Modeling brain responses.

Karl J. Friston, William Penny, Olivier David

Modelling brain responses

Karl J Friston, William Penny and Olivier David

The Wellcome Dept. of Cognitive Neurology,
University College London
Queen Square, London, UK WC1N 3BG
Tel (44) 020 7833 7456
Fax (44) 020 7813 1445
Email k.friston@fil.ion.ucl.ac.uk

Short title: Modelling brain responses.

Keywords: Models, Dynamic, Inference, Causal, and Neuroimaging
Word Count: 12,641
Figures: 9

ABSTRACT

Inferences about brain function, using neuroimaging data, rest on models of how the data were caused. These models can be quite diverse, ranging from conceptual models of functional anatomy to nonlinear mathematical models of neuronal and hemodynamics. The aim of this review is to introduce the key models used in imaging neuroscience and how they relate to each other. We start with anatomical models of functional brain architectures, which motivate some of the fundaments of neuroimaging. We then turn to basic statistical models (*e.g.* the general linear model) used for making classical and Bayesian inferences about *where* neuronal responses are expressed. By incorporating biophysical constraints, these basic models can be finessed and, in a dynamic setting, rendered causal. This allows us to infer *how* interactions among brain regions are mediated.

We will cover models of brain responses, starting with general linear models of functional magnetic resonance imaging (fMRI) data, used for classical inference about regionally specific responses. This model is successively refined until we arrive at neuronal mass models of electroencephalographic (EEG) responses. The latter models afford mechanistic inferences about how evoked responses are caused, at the level of neuronal subpopulations and the coupling among them.

I INTRODUCTION

Neuroscience depends on conceptual, anatomical, statistical and causal models that link ideas about how the brain works to observed neuronal responses. The aim of this review is to highlight the relationships among the sorts of models that are employed in imaging. We will show how simple statistical models, used to identify *where* evoked brain responses are expressed (*c.f.* neo-phrenology) can be elaborated to provide models of *how* neuronal responses are caused (*e.g.* dynamic causal modelling). These successive elaborations rely, increasingly, on biological mechanisms. We will review a series of models that cover conceptual models, motivating experimental design, to detailed biophysical models of coupled neuronal ensembles that enable questions to be asked, at a physiological and computational level.

Anatomical models of functional brain architectures motivate the fundaments of neuroimaging. In Section II we start by reviewing the distinction between functional *specialisation* and *integration* and how these principles serve as the basis for most models of neuroimaging data. In section III, we turn to simple statistical models (*e.g.* the general linear model) used for making classical and Bayesian inferences about functional specialisation, in terms of *where* neuronal responses are expressed. Characterising a region-specific effect rests on estimation and inference. Inferences in neuroimaging may be about differences seen when comparing one group of subjects to another or, within subjects, changes over a sequence of observations. They may pertain to structural differences (*e.g.* in voxel-based morphometry - Ashburner and Friston 2000) or neurophysiological indices of brain functions (*e.g.* fMRI or EEG). The principles of data analysis are very similar for all these applications. We will focus on the analysis of fMRI time-series, because this covers most of the issues encountered in other modalities. By incorporating biological constraints, simple observation models can be made more realistic and, in a dynamic framework, causal. This section concludes by considering some of the recent advances in biophysical modelling of hemodynamic responses. All the models considered in this section pertain to regional responses. In the final section, we focus on models of distributed responses, where the interactions among cortical areas or neuronal subpopulations are modelled explicitly. This section covers the distinction between *functional and effective connectivity* and reviews dynamic causal modelling of functional integration, using fMRI and EEG. We conclude with an example from ERP (event-related potential) research and show how the P300 can be explained by changes in coupling among neuronal sources that may underlie perceptual learning.

II ANATOMIC MODELS

1. Functional specialisation and integration

The brain appears to adhere to two key principles of functional organisation, *functional specialisation* and *functional integration,* 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 nineteenth 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 *et al* 1984). This meeting was entitled "Localisation of function in the cortex cerebri". Goltz (1881) although accepting the results of electrical stimulation in dog and monkey cortex, considered that the excitation method was inconclusive, in that movements 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 observations on patients with brain lesions some years later (see Absher and Benson 1993) that led to the concept of *disconnection syndromes* and the refutation of localisationism as a complete or sufficient explanation of cortical organisation. 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, and that this specialisation is 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. In this view functional specialisation is only meaningful in the context of functional integration and *vice versa*.

2. Functional specialisation and segregation

The functional role of any component (*e.g.* cortical area, subarea or neuronal population) 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. This architectural constraint necessitates both convergence and divergence of cortical connections. Extrinsic connections among 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, 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 notion that patchy connectivity is the anatomical infrastructure that mediates functional segregation and specialisation. If it is the case that neurons in a given cortical area share a common responsiveness, by virtue of their extrinsic connectivity, to some sensorimotor or cognitive attribute, then this functional segregation is also an anatomical one.

In summary, functional specialisation suggests that challenging a subject with the appropriate sensorimotor attribute or cognitive process should lead to activity changes in, and only in, the specialised areas. This is the anatomical and physiological model upon which the search for regionally specific effects is based. We will deal first with models of regionally specific responses and return to models of functional integration later.

III STATISTICAL MODELS OF REGIONAL RESPONSES

1. Statistical parametric mapping

Functional mapping studies are usually analysed with some form of statistical parametric mapping. Statistical parametric mapping entails the construction of spatially extended statistical processes to test hypotheses about regionally specific effects (Friston *et al* 1991). Statistical parametric maps (SPMs) are image processes with voxel values that are, under the null hypothesis, distributed according to a known probability density function, usually the Student's T or F distributions. These are known colloquially as T- or F-maps. The success of statistical parametric mapping is due largely to the simplicity of the idea. Namely, one analyses each and every voxel using any standard (univariate) statistical test. These usually test for activation, or regression on some explanatory variable. The resulting statistical parameters are assembled into an image - the SPM. SPMs are interpreted as statistical processes that are continuous in space (or sometimes time) by referring to the probabilistic behaviour of random fields (Adler 1981, Friston *et al* 1991, Worsley *et al* 1992, Worsley *et al* 1996). Random fields model both the univariate probabilistic characteristics of a SPM and any non-stationary spatial covariance structure under the null hypothesis. 'Unlikely' excursions of the SPM are interpreted as regionally specific effects, attributable to the sensorimotor or cognitive process that has been manipulated experimentally.

Over the years statistical parametric mapping (Friston *et al* 1995a) has come to refer to the conjoint use of *the general linear model* (GLM) and r*andom field* theory (RFT) to analyse and make classical inferences about spatially extended data through statistical parametric maps. The GLM is used to *estimate* some parameters that could explain the spatially continuous data in exactly the same way as in conventional analysis of discrete data. RFT is used to resolve the multiple-comparisons problem that ensues when making *inferences* over a volume of the brain. RFT provides a method for adjusting *p* values for the search volume of an SPM to control false positive rates. It plays the same role for *continuous* data (*i.e.* images or time-series) as the Bonferonni correction for a family of discontinuous or *discrete* statistical tests.

Later we will consider the Bayesian alternative to classical inference with SPMs. This rests on conditional inferences about an effect, given the data, as opposed to classical inferences about the data, given the effect is zero. Bayesian inferences about effects that are continuous in space use Posterior Probability Maps (PPMs). Although less established than SPMs, PPMs are potentially very useful, not least because they do not have to contend with the multiple-comparisons problem induced by classical inference (see Berry and Hochberg 1999). In contradistinction to SPM, this means that inferences about a given regional response do not depend on inferences about responses elsewhere. Before looking at the models underlying Bayesian inference we first consider estimation and classical inference in the context of the GLM.

2. The general linear model

Statistical analysis of imaging data corresponds to (i) modelling the data to partition observed neurophysiological responses into components of interest, confounds and error and (ii) making inferences, about interesting effects, using the variances of the partitions. A brief review of the literature may give the impression that there are numerous ways to analyse PET and fMRI time-series, with a diversity of statistical and

conceptual approaches. This is not the case. With very few exceptions, every analysis is a variant of the general linear model. These include; (i) Simple T-tests on scans assigned to one condition or another. (ii) Correlation coefficients between observed responses and boxcar stimulus functions in fMRI (Bandettini *et al* 1993). (iii) Inferences made using multiple linear regression. (iv) Evoked responses estimated using linear time invariant models and (v) selective averaging to estimate event-related responses. Mathematically, they are all identical and can be implemented with the same equations and algorithms. The only thing that distinguishes among them is the design matrix encoding the experimental design.

The general linear model is an equation

$$y = X\beta + \varepsilon \qquad 1$$

expressing the observed response y in terms of a linear combination of explanatory variables in the matrix X plus a well-behaved error term. The general linear model is variously known as 'analysis of [co]variance' or 'multiple regression ' and subsumes simpler variants, like the 'T-test' for a difference in means, to more elaborate linear convolution models such as finite impulse response (FIR) models. The matrix X that contains the explanatory variables (*e.g.* designed effects or confounds) is called the *design matrix*. Each column of the design matrix corresponds to some effect one has built into the experiment or that may confound the results. These are referred to as explanatory variables, covariates or regressors. Sometimes the design matrix contains covariates or indicator variables that take values of 0 or 1, to indicate the presence of a particular level of an experimental factor (*c.f.* analysis of variance - ANOVA). The example in Figure 1 relates to a fMRI study of visual stimulation under four conditions. The effects on the response variable are modelled in terms of functions of the presence of these conditions (*i.e.* box or stick functions smoothed with components of a hemodynamic response function). Note that this is more complicated than a simple ANOVA, because the design matrix is modelling a time-series, as opposed to discrete observations (see below). The relative contribution of each of these columns to the response is controlled by the parameters β. These are estimated using standard least squares. Inferences about the parameter estimates are made using T or F statistics, depending upon whether one is looking at one, or more, linear combinations of them.

In simple analyses the design matrix contains indicator variables or parametric variables encoding the experimental manipulations. These are formally identical to classical ANOVA or multiple linear regression models. However, when the observations correspond to time-series, convolution models are often used: An important instance of the GLM, from the perspective of fMRI, is the linear time invariant (LTI) convolution model. Mathematically this is no different from any other GLM. However, it explicitly treats the data-sequence as an ordered time-series and enables a signal processing perspective that can be useful.

Each column of X has an associated but unknown parameter. Some of these parameters will be of interest (*e.g.* the effect of a sensorimotor or cognitive condition or the regression coefficient of hemodynamic responses on reaction time). The remaining parameters will be of no interest and pertain to nuisance or confounding effects (*e.g.* the effect of being a particular subject or the regression slope of regional activity on global activity). The statistical test is directed to interesting effects by specifying the hull hypothesis with a *contrast*. A contrast is simply a linear mixture of parameter estimates. The T statistic allows one to test the null hypothesis that some contrast (*e.g.* a subtraction) of the estimates is zero. The T statistic obtains by dividing the

contrast (specified by contrast weights) of the parameter estimates, by its standard error. Sometimes, several contrasts are tested jointly. For example, when using polynomial (Büchel *et al* 1996) or basis function expansions of some experimental factor. In these instances, the SPM{F} is used and is specified with a *matrix* of contrast weights that can be thought of as a collection of 'T contrasts' that one wants to test *en masse*.

Having computed the statistic, RFT is used to assign adjusted *p*-values to topological features of the SPM, such as the height of peaks or the spatial extent of regions above a threshold. This *p*-value is a function of the search volume and smoothness of the residuals (see Figure 1). The intuition behind RFT is that it allows one to control the false positive rate of peaks or 'blobs' corresponding to regional effects. A Bonferonni correction would control the false positive rate of voxels, which is inexact and unnecessarily severe. The *p*-value is the probability of getting a peak in the SPM, or higher, by chance over the search volume. If sufficiently small (usually less than 0.05) the regional effect can be declared significant.

The equations summarised in Figure 1 can be used to implement a vast range of statistical analyses. The issue is therefore not so much the mathematics but the formulation of a design matrix X appropriate to the study design and inferences that are sought. Before considering general linear models as biophysical or causal models of brain responses we will focus on the design matrix as a device to specify experimental design. Here the explanatory variables encode treatment effects that we assume are expressed in a linear and instantaneous fashion in the data, without reference to any particular mechanism.

Figure 1 about here

3. Experimental design

This section considers the different sorts of designs employed in neuroimaging studies. Experimental designs can be classified as *single factor* or *multifactorial* designs, within this classification the levels of each factor can be *categorical* or *parametric*.

3.1 Categorical designs, cognitive subtraction and conjunctions

The tenet of cognitive subtraction is that the difference between two tasks can be formulated as a separable cognitive or sensorimotor component. Regionally specific differences, in the responses evoked by the two tasks, identify the corresponding functionally specialised area. Early applications of subtraction range from the functional anatomy of word processing (Petersen *et al* 1989) to functional specialisation in extrastriate cortex (Lueck *et al* 1989). The latter studies involved presenting visual stimuli with and without some sensory attribute (*e.g.* colour, motion *etc*). The areas highlighted by subtraction were identified with homologous areas in monkeys that showed selective electrophysiological responses to equivalent visual stimuli.

Cognitive conjunctions (Price and Friston 1997) can be thought of as an extension of the subtraction technique, in the sense that they combine a series of subtractions. In subtraction ones tests a *single* hypothesis pertaining to the activation in one task relative to another. In conjunction analyses *several* contrasts are tested, asking whether all the activations, in a series of task pairs, are expressed conjointly. Consider the problem of identifying regionally specific activations due to a particular cognitive component (*e.g.* object recognition). If one can identify a series of task pairs whose differences have only that component in common, then the region

which activates, in all the corresponding subtractions, can be associated with the common component. In short, conjunction analyses allow one to disclose context-invariant regional responses.

3.2 Parametric designs

The premise behind parametric designs is that regional physiology will vary systematically with the degree of cognitive or sensorimotor processing or deficits thereof. Examples of this approach include the PET experiments of Grafton *et al* (1992) that demonstrated significant correlations between hemodynamic responses and the performance of a visually guided motor tracking task. In sensory studies Price *et al* (1992) demonstrated a remarkable linear relationship between perfusion in peri-auditory regions and frequency of aural word presentation. This correlation was not observed in Wernicke's area, where perfusion appeared to correlate, not with the discriminative attributes of the stimulus, but with the presence or absence of semantic content. These relationships or *neurometric functions* may be linear or nonlinear. Using polynomial regression, in the context of the GLM, one can identify nonlinear relationships between stimulus parameters (*e.g.* stimulus duration or presentation rate) and evoked responses. To do this one usually uses a SPM{F} (see Büchel *et al* 1996).

The example provided in Figure 2 illustrates both categorical and parametric aspects of design and analysis. These data were obtained from an fMRI study of visual motion processing using radially moving dots. The stimuli were presented over a range of speeds using *isoluminant* and *isochromatic* stimuli. To identify areas involved in visual motion a stationary dots condition was subtracted from a moving dots conditions (see the contrast weights on the upper right). To ensure significant motion-sensitive responses, using both colour and luminance cues, a conjunction of the equivalent subtractions was assessed under both viewing contexts. Areas V5 and V3a are seen in the ensuing SPM{T}. The T values in this SPM are simply the minimum of the T values for each subtraction. Thresholding this SPM ensures that all voxels survive a threshold in each subtraction separately. This *conjunction* SPM has an equivalent interpretation; it represents the intersection of the excursion sets, defined by the threshold of each *component* SPM. This intersection is the essence of a conjunction.

The responses in left V5 are shown in the lower panel of Figure 2 and speak to a compelling inverted 'U' relationship between speed and evoked response that peaks at around six degrees per second. It is this sort of relationship that parametric designs try to characterise. Interestingly, the form of these speed-dependent responses was similar using both stimulus types, although luminance cues are seen to elicit a greater response. From the point of view of a factorial design there is a *main effect* of cue (isoluminant vs. isochromatic), a main [nonlinear] effect of speed, but no speed by cue *interaction*.

Figure 2 about here

3.3 Multifactorial designs

Factorial designs are more prevalent than single factor designs because they enable inferences about interactions. At its simplest, an interaction represents a change in a change. Interactions are associated with factorial designs where two or more factors are combined in the same experiment. The effect of one factor, on the effect of the other, is assessed by the interaction term. Factorial designs have a wide range of applications.

An early application, in neuroimaging, examined physiological adaptation and plasticity during motor performance, by assessing time by condition interactions (Friston *et al* 1992). Factorial designs have an important role in the context of cognitive subtraction and additive factor logic by virtue of being able to test for interactions, or context-sensitive activations (*i.e.* to demonstrate the fallacy of 'pure insertion'. See Friston *et al* 1996a). These interaction effects can sometimes be interpreted as (i) the integration of the two or more [cognitive] processes or (ii) the modulation of one [perceptual] process by another.

In summary, the design matrix encodes the causes of observed data and, in particular, treatment effects caused by changes in the level of various experimental factors. These factors can have categorical or parametric levels and most experiments nowadays use multiple factors to test for both main effects and interactions. Before turning to mechanistically more informed formulations of the general linear model we will consider briefly the two sorts of inferences that can be made about the parameter estimates.

4. Classical and Bayesian inference

To date, inference in neuroimaging has been restricted largely to classical inferences based upon statistical parametric maps. The statistics that comprise these SPMs are essentially functions of the data. The probability distribution of the chosen statistic, under the null hypothesis (*i.e.* the null distribution) is used to compute a *p*-value. This *p*-value is the probability of obtaining the statistic, or the data, given that the null hypothesis is true. If sufficiently small, the null hypothesis can be rejected and an inference is made. The alternative approach is to use Bayesian or conditional inference based upon the posterior distribution of the activation given the data (Holmes & Ford 1993). This necessitates the specification of priors (*i.e.* the probability distribution of the activation). Bayesian inference requires the posterior distribution and therefore rests upon a posterior density analysis. A useful way to summarise this posterior density is to compute the probability that the activation exceeds some threshold. This represents a Bayesian inference about the effect, in relation to the specified threshold. By computing posterior probability for each voxel we can construct posterior probability maps or PPMs that are a useful complement to classical SPMs.

The motivation for using conditional or Bayesian inference is that it has high face validity. This is because the inference is about an effect, or activation, being greater than some specified size that has some meaning in relation to underlying neurophysiology. This contrasts with classical inference, in which the inference is about the effect being significantly different than zero. The problem for classical inference is that trivial departures from the null hypothesis can be declared significant, with sufficient data or sensitivity. From the point of view of neuroimaging, posterior inference is especially useful because it eschews the multiple-comparisons problem. In classical inference, one tries to ensure that the probability of rejecting the null hypothesis incorrectly is maintained at a small rate, despite making inferences over large volumes of the brain. This induces a multiple-comparisons problem that, for spatially continuous data, requires an adjustment or correction to the *p*-value using RFT as mentioned above. This correction means that classical inference becomes less sensitive or powerful with large search volumes. In contradistinction, posterior inference does not have to contend with the multiple-comparisons problem because there are no false-positives. The probability that activation has occurred, given the data, at any particular voxel is the same, irrespective of whether one has analysed that voxel

or the entire brain. For this reason, posterior inference using PPMs represents a relatively more powerful approach than classical inference in neuroimaging.

4.1 Hierarchical models and empirical Bayes

PPMs require the posterior distribution or conditional distribution of the activation (a contrast of conditional parameter estimates) given the data. This posterior density can be computed, under Gaussian assumptions, using Bayes rule. Bayes rule requires the specification of a likelihood function and the prior density of the model's parameters. The models used to form PPMs and the likelihood functions are exactly the same as in classical SPM analyses, namely the GLM. The only extra bit of information that is required is the prior probability distribution of the parameters. Although it would be possible to specify them using independent data or some plausible physiological constraints, there is an alternative to this fully Bayesian approach. The alternative is *empirical Bayes* in which the prior distributions are estimated from the data. Empirical Bayes requires a *hierarchical observation model* where the parameters and hyper-parameters at any particular level can be treated as priors on the level below. There are numerous examples of hierarchical observation models in neuroimaging. For example, the distinction between fixed- and mixed-effects analyses of multi-subject studies relies upon a two-level hierarchical model. However, in neuroimaging there is a natural hierarchical observation model that is common to all brain mapping experiments. This is the hierarchy induced by looking for the same effects at every voxel within the brain (or grey matter). The first level of the hierarchy corresponds to the experimental effects at any particular voxel and the second level comprises the effects over voxels. Put simply, the variation in a contrast, over voxels, can be used as the prior variance of that contrast at any particular voxel. Hierarchical linear models have the following form

$$\begin{aligned} y &= X^{(1)}\beta^{(1)} + \varepsilon^{(1)} \\ \beta^{(1)} &= X^{(2)}\beta^{(2)} + \varepsilon^{(2)} \\ \beta^{(2)} &= \ldots \end{aligned} \qquad 2$$

This is exactly the same as Eq(1) but now the parameters of the first level are generated by a supraordinate linear model and so on to any hierarchical depth required. These hierarchical observation models are an important extension of the GLM and are usually estimated using Expectation Maximisation (EM) (Dempster *et al* 1977). In the present context, the response variables comprise the responses at all voxels and $\beta^{(1)}$ are the treatment effects we want to make an inference about. Because we have invoked a second level the first-level parameters embody random effects and are generated by a second level linear model. At the second level $\beta^{(2)}$ is the average effect over voxels and $\varepsilon^{(2)}$ its voxel-to-voxel variation. By estimating the variance of $\varepsilon^{(2)}$ one is implicitly estimating an empirical prior on the first-level parameters at each voxel. This prior can then be used to estimate the posterior probability of $\beta^{(1)}$ being greater than some threshold at each voxel. An example of the ensuing PPM is provided in Figure 3 along with the classical SPM.

Figure 3 about here

In this section we have seen how the GLM can be used to test hypotheses about brain responses and how, in a hierarchical form, it enables empirical Bayesian or conditional inference. In the next section we will deal with dynamic systems and how they can be formulated as GLMs. These dynamic models take us closer to how brain responses are actually caused by experimental manipulations and represent the next step in working toward causal models of brain responses.

5. Dynamic models

5.1 Convolution models and temporal basis functions

In Friston *et al* (1994) the form of the hemodynamic impulse response function (HRF) was estimated using a least squares de-convolution and a time invariant model, where evoked neuronal responses are *convolved* or smoothed with an HRF to give the measured hemodynamic response (see also Boynton *et al* 1996). This simple linear convolution model is the cornerstone for making statistical inferences about activations in fMRI with the GLM. An impulse response function is the response to a single impulse, measured at a series of times after the input. It characterises the input-output behaviour of the system (*i.e.* voxel) and places important constraints on the sorts of inputs that will excite a response.

Knowing the forms that the HRF can take is important for several reasons, not least because it allows for better statistical models of the data. The HRF may vary from voxel to voxel and this has to be accommodated in the GLM. To allow for different HRFs in different brain regions the notion of temporal basis functions, to model evoked responses in fMRI, was introduced (Friston *et al* 1995b) and applied to event-related responses in Josephs *et al* (1997) (see also Lange and Zeger 1997). The basic idea behind temporal basis functions is that the hemodynamic response, induced by any given trial type, can be expressed as the linear combination of several [basis] functions of peristimulus time. The convolution model for fMRI responses takes a stimulus function encoding the neuronal responses and convolves it with an HRF to give a regressor that enters the design matrix. When using basis functions, the stimulus function is convolved with all the basis functions to give a series of regressors (in Figure 1 we used four stimulus functions and two basis functions to give eight regressors). Mathematically we can express this model as

$$\begin{aligned} y(t) &= X\beta + \varepsilon \\ X_i &= T_i(t) \otimes u(t) \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} y(t) &= u(t) \otimes h(t) \\ h(t) &= \beta_1 T_1(t) + \beta_2 T_2(t) + \ldots \end{aligned} \qquad 3$$

where $\otimes$ means convolution. This equivalence illustrates how temporal basis functions allow one to take any convolution model (right) and convert it into a GLM (left). The parameter estimates β_i are the coefficients or weights that determine the mixture of basis functions of time $T_i(t)$ that best models $h(t)$, the HRF for the trial type and voxel in question. We find the most useful basis set to be a canonical HRF and its derivatives with respect to the key parameters that determine its form (see below). Temporal basis functions are important

because they enable a graceful transition between conventional multi-linear regression models with one stimulus function per condition and FIR models with a parameter for each time point following the onset of a condition or trial type. Figure 4 illustrates this graphically (see Figure legend). In short, temporal basis functions offer useful constraints on the form of the estimated response that retain the flexibility of FIR models and the efficiency of single regressor models.

Figure 4 about here

6. Biophysical models

6.1 Input-state-output systems

By adopting a convolution model for brain responses in fMRI we are implicitly positing some underlying dynamic system that converts neuronal responses into observed hemodynamic responses. Our understanding of the biophysical and physiological mechanisms that underpin the HRF has grown considerably in the past few years (*e.g.* Buxton and Frank 1997, Mandeville *et al* 1999, Hoge *et al* 1999). Figure 5 shows some simulations based on the hemodynamic model described in Friston *et al* (2000). Here, neuronal activity induces some auto-regulated vasoactive signal that causes transient increases in regional cerebral blood flow (rCBF). The resulting flow increases dilate a venous balloon, increasing its volume (v) and diluting venous blood to decrease deoxyhemoglobin content (q). The blood oxygenation-level-dependent (BOLD) signal is roughly proportional to the concentration of deoxyhemoglobin (q/v) and follows the rCBF response with about a one second delay. The model is framed in terms of differential equations, examples of which are provided in Figure 5.

Figure 5 about here

Notice that we have introduced variables, like volume and deoxyhemoglobin concentrations that are not actually observed. These are referred to as the *hidden states* of input-state-output models. The state and output equations of any analytic dynamical system are

$$\begin{aligned} \dot{x}(t) &= f(x,u,\theta) \\ y(t) &= g(x,u,\theta) + \varepsilon \end{aligned} \qquad 4$$

The first line is an ordinary differential equation and expresses the rate of change of the states as a parameterised function of the states and inputs. Typically the inputs $u(t)$ correspond to designed experimental effects (*e.g.* the stimulus function in fMRI). There is a fundamental and causal relationship (Fliess *et al* 1983) between the outputs and the history of the inputs in Eq(4). This relationship conforms to a Volterra series, which expresses the output $y(t)$ as a generalised convolution of the input $u(t)$, critically without reference to the

hidden states $x(t)$. This series is simply a functional Taylor expansion of the outputs with respect to the inputs (Bendat 1990). The reason it is a *functional* expansion is that the inputs are a function of time[1].

$$y(t) = \sum_i \int_0^t \cdots \int_0^t \kappa_i(\sigma_1, \ldots, \sigma_i) u(t-\sigma_1), \ldots, u(t-\sigma_i) d\sigma_1, \ldots, d\sigma_i$$
$$\kappa_i(\sigma_1, \ldots, \sigma_i) = \frac{\partial^i y(t)}{\partial u(t-\sigma_1), \ldots, \partial u(t-\sigma_i)} \qquad 5$$

were $\kappa_i(\sigma_1, \ldots \sigma_i)$ is the ith order kernel. In Eq (5) the integrals are restricted to the past. This renders Eq (5) causal. The key thing here is that Eq(5) is simply a convolution and can be expressed as a GLM as in Eq(3). This means that we can take a neurophysiologically realistic model of hemodynamic responses and use it as an observation model to estimate parameters using observed data. Here the model is parameterised in terms of kernels that have a direct analytic relation to the original parameters θ of the biophysical system (through equation 5). The first-order kernel is simply the conventional HRF. High-order kernels correspond to high-order HRFs and can be estimated using basis functions as described above. In fact, by choosing basis functions according to

$$A(\sigma)_i = \frac{\partial \kappa(\sigma)_1}{\partial \theta_i} \qquad 6$$

one can estimate the biophysical parameters because, to a first order approximation, $\beta_i = \theta_i$. The critical step we have taken here is to start with a causal dynamic model of how responses are generated and construct a general linear observation model that allows us to estimate and infer things about the parameters of that model. This is in contrast to the conventional use of the GLM with design matrices that are not informed by a forward model of how data are caused. This approach to modelling brain responses has a much more direct connection with underlying physiology and rests upon an understanding of the underlying system.

6.2 Nonlinear system identification

Once a suitable causal model has been established (*e.g.* Figure 5), we can estimate second-order kernels. These kernels represent a nonlinear characterisation of the HRF that can model interactions among stimuli in causing responses. One important manifestation of the nonlinear effects, captured by the second order kernels, is a modulation of stimulus-specific responses by preceding stimuli that are proximate in time. This means that responses at high stimulus presentation rates saturate and, in some instances, show an inverted U behaviour. This behaviour appears to be specific to BOLD effects (as distinct from evoked changes in cerebral blood flow) and may represent a *hemodynamic refractoriness*. This effect has important implications for event-related fMRI, where one may want to present trials in quick succession.

[1] For simplicity, here and in and Eq(7), we deal with only one experimental input.

The results of a typical nonlinear analysis are given in Figure 6. The results in the right panel represent the average response, integrated over a 32-second train of stimuli as a function of stimulus onset asynchrony (SOA). These responses are based on the kernel estimates (left hand panels) using data from a voxel in the left posterior temporal region of a subject obtained during the presentation of single words at different rates. The solid line represents the estimated response and shows a clear maximum at just less than one second. The dots are responses based on empirical data from the same experiment. The broken line shows the expected response in the absence of nonlinear effects (*i.e.* that predicted by setting the second order kernel to zero). It is clear that nonlinearities become important at around two seconds, leading to an actual diminution of the integrated response at sub-second SOAs. The implication of this sort of result is that the assumptions of the linear convolution models discussed above are violated with sub-second SOAs (see also Buckner *et al* 1996 and Burock *et al* 1998)

Figure 6 about here

In summary, we started with models of regionally specific responses, framed in terms of the general linear model, in which responses were modelled as linear mixtures of designed changes in explanatory variables. Hierarchical extensions to linear observation models enable random-effects analyses and, in particular an empirical Bayesian approach. The mechanistic utility of these models is realised though the use of forward models that embody causal dynamics. Simple variants of these are the linear convolution models used to construct explanatory variables in conventional analyses of fMRI data. These are a special case of generalised convolution models that are mathematically equivalent to input-state-output systems comprising hidden states. Estimation and inference with these dynamic models tells us something about *how* the response was caused, but only at the level of a single voxel. The next section retains the same perspective on models, but in the context of distributed responses and functional integration.

IV MODELS OF FUNCTIONAL INTEGRATION

1. Functional and Effective connectivity

Imaging neuroscience has firmly established functional specialisation as a principle of brain organisation in man. The integration of specialised areas has proven more difficult to assess. Functional integration is usually inferred on the basis of correlations among measurements of neuronal activity. Functional connectivity has been defined as statistical dependencies or correlations *among remote neurophysiological events*. However correlations can arise in a variety of ways: For example in multi-unit electrode recordings they can result from stimulus-locked transients evoked by a common input or reflect stimulus-induced oscillations mediated by synaptic connections (Gerstein and Perkel 1969). Integration within a distributed system is usually better understood in terms of effective connectivity: Effective connectivity refers explicitly to *the influence that one neural system exerts over another*, either at a synaptic (*i.e.* synaptic efficacy) or population level. It has been proposed that "the [electrophysiological] notion of effective connectivity should be understood as the experiment- and time-dependent, simplest possible circuit diagram that would replicate the observed timing

relationships between the recorded neurons" (Aertsen and Preißl 1991). This speaks to two important points: (i) Effective connectivity is dynamic, *i.e.* activity- and time-dependent and (ii) it depends upon a model of the interactions. The estimation procedures employed in functional neuroimaging can be divided into those based on (i) linear regression models (*e.g.* McIntosh and Gonzalez-Lima 1994) or (ii) nonlinear dynamic causal models.

There is a necessary link between functional integration and multivariate analyses because the latter are necessary to model interactions among brain regions. Multivariate approaches can be divided into those that are inferential in nature and those that are data-led or exploratory. We will first consider multivariate approaches that are generally based on functional connectivity or covariance patterns (and are generally exploratory) and then turn to models of effective connectivity (that allow for some form of inference).

1.1 Eigenimage analysis and related approaches

In Friston *et al* (1993) we introduced voxel-based principal component analysis (PCA) of neuroimaging time-series to characterise distributed brain systems implicated in sensorimotor, perceptual or cognitive processes. These distributed systems are identified with principal components or *eigenimages* that correspond to spatial modes of coherent brain activity. This approach represents one of the simplest multivariate characterisations of functional neuroimaging time-series and falls into the class of exploratory analyses. Principal component or eigenimage analysis generally uses singular value decomposition (SVD) to identify a set of orthogonal spatial modes that capture the greatest amount of variance expressed over time. As such the ensuing modes embody the most prominent aspects of the variance-covariance structure of a given time-series. Noting that covariance among brain regions is equivalent to functional connectivity renders eigenimage analysis particularly interesting because it was among the first ways of addressing functional integration (*i.e.* connectivity) with neuroimaging data. Subsequently, eigenimage analysis has been elaborated in a number of ways. Notable among these is canonical variate analysis (CVA) and multidimensional scaling (Friston et al 1996b, 1996c). Canonical variate analysis was introduced in the context of ManCova (multiple analysis of covariance) and uses the generalised eigenvector solution to maximise the variance that can be explained by some explanatory variables relative to error. CVA can be thought of as an extension of eigenimage analysis that refers explicitly to some explanatory variables and allows for statistical inference.

In fMRI, eigenimage analysis (*e.g.* Sychra *et al* 1994) is generally used as an exploratory device to characterise coherent brain activity. These variance components may, or may not be, related to experimental design. For example, endogenous coherent dynamics have been observed in the motor system at very low frequencies (Biswal *et al* 1995). Despite its exploratory power, eigenimage analysis is fundamentally limited for two reasons. Firstly, it offers only a linear decomposition of any set of neurophysiological measurements and second, the particular set of eigenimages or spatial modes obtained is uniquely determined by constraints that are biologically implausible. These aspects of PCA confer inherent limitations on the interpretability and usefulness of eigenimage analysis of biological time-series and have motivated the exploration of nonlinear PCA and neural network approaches.

Two other important approaches deserve mention here. The first is independent component analysis (ICA). ICA uses entropy maximisation to find, using iterative schemes, spatial modes or their dynamics that are approximately *independent*. This is a stronger requirement than *orthogonality* in PCA and involves removing

high order correlations among the modes (or dynamics). It was initially introduced as *spatial* ICA (McKeown *et al* 1998) in which the independence constraint was applied to the modes (with no constraints on their temporal expression). More recent approaches use, by analogy with magneto- and electrophysiological time-series analysis, *temporal* ICA where the dynamics are enforced to be independent. This requires an initial dimension reduction (usually using conventional eigenimage analysis). Finally, there has been an interest in cluster analysis (Baumgartner et al 1997). Conceptually, this can be related to eigenimage analysis through multidimensional scaling and principal co-ordinate analysis.

All these approaches are interesting, but hardly anyone uses them. This is largely because they tell you nothing about how the brain works or allow one to ask specific questions. Simply demonstrating statistical dependencies among regional brain responses (*i.e.* demonstrating functional connectivity) does not address how these responses were caused. To address this one needs explicit models of integration or more precisely, effective connectivity.

2. Dynamic causal modelling with bilinear models

This section is about modelling interactions among neuronal populations, at a cortical level, using neuroimaging time series. The aim of these dynamic causal models (DCMs) is to estimate, and make inferences about, the coupling among brain areas and how that coupling is influenced by changes in experimental context (*e.g.* time or cognitive set). The basic idea is to construct a reasonably realistic neuronal model of interacting cortical regions or nodes. This model is then supplemented with a forward model of how neuronal or synaptic activity translates into a measured response (see previous section). This enables the parameters of the neuronal model (*i.e.* effective connectivity) to be estimated from observed data.

Intuitively, this approach regards an experiment as a designed perturbation of neuronal dynamics that are promulgated and distributed throughout a system of coupled anatomical nodes to change region-specific neuronal activity. These changes engender, through a measurement-specific forward model, responses that are used to identify the architecture and time constants of the system at a neuronal level. This represents a departure from conventional approaches (*e.g.* structural equation modelling and autoregression models; McIntosh & Gonzalez-Lima 1994; Büchel & Friston 1997; Harrison *et al* 2003), in which one assumes the observed responses are driven by endogenous or intrinsic noise (*i.e.* innovations). In contradistinction, dynamic causal models assume the responses are driven by designed changes in inputs. An important conceptual aspect of dynamic causal models pertains to how the experimental inputs enter the model and cause neuronal responses. Experimental variables can illicit responses in one of two ways. First, they can elicit responses through direct influences on specific anatomical nodes. This would be appropriate, for example, in modelling sensory evoked responses in early visual cortices. The second class of input exerts its effect vicariously, through a modulation of the coupling among nodes. These sorts of experimental variables would normally be more enduring; for example attention to a particular attribute or the maintenance of some perceptual set. These distinctions are seen most clearly in relation to particular forms of causal models used for estimation, for example the bilinear approximation

$$\begin{aligned}\dot{x} &= f(x,u) \\ &= Ax + uBx + Cu \\ y &= g(x) + \varepsilon\end{aligned} \qquad 7$$

$$A = \frac{\partial f}{\partial x} \qquad B = \frac{\partial^2 f}{\partial x \partial u} \qquad C = \frac{\partial f}{\partial u}$$

where $\dot{x} = \partial x / \partial t$. This is an approximation to any model of how changes in neuronal activity in one region x_i are caused by activity in the other regions. Here the output function $g(x)$ embodies a hemodynamic model, linking neuronal activity to BOLD, for each region (*e.g.* that in Figure 5). The matrix A represents the connectivity among the regions in the absence of input u. Effective connectivity is the influence that one neuronal system exerts over another in terms of inducing a response $\partial \dot{x} / \partial x$. This latent connectivity can be thought of as the intrinsic coupling in the absence of experimental perturbations. The matrix B is effectively the change in latent coupling induced by the input. It encodes the input-sensitive changes in A or, equivalently, the modulation of effective connectivity by experimental manipulations. Because B is a second-order derivative it is referred to as *bilinear*. Finally, the matrix C embodies the extrinsic influences of inputs on neuronal activity. The parameters $\theta = \{A, B, C\}$ are the connectivity or coupling matrices that we wish to identify and define the functional architecture and interactions among brain regions at a neuronal level.

Because Eq(7) has exactly the same form as Eq(4) we can express it as a GLM and estimate the parameters using **EM** in the usual way (see Friston *et al* 2003). Generally estimation in the context of highly parameterised models like DCMs requires constraints in the form of priors. These priors enable conditional inference about the connectivity estimates. The sorts of questions that can be addressed with DCMs are now illustrated by looking at how attentional modulation might be mediated in sensory processing hierarchies in the brain.

2.1 DCM and attentional modulation

It has been established that the superior posterior parietal cortex (**SPC**) exerts a modulatory role on **V5** responses using Volterra-based regression models (Friston and Büchel 2000) and that the inferior frontal gyrus (**IFG**) exerts a similar influence on **SPC** using structural equation modelling (Büchel and Friston 1997). The example here shows that DCM leads to the same conclusions but starting from a completely different construct. The experimental paradigm and data acquisition parameters are described in the legend to Figure 7. This Figure also shows the location of the regions that entered into the DCM. These regions were based on maxima from conventional SPMs testing for the effects of photic stimulation, motion and attention. Regional time courses were taken as the first eigenvariate of 8mm spherical volumes of interest centred on the maxima shown in the figure. The inputs, in this example, comprise one sensory perturbation and two contextual inputs. The sensory input was simply the presence of photic stimulation and the first contextual one was presence of motion in the visual field. The second contextual input, encoding attentional set, was unity during attention to speed changes and zero otherwise. The outputs corresponded to the four regional eigenvariates in (Figure 7, left panel). The intrinsic connections were constrained to conform to a hierarchical pattern in which each area was reciprocally

connected to its supraordinate area. Photic stimulation entered at, and only at, **V1**. The effect of motion in the visual field was modelled as a bilinear modulation of the **V1** to **V5** connectivity and attention was allowed to modulate the backward connections from **IFG** and **SPC**.

Figure 7 about here

The results of the DCM are shown in Figure 7 (right panel). Of primary interest here is the modulatory effect of attention that is expressed in terms of the bilinear coupling parameters for this input. As expected, we can be highly confident that attention modulates the backward connections from **IFG** to **SPC** and from **SPC** to **V5**. Indeed, the influences of **IFG** on **SPC** are negligible in the absence of attention (dotted connection). It is important to note that the only way that attentional manipulation can effect brain responses was through this bilinear effect. Attention-related responses are seen throughout the system (attention epochs are marked with arrows in the plot of **IFG** responses in Figure 7). This attentional modulation is accounted for, sufficiently, by changing just two connections. This change is, presumably, instantiated by instructional set at the beginning of each epoch.

The second thing, this analysis illustrates, is how functional segregation is modelled in DCM. Here one can regard **V1** as 'segregating' motion from other visual information and distributing it to the motion-sensitive area **V5**. This segregation is modelled as a bilinear 'enabling' of **V1** to **V5** connections when, and only when, motion is present. Note that in the absence of motion the intrinsic **V1** to **V5** connection was trivially small (in fact the estimate was -0.04). The key advantage of entering motion through a bilinear effect, as opposed to a direct effect on **V5**, is that we can finesse the inference that **V5** shows motion-selective responses with the assertion that these responses are mediated by afferents from **V1**. The two bilinear effects above represent two important aspects of functional integration that DCM is able characterise.

2.2 Structural equation modelling

The central idea, behind dynamic causal modelling (DCM), is to treat the brain as a deterministic nonlinear dynamic system that is subject to inputs and produces outputs. Effective connectivity is parameterised in terms of coupling among unobserved brain states (*e.g.* neuronal activity in different regions). The objective is to estimate these parameters by perturbing the system and measuring the response. This is in contradistinction to established methods for estimating effective connectivity from neurophysiological time-series, which include structural equation modelling and models based on multivariate auto-regressive processes. In these models, there is no designed perturbation and the inputs are treated as unknown and stochastic. Furthermore, the inputs are often assumed to express themselves instantaneously such that, at the point of observation the change in states will be zero. From Eq(7), in the absence of bilinear effects we have

$$\begin{aligned} \dot{x} &= 0 \\ &= Ax + Cu \\ x &= -A^{-1}Cu \end{aligned} \qquad 8$$

This is the regression equation used in SEM where $A = A' - I$ and A' contains the off-diagonal connections among regions. The key point here is that A is estimated by assuming u is some random innovation with known covariance. This is not really tenable for designed experiments when u represent carefully structured experimental inputs. Although SEM and related autoregressive techniques are useful for establishing dependence among regional responses, they are not surrogates for informed causal models based on the underlying dynamics of these responses.

In this section we have covered multivariate techniques ranging from eigenimage analysis that does not have an explicit forward or causal model to DCM that that does. The bilinear approximation to any DCM has been illustrated though its use with fMRI to study attentional modulation. The parameters of bilinear DCMs include first-order effective connectivity A and its experimentally induced changes B. Although the Bilinear approximation is useful it is possible to model effective connectivity among neuronal subpopulations explicitly. We conclude with a DCM that embraces a number of neurobiological facts and takes us much closer to a mechanistic understanding of how brain responses are generated. This example uses responses measured with EEG.

3. Dynamic Causal modelling with neural mass models

Event-related potentials (ERPs) have been used for decades as electrophysiological correlates of perceptual and cognitive operations. However, the exact neurobiological mechanisms underlying their generation are largely unknown. In this section we use neuronally plausible models to understand event-related responses. The example used in this section shows that changes in connectivity are sufficient to explain certain ERP components. Specifically we will look at the P300, a late component associated with rare or unexpected events. If the unexpected nature of rare stimuli depends on learning which stimuli are frequent, then the P300 must be due to plastic changes in connectivity that mediate perceptual learning. We conclude by showing that recent advances in the modelling of evoked responses now afford measures of connectivity among cortical sources that can be used to quantify the effects of perceptual learning.

Figure 8 about here

3.1 Hierarchical neural mass models

The minimal model we have developed (David *et al*, in press), uses the connectivity rules described in Felleman and Van Essen (1991) to assemble a network of coupled sources. These rules are based on a partitioning of the cortical sheet into supra-, infra-granular layers and granular layer (layer 4). Bottom-up or forward connections originate in agranular layers and terminate in layer 4. Top-down or backward connections target agranular layers. Lateral connections originate in agranular layers and target all layers. These long-range or extrinsic cortico-cortical connections are excitatory and arise from pyramidal cells.

Each region or source is modelled using a neural mass model decribed in David and Friston (2003), based on the model of Jansen and Rit (1995). This model emulates the activity of a cortical area using three neuronal subpopulations, assigned to granular and agranular layers. A population of excitatory pyramidal (output) cells receives inputs from inhibitory and excitatory populations of interneurons, via intrinsic connections (intrinsic

connections are confined to the cortical sheet). Within this model, excitatory interneurons can be regarded as spiny stellate cells found predominantly in layer 4 and in receipt of forward connections. Excitatory pyramidal cells and inhibitory interneurons will be considered to occupy agranular layers and receive backward and lateral inputs (see Figure 8).

To model event-related responses, the network receives inputs via input connections. These connections are exactly the same as forward connections and deliver inputs u to the spiny stellate cells in layer 4. In the present context, inputs u model subcortical auditory inputs. The vector C controls the influence of the input on each source. The lower, upper and leading diagonal matrices A^F, A^B, A^L encode forward, backward and lateral connections respectively. The DCM here is specified in terms of the state equations shown in Figure 8 and a linear output equation

$$\begin{aligned} \dot{x} &= f(x,u) \\ y &= Lx_0 + \varepsilon \end{aligned} \qquad 9$$

where x_0 represents the transmembrane potential of pyramidal cells and L is a lead field matrix coupling electrical sources to the EEG channels. This should be compared to the DCM above for hemodynamics. Here the equations governing the evolution of neuronal states are much more complicated and realistic, as opposed to the bilinear approximation in Eq(7). Conversely, the output equation is a simple linearity, as opposed to the nonlinear observer used for fMRI. As an example, the state equation for the inhibitory subpopulation is[2]

$$\begin{aligned} \dot{x}_7 &= x_8 \\ \dot{x}_8 &= \frac{H_e}{\tau_e}((A^B + A^L + \gamma_3 I)S(x_0)) - \frac{2x_8}{\tau_e} - \frac{x_7}{\tau_e^2} \end{aligned} \qquad 10$$

Within each subpopulation, the evolution of neuronal states rests on two operators. The first transforms the average density of pre-synaptic inputs into the average postsynaptic membrane potential. This is modelled by a linear transformation with excitatory (e) and inhibitory (i) kernels parameterised by $H_{e,i}$ and $\tau_{e,i}$. $H_{e,i}$ control the maximum post-synaptic potential and $\tau_{e,i}$ represent a lumped rate constant. The second operator S transforms the average potential of each subpopulation into an average firing rate. This is assumed to be instantaneous and is a sigmoid function. Interactions, among the subpopulations, depend on constants $\gamma_{1,2,3,4}$, which control the strength of intrinsic connections and reflect the total number of synapses expressed by each subpopulation. In Eq(10), the top line expresses the rate of change of voltage as a function of current. The second line specifies how current changes as a function of voltage, current and presynaptic input from extrinsic and intrinsic sources. Having specified the DCM one can estimate the coupling parameters from empirical data using **EM** as described above.

[2] Propagation delays on the extrinsic connections have been omitted for clarity here and in Figure 8.

3.2 Perceptual learning and the P300

The example shown in Figure 9 is an attempt to model the P300 in terms of changes in backward and lateral connections among cortical sources. In this example, two [averaged] channels of EEG data were modelled with three cortical sources. Using this generative or forward model we estimated differences in the strength of these connections for rare and frequent stimuli. As expected, we could account for detailed differences in the ERPs (the P300) by changes in connectivity (see figure legend for details). Interestingly these differences were expressed selectively in the lateral connections. If this model is a sufficient approximation to the real sources, these changes are a non-invasive measure of plasticity, mediating perceptual learning, in the human brain.

Figure 9 about here

CONCLUSION

In this article we have reviewed some key models that underpin image analysis and have touched briefly on ways of assessing specialisation and integration in the brain. These models can be regarded as a succession of modelling endeavours, drawing more and more on our understanding of how brain-imaging signals are generated, both in terms of biophysics and the underlying neuronal interactions. We have seen how hierarchical linear observation models encode the treatment effects elicited by experimental design. General linear models based on convolution models imply an underlying dynamic input-state-output system. The form of these systems can be used to constrain convolution models and explore some of their simpler nonlinear properties. By creating observation models based on explicit forward models of neuronal interactions, one can now start to model and assess interactions among distributed cortical areas and make inferences about coupling at the neuronal level. The next years will probably see an increasing realism in the dynamic causal models introduced above (see Horwitz *et al* 2001). There are already attempts to use plausible models of neuronal ensembles to estimate network parameters of evoked responses in EEG. These endeavours are likely to encompass fMRI signals in the near future enabling the conjoint modelling, or fusion, of different modalities and the marriage of computational neuroscience with the modelling of brain responses.

Acknowledgements:

The Wellcome Trust funded this work. We would like to thank all my colleagues at the Wellcome Department of Imaging Neuroscience for their help with this review.

References

Absher JR and Benson DF. (1993). Disconnection syndromes: an overview of Geschwind's contributions. *Neurology* **43**:862-867

Adler RJ. (1981). In "*The geometry of random fields*". Wiley New York

Aertsen A and Preißl H. (1991). Dynamics of activity and connectivity in physiological neuronal Networks. In *Non Linear Dynamics and Neuronal Networks.* Ed Schuster HG VCH publishers Inc. New York NY USA p281-302

Ashburner J and Friston KJ. (2000). Voxel-based morphometry - the methods. *NeuroImage.* **11**:805-21.

Bandettini PA, Jesmanowicz A, Wong EC and Hyde JS. (1993) Processing strategies for time course data sets in functional MRI of the human brain. *Mag. Res. Med.* **30**:161-173

Baumgartner R, Scarth G, Teichtmeister C, Somorjai R and Moser E. (1997). Fuzzy clustering of gradient-echo functional MRI in the human visual cortex Part 1: reproducibility. *J Mag. Res. Imaging* **7**:1094-1101

JS Bendat. (1990). *Nonlinear System Analysis and Identification from Random Data.* John Wiley and Sons, New York USA

Berry DA and Hochberg Y. (1999). Bayesian perspectives on multiple comparisons. *J. Statistical Planning and Inference.* **82**:215-227

Biswal B, Yetkin FZ, Haughton VM and Hyde JS. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Mag. Res. Med.* **34**:537-541

Boynton GM, Engel SA, Glover GH and Heeger DJ. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci.* **16**:4207-4221

Büchel C, Wise RJS, Mummery CJ, Poline J-B, and Friston KJ. (1996). Nonlinear regression in parametric activation studies. *NeuroImage* **4**:60-66

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

Buckner R, Bandettini P, O'Craven K, Savoy R, Petersen S, Raichle M and Rosen B. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proc. Natl. Acad. Sci. USA* **93**:14878-14883

Burock MA, Buckner RL, Woldorff MG, Rosen BR and Dale AM. (1998). Randomized Event-Related Experimental Designs Allow for Extremely Rapid Presentation Rates Using Functional MRI. *NeuroReport* **9**:3735-3739

Buxton RB and Frank LR. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J. Cereb. Blood. Flow Metab.* **17**:64-72.

Dale A and Buckner R. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapp.* **5**:329-340

David O, Friston KJ. (2003) A neural mass model for MEG/EEG: coupling and neuronal dynamics. *NeuroImage* **20**:1743-55.

David, O., Harrison, L., Kilner, J., Penny, W., and Friston, K. J. (2004) Studying effective connectivity with a neural mass model of evoked MEG/EEG responses. *BioMag* –in press

Dempster AP, Laird NM and Rubin DR (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Series B* **39**:1-38

M Fliess, M Lamnabhi and F Lamnabhi-Lagarrigue (1983). An algebraic approach to nonlinear functional expansions. *IEEE Trans. Circuits Syst.* **30**:554-570

Felleman DJ, Van Essen DC. (1992).Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**:1-47.

Friston KJ, Frith CD, Liddle PF and Frackowiak RSJ. (1991). Comparing functional (PET) images: the assessment of significant change. *J. Cereb. Blood Flow Metab.* **11**:690-699

Friston KJ, Frith CD, Passingham RE, Liddle PF and Frackowiak RSJ. (1992). Motor practice and neurophysiological adaptation in the cerebellum: A positron tomography study. *Proc. Roy. Soc. Lon. Series B* **248**:223-228

Friston KJ, Frith CD, Liddle PF and Frackowiak RSJ. (1993). Functional Connectivity: The principal component analysis of large data sets. *J Cereb. Blood Flow Metab.* **13**:5-14

Friston KJ, Jezzard PJ and Turner R. (1994). Analysis of functional MRI time-*series Hum. Brain Mapp.* **1**:153-171

Friston KJ, Holmes AP, Worsley KJ, Poline JB, Frith CD and Frackowiak RSJ. (1995a). Statistical Parametric Maps in functional imaging: A general linear approach *Hum. Brain Mapp.* **2**:189-210

Friston KJ, Frith CD, Turner R and Frackowiak RSJ. (1995b). Characterising evoked hemodynamics with fMRI. *NeuroImage* **2**:157-165

Friston KJ, Price CJ, Fletcher P, Moore C, Frackowiak RSJ and Dolan RJ. (1996a). The trouble with cognitive subtraction. *NeuroImage* **4**:97-104

Friston KJ, Poline J-B, Holmes AP, Frith CD and Frackowiak RSJ. (1996b). A multivariate analysis of PET activation studies. *Hum. Brain Mapp.* **4**:140-151

Friston KJ, Frith CD, Fletcher P, Liddle PF and Frackowiak RSJ. (1996c). Functional topography: multidimensional scaling and functional connectivity in the brain. *Cerebral Cortex* **6**:156-164

Friston KJ, Mechelli A, Turner R, Price CJ. (2000). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage.* **12**:466-77.

Friston KJ and Büchel C. (2000). Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proc Natl Acad Sci U S A.* **97**:7591-6.

Friston KJ, Harrison L, Penny W. (2003) Dynamic causal modelling. *NeuroImage* **19**:1273-302

Gerstein GL and Perkel DH. (1969). Simultaneously recorded trains of action potentials: Analysis and functional interpretation. *Science* **164**:828-830

Goltz F. (1881). In "*Transactions of the 7th international medical congress*" (W MacCormac Ed) Vol. I JW Kolkmann London. p218-228

Grafton S, Mazziotta J, Presty S, Friston KJ, Frackowiak RSJ and Phelps M. (1992). Functional anatomy of human procedural learning determined with regional cerebral blood flow and PET. *J Neurosci.* **12**:2542-2548

Harrison LM, Penny W. and Friston KJ. (2003). Multivariate autoregressive modelling of fMRI time series. *NeuroImage.* **19**:1477-91.

Hoge RD, Atkinson J, Gill B, Crelier GR, Marrett S and Pike GB. (1999). Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. *Proc. Natl. Acad. Sci.* **96**:9403-9408

Holmes A and Ford I. (1993). A Bayesian approach to significance testing for statistic images from PET. In Quantification of Brain function, Tracer kinetics and Image analysis in brain PET. Eds. K. Uemura, NA. Lassen, T. Jones and I. Kanno. *Excerpta Medica, Int. Cong.* Series No. **1030**: 521-534

Horwitz B, Friston KJ and Taylor JG. (2001). Neural modelling and functional brain imaging: an overview. *Neural Networks* **13**:829-846

Jansen BH, Rit VG. (1995). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biol. Cybern* **73**:357-66.

Lange N and Zeger SL. (1997). Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion) J *Roy. Stat. Soc. Ser. C.* **46**:1-29

Lueck CJ, Zeki S, Friston KJ, Deiber MP, Cope NO et al (1989). The colour centre in the cerebral cortex of man. *Nature* **340**:386-389

Mandeville JB, Marota JJ, Ayata C, Zararchuk G, Moskowitz MA, B Rosen B, and Weisskoff RM. (1999). Evidence of a cerebrovascular postarteriole Windkessel with delayed compliance. *J. Cereb. Blood Flow Metab.* **19**:679-689

McIntosh AR and Gonzalez-Lima F. (1994). Structural equation modelling and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.* **2**:2-22

McKeown M, Jung T-P, Makeig S, Brown G, Kinderman S, Lee T-W and Sejnowski T. (1998). Spatially independent activity patterns in functional MRI data during the Stroop colour naming task. *Proc. Natl. Acad. Sci.* **95**:803-810

Petersen SE, Fox PT, Posner MI, Mintun M and Raichle ME. (1989). Positron emission tomographic studies of the processing of single words. *J Cog. Neurosci.* **1**:153-170

Phillips CG Zeki S and HB Barlow HB. (1984). Localisation of function in the cerebral cortex Past present and future. *Brain* **107**:327-361

Price CJ, Wise RJS, Ramsay S, Friston KJ, Howard D, et al. (1992). Regional response differences within the human auditory cortex when listening to words. *Neurosci. Lett.* **146**:179-182

Price CJ and Friston KJ. (1997). Cognitive Conjunction: A new approach to brain activation experiments. *NeuroImage* **5**:261-270

Sychra JJ, Bandettini PA, Bhattacharya N and Lin Q. (1994). Synthetic images by subspace transforms I Principal component images and related filters. *Med. Physics* **21**:193-201

Talairach P and Tournoux J. (1988) "*A Stereotactic coplanar atlas of the human brain*" Stuttgart Thieme

Worsley KJ, Evans AC, Marrett S and Neelin P. (1992). A three-dimensional statistical analysis for rCBF activation studies in human brain. *J Cereb. Blood Flow Metab.* **12**:900-918

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ and Evans AC. (1996). A unified statistical approach or determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4**:58-73

Zeki S. (1990). The motion pathways of the visual cortex. In "*Vision: coding and efficiency*" C Blakemore Ed. Cambridge University Press UK p321-345

Figure Legends

Figure 1

The general linear model: The general linear model is an equation expressing the response variable Y in terms of a linear combination of explanatory variables in a design matrix X and an error term with assumed or known autocorrelation Σ. In fMRI the data can be filtered with a convolution or residual forming matrix (or a combination) S, leading to a generalised linear model that includes [intrinsic] serial correlations and applied [extrinsic] filtering. Different choices of S correspond to different estimation schemes as indicated on the upper left. The parameter estimates obtain in a least squares sense using the pseudoinverse (denoted by +) of the filtered design matrix. An effect of interest is specified by a vector of contrast weights c that give a weighted sum or compound of parameter estimates, referred to as a *contrast*. The T statistic is simply this contrast divided by the standard error (*i.e.* square root of its estimated variance). The ensuing T statistic is distributed with v degrees of freedom. The equations for estimating the variance of the contrast and the degrees of freedom are provided in the right-hand panel and accommodate the non-sphericity implied by Σ.

Figure 2

Top right: Design matrix: This is an image representation of the design matrix. Contrasts: These are the vectors of contrast weights defining the linear compounds of parameters tested. The contrast weights are displayed over the column of the design matrix encoding the effects tested. The design matrix here includes condition-specific effects (boxcars convolved with a hemodynamic response function). Odd columns correspond to stimuli shown under isochromatic conditions and even columns model responses to isoluminant stimuli. The first two columns are for stationary stimuli and the remaining columns are for stimuli of increasing speed. The final column is a constant term. Top left: SPM{T}: This is a maximum intensity projection of the SPM{T} conforming to the standard anatomical space of Talairach and Tournoux (1988). The T values here are the minimum T values from both contrasts, thresholded at $p = 0.001$ uncorrected. The most significant conjunction is seen in left V5. Lower panel: Plot of the condition-specific parameter estimates for this voxel. The T value was 9.25 ($p<0.001$ adjusted according to RFT).

Figure 3

SPM and PPM for a fMRI study of attention to visual motion. The display format in the lower panel uses an axial slice through extrastriate regions but the thresholds are the same as employed in maximum intensity projections (upper panels). Upper right: The activation threshold for the PPM was 0.7 a.u., meaning that all voxels shown had a 90% chance of an activation of 0.7% or more. Upper left: The corresponding SPM using an adjusted threshold at $p = 0.05$. Note the bilateral foci of motion-related responses in the PPM that are not seen in the SPM (grey arrows). As can be imputed from the design matrix (upper middle panel), the statistical model of evoked responses comprised boxcar regressors convolved with a canonical hemodynamic response function. The middle column corresponds to the presentation of moving dots and was the stimulus property tested by the contrast.

Figure 4

Temporal basis functions offer useful constraints on the form of the estimated response that retain the flexibility of FIR models and the efficiency of single regressor models. The specification of these constrained FIR models

involves setting up stimulus functions $u(t)$ that model expected neuronal changes [*e.g.* boxcars of epoch-related responses or spikes (delta functions) at the onset of specific events or trials]. These stimulus functions are then convolved with a set of basis functions $T_i(t)$ of peri-stimulus time that, in some linear combination, model the HRF. The ensuing regressors are assembled into the design matrix. The basis functions can be as simple as a single canonical HRF (middle), through to a series of delayed delta functions (bottom). The latter case corresponds to a FIR model and the coefficients constitute estimates of the impulse response function at a finite number of discrete sampling times. Selective averaging in event-related fMRI (Dale and Buckner 1997) is mathematically equivalent to this limiting case.

Figure 5

Right: Hemodynamics elicited by an impulse of neuronal activity as predicted by a dynamical biophysical model (left). A burst of neuronal activity causes an increase in flow-inducing signal that decays with first order kinetics and is down regulated by local flow. This signal increases rCBF, which dilates the venous capillaries, increasing volume (v). Concurrently, venous blood is expelled from the venous pool decreasing deoxyhemoglobin content (q). The resulting fall in deoxyhemoglobin concentration leads to a transient increases in BOLD (blood oxygenation level dependent) signal and a subsequent undershoot. Left: Hemodynamic model on which these simulations were based (see Friston *et al* 2000 for details).

Figure 6

Left panels: Volterra kernels from a voxel in the left superior temporal gyrus at -56, -28, 12mm. These kernel estimates were based on a single-subject study of aural word presentation at different rates (from 0 to 90 words per minute) using a second order approximation to a Volterra series expansion modelling the observed hemodynamic response to stimulus input (a delta function for each word). These kernels can be thought of as a characterisation of the second order hemodynamic response function. The first order kernel κ_1 (upper panel) represents the (first-order) component usually presented in linear analyses. The second-order kernel (lower panel) is presented in image format. The colour scale is arbitrary; white is positive and black is negative. The insert on the right represents $\kappa_1\kappa_1^T$, the second-order kernel predicted by a simple model that involved a linear convolution with κ_1 followed by some static nonlinearity. Right panel: Integrated responses over a 32-second stimulus train as a function of SOA. Solid line: Estimates based on the nonlinear convolution model parameterised by the kernels on the left. Broken line: The responses expected in the absence of second-order effects (*i.e.* in a truly linear system). Dots: Empirical averages based on the presentation of actual stimulus trains.

Figure 7

Results of a DCM analysis of attention to visual motion with fMRI. Right panel: Functional architecture based upon the conditional estimates shown alongside their connections, with the percent confidence that they exceeded threshold in brackets. The most interesting aspects of this architecture involve the role of motion and attention in exerting bilinear effects. Critically, the influence of motion is to enable connections from **V1** to the motion-sensitive area **V5**. The influence of attention is to enable backward connections from the inferior frontal

gyrus (**IFG**) to the superior parietal cortex (**SPC**). Furthermore, attention increases the influence of **SPC** on **V5**. Dotted arrows connecting regions represent significant bilinear effects in the absence of a significant intrinsic coupling. Left Panel: Fitted responses based upon the conditional estimates and the adjusted data are shown for each region in the DCM. The insert (upper left) shows the location of the regions.

Subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes). The data were acquired from a normal subject at 2 Tesla using a Magnetom VISION (Siemens, Erlangen) whole body MRI system, equipped with a head volume coil. Contiguous multi-slice T2*-weighted fMRI images were obtained with a gradient echo-planar sequence (TE = 40ms, TR = 3.22 seconds, matrix size = 64x64x32, voxel size 3x3x3mm). Each subject had 4 consecutive 100-scan sessions comprising a series of 10-scan blocks under 5 different conditions D F A F N F A F N S. The first condition (D) was a dummy condition to allow for magnetic saturation effects. F (Fixation) corresponds to a low-level baseline where the subjects viewed a fixation point at the centre of a screen. In condition A (Attention) subjects viewed 250 dots moving radially from the centre at 4.7 degrees per second and were asked to detect changes in radial velocity. In condition N (No attention) the subjects were asked simply to view the moving dots. In condition S (Stationary) subjects viewed stationary dots. The order of A and N was swapped for the last two sessions. In all conditions subjects fixated the centre of the screen. During scanning there were no speed changes. No overt response was required in any condition.

Figure 8

Schematic of the DCM used to model electrical responses. This schematic shows the state equation describing the dynamics of sources or regions. Each source is modelled with three subpopulations (pyramidal, spiny stellate and inhibitory interneurons) as described in Jansen and Rit (1995) and David and Friston (2003). These have been assigned to granular and agranular cortical layers which receive forward and backward connection respectively.

Figure 9

Summary of a dynamic causal modelling of ERPs elicited during an auditory P300 paradigm, employing rare and frequent pure tones. Upper panel: Schematic showing the architecture of the neuronal model used to explain the empirical data. Sources were coupled with extrinsic cortico-cortical connections following the rules of Felleman and van Essen (1991). The free parameters of this model included intrinsic and extrinsic connection strengths that were adjusted to best explain the observed ERPs. In this example the lead field was also estimated, with no spatial constraints. The parameters were estimated for ERPs recorded during the presentation of rare and frequent tones and are reported beside their corresponding connection (frequent/rare). The most notable finding was that the P300 could be explained by a selective increase in lateral connection strength (strengths highlighted in bold). Lower panel: The channel positions (left) and ERPs (right) averaged over two subsets of channels (circled on the left). Note the correspondence between the measured ERPs and those generated by the model. See David *et al* (2004) for details.

We modelled event-related potentials that exhibited a strong modulation of the P300 component, on comparing responses to frequent and rare stimuli using an auditory oddball paradigm. Auditory stimuli, 1000 or

2000 Hz tones with 5 ms rise and fall times and 80 ms duration, were presented binaurally. The tones were presented for 15 minutes, every 2 seconds in a pseudo-random sequence with 2000-Hz tones occurring 20% of the time and 1000-Hz tones occurring 80% of the time. The subject was instructed to keep a mental record of the number of 2000-Hz tones (non-frequent target tones). Data were acquired using 128 EEG electrodes with 1000 Hz sample frequency. Before averaging, data were referenced to mean earlobe activity and band-pass filtered between 1 and 30 Hz. Trials showing ocular artefacts and bad channels were removed from further analysis.