions about the distribution of causes. However, empirical Bayes suggests that these assumptions can be relaxed and that priors can be learned in a hierarchical context. We have tried to show that this hierarchical prediction can be implemented in brain-like architectures and in a biologically plausible fashion.

A key point, made above, is that backward connections, mediating internal or generative models of how sensory inputs are caused, are essential if the processes generating inputs are difficult to invert. This non-invertibility demands an explicit parameterisation of both the generative model (backward connections) and approximate recognition (forward connections). This suggests that feedforward architectures are not sufficient for representational learning or perception. Moreover, non-linearities in generative models, that make backward connections necessary, require these connections to be modulatory, so that estimated causes in higher cortical levels can interact to predict responses in lower levels. This is important in relation to asymmetries in forward and backward connections that have been characterised empirically.

The arguments in this article were developed under hierarchical models of brain function, where high-level systems provide a prediction of the inputs to lower-levels. Conflict between the two is resolved by changes in the high-level representations, which are driven by the ensuing error in lower regions, until the mismatch is 'cancelled'. From this perspective the specialisation of any region is determined both by bottom-up driving inputs and by top-down predictions. Specialisation is therefore not an intrinsic property of any region but depends on both forward and backward connections with other areas. Because the latter have access to the context in which the inputs are generated they are in a position to modulate the selectivity or specialisation of lower areas. The implications for classical models (e.g. classical receptive fields in electrophysiology, classical specialisation in neuroimaging and connectionism in cognitive models) are severe and suggest these models may provide incomplete accounts of real brain architectures. On the other hand, representational learning, in the context of hierarchical generative models not only accounts for extra-classical phenomena seen empirically but enforces a view of the brain as an inferential machine through its empirical Bayesian motivation.

Acknowledgements

The Wellcome Trust funded this work. I would like to thank my colleagues for help in writing this paper and developing the ideas, especially Cathy Price for the psychological components, Peter Dayan for the computational neuroscience and three anonymous reviewers for extremely helpful suggestions.

References

Abbot, L. F., Varela, J. A., Sen, K., & Nelson, S. B. (1997). Synaptic depression and cortical gain control. *Science*, *275*, 220–223.

Absher, J. R., & Benson, D. F. (1993). Disconnection syndromes: An overview of Geschwind's contributions. *Neurology*, *43*, 862–867.

Angelucci, A., Levitt, J. B., & Lund, J. S. (2002a). Anatomical origins of the classical receptive field and modulatory surround field of single neurons in macaque visual cortical area V1. *Progress in Brain Research*, *136*, 373–388.

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J. M., Bullier, J., & Lund, J. S. (2002b). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, *22*, 8633–8646.

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

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature*, *306*, 21–26.

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

Batardiere, A., Barone, P., Knoblauch, K., Giroud, P., Berland, M., Dumas, A. M., & Kennedy, H. (2002). Early specification of the hierarchical organization of visual cortical areas in the macaque monkey. *Cerebral Cortex*, *12*, 453–465.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind de-convolution. *Neural Computation*, *7*, 1129–1159.

Büchel, C., & Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, *7*, 768–778.

Buonomano, D. V., & Merzenich, M. M. (1998). Cortical plasticity: From synapses to maps. *Annual Review of Neuroscience*, *21*, 149–186.

Common, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314.

Crick, F., & Koch, C. (1998). Constraints on cortical and thalamic projections: The no-strong-loops hypothesis. *Nature*, *391*, 245–250.

Dayan, P., & Abbot, L. F. (2001). *Theoretical neuroscience. Computational and mathematical modelling of neural systems*. Cambridge, MA: MIT Press.

Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholtz machine. *Neural Computation*, *7*, 889–904.

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

Dong, D., & McAvoy, T. J. (1996). Nonlinear principal component analysis—based on principal curves and neural networks. *Computers and Chemical Engineering*, *20*, 65–78.

Edelman, G. M. (1993). Neural Darwinism: Selection and reentrant signalling in higher brain function. *Neuron*, *10*, 115–125.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *Journal of the American Statistical Association*, *68*, 117–130.

Farah, M., & McClelland, J. (1991). A computational model of semantic memory impairments: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*, 339–357.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Foldiak, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, *64*, 165–170.

Frey, U., & Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, *385*, 533–536.

Friston, K. J. (1995). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, *2*, 56–78.

Friston, K. J. (2000). The labile brain. III. Transients and spatio-temporal receptive fields. *Philosophical Transactions of the Royal Society of London B*, *355*, 253–265.

Friston, K. J. (2002a). Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience*, *25*, 221–250.

Friston, K. J. (2002b). Functional integration and inference in the brain. *Progress in Neurobiology*, *68*, 113–143.

Friston, K. J. (2002c). Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage*, *16*, 513–530.

Friston, K. J., Frith, C. D., & Frackowiak, R. S. J. (1993). Principal component analysis learning algorithms: A neurobiological analysis. *Proceedings of the Royal Society B*, *254*, 47–54.

Friston, K. J., Frith, C., Passingham, R. E., Dolan, R., Liddle, P., & Frackowiak, R. S. J. (1992). Entropy and cortical activity: Information theory and PET findings. *Cerebral Cortex*, *3*, 259–267.

Friston, K. J., Phillips, J., Chawla, D., & Büchel, C. (2000). Nonlinear PCA: Characterising interactions between modes of brain activity. *Philosophical Transactions of the Royal Society of London B*, *355*, 135–146.

Fuhrmann, G., Segev, I., Markram, H., & Tsodyks, M. (2002). Coding of temporal information by activity-dependent synapses. *Journal of Neurophysiology*, *87*, 140–148.

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

Girard, P., & Bullier, J. (1989). Visual activity in area V2 during reversible inactivation of area 17 in the macaque monkey. *Journal of Neurophysiology*, *62*, 1287–1301.

Harth, E., Unnikrishnan, K. P., & Pandya, A. S. (1987). The inversion of sensory processing by feedback pathways: A model of visual cognitive functions. *Science*, *237*, 184–187.

Hilgetag, C. C., O'Neill, M. A., & Young, M. P. (2000). Hierarchical organisation of macaque and cat cortical sensory systems explored with a novel network processor. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, *355*, 71–89.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The wake–sleep algorithm for unsupervised neural networks. *Science*, *268*, 1158–1161.

Hinton, G. T., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*, 74–95.

Hirsch, J. A., & Gilbert, C. D. (1991). Synaptic physiology of horizontal connections in the cat's visual cortex. *Journal of Neuroscience*, *11*, 1800–1809.

Karhunen, J., & Joutsensalo, J. (1994). Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, *7*, 113–127.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, *407*, 717–726.

Kawato, M., Hayakawa, H., & Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network*, *4*, 415–422.

Kay, J., & Phillips, W. A. (1996). Activation functions, computational goals and learning rules for local processors with contextual guidance. *Neural Computation*, *9*, 895–910.

Kisvarday, Z. F., Crook, J. M., Buzas, P., & Eysel, U. T. (2000). Combined physiological–anatomical approaches to study lateral inhibition. *Journal of Neuroscience Methods*, *103*, 91–106.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, *37*, 233–243.

Linsker, R. (1990). Perceptual neural organisation: Some approaches based on network models and information theory. *Annual Review of Neuroscience*, *13*, 257–281.

Liu, J., & Newsome, W. T. (2003). Functional organization of speed tuned neurons in visual area MT. *Journal of Neurophysiology*, *89*, 246–256.

MacKay, D. M. (1956). The epistemological probalem for automata. In Automomata Studies. Princeton, NJ. Princeton University Press. pp. 235–251.

Martin, S. J., Grimwood, P. D., & Morris, R. G. (2000). Synaptic plasticity and memory: An evaluation of the hypothesis. *Annual Review of Neuroscience*, *23*, 649–711.

McIntosh, A. R. (2000). Towards a network theory of cognition. *Neural Networks*, *13*, 861–870.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*, 1013–1052.

Mumford, D. (1992). On the computational architecture of the neocortex II. The role of cortico-cortical loops. *Biological Cybernetics*, *66*, 241–251.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Science, USA*, *99*, 15164–15169.

Nebes, R. (1989). Semantic memory in Alzheimer's disease. *Psychological Bulletin*, *106*, 377–394.

Neisser, U. (1967). Cognitive psychology. New York: Appleton-Century-Crofts.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, *1*, 61–68.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Optican, L., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II. Information theoretic analysis. *Journal of Neurophysiology*, *57*, 132–146.

Pack, C. C., & Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, *409*, 1040–1042.

Phillips, W. A., & Singer, W. (1997). In search of common foundations for cortical computation. *Behavioural and Brain Sciences*, *20*, 57–83.

Phillips, C. G., Zeki, S. H. B., & Barlow, H. B. (1984). Localisation of function in the cerebral cortex. Past present and future. *Brain*, *107*, 327–361.

Plaut, D., & Shallice, T. (1993). Deep dyslexia—A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularisation theory. *Nature*, *317*, 314–319.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, in press.

Pouget, A., Deneve, S., & Duhamel, J. R. (2002). A computational perspective on the neural basis of multisensory spatial. *Natural Review of Neuroscience*, *3*, 741–747.

Rao, R. P. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, *39*, 1963–1989.

Rao, R. P., & Ballard, D. H. (1998). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience*, *2*, 79–87.

Rivadulla, C., Martinez, L. M., Varela, C., & Cudeiro, J. (2002). Completing the corticofugal loop: A visual role for the corticogeniculate type 1 metabotropic glutamate receptor. *Journal of Neuroscience*, *22*, 2956–2962.

Rizzo, M., Nawrot, M., & Zihl, J. (1995). Motion and shape perception in cerebral akinetopsia. *Brain*, *118*, 1105–1127.

Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research*, *179*, 3–20.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Salin, P.-A., & Bullier, J. (1995). Corticocortical connections in the visual system: Structure and function. *Psychological Bulletin*, *75*, 107–154.

Sandell, J. H., & Schiller, P. H. (1982). Effect of cooling area 18 on striate cortex cells in the squirrel monkey. *Journal of Neurophysiology*, *48*, 38–48.

Sherman, S. M., & Guillery, R. W. (1998). On the actions that one nerve cell can have on another: Distinguishing ‘drivers’ from ‘modulators’. *Proceedings of the National Academy of Science USA*, *95*, 7121–7126.

Sincich, L. C., & Horton, J. C. (2002). Divided by cytochrome oxidase: A map of the projections from V1 to V2 in macaques. *Science*, *295*, 1734–1737.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*, 869–873.

Taleb, A., & Jutten, C. (1997). Nonlinear source separation: The post-nonlinear mixtures. *ESANN'97 Bruges*, *April*, 279–284.

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Science*, *91*, 5033–5037.

Tovee, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the response of single neurons in the primate temporal visual cortex. *Journal of Neurophysiology*, *70*, 640–654.

Treue, S., & Maunsell, H. R. (1996). Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature*, *382*, 539–541.

Warrington, E. K., & McCarthy, R. (1983). Category specific access dysphasia. *Brain*, *106*, 859–878.

Warrington, E. K., & McCarthy, R. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, *110*, 1273–1296.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–853.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, *358*, 593–602.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317–1329.

Zeki, S. (1990). The motion pathways of the visual cortex. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 321–345). UK: Cambridge University Press.

Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature*, *335*, 311–317.