PERGAMON

Neural Networks 13 (2000) 871–882

Neural Networks

www.elsevier.com/locate/neunet

2000 Special Issue

Assessing interactions among neuronal systems using functional neuroimaging

C. Büchel[a,*], K. Friston[b]

[a]*Cognitive Neuroscience Laboratory, Department of Neurology, University of Hamburg, Martin Str. 52, 20246 Hamburg, Germany*
[b]*Wellcome Department of Cognitive Neurology, 12 Queen Square, London WC1N 3BG, UK*

Received 1 July 2000; accepted 12 July 2000

Abstract

We show that new methods for measuring effective connectivity allow us to characterise the interactions between brain regions that underlie the complex interactions among different processing stages of functional architectures. © 2000 Published by Elsevier Science Ltd.

1. Introduction

In the late 19th century the early investigations of brain function were dominated by the concept of functional segregation. This approach was driven largely by the data available to scientists of that era. Patients with circumscribed lesions were found who were impaired in one particular ability while other abilities remained largely intact. Indeed, descriptions of patients with different kinds of aphasia (an impairment of the ability to use or comprehend words), made at this time, have left a permanent legacy in the contrast between Broca's and Wernicke's aphasia. These syndromes were thought to result from damage to anterior or posterior regions of the left hemisphere. In the first part of the 20th century the idea of functional segregation fell into disrepute and the doctrine of "mass action" held sway, proposing that higher abilities depended on the function of the brain "as a whole" (Lashley, 1929). This doctrine was always going to be unsatisfying. However, with the resources available at the time it was simply not possible to make any progress studying the function of the "brain as a whole". By the end of the 20th century the concept of functional segregation has returned to domination.

The doctrine is now particularly associated with cognitive neuropsychology and is enshrined in the concept of double dissociation (see Shallice, 1988, chap. 10). A double dissociation is demonstrated when neurological patients can be found with "mirror" abnormalities. For example, many patients have been described who have severe impairments of long-term memory while their short-term memory is intact. Warrington and Shallice (1969) described the first of a series of patients who had severe impairments of phonological short-term memory, but no impairments of long-term memory. This is a particularly striking example of double dissociation. It demonstrates that different brain regions are involved in short- and long-term memory. Furthermore, it shows that these regions can function in a largely independent fashion. This observation caused major problems for theories of memory, extant at the time, which supposed that inputs to long-term memory emanated from short-term memory systems (e.g. Atkinson & Shiffrin, 1968).

Functional brain imaging avoids many of the problems of lesion studies, but here too, the field has been dominated by the doctrine of functional segregation. Nevertheless, it is implicit in the subtraction method that brain regions communicate with each other. If we want to distinguish between brain regions associated with certain central processes for example, then we will design an experiment in which the sensory input and motor output is the same across all conditions. In this way activity associated with sensory input and motor output will cancel out. The early studies of reading by Posner and his colleagues are still among the best examples of this approach (Petersen, Fox, & Snyder, 1990; Posner, Petersen, Fox, & Raichle, 1988). The design of these studies was based on the assumption that reading goes through a single series of discrete and independent stages; visual shapes are analysed to form letters, letters are put together to form words, the visual word form is translated into sound, the sound form is translated into articulation, and so on. By comparison of suitable

* Corresponding author.
E-mail address: buechel@uke.uni-hamburg.de (C. Büchel).

0893-6080/00/$ - see front matter © 2000 Published by Elsevier Science Ltd.
PII: S0893-6080(00)00066-6

tasks (e.g. letters vs false font, words vs letters, etc.), each stage can be isolated and the associated brain region identified. Although subsequent studies have shown that this characterisation of the brain activity associated with reading is a considerable oversimplification, the original report still captures the essence of most functional imaging studies; a number of discrete cognitive stages are mapped onto discrete brain areas. Nothing is revealed about how the cognitive processes interact or how the brain regions communicate with each other. If word recognition really did depend on the passage of information through a single series of discrete stages we would at least like to know the temporal order in which the associated brain regions were engaged. Some evidence comes from EEG and MEG studies. In fact, we know that word recognition depends upon at least two parallel routes; one via meaning and one via phonology (Marshall & Newcombe, 1973). Given this model we would like to be able to specify the brain regions associated with each route and have some measure of the strengths of the connections between these different regions.

In this article we shall show that new methods for measuring effective connectivity allow us to characterise the interactions between brain regions which underlie the complex interactions among different processing stages of functional architectures.

2. Definitions

In the analysis of neuroimaging time-series (i.e. signal-changes in a set of voxels, expressed as a function of time), functional connectivity is defined as the *temporal correlations between spatially remote neurophysiological events* (Friston, Frith, Liddle, & Frackowiak, 1993). This definition provides a simple characterisation of functional interactions. The alternative is effective connectivity (i.e. *the influence one neuronal system exerts over another*) (Friston, Frith, & Frackowiak, 1993). These concepts originated in the analysis of separable spike trains obtained from multiunit electrode recordings (Aertsen & Preissl, 1991; Gerstein & Perkel, 1969). Functional connectivity is simply a statement about the observed correlations; it does not comment on how these correlations are mediated. For example, at the level of multiunit micro-electrode recordings, correlations can result from *stimulus-locked transients*, evoked by a common afferent[1] input, or reflect *stimulus-induced oscillations* and phasic coupling of neural assemblies, mediated by synaptic connections (Gerstein, Bedenbaugh, & Aertsen, 1989). Effective connectivity is closer to the notion of a connection, either at a synaptic (cf. synaptic efficacy) or cortical level. Although functional and effective connectivity can be invoked at a conceptual level in both neuroimaging and electrophysiology they differ fundamentally at a practical level. This is because the time-scales and nature of neurophysiological measurements are very different (seconds vs milliseconds and hemodynamic vs spike trains). In electrophysiology it is often necessary to remove the confounding effects of stimulus-locked transients (that introduce correlations *not* causally mediated by direct neural interactions) in order to reveal an underlying connectivity. The confounding effect of stimulus-evoked transients is less problematic in neuroimaging because propagation of signals from primary sensory areas onwards is mediated by neuronal connections (usually reciprocal and interconnecting). However, it should be remembered that functional connectivity is not necessarily due to effective connectivity (e.g. common neuromodulatory input from ascending aminergic neurotransmitter systems or thalamo-cortical afferents) and, where it is, effective influences may be indirect (e.g. polysynaptic relays through multiple areas). In this article we will only focus on effective connectivity. More details about functional connectivity can be found in Friston, Frith, Liddle et al. (1993).

[1] That is, signal input into the neural system as a result of external stimulation.

3. Effective connectivity

3.1. A simple model

Effective connectivity depends on two models: a mathematical model, describing "how" areas are connected and a neuroanatomical model describing "which" areas are connected. We shall consider linear and non-linear models. Perhaps the simplest model of effective connectivity expresses the hemodynamic change at one voxel as a weighted sum of changes elsewhere. This can be regarded as a multiple linear regression, where the effective connectivity reflects the amount of regional cerebral blood flow (rCBF) variability, at the target region, attributable to rCBF changes at a source region. As an example, consider the influence of other areas M on area $V1$. This can be framed in a simple equation:

$$V1 = Mc + e \tag{1}$$

where $V1$ is a $n \times 1$ column vector with n scans, M is a $n \times m$ matrix with m regions and n observations (scans), c is a $m \times 1$ column vector with a parameter estimate for each region and e is a vector of error terms.

Implicit in this interpretation is a mediation of the influence among brain regions by neuronal connections with an effective strength equal to the (regression) coefficients c. This highlights the fact that the linear model assumes that the connectivity is constant over the whole range of activation and does not depend on input from other sources.

Experience suggests that the linear model can give fairly robust results. One explanation is that the dimensionality (the number of things that are going on) of the physiological changes can be small by experimental design. In other words the brain responds to simple and well-organised

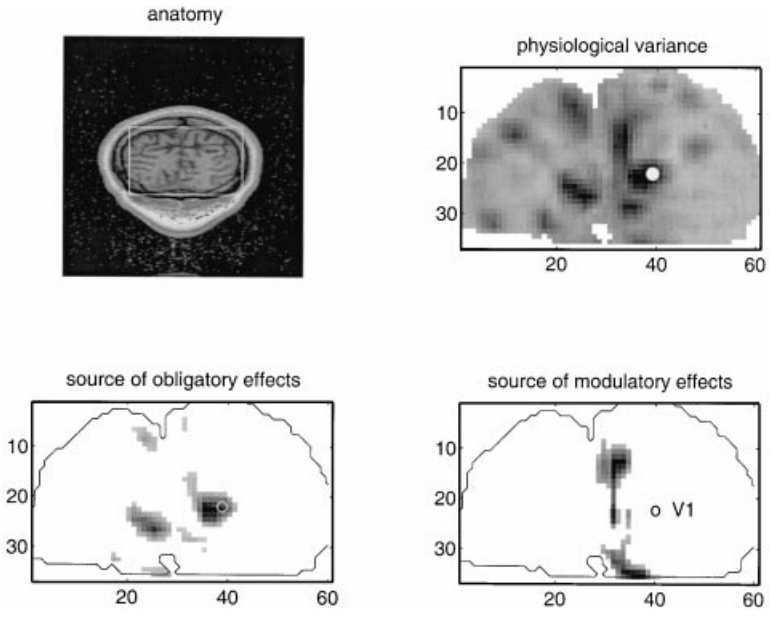

Fig. 1. Maps of the estimates of obligatory and modulatory connection strengths to right *V*1. Top left: anatomical features of the coronal data used. This image is a high-resolution anatomical MRI scan of the subject that corresponds to the fMRI slices. The box defines the position of a sub-partition of the fMRI time-series selected for analysis. Top right: the location of the reference voxel designated as right *V*1 (white dot). This location is shown on a statistical parametric map of physiological variance (calculated for each voxel from the time-series of 60 scans). Lower left and lower right: maps of c_O and c_M. The images have been scaled to unit variance and thresholded at $p = 0.05$ (assuming, under the null hypothesis of no effective connectivity, the estimates have a Gaussian distribution). The reference voxel in *V*1 is depicted by a circle. The key thing to note is that *V*1 is subject to modulatory influences from ipsilateral and extensive regions of *V*2.

experiments in a simple and well-organised way. Generally, however, neurophysiological interactions are non-linear and the adequacy of linear models must be questioned (or at least qualified). Consequently we will focus on a non-linear model of effective connectivity (Friston, Ungerleider, Jezzard, & Turner, 1995).

Reversible cooling experiments in monkey visual cortex, during visual stimulation, have demonstrated that neuronal activity in *V*2 depends on forward inputs from *V*1. Conversely neuronal activity in *V*1 is *modulated* by backward or re-entrant connections from *V*2 to *V*1 (Girard & Bullier, 1988; Sandell & Schiller, 1982; Schiller & Malpeli, 1977). Retinotopically corresponding regions of *V*1 and *V*2 are reciprocally connected in the monkey. *V*1 provides a crucial input to *V*2, in the sense that visual activation of *V*2 cells depends on input from *V*1. This dependency has been demonstrated by deactivating (reversibly cooling) *V*1 while recording from *V*2 during visual stimulation. In contrast, cooling *V*2 has a more *modulatory* effect on *V*1 activity. The cells in *V*1 that were most affected by *V*2 deactivation were in the infragranular layers, suggesting *V*2 may use this pathway to modulate the output from *V*1 (Sandell & Schiller, 1982). Because, in the absence of *V*1 input, these re-entrant connections do not constitute an efficient drive to *V*2 cells, their role is most likely "to modulate the information relayed through area 17".

To examine the interactions between *V*1 and *V*2, using fMRI in humans, it is possible to use a non-linear model of effective connectivity, extended to include a modulatory interaction (Eq. (1)):

$$V1 = M{\cdot}c_O + \mathrm{diag}(V1)Mc_M + e \quad (2)$$

where diag(*V*1) refers to a diagonal matrix with elements in the vector *V*1; this premultiplies the (scan × region) matrix *M* so that each region's contribution to the model is affected by the activity in *V*1. This model has two terms that allow for the activity in area *V*1 to be influenced by the activity in other areas *M* (our hypothesis being that *V*2 is prominent amongst those areas). The first represents an effect that depends only on afferent input from other areas *M*. This is the activity in *M* scaled by c_O. The coefficients in c_O are referred to as *obligatory* or driving connection strengths, in the sense that a change in areas *M* results in an obligatory response in area *V*1. This is similar to *c* in the simple linear model above. Conversely, the second term reflects a *modulatory* influence of areas *M* on area *V*1. The coefficient determining the size of this effect (c_M) is referred to as a modulatory connection strength, because the overall effect depends on both the afferent input ($M{\cdot}c_M$) and intrinsic activity in *V*1. This effect can be considered as a greater responsiveness of *V*1 to inputs with higher intrinsic activation of *V*1.

This intrinsic activity-dependent effect, determined by the value of c_M, provides an intuitive sense of how to estimate c_M. Imagine one were able to "fix" the activity in *V*1 at a *low* level and measure the connectivity between the regions in *M* and *V*1 assuming a simple linear relationship [Eq. (1)]: a value for the sensitivity of *V*1 to changes

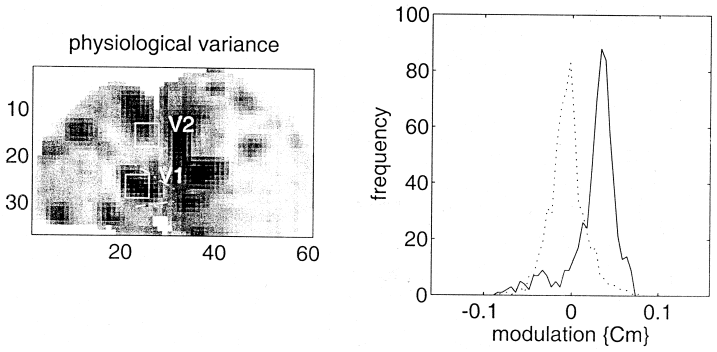

Fig. 2. Graphical presentation of a direct test of the hypothesis concerning the asymmetry between forward and backward *V*1–*V*2 interactions. Left: a map of physiological variance showing the positions of two boxes defining regions in left *V*1 and *V*2. The broken lines correspond (roughly) to the position of the *V*1/*V*2 border according to the atlas of Talairach and Tournoux (1988). The value of c_M was computed for all voxels in each box and normalized to unity over the image. The frequency distribution of c_M connecting the two regions is presented on the right. The modulatory backward connections (*V*2–*V*1, solid line) are clearly higher than the modulatory forward connections (*V*1–*V*2, broken line).

elsewhere could be obtained, say c_1. Now, if the procedure were repeated with *V*1 activity fixed at a *high* level, a second (linear) estimate would be obtained, say c_2. In the presence of a substantial modulatory interaction between regions in *M* and *V*1 the second estimate (c_2) will be higher than the first (c_1). This is because the activity intrinsic to *V*1 is higher and *V*1 should be more sensitive to inputs. In short $c_2 - c_1$ provides an estimate of the modulatory influence on *V*1 (similarly $c_1 + c_2$ is related to c_O). By analogy to reversible cooling which allows one to remove the effects of isolated cortical regions, we "fix" activity *post hoc* by simply selecting a subset of data in which the *V*1 activity is confined to some small range (high or low activity).

A relevant example analysis is now described. The data used in this analysis were a time-series of 64 gradient-echo EPI 5 mm coronal slices through the calcarine sulcus and extrastriate areas. Images were obtained every 3 s from a normal male subject using a 4 T whole body system. Photic stimulation (at 16 Hz) was provided by goggles fitted with light-emitting diodes. The stimulation was off for the first 30 s, on for the second 30 s, off for the third, and so on. The first four scans were removed to eliminate magnetic saturation effects and the remainder were realigned.

A reference voxel was chosen in right *V*1 and the effective connection strengths c_M were estimated allowing a map of c_M (and c_O) to be constructed. This map provides a direct test of the hypothesis concerning the topography and regional specificity of modulatory influences on *V*1. The lower row in Fig. 1 shows maps of c_O and c_M (neurological convention — right *V*1 is marked); these reflect the degree to which the area exerts an obligatory (left) or modulatory (right) effect on *V*1 activity. These maps have been thresholded at 1.64 after normalization to a standard deviation of unity. This corresponds to an uncorrected threshold of $p < 0.05$.

The obligatory connections to the reference voxel derive mainly from *V*1 itself, both ipsilaterally and contralaterally with a small contribution from contiguous portions of *V*2. The effective connectivity from contralateral *V*1 should not be over-interpreted given that: (i) the source of many afferents to *V*1 (the lateral geniculate nuclei) were not included in the field of view; and that (ii) this finding can be more parsimoniously explained by "common input". As predicted, and with remarkable regional specificity, the modulatory connections were most marked from ipsilateral *V*2, dorsal and ventral to the calcarine fissure (note that "common input" cannot explain interactions between *V*1 and *V*2 because the geniculate inputs are largely restricted to *V*1).

3.1.1. Functional asymmetry in V2–V1 and V1–V2 modulatory connections

To address functional asymmetry[2] in terms of forwards and backwards modulatory influences the modulatory connection strengths between two extended regions (two 5×5 voxel squares) in ipsilateral *V*1 and *V*2 were examined. The estimates of effective connection strengths were based on hemodynamic changes in all areas and the subset of connections between the two regions were selected to compare the distributions of forward and backward modulatory influences. Fig. 2 shows the location of the two regions (this time on the left) and the frequency distribution (i.e. histogram) of the estimates for connections from the voxels in the *V*1 box to the *V*2 box (broken line) and the corresponding estimates for connections from voxels in the *V*2 box to *V*1 (solid line). There is a remarkable dissociation, with backward modulatory effects (*V*2–*V*1) being much greater than forward effects (*V*1–*V*2). This can be considered a confirmation of the functional asymmetry hypothesis.

3.2. Structural equation modelling

The simple model above was sufficient to analyse effective connectivity to one region at a time (e.g. *V*1 or *V*2). We will now introduce structural equation modelling as a tool

[2] Here we are referring to asymmetry in connections, not asymmetry between hemispheres.

allowing for more complicated models comprising many regions of interest and demonstrate how non-linear interactions are dealt with in this context. The basic idea behind structural equation modelling (SEM) differs from the usual statistical approach of modelling individual observations. In multiple regression or AnCova models the regression coefficients derive from the minimisation of the sum of squared differences of the predicted and observed dependent variables (i.e. activity in the target region). Structural equation modelling approaches the data from a different perspective; instead of considering variables individually the emphasis lies on the variance–covariance structure.[3] Thus models are solved in structural equation modelling by minimising the difference between the observed variance-covariance structure and the one implied by a structural or path model. In the past few years structural equation modelling has been applied to functional brain imaging. For example McIntosh et al. (1994) demonstrated the dissociation between ventral and dorsal visual pathways for object and spatial vision using structural equation modelling of PET data in the human. In this section we will focus on the theoretical background of structural equation modelling and demonstrate this technique using fMRI.

In terms of neuronal systems a measure of covariance represents the degree to which the activities of two or more regions are related (i.e. functional connectivity). The study of variance–covariance structures here is much simpler than in many other fields; the interconnection of the dependent variables (regional activity of brain areas) is anatomically determined and the activation of each region can be directly measured with functional brain imaging. This represents a major difference to "classical" structural equation modelling in the behavioural sciences, where models are often hypothetical and include latent variables denoting rather abstract concepts like intelligence.

As mentioned above, structural equation modelling minimises the difference between the observed or measured covariance matrix and the one that is implied by the structure of the model. The free parameters (path coefficients or connection strengths c above) are adjusted to minimise the difference[4] between the measured and modelled covariance matrix (see Büchel & Friston, 1997) for details).

An important issue in structural equation modelling is the determination of the participating regions and the underlying anatomical model. Several approaches to this issue can be adopted. These include categorical comparisons between different conditions, statistical images highlighting structures of functional connectivity and non-human electrophysiological and anatomical studies (McIntosh & Gonzalez-Lima, 1994).

With respect to anatomical connectivity in humans the advent of new MR techniques promises a better characterisation of neuronal connectivity in humans. Diffusion tensor imaging (DTI) measures the anisotropy of diffusion in the brain. The main anisotropy exists in the white matter because the orientation of neuronal fibres (axons) allows molecules to diffuse easier along the fibre than in other directions. Therefore the main direction of the diffusion tensor reflects the underlying orientation of white matter tracts. Through tracing algorithms it is now possible to infer the connectivity of individual regions (e.g. activations derived from an fMRI study) in an individual brain.

A model is always a simplification of reality: exhaustively correct models either do not exist or would be too complicated to understand. In the context of effective connectivity one has to find a compromise between complexity, anatomical accuracy and interpretability. There are also mathematical constraints on the model; if the number of free parameters exceeds the number of observed covariances the system is underdetermined and no single solution exists.

Each estimated model can be analysed to give an overall goodness-of-fit measure, for use when comparing different models with each other. A "nested model" approach can be used to compare different models (e.g. data from different groups or conditions) in the context of structural equation modelling. A so-called "null-model" is constructed where the estimates of the free parameters are constrained to be the same for both groups. The alternative model allows free parameters to differ between groups. The significance of the differences between the models is expressed by the difference of the goodness-of-fit statistic. Consider the following hypothetical example. Subjects are scanned under two different conditions, e.g. "attention" and "no attention". The hypothesis might be that within a system of regions A, B, C and D, the connectivity between A and B is different under the two attentional conditions. To determine whether the difference in connectivity is statistically significant, we estimate the goodness-of-fit measure for two models: model 1 allows the connectivity between A and B to take different values for both conditions. Model 2 constrains the path coefficient between A and B to be equal for "attention" and "no attention". If the change of connectivity between "attention" and "no attention" for the connection of A and B is negligible, the constrained model (Model 2) should fit the data equally well compared to the free model (Model 1). We can now infer whether the difference of the two goodness-of-fit measures is significant. Non-linear models can also be accommodated in the framework of SEM by introducing additional variables containing a non-linear function (e.g. $f(x) = x^2$) of the original variables (Kenny & Judd, 1984). Interactions of variables can be incorporated in a similar fashion, wherein a new variable, containing the product of the two interacting variables, is

[3] The variance–covariance structure describes in detail the dependencies between the different variables (in this case, the measured regional responses to stimulation).

[4] The free parameters are estimated by minimising a function of the observed and implied covariance matrix. To date the most widely used objective function in structural equation modelling is the maximum likelihood (ML) function.

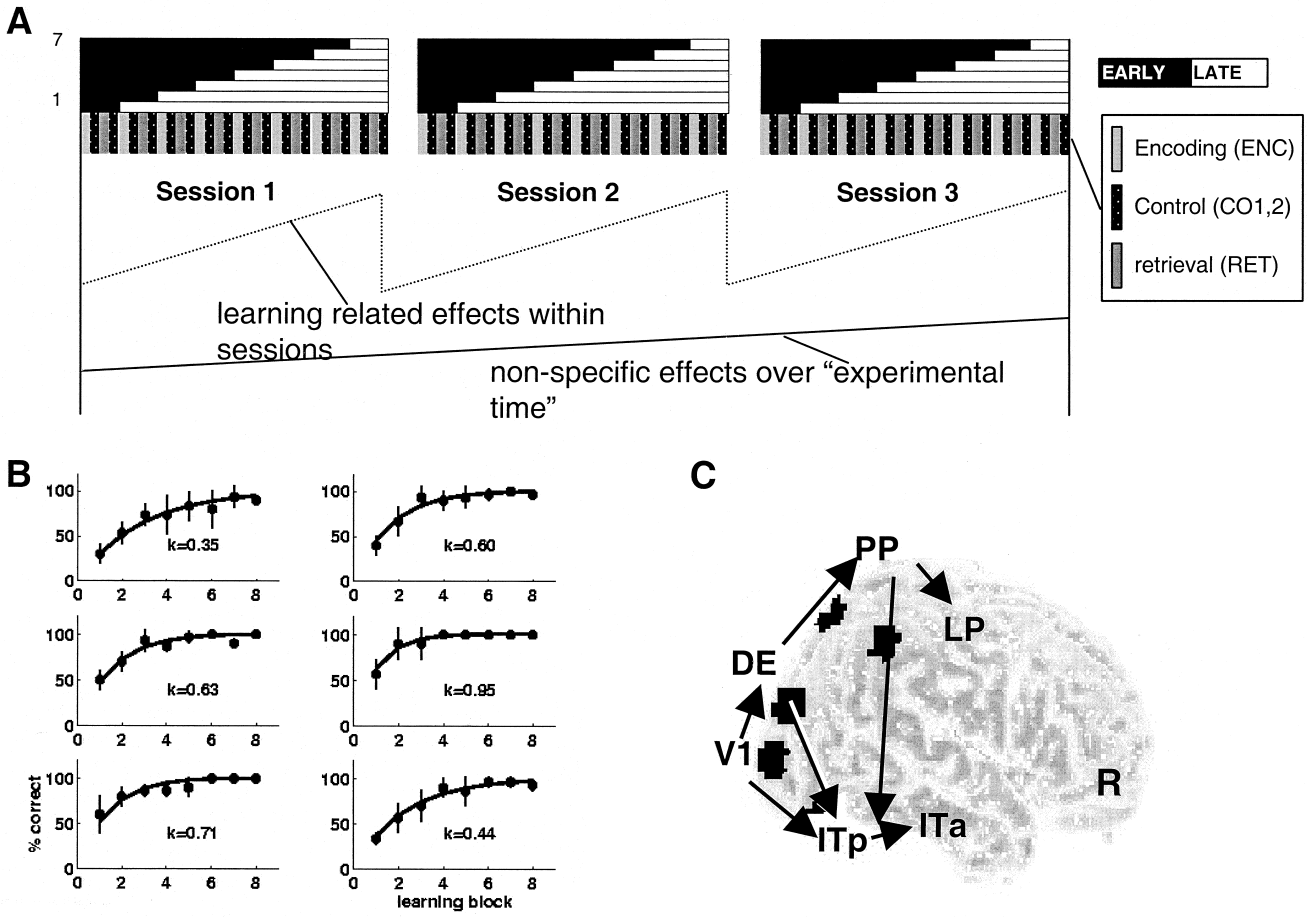

Fig. 3. Changes in effective connectivity over time in paired associates learning: (a) the design of the study. Blocks of "encoding" and "retrieval" were alternated by control conditions. Subjects had to perform three individual learning sessions, to avoid the confounding effect of time. (b) The behavioural performance data for each of the six subjects averaged across all three learning sessions. (c) The anatomical model. Processing of object identity is mainly a property of the ventral visual pathway, whereas object location is a property of the dorsal stream. We focussed on the interstream connections (mainly PP to ITp) based on the hypothesis that learning the association of object identity and spatial location will lead to an increase in effective connectivity between the ventral and the dorsal stream.

introduced as an additional influence. This is similar to the approach used in the previous section, where the interaction was expressed by the influence of the product of *V*1 and *V*2 on *V*1. We will now demonstrate these ideas using an example. More details of structural equation modelling, including the operational equations can be found in Büchel and Friston (1997).

3.2.1. Example — learning

In this first example we were interested in changes in effective connectivity over time as expected during paired associates learning (Büchel, Coull, & Friston, 1999). In the case of object-location memory several functional studies have demonstrated activation of ventral occipital and temporal regions during the retrieval of object identity and, conversely, increased responses in dorsal parietal areas during the retrieval of spatial location (Milner, Johnsrude, & Crane, 1997). These results suggest domain-specific representations in posterior neocortical structures, closely related to those involved in perception, a finding that accords with the segregation of ventral and dorsal pathways in processing categorical or spatial stimulus features, respectively. Another phenomenon observed in some learning studies is a decrease of neural responses (i.e. adaptation) to repeated stimulus presentations. This repetition suppression has been replicated consistently in primate electrophysiological and human functional imaging studies (Desimone, 1996). For object-location learning, it is intuitively likely that two specialised systems need to interact to establish an association. Domain-specific representations or repetition suppression are not sufficient to account for this associative component. In other words, functional segregation and localised response properties cannot account for associative learning alone.

In our fMRI experiment, decreases in activation during learning, indicative of repetition suppression, were observed in several cortical regions in the ventral and dorsal visual pathway. Within the framework of repetition suppression it has been hypothesised that decreases in neural responses are a secondary result of enhanced response selectivity (Wiggs

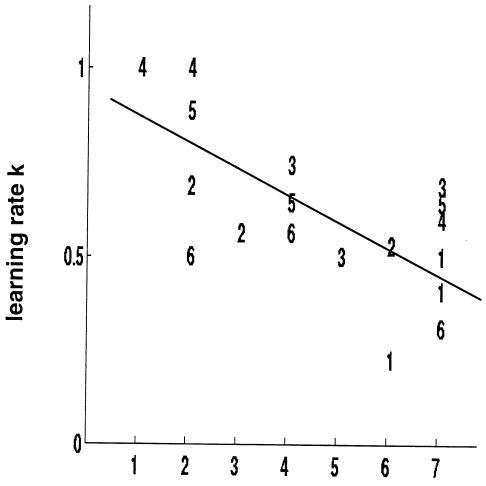

Fig. 4. Changes in effective connectivity predict learning. This graph shows the correlation between the temporal index of changes in effective connectivity and learning. The temporal index is defined as the time of a maximum increase in effective connectivity between PP and ITp, e.g. a temporal index of 3 indicates that the maximum increase in effective connectivity occurred between the third and the fourth block. The numbers denote the subject from which this temporal index of effective connectivity was obtained. Each subject was scanned during three independent learning sessions, therefore each number appears three times. A negative slope means that the maximum increase in effective connectivity occurs earlier in fast learning.

& Martin, 1998). By analogy to the development and plasticity of cortical architectures, this refined selectivity is likely to be due to changes in effective connectivity within the system at a synaptic level. We explicitly addressed this notion by characterising time-dependent changes in effective connectivity during learning.

The experiment was performed on a 2 T MRI system equipped with a head volume coil. fMRI images were obtained every 4.1 s with echo-planar imaging (48 slices in each volume). Six subjects had to learn and recall the association between 10 simple line drawings of real-world objects and 10 locations on a screen during fMRI. Each learning trial consisted of four conditions, “Encoding”, “Control”, “Retrieval” and “Control” (Fig. 3a). The behavioural data acquired during “Retrieval” demonstrated that all six subjects were able to learn the association between object identity and spatial location, for all 10 objects, within eight learning blocks, as indicated by the ensuing asymptotic learning curves (Fig. 3b).

The structural model used in the analysis embodies connections within and across ventral and dorsal visual pathways and was based on anatomical studies in primates (Fig. 3c). Primary visual cortex was modelled as the origin of both pathways. In addition to “interstream” connections between dorsal extrastriate cortex (DE) and the fusiform region (ITp) and between the posterior parietal cortex (PP) and ITp, we included direct connections based on a hierarchical cortical organisation. Given our hypothesis relating to changes in effective connectivity between dorsal and ventral pathways, the path analysis focused on the connection between posterior parietal cortex (PP, dorsal stream) and posterior inferotemporal cortex (ITp, ventral stream). We divided each learning session into EARLY (first part) and LATE observations (second part) and estimated separate path coefficients for each partition.

The path coefficient between PP and ITp increased significantly during learning in the group ($p < 0.05$) and was confirmed by an analysis of individual subjects showing an increase in effective connectivity between PP and ITp of 0.27. In contrast to the connections between streams, connections within the dorsal pathway decreased over time.

The estimated change in connectivity from PP to ITp clearly depended on the cut-off point between EARLY and LATE. To unequivocally establish a relationship between neurophysiologically mediated changes in connectivity and behavioural learning, we examined the relationship between the temporal pattern of effective connectivity changes and learning speed for all sessions and subjects. We estimated the differences in effective connectivity for seven EARLY and LATE partitions, by successively shifting the cut-off. The cut-off time at which the connectivity change peaked was used as a temporal index of changes in effective connectivity (i.e. plasticity). The significant regression of k, a measure of learning speed,[5] on this plasticity index indicated that for sessions showing fast learning (i.e. high k) the maximum difference in path coefficients between PP and ITp was achieved earlier in the session (i.e. EARLY comprises less scans relative to LATE) (Fig. 4). In other words, the temporal pattern of changes in effective connectivity strongly predicted learning or acquisition.

3.2.2. Example — attention

Electrophysiological and neuroimaging studies have shown that attention to visual motion can increase the responsiveness of the motion-selective cortical area *V*5 (O’Craven & Savoy, 1995; Treue & Maunsell, 1996) and the posterior parietal cortex (PP) (Assad & Maunsell, 1995). Increased or decreased activation in a cortical area is often attributed to attentional modulation of the cortical projections to that area. This leads to the notion that attention is associated with changes in connectivity.

Here we present fMRI data from an individual subject, scanned under identical visual motion stimulus conditions, while changing only the attentional component of the tasks employed. First, we identify regions that show differential activations in relation to attentional set. In the second stage, changes in effective connectivity to these areas are assessed using structural equation modelling. Finally, we show how these attention-dependent changes in effective connectivity can be explained by the modulatory influence of parietal areas using a non-linear extension of structural equation

[5] All individual behavioural learning curves were well approximated by the function $1 - e^{-kx}$ where $0 < k < 1$ indexes learning speed. Small values of k indicate slower learning.

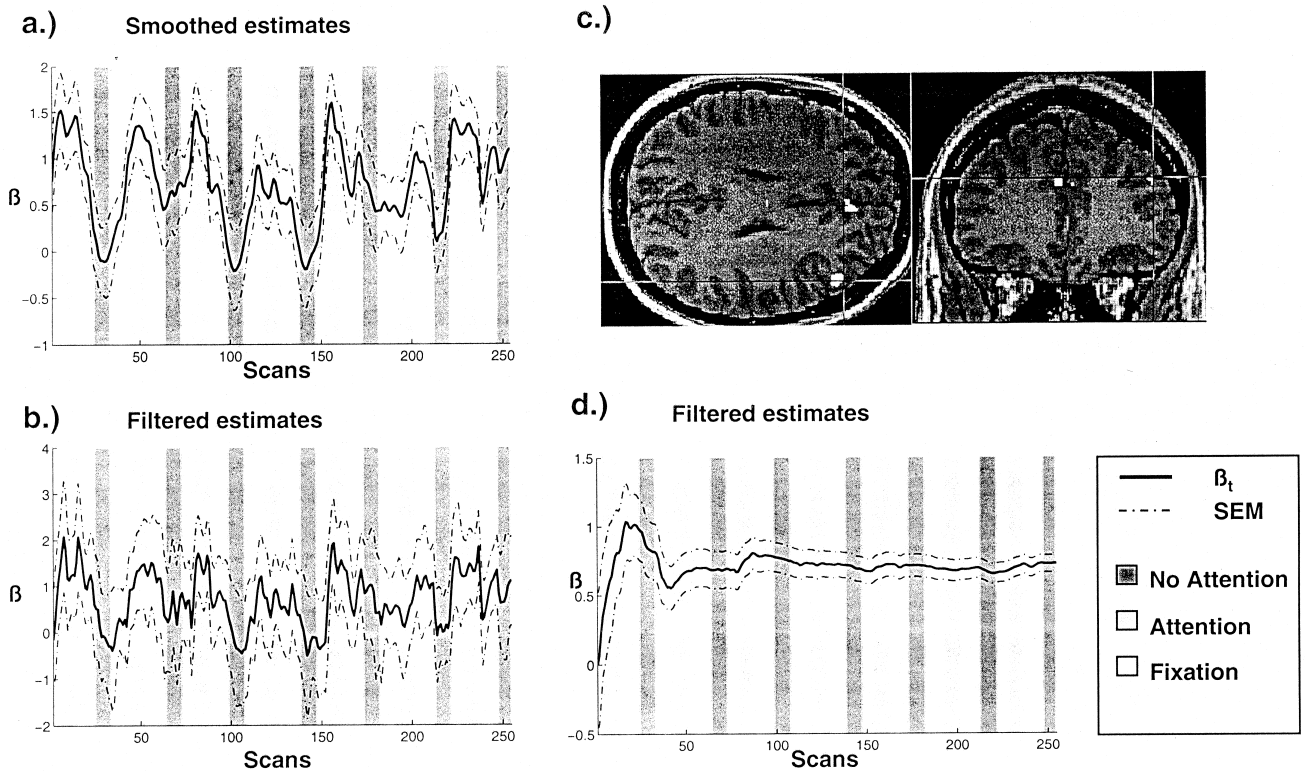

Fig. 5. Structural equation model of the dorsal visual pathway, comparing "attention" and "no attention". Connectivity between right *V*1 and *V*5 is increased during "attention" relative to "no attention". This is also shown for the connection between *V*5 and PP.

modelling. The specific hypothesis we addressed was that parietal cortex could modulate the inputs from *V*1 to *V*5.

The experiment was performed on a 2 T MRI system equipped with a head volume coil. fMRI images were obtained every 3.2 s with echo-planar imaging (32 slices in each volume). The subject was scanned during four different conditions: "fixation", "attention", "no attention" and "stationary". Each condition lasted 32 s giving 10 volumes per condition. We acquired a total of 360 images. During all conditions the subjects looked at a fixation point in the middle of a screen. In this section we are only interested in the two conditions with visual motion ("attention" and "no attention"), where 250 small white dots moved radially from the fixation point, in random directions, towards the border of the screen, at a constant speed of 4.7° per second. The difference between "attention" and "no attention" lay in the explicit command given to the subject shortly before the condition: "just look" indicated "no attention" and "detect changes" the "attention" condition. Both visual motion conditions were interleaved with "fixation". No response was required.

Regions of interest (ROI) were defined by categorical comparisons using an output statistical image ("SPM{Z}") comparing "attention" and "no attention" and comparing "no attention" and "fixation". As predicted, given a stimulus consisting of radially moving dots, we found activation of the lateral geniculate nucleus (LGN), primary visual cortex (*V*1), motion sensitive area *V*5 and the posterior parietal complex (PP). For the subsequent analysis of effective connectivity, we defined ROI with a diameter of 8 mm, centred around the most significant voxel as revealed by the categorical comparison. A single time-series, representative of this region, was defined by the first eigenvector of all the voxels in the ROI (Büchel & Friston, 1997).

Our model of the dorsal visual stream included the LGN, primary visual cortex (*V*1), *V*5 and the posterior parietal complex (PP). Although connections between regions are generally reciprocal, for simplicity we only modelled unidirectional paths.

To assess effective connectivity in a condition-specific fashion, we used time-series that comprised observations during the condition in question. Path coefficients for both conditions ("attention" and "no attention") were estimated using a maximum likelihood function. To test for the impact of changes in effective connectivity between "attention" and "no attention", we defined a free model (allowing different path coefficients between *V*1 and *V*5 for attention and no attention) and a constrained model (constraining the $V1 \rightarrow V5$ coefficients to be equal). Fig. 5 shows the free model and the estimated path coefficients. The connectivity between *V*1 and *V*5 increases significantly during attention. Note that there is also a significant difference in connectivity between *V*5 and PP.

The linear path model comparing "attention" and "no attention" revealed increased effective connectivity in the dorsal visual pathway in relation to attention. The question

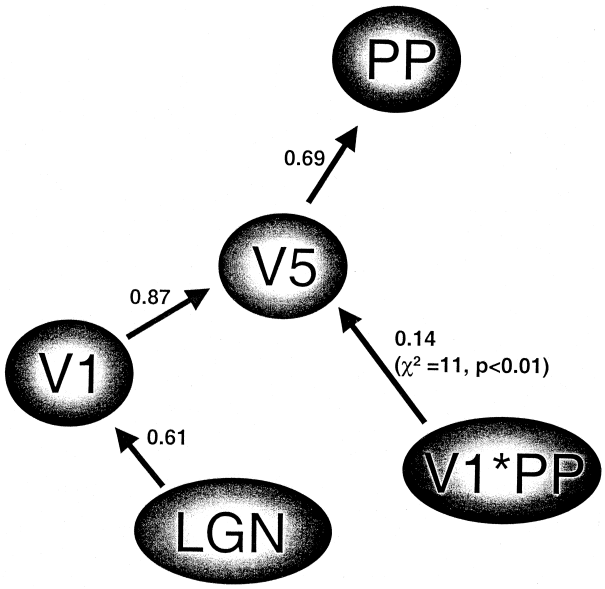

Fig. 6. Structural equation model of the dorsal visual pathway incorporating the interaction of right PP on the connection from right *V*1 to *V*5.

that arises is — which part of the brain is capable of modulating this pathway? Based on lesion studies (Lawler & Cowey, 1987) and on the system for directed attention as described in Mesulam (1990), the posterior parietal cortex is hypothesised to play such a modulatory role.

We extended our model accordingly to allow for non-linear interactions, testing the hypothesis that the PP acts as a moderator of the connectivity between *V*1 and *V*5. Assuming a non-linear modulation of this connection, we constructed a new variable "*V*1PP" in our analysis. This variable, mediating the interaction, is simply the time-series from region *V*1 multiplied (element by element) by the time-series of the right posterior parietal region.

The influence of this new variable on *V*5 corresponds to the influence of the posterior parietal cortex on the connection between *V*1 and *V*5 (i.e. the influence of *V*1 on *V*5 is greater when activity in PP is high). The model is shown in Fig. 6. Because our non-linear model could accommodate changes in connectivity between "attention" and "no attention", the entire time-series was analysed (i.e. attention-specific changes are now explicitly modelled by the interaction term).

As in the linear model, we tested for the significance of the interaction effect by comparing a restricted and free model. In the restricted model the interaction term (i.e. path from *V*1PP to *V*5) was set to zero. Omitting the interaction term led to a significantly reduced model fit ($p < 0.01$), indicating the predictive value of the interaction term.

The presence of an interaction effect of the PP on the connection between *V*1 and *V*5 can also be illustrated by a simple regression analysis. If PP shows a positive modulatory influence on the path between *V*1 and *V*5, the influence of *V*1 on *V*5 should depend on the activity of PP. This can be tested, by splitting the observations into two sets, one containing observations in which PP activity is high and another one in which PP activity is low. It is now possible to perform separate regressions of *V*5 on *V*1 using both sets. If the hypothesis of positive modulation is true, the slope of the regression of *V*5 on *V*1 should be steeper under high values of PP. This approach is comparable to the one outlined in the first section, where we used high and low values to demonstrate a modulatory effect of activity intrinsic to *V*1 on the influence *V*2 has over *V*1.

3.3. Variable parameter regression

As demonstrated in previous sections, the basic linear model can be seen as a linear regression. The regression coefficient is then interpreted as a measure of the connectivity between areas. This interpretation of course implies that the influence is mediated by neural connections with an effective strength equal to the regression coefficient. Using this approach one immediately makes the assumption that the effective connectivity does not change over observations, because only a single regression coefficient for the whole time-series is estimated. This is unsuitable for the assessment of effective connectivity in functional imaging, as the goal in some experiments is to demonstrate changes in effective connectivity, for instance as a function of different conditions (e.g. "attention" and "no attention") or simply time itself. In the framework of regression analysis there are three ways around this problem. Firstly, one could split up the data in different groups according to the experimental condition (e.g. "attention" and "no attention") and then test for the difference of the regression coefficients. However, we may not know a priori the time-course of the changes that allow us to split the data in this way. A second, more general solution is to expand the explanatory variable in terms of a set of basis functions to account for changes in connectivity. Here we will present another alternative, variable parameter regression (VPR), that allows one to characterise the variation of the regression coefficient using the framework of state-space models and the Kalman filter (Kalman, 1960).

3.3.1. Mathematical background

Consider the classical regression model

$$y = x\beta + u \tag{3}$$

where y is the measured data vector, x is a vector of explanatory variables and β is the unknown parameter. Usually β is estimated as

$$\hat{\beta} = pinv(x)y \tag{4}$$

However β can also be estimated recursively with the advantage that inversion of a smaller matrix is necessary. This approach is known as recursive least squares (Harvey, 1993). This basic model is now extended to allow β to evolve over time. Variable parameter regression assumes T ordered scalar observations $(y_1,\ldots,y_T)$ generated by the

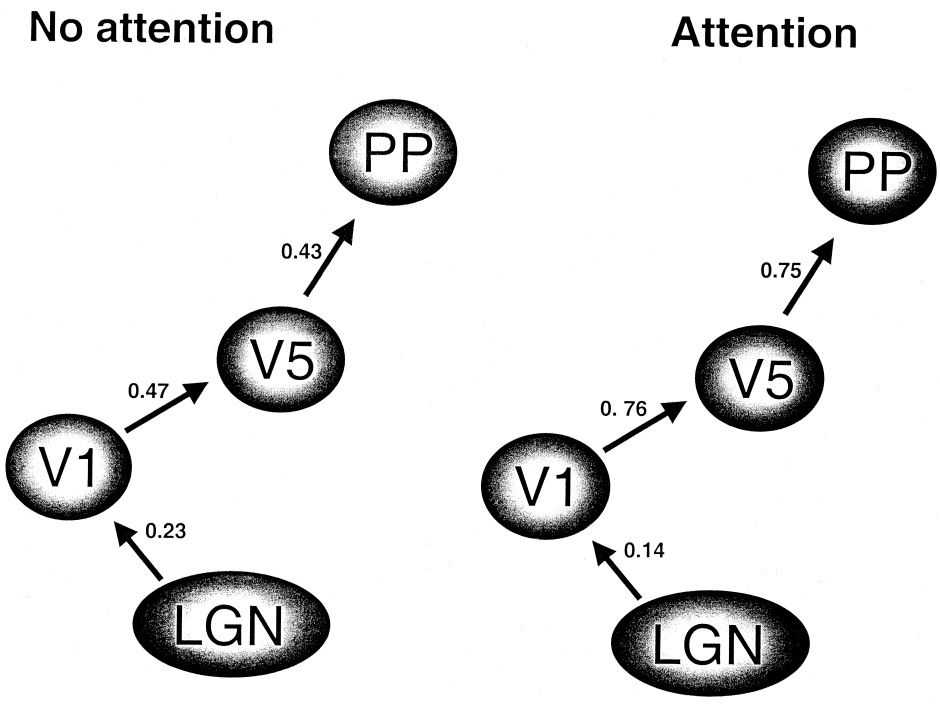

Fig. 7. (a), (b) The trajectory of the smoothed and filtered estimates $\hat{\beta}_t(T)$ together with the associated standard errors for the variable parameter estimation of effective connectivity between *V*5 and PP. It is evident that $\hat{\beta}_t$ (the dynamic regression coefficient) is higher during the "attention" conditions relative to the "no attention" conditions. (c) Areas that significantly covaried with the time-dependent measure of effective connectivity between *V*5 and PP [i.e. $\hat{\beta}_t(T)$]. SPM{Z} thresholded at $p < 0.001$ (uncorrected) overlaid on coronal and axial slices of the subject's structural MRI. The maximum under the cross-hairs was at 45, 21, 39 mm, $Z = 4$. (d) The relationship between our technique and an ordinary regression analysis. In this analysis the variance term *P* was set to zero (i.e. fixed regression model). The trajectory of $\hat{\beta}_t$ now converges to $\beta(= 0.73)$, the regression coefficient of the model $y = x\beta + u$.

model:

$$y_t = x_t\beta_t + u_t, \qquad t = 1,\ldots,T, \tag{5}$$

$$u_t \sim N(0), \sigma^2 \tag{6}$$

where x_t is an *n*-dimensional row vector of known regressors and β_t is an *n*-dimensional column vector of unknown coefficients that corresponds to estimates of effective connectivity. u_t is drawn from a Gaussian distribution. All observations are expressed as deviations from the mean.

A recursive algorithm known as the Kalman filter (Kalman, 1960) can now be applied to estimate the state-variable (β) at each point in time and also allows one to estimate the log-likelihood function of the model. A numerical optimisation algorithm is then employed to maximize the likelihood function with respect to *P*. As the Kalman filter is a recursive procedure, the estimation of β_t is based on all observations up to time *t*. Therefore, the filtered estimates will be more accurate towards the end of the sample. This fact is corrected for by the Kalman smoothing algorithm which is employed post hoc and runs backwards in time, taking account of the information made available after time *t*. Details of the Kalman filter and smoothing recursions can be found in standard textbooks of time-series analysis and econometrics (e.g. Chow, 1983; Harvey, 1990).

3.3.2. Example — attention to visual motion

To illustrate VPR we use the single-subject data set from the attention-to-visual-motion study. We concentrate on the effect of attention on the connection between the motion sensitive area *V*5 and the PP cortex in the right hemisphere. Using structural equation modelling, we have demonstrated that it is principally this connection, in the dorsal visual stream, that is modulated by attention (Büchel & Friston, 1997). In the current analysis we were interested whether variable parameter regression was capable of reproducing these findings. We therefore assessed the effective connectivity β_t by regressing PP on *V*5. An alternate direction search, numerical optimisation gave a chi-squared statistic of 56.4. We therefore had to reject the null-hypothesis of no variation at the 5% level. *P* was estimated to be 0.074 and σ^2 was 0.23. The ordinary regression coefficient β for the model $y = x\beta + u$ was estimated at 0.73. Fig. 7a and b shows the trajectories of the smoothed and filtered estimates $\hat{\beta}_t(T)$ together with the associated standard errors. It is clearly evident that $\hat{\beta}_t$ is higher during the "attention" conditions relative to the "no attention" conditions. Fig. 7d relates our technique to an ordinary regression. In this analysis we constrained the variance term *P* to zero and re-estimated β_t. The trajectory of $\hat{\beta}_t$ now converges to β, the ordinary regression coefficient of the model $y = x\beta + u$. As expected the smoothed estimates are simply a constant (i.e. $\beta = 0.73$).

We interpret $\hat{\beta}_t$ as an index of effective connectivity between area *V*5 and the posterior parietal cortex. In our example, the connection between *V*5 and PP resembles the *site* of attention modulation. This leads to an interesting extension, where one might hypothesise that a third region is responsible for the observed variation in effective connectivity indicated by the trajectory of $\hat{\beta}_t(T)$ In other words after specifying the *site* and nature of attentional modulation we now want to know the location of the *source*. We addressed this by using $\hat{\beta}_t(T)$ as an explanatory variable in an ordinary regression analysis to identify voxels that covaried with this measure of effective connectivity. Fig. 7c shows the result of this analysis. Among areas with statistically significant ($p < 0.001$, uncorrected) positive covariation was the dorsolateral prefrontal cortex and the anterior cingulate cortex. This result confirms the putative modulatory role of the dorsolateral prefrontal cortex in attention to visual motion, as suggested by previous analyses (Büchel & Friston, 1997).

3.4. Effective connectivity vs categorical comparisons

One obvious advantage of the assessment of effective connectivity is that it allows one to test hypotheses about the integration of cortical areas. For example, in the presence of modulation, the categorical comparison between "attention" and "no attention" might reveal prestriate, parietal and frontal activations. However, the only statement possible is that these areas show higher cortical activity during the "attention" condition as opposed to the "no attention" condition. The analysis of effective connectivity revealed two additional results. Firstly, attention affects the pathway from *V*1 to *V*5 and from *V*5 to PP. Secondly, the introduction of non-linear interaction terms allowed us to test a hypothesis about how these modulations are mediated. The latter analysis suggested that the posterior parietal cortex exerts a modulatory influence on area *V*5.

The measurements used in all examples in this work were *hemodynamic* in nature. This limits an interpretation at the level of *neuronal* interactions. However, the analogy between the form of the non-linear interactions described above and voltage-dependent (i.e. modulatory) connections is a strong one. It is possible that the modulatory impact of *V*2 on *V*1 (and of PP on *V*5) is mediated by predominantly voltage-dependent connections. The presence of horizontal voltage-dependent connections within *V*1 has been established in cat striate cortex (Hirsch & Gilbert, 1991). We know of no direct electrophysiological evidence to suggest that extrinsic backward *V*2 to *V*1 connections are voltage-dependent; however our results are consistent with this. An alternative explanation for modulatory effects, which does not necessarily involve voltage-dependent connections, can be found in the work of Aertsen and Preissl (1991). These authors show that effective connectivity varies strongly with, or is modulated by, background neuronal activity. The mechanism relates to the efficacy of subthreshold EPSPs in establishing dynamic interactions. This efficacy is a function of post-synaptic depolarisation, which in turn depends on the tonic background of activity.

4. Conclusions

This work has reviewed the basic concepts of effective connectivity in neuroimaging. We have introduced several methods to assess effective connectivity, i.e. multiple linear regression, covariance structural equation modelling and variable parameter regression. The first example demonstrated that non-linear interactions can be characterised using simple extensions of linear models. In the second example structural equation modelling was introduced as a device that allows one to combine observed changes in cortical activity and anatomical models. The first example of this technique revealed changes in effective connectivity between the dorsal and the ventral stream over time in a paired-associates learning paradigm. The temporal pattern of these changes was highly correlated with individual learning performance and therefore changes in effective connectivity predicted learning speed. The second example of structural equation modelling focused on backwards modulatory influences of high order areas on connections among lower order areas. Both examples concentrated on changes in effective connectivity and allowed us to characterise the interacting areas of the network at a functional level. Variable parameter regression was then introduced as a flexible regression technique, allowing the regression coefficient to smoothly vary over time. Again we confirmed the backwards modulatory effect of higher cortical areas on those areas situated lower in the cortical hierarchy. Although less than a mature field, the approach to neuroimaging data, and regional interactions, discussed above is an exciting endeavour that is starting to attract more and more attention.

References

Aertsen, A., & Preissl, H. (1991). *Dynamics of activity and connectivity in physiological neuronal networks*, New York: VCH.

Assad, J. A., & Maunsell, J. H. (1995). Neuronal correlates of inferred motion in primate posterior parietal cortex. *Nature*, *373*, 518–521.

Atkinson, R. C., & Shiffrin, R. M. (1968). In K. W. Spence & J. T. Spence, *Human memory: a proposed system and its control processes. The psychology of learning and motivation: advances in research and theory*, vol. 2. New York: Academic Press.

Büchel, C., Coull, J. T., & Friston, K. J. (1999). The predictive value of changes in effective connectivity for human learning. *Science*, *283*, 1538–1541.

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.

Chow, G. C. (1983). *Econometrics*, New York: McGraw Hill.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *PNAS*, *93*, 13494–13499.

Friston, K. J., Frith, C. D., & Frackowiak, R. S. J. (1993). Time-dependent

changes in effective connectivity measured with PET. *Human Brain Mapping, 1*, 69–80.

Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. J. (1993). Functional connectivity: the principal component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow and Metabolism, 13*, 5–14.

Friston, K. J., Ungerleider, L. G., Jezzard, P., & Turner, R. (1995). Characterizing modulatory interactions between V1 and V2 in human cortex with fMRI. *Human Brain Mapping, 2*, 211–224.

Gerstein, G. L., Bedenbaugh, P., & Aertsen, A. (1989). Neuronal assemblies. *IEEE Transactions on Biomedical Engineering, 36*, 4–14.

Gerstein, G. L., & Perkel, D. H. (1969). Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science, 164*, 828–830.

Girard, P., & Bullier, J. (1988). Visual activity in area V2 during reversible inactivation of area 17 in the Macaque monkey. *Journal of Neurophysiology, 62*, 1287–1301.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*, Cambridge, MA: Cambridge University Press.

Harvey, A. C. (1993). *Time series models*, London: Harvester & Wheatsheaf.

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

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering D, 82*, 35–45.

Kenny, D. A., & Judd, C. M. (1984). Estimating nonlinear and interactive effects of latent variables. *Psychology Bulletin, 96*, 201–210.

Lashley, K. S. (1929). *Brain mechanisms and intelligence*, Chicago, IL: University of Chicago Press.

Lawler, K. A., & Cowey, A. (1987). On the role of posterior parietal and prefrontal cortex in visuo-spatial perception and attention. *Experimental Brain Research, 65*, 695–698.

Marshall, J. C., & Newcombe, F. (1973). Patterns of Paralexia: a neurolinguistic approach. *Journal of Psycholinguistic Research, 2*, 175–199.

McIntosh, A. R., & Gonzalez-Lima, F. (1994). Structural equation modelling and its application to network analysis in functional brain imaging. *Human Brain Mapping, 2*, 2–22.

McIntosh, A. R., Grady, C. L., Ungerleider, L. G., Haxby, J. V., Rapoport, S. I., & Horwitz, B. (1994). Network analysis of cortical visual pathways mapped with PET. *Journal of Neuroscience, 14*, 655–666.

Mesulam, M. M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology, 28*, 597–613.

Milner, B., Johnsrude, I., & Crane, J. (1997). Right medial temporal-lobe contribution to object-location memory. *Philosophical Transactions of the Royal Society of London, Series B, 352*, 1469–1474.

O'Craven, K. M., & Savoy, R. L. (1995). Voluntary attention can modulate fMRI activity in human MT/MST. *Investigational Ophthalmological Vision Science, 36*, S856 (suppl.).

Petersen, S. E., Fox, P. T., Snyder, A. Z., & Raichle, M. E. (1990). Activation of extrastriate and frontal cortical areas by words and word-like stimuli. *Science, 249*, 1041–1044.

Posner, M. I., Petersen, S. E., Fox, P. T., & Raichle, M. E. (1988). Localization of cognitive operations in the human brain. *Science, 240*, 1627–1631.

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

Schiller, P. H., & Malpeli, J. G. (1977). The effect of striate cortex cooling on area 18 cells in the monkey. *Brain Research, 126*, 366–369.

Shallice, T. (1988). *From neuropsychology to mental structure*, Cambridge, MA: Cambridge University Press.

Talairach, P., & Tournoux, J. (1988). *A Stereotactic coplanar atlas of the human brain*, Stuttgart: Thieme.

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., & Shallice, T. (1969). The selective impairment of auditory short term memory. *Brain, 92*, 885–896.

Wiggs, C. L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinions in Neurobiology, 8*, 227–233.