On the Relation Between Brain Images and Brain Neural Networks

J.G. Taylor,[1,2]* B. Krause,[2,3] N.J. Shah,[2] B. Horwitz,[4] and H.-W. Mueller-Gaertner[2,3]

[1]*Department of Mathematics, King's College, Strand, London, UK*
[2]*Institute of Medicine, Research Centre Juelich, Germany*
[3]*Department of Nuclear Medicine, Heinrich-Heine University Hospital, Duesseldorf, Germany*
[4]*Voice, Speech and Language Branch, National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, Maryland*

Abstract: The relationship between brain images observed by PET and fMRI and the underlying neural activity is analysed using recent results on the detailed nature of averaged and synchronised activity of coupled neural networks and on a simplifying model of the level of blood flow caused by neural activity. The conditions on the coupled neural systems are specified that lead to structural equation models, giving support to analysis of the covariance structural equation modelling of brain imaging data. The relation between the resulting models and possible neural codes are analysed. Furthermore, a new form of structural equation model is derived, in which all neuronal activity arises as hidden variables. We discuss how the results of such analyses can be transported back to the domain of coupled temporally dynamic neural systems in the brain appropriate to EEG and MEG observations. *Hum. Brain Mapping 9:165–182, 2000.* © 2000 Wiley-Liss, Inc.

Key words: fMRI; brain; PET; neural activity

INTRODUCTION

There is currently much activity attempting to probe the results of brain imaging data arising from PET and fMRI machines. In particular, there is considerable interest in the use of structural equation modelling [Bollen, 1989; Loehlm, 1992] of the correlated activities of regions in the brain while a subject is performing a given task. This analyses the manner in which the coupling between different random variables can be assessed by means of using their cross-correlations, based on simple assumptions of linear dependence between the variables (together with the possible existence of "hidden" or unobserved variables). The approach, when applied to the brain, will be termed [following McIntosh and Gonzalez-Lima, 1994] covariance structural equation modelling (CSEM), to distinguish it from modelling of anatomical aspects of the brain (gyri, gray-white matter, and so on).

This method allows us in principle to uncover the manner in which various regions of the brain are coupled together in what is hoped is a causal fashion [Krause et al., 1999; McIntosh and Gonzalez-Lima, 1994; McIntosh et al., 1994]. In particular, it permits analysis to be performed without direct subtraction of brain imaging data obtained under two different conditions. This can otherwise cause disappearance of an area of importance that has the same level of activity but different functional links with other areas across the two conditions. Such a case was clearly demon-

*Correspondence to: J.G. Taylor, Department of Mathematics, King's College, Strand, London WC2R2LS, United Kingdom.

Received for publication 29 December 1998; accepted 2 December 1999

strated by the manner in which peristriate cortex was differently involved with other brain areas in object and spatial vision tasks while being activated at about the same level in both tasks [Horwitz et al., 1992b]. Furthermore, there is the hope that the resulting structural models capture important features of the underlying neural activity itself. Because such neural activity is the final target of the experimentation, the use of such structural models in analysis of PET and fMRI data is very attractive.

Inferring casual relations between covariance data alone is not easily accomplished, because there may be many ways of modelling the same covariance data structurally. Such ambiguity in the derivation of structural models is another reason why derivation of any CSEM for brain imaging data from the underpinning neural activity is important. It may be that ambiguity will still be present even when a neural underpinning of the CSEM is obtained. However, given such a neural framework, it will then be possible to include structural constraints of a neuroanatomical kind into the solutions with more confidence than without it. Thus, the program being embarked on in this paper can be seen as part of giving the boundary conditions needed to narrow down the set of CSEM solutions to covariating modular activity.

There is a basic question that must be answered before a CSEM of brain imaging data can be properly understood as a true mirror of the underlying neural activity. It stems from the considerable controversy that these models have aroused by their use in the social sciences and psychology (see, for example, the discussion in Bollen, 1989, and references therein). This controversy is partly because of some strong but unsubstantiated claims made by various practitioners that they were detecting underlying social or psychological variables. There was no proof that they were. The question we have to answer is, therefore, what justification can be given for the use of covariance structural equation models in brain imaging? Related to this is the problem of interpretation of the connection strengths resulting from a CSEM for a particular task. Are they simply the connection weights between the underpinning interacting neural modules?

A particular aspect of the above problems involves the difference between observable and latent variables. The latter are only detectable by their common presence in a number of observable variables, but are not directly observable themselves. As is to be expected, their existence is harder to justify than that of directly observable variables. What is the situation for brain imaging; are there hidden variables, and if so, what are they, and how are they to be justified?

In all, then, there are three questions to be answered: (1) Is it possible to justify CSEMs as used in brain imaging from suitably averaged underpinning neural activities in the relevant modules? (2) How are the connection strengths of the resulting CSEMs related to the underlying synaptic weights for the connections between the various modules involved? (3) Are there any hidden variables which need to be introduced in the CSEMs as derived from the underpinning neural activities? This paper is dedicated to presenting tentative answers to these three questions.

Answers to these questions are obtained by deriving the covariance structural equation models from the underlying neural activity, which in the first place gives rise to the correlated brain imaging data. This can be achieved by discovering a general set of assumptions on the nature of the underlying neural activity, which leads to the covariance structural equation models appropriate for given tasks. Such assumptions would correspond to the discovery of "bridging" laws [Churchland, 1986] between the neural networks of the brain and the covariance structural equation models deduced from PET and fMRI data. The gap such bridging laws fill is an important one: without them, the implications of the structural models of brain imaging data are unclear. With them, a direct neural interpretation of the parameters obtained by path analysis in terms of synaptic strengths would be possible, thereby allowing the carrying over of PET and fMRI results to the arena of MEG and EEG (where also structural equation modelling can be performed, although now involving time more strongly, in the coupling between brain areas).

The purpose of this paper is to attempt to develop such bridging laws. We present a set of assumptions about the underlying neural activity, which we argue are reasonably well biologically justified and which lead to covariance structural equation models for the correlation of activity of brain areas observed by brain imaging techniques. We will show how the path coefficients of such models relate to the synaptic connection weights between the relevant areas. Moreover, we will indicate how it is possible to build back up from covariance structural equation models to predict some of the temporal features of neural activity which is measurable by EEG and MEG techniques. There are other features, such as that of fast oscillatory activity, which in general will only be able to be included by direct measurements using these techniques themselves.

In the process, we will find that the basic assumptions of the CSEM approach, which uses the blood flow level directly, or the *z*- or *t*-scores of SPM maps of activity, observed by PET and fMRI as the random variables in the CSEM approach, cannot be justified. This is so in spite of taking the simplest model of the relation between blood flow and underlying neural activity. In spite of the simplicity of the idea that "neurons who play together brain image together" [Horwitz, 1991; see also Tononi et al., 1992], the manner in which the various steps need to be taken to go from "playing-together" neurons to brain imaging results is not trivial, as we will see. It is the purpose of this paper to see what steps can actually be taken, and what results arise. In particular, we will attempt to obtain forms of CSEMs which are derivable from underlying neural activity under the simplest of assumptions.

A further question we will consider, as a result of the analysis, is as to the steps needed to go from the resulting CSEMs to MEG and EEG measurements. There are extra features which have to be considered in going from blood flow, and the related spatial and temporal averaging, to neural activity of localised synchronised clumps of neurons in which there is no temporal averaging. This raises the important question: what more does MEG/EEG results provide beyond PET/fMRI, and what are the bridging laws between the blood flow measurements and the former data? We will attempt to attack this problem toward the end of the paper.

We start the paper by giving a review of covariance structural equation models in the next section and a review of neural network models in the following one. We present the underpinning neural network model in the Section on Deduction of Structural Models, and then develop there a set of assumptions on neural coding in separate modules, which allow averaging over neural activity to be performed. The resulting equations are identified preliminarily with covariance structural equation models in which the path strengths are directly related to suitably restricted sets of synaptic weights relevant to the task at hand. This preliminary model is then developed further in the following section, in which we start from the equations relating localised spatial averages of neural activity to each other and derive the resulting blood flow. The necessary assumptions being made are properly expressible only in terms of separate populations of excitatory and inhibitory neurons. Covariance structural equation models are obtained, with well-defined bridging laws, for suitable interactions between these populations; the models are an extension of the standard ones previously used to analyse brain imaging data, in which the neuron activity enters only as hidden variables. We then consider how we may climb back up to the temporal domain, so allowing predictions to be made for MEG and EEG analysis. Extensions of the simple neural networks used in the relation to PET and fMRI needed for this new domain are considered as part of this extension. A discussion as to what are some of the sources of variability in the CSEM approach to PET and fMRI is then given. The paper finishes with conclusions and a discussion of various unanswered questions.

COVARIANCE STRUCTURAL EQUATION MODELLING

There are numerous good accounts of this approach to the analysis of statistical data [Basilevsky, 1994; Bollen, 1989; Loehlin, 1992]. As noted in Bollen [1989], there are three components of the approach: (1) path analysis, (2) the synthesis of latent variable and measurement models, and (3) estimation procedures. We will not be concerned directly with the third aspect. It is the first and second points which are of importance to us, especially the first. We will, however, seriously have to consider possible latent (hidden) variables in due course.

Path analysis involves first the construction of a path diagram, then the determination of the relation between the path coefficients in terms of the correlation coefficients between the underlying random variables. Secondly, the decomposition of the effects of one variable upon another is achieved as either direct, indirect, and total. Thus, in the path diagram example of Figure 1, there are the observable random variables A, C, and D (denoted by square boxes) whose effects on the latent variable B (denoted by a circle), as well as the effects of B on C and D, are denoted by the small letters a, b, and c. These path strengths give the effects of a variable on the other one at the end of the path to which it is connected. This implies that, for the variables of Figure 1,

$$A = \eta_A$$

$$B = aA + Z$$

$$C = bB + X$$

$$D = cB + Y \tag{1}$$

It is assumed that all variables have zero mean, that the variables X, Y, Z, η_A are uncorrelated with each other, and that A, C, and D have unit variance. From

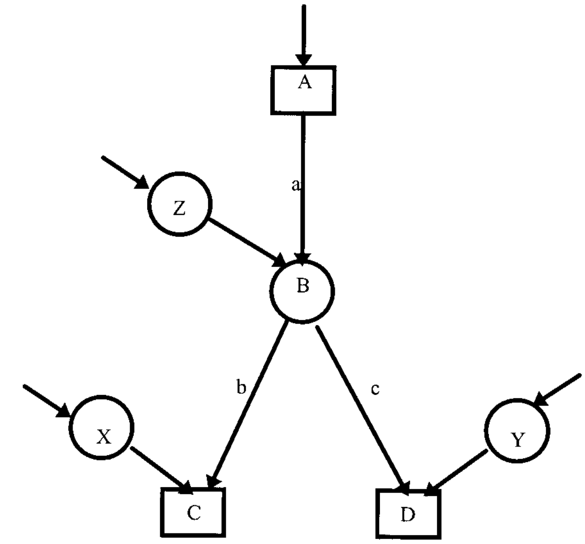

Figure 1.
Example of a path model in which the random variables A, C, D are observable (they are all in square boxes) together with latent or unobservable variable B, placed in a circle. Path strengths are denoted by small letters a, b, c and shown in the diagram. There are also residual variables X, Y, Z in the figure; these are also denoted by circles, but have their effects shown by arrows with no strength attached (because they are taken to be normalised variables).

Eq. (1) we can obtain immediately the correlation equations

$$r_{AC} = ab$$

$$r_{AD} = ac$$

$$r_{CD} = bc \qquad (2)$$

which may immediately be solved for a, b, and c. We may write Eq. (1) in matrix form as

$$\mathbf{v} = \mathbf{M}\mathbf{v} + \eta \qquad (3)$$

where $\mathbf{v}$ is the vector of observables $\mathbf{v}^T = (A, B, C, D)$, η is the vector of residuals $\eta^T = (\eta_A, Z, X, Y)$, and $\mathbf{M}$ is the matrix

(0000)

(a000)

(0b00)

(0c00)

Then the solution to (3) is that

$$\mathbf{v} = [\mathbf{1} - \mathbf{M}]^{-1}\eta \qquad (4)$$

and the calculated correlations for the components of $\mathbf{v}$ are

$$\langle \mathbf{vv} \rangle = [\mathbf{1} - \mathbf{M}]^{-1}\mathbf{C}[\mathbf{1} - \mathbf{M}]^{-1} \qquad (5)$$

where $\mathbf{C}$ is the correlation matrix for the residuals; with the assumptions we have made this is the unit matrix. We may then use the observed correlation matrix $\langle \mathbf{vv} \rangle_{\text{obs}}$ to attempt to solve Eq. (5) for the unknown coefficients of $\mathbf{M}$. This may be done by the use of a mean-square or similar error function (possibly more robust against the effects of outliers) constructed from the error difference between the two correlation matrices [one being the data, the other constructed from the path coefficients using Eqs. (2)]. This error term is a function of the unknown parameters in the matrix $\mathbf{M}$, so minimising the error in these variables will lead to the optimal solution to Eq. (5) (although care must be taken to distinguish between possible local minima and the global minimum). Significance levels of the resulting model can also be attached to this error score since in the case of normally distributed matrix elements the error will be distributed as the chi-squared distribution [Loehlin, 1992].

The decomposition of effects in the model is clear from Figure 1. The effects on A are purely residual, those on C and D are both residual and from the latent variable B. This latter has effects arising from A and its residual Z, so that in total C and D have effects from their own residuals and from A and the residual of B through the paths of strengths b and c, respectively.

Finally, it would be necessary to test other possible models of effects on A, C, and D and compare them with the model of Figure 1. This can be done by subtraction of the relevant chi-squared from each other for two different models.

NEURAL NETWORKS

The dynamics of neural networks depends on the nature of the neurons being considered as the "atoms" of the system. There are numerous possible features of neurons that have to be decided upon before a neural model can be constructed. The more complex (and complete) the model the more difficult it will be to bridge the gap to structural models of brain imaging data. We have to choose between the possibilities:

- graded vs. spiking neurons,
- deterministic firing vs. stochastic firing from threshold fluctuations,
- stochastic neuron effects from stochastic quantal release (corresponding to stochastic connection strengths),
- complex vs. simple cell geometry (many compartments vs. one),
- temporal features of ionic channels,
- temporal features of synaptic responses (alpha functions), and more generally effects from depressive and facilitatory synapses [Abbott et al., 1997; Markram and Tsodyks, 1996]
- linear vs. nonlinear cell response,
- simple one-layer vs. multilayer modules,
- learning as Hebbian vs. reinforcement vs. supervised,
- the presence of neuromodulators as a global signal or local signal,
- network dynamics as associators vs. relaxors to attractors,
- intrinsic oscillatory membrane potentials.

There are numerous further complexities which could also be added; however, we remarked that we should keep to the simplest model to start with; we will discuss extensions later. Thus, to start with, we consider:

- rate coding,
- leaky integration,
- synaptic transfer without temporal effects,
- point neurons with no complex geometry,
- no learning,
- no neuromodulators,
- no oscillations.

There is support for such simplifying assumptions from the overall feature of fMRI and PET activity as coupled to temporally slow haemodynamics (compared to neuronal dynamics). Thus, spikes, oscillations, and other temporal features above are averaged over by the temporal character of blood flow. Similarly, the nonpoint like aspects of neurons are lost in the spatial averaging occurring. The variables for each neuron will therefore be its single membrane potential $u_{\alpha i}$, where α is the label of the module to which the neuron belongs and i is the label of the neuron in its module. The potentials $u_{\alpha i}$ will each satisfy the standard Hodgkin-Huxley equation without shunting terms, so is of form

$$\tau_{\alpha i} du_{\alpha i}/dt = -u_{\alpha i} + \sum w_{\alpha i,,\beta j} f_{\beta j}(u_{\beta j}) + \text{nonlinear terms} + I_{\alpha i}(t) \quad (6)$$

where the quantity $\tau_{\alpha I}$ is the time-constant of the neuron, $I_{\alpha i}(t)$ is the external input arriving on it from outside the set of coupled modules, $w_{\alpha i,,\beta j}$ is the strength of the connection between the jth neuron in module β to the neuron under consideration and $f_{\beta j}$ is the graded response function of the neuron feeding to the neuron of interest. It could be questioned why we include the capacitative term on the left-hand side of Eq. (6), because that is expected to be washed out by the averaging by blood flow. We include it here as a link to the underlying faster neural activity. This term will be especially helpful for us to rebuild the fast-time MEG and EEG equations from the CSEM forms we obtain. Finally, the nonlinear term corresponds to the coupling between inputs from more than one module to the given one. Such terms are termed "sigma-pi" terms in neural network terminology [McClelland et al., 1986], and are of the form

$$\sum w_{\alpha i,,\beta j\gamma k} f_{\beta j}(u_{\beta j}) f_{\gamma k}(u_{\gamma k}) + \text{higher order terms} \quad (6a)$$

where the higher order terms involve inputs from at least three modules. We will neglect this nonlinear sigma-pi for the moment, but will return to it later in consideration of nonlinearities being noted as arising in CSEM effects [Friston et al., 1998; Glover, 1999].

We may interpret the various terms in (6) as arising from different forms of current flow. The left-hand side describes a capacitative current flow term (with τ = 1/RC, C being the capacity and R the resistance of the leaky cell membrane). The various terms on the right-hand side describe a leaky resistive membrane current, ionic currents from various other neuronal inputs, and an external current injection term.

It is usual to take $f_{\beta j}$ as independent of the labels β and j, being of the form of

$$f(x) = [1 + e^{-x/T}]^{-1}$$

where T denotes some effective "temperature." The function grows monotonically to a saturating value, as is known from experiment. When T is very large, then f is roughly constant, whereas when $T \approx 0$, then f approaches the unit step function or Heaviside function. The connection weights in general have no constraints other than Dale's law, that connections from a given neuron usually have the same sign.

We are thus at a stage to attempt to build the bridge from the neural networks of the brain to the structural equations describing averaged activity.

PRELIMINARY DEDUCTION OF STRUCTURAL MODELS FROM NEURAL ACTIVITY

In this section, we propose to consider possible simplifying assumptions which will allow us to obtain CSEMs from the Eqs. (6) of the underlying neural networks. We consider several assumptions lead to CSEMs; these provide possible neural underpinnings of CSEMs and lead to associated bridging laws.

Common assumptions

Let us start from the assumed neural systems of the previous section: the responses of the neurons are their mean firing rates, their dynamics being determined by Hodgkin-Huxley equations with a sigmoidal response of each neuron as a function of the membrane potential. We will further assume that there are a set of anatomically-defined modules M_α, where the label α runs over a finite index set A. The neurons in the module M_α will be labeled by the further index i, which will run over the set $\{1, 2, \ldots, N_\alpha\}$ where all of the numbers N_α are much larger than unity, $N_\alpha \gg 1$. This situation corresponds to vertebrate brains, where values of N_α of about a million are the average. The membrane potential $u_{\alpha i}$ of the ith neuron in M_α then satisfies the first-order differential Eq. (6) above. The parameter $\tau_{\alpha i}$ (the decay constant of the membrane for the ith neuron in module α) and its response function $f_{\alpha i}$ are all in general dependent on the labels α and i for the particular neuron. We now assume that these quantities are the same for all neurons being considered in a given module, although we must realise that they provide a source of variation which will ultimately have to be taken into account.

The quantities $I_{\alpha i}(t)$ are a set of external inputs assumed accessing each of the neurons directly. Most neurons in the brain do not have such external input. However, we should still keep it present as an averaged effect of input from modules not being considered. These can be: (a) cortical modules not included to avoid making the path model too complex, and (b) cortical regions not included because their input is not detectable by the functional imaging being used. We will assume that the modulation from these two sources is independent of each neuron i but is the same (on all neurons of a given module being considered) from the background of these other modules. This may not be defensible, but is at least a first approximation whose modification can be considered later: (c) noncortical modules whose output is not being measured, such as in the midbrain or brain stem. These can play an important modulatory role which could vary during the total period of an experiment. We will assume that such modulation is the same for all neurons in a given module, an assumption that is reasonable at a first approximation as for the previous assumption, and (d) external stimulus-led input to primary sensory regions; this will be expected to be dependent on particular neurons in the primary sensory regions which feed to it, so will be taken explicitly into account for these modules only, which we will separate out from the higher level ones involved in various cognitive tasks. The inputs in this case will be termed "external." Under these assumptions, the coupled Eqs. (6) reduce to the set

$$\tau_\alpha du_{\alpha i}/dt = -u_{\alpha i} + \sum w_{\alpha i,,\beta j} f_\beta(u_{\beta j}) + I_\alpha(t) \qquad (7)$$

We have not differentiated in the inputs $I_\alpha(t)$, which are appropriate to be chosen from the four classes (a), (b), (c), and (d); these will be considered at the relevant time. The time constants in (7) are at least 100 times smaller than those arising from blood flow dynamics, so may be neglected for fMRI and PET; the resulting version of (7) becomes:

$$u_{\alpha i} = \sum w_{\alpha i,,\beta j} f_\beta(u_{\beta j}) + I_\alpha(t) \qquad (7a)$$

where the time-dependence of the inputs is immediately picked up by the membrane potentials, with no time lags.

Equations (7a) already have the form of CSEMs, although in a nonlinear version. However, they involve one variable for each neuron in each module. The task ahead is to reduce such a large set of descriptors to a small number (if possible one) per module. Therefore, additional properties of the coupled modules must be introduced to remove the dependence on particular neurons by suitable averaging or otherwise, and so allowing bridging laws to be deduced.

Various approaches are possible, which we will now consider. Initially, we entertain the simplest assumption, that of random neural coupling between modules; however, this is found to lead to unwanted simplicity of connection strengths in the resulting CSEM. More input-sensitive forms of CSEM are then derived which depend on the manner in which past experience is coded into the neural weights of the modules. Different assumptions lead to different interpretations of the connection strengths of the resulting CSEM models; these are delineated.

Randomly coupled modules

A powerful reduction of the high-dimensional activity of the many neurons in each of the coupled modules is possible under the "random coupling" assumption [Geman, 1982; Wennekers and Pasemann, 1996]. This requires that each of the elements of the matrix $W_{\alpha\beta}$, defined as $(\mathrm{W}_{\alpha\beta}) = N_{\beta} w_{\alpha\beta ij}$, is randomly chosen using a specific probability distribution $p_{\alpha\beta}$ with mean $w_{\alpha\beta}$ and characteristic function analytic at the origin. This we term the "strong randomness assumption." Then it can be proved [Geman, 1982], using the properties of random matrices, that all of the neurons in the module M_{α} converge to the same activity level asymptotically in time, equal to their average value. The set of mean activity levels $\{u_{\alpha}(t)\}$ in the αth module satisfy the reduced coupled equations

$$\tau_{\alpha} du_{\alpha}/dt = -u_{\alpha} + \sum w_{\alpha,,\beta} f_{\beta}(u_{\beta}) + I_{\alpha}(t) \qquad (8a)$$

and for each i,

$$u_{\alpha i}(t) \rightarrow u_{\alpha}(t) \qquad (8b)$$

as $t \rightarrow \infty$. We can understand (8a) and (8b) informally as arising from the fact that there is one eigenvalue of the matrix $W_{\alpha\beta}$ of order N_{β}, and all the others remain finite as N_{β} increases. If Eqs. (8) are considered in the asymptotic time domain, they then reduce to the system

$$u_{\alpha} = \sum w_{\alpha,,\beta} f_{\beta}(u_{\beta}) + I_{\alpha} \qquad (9)$$

These equations are identical to those suggested in [Horwitz, 1990] to simulate the correlation features obtained in earlier studies of Horwitz and his colleagues. Thus, we see that one form of bridging assumption is that of random coupling both between modules and internally in each module (plus a further technical assumption).

This result is satisfactory for a particular task, but gives no hint as to how different modules are recruited separately into the networks of modules used for different tasks. Thus, ventral and dorsal visual modules are involved in object and spatial visual tasks, respectively [McIntosh et al., 1994]. Equations (9) would not be able to distinguish between this differential involvement of each of these areas. Another example is the asymmetry between the modules used in encoding and retrieval in various memory tasks [Krause et al., 1999; Tulving et al., 1994]. Such selectivity of different modules across tasks is not able to be obtained by such a strong postulate as random connectivity; connection strengths are not task-independent. Of course different inputs to the modules can help in resolving such "routing" problems, although that will not help in the case of V1, for example, which has visual inputs that are processed differently for spatial processing (into the dorsal path) and object processing (into the ventral path) [Horwitz et al., 1992b]. Thus, numerous connection strengths vary across these two paradigms, in spite of similar visual inputs.

The implications of the postulate are even less appealing on closer inspection. The system of Eqs. (8) has the activity of each module reduced to that of a single equivalent "mean neuron." There is no other information contained in the total activity. Such reduction is too strong; it destroys the selective response of each module to information it is receiving from earlier (and later) modules, to which it can respond selectively according to the overall task and the inputs. It may be true that the mean activity of a module is roughly independent of which neurons are active, so that averaging over the active neurons in a net leads to Eq. (8) [and hence (9)]. But that is little justification for lumping all of the activities of the neurons together in the first place. In any case, we accept that there is no evidence for such random connectivity between cortical modules; it is known that there is, however, a rough preservation of topography in several modules. Thus, a limited form of random connectivity may be allowable. leading to averaging over localised neural activity. We will not consider that in any detail further, but assume that such averaging as is allowed has been performed.

The result (9) encapsulates, modulo the question of the relation between neural activity and blood flow (to which we turn in the next section), all that is contained in present covariance structural equation models of brain imaging data. However, as noted above, the strengths of connections between modules are modulated by the task at hand. How can that be achieved without loss of the insight provided by Eq. (9)? We turn to discuss weakening of the assumptions leading from (6) to (9).

Mean field equations

One of the important ingredients of the above reduction of randomly coupled neural modules to the structural Eqs. (9) is that the weights are scaled by the factor $(1/N_\beta)$ and the largest eigenvalue of such random matrices is the value $w_{\alpha\beta}$ with associated eigenvector **1,** the vector with unity in all its components. This was used in Geman [1982] to prove almost surely the existence of a strong form of "local chaos," that for large modules there is independence of the activity variables $u_{\alpha i}$ among each other and with the connection strengths $w_{\alpha i\beta j}$. Local chaos has also been proved by using a dynamical mean-field theory in the analytical framework of functional integration [Cessac et al., 1994]. However, local chaos is still too strong for the assignment problem we raised above, since it allows ensemble-average equations to be deduced which are independent of the detailed task in which the modules are involved. They only incorporate the mean and variance of the distributions by which the connection weights were picked, as in the previous case. This is still true in the case of zero mean, when the variance of the distributions must behave like $1/N$ for large number N of neurons in each module in order to have sensible results [Cessac et al., 1994].

In total then, the mean field or average approach also does not avoid the assignment problem we face.

Pattern storage networks

Different tasks are specified by the different patterns which activate a given module. This is a feature we term "pattern separation." Inclusion will now be made in the connection matrices of the patterns that have been stored by the module in the past so as to act as a filter on inputs. To do this initially, only the linearised version of the response function f of the neurons will be taken, so that Eq. (6) is a linear one in the unknowns u. The form of the weight matrix will be assumed of Hebbian form [Hebb, 1949], for purpose of analysis,

$$w_{\alpha i\beta j} = \sum w^{\mu}_{\alpha\beta} u^{\mu}_{\alpha i} u^{\mu}_{\beta j} \tag{10}$$

where the summation in (10) is over the pattern labels μ, assumed finite in number, and the quantities $w^{\mu}_{\alpha\beta}$ are constants. The pattern vectors u^{μ}_{j} are assumed to be binary patterns. The form (10) arises, as is well known, from a large class of Hebbian learning laws, and is the simplest form of connectivity in which pattern separation occurs. We will discuss later a more general approach to pattern separation. We now assume that a given task involves a certain subset of the patterns that have been stored in the connection weight matrices, different tasks having different subsets of the patterns crucially involved in their solution. The quantity of interest for the neuronal activations in a given module being measured by brain imaging systems, in the task with pattern set μ, is the projection of total activity in the module onto a given pattern:

$$\textstyle\sum_i u_{\alpha i} u^{\mu}_{\alpha i} = v^{\mu}_{\alpha} \tag{11}$$

where the summation is over the neurons i of the module α. This quantity is the coordinate of the neuronal activity in the coordinate system given by the input patterns. For only one pattern μ, then only v^{μ}_{α} will be measured by brain imaging, for a given module α. If a number of patterns are involved in a given task, then the relevant quantity of interest is the mean of the v^{μ}_{α} over the set of relevant values of μ.

The resulting network equations for the quantities v^{μ}_{α}, from (7) and (11), is

$$v^{\mu}_{\alpha} = \sum_{\beta,v,j} w^{v}_{\alpha\beta} (\mathbf{u}^{\mu}_{\alpha} \mathbf{u}^{v}_{\alpha}) v^{v}_{\beta j} f(u_{\beta j}) + I^{\mu}_{\alpha} \tag{12}$$

where $\mathbf{u}^{\mu}_{\alpha}$ is the vector with components $u^{\mu}_{\alpha i}$. If we assume (1) that the patterns are orthogonal, and (2) the neuronal response function $f()$ is approximately linear, then Eq. (12) reduces to the previous form of Eq. (9),

$$v^{\mu}_{\alpha} = \sum_{\beta} w^{\mu}_{\alpha\beta} v^{\mu}_{\beta} + I^{\mu}_{\alpha} \tag{13}$$

where (13) is now in the form of a linearised structural model. Assumption (1) is not drastic; assumption (2) may be so, but without it the reduction to (13) does not seem possible, nor a closed form of CSEM result.

The connection matrix $w^{\mu}_{\alpha\beta}$ now has the pattern label attached, as needed to solve the module assignment problem present in Eq. (13), which we already raised. For different tasks, there is freedom in the manner in which the connection matrix may be chosen. However, if tasks involve the same patterns, the connection matrices must be identical across the tasks; if tasks involve some common patterns, then conditions are imposed on the possible values of the connection matrices.

We may weaken the orthogonality assumption by only requiring that the patterns are mutually orthogonal between pattern subsets required for different tasks. Further analysis can be performed to obtain a similar equation to (13); we will not pursue this matter further here.

The derivation of Eq. (13) is more satisfactory than by use of random connectivity, but it involves an assumption that we would like to remove: that we can divide patterns into subsets, with all inputs belonging to one or other of these subsets. That is too limiting; patterns may be divided up in this way but will contain individual features, such as colour or shape, which are also possessed by patterns in other categories. Modules in the brain, such as V4, code for such more general features. Thus, we need to remove such a Procrustean form of pattern separation.

Hierarchical processing

It is now assumed, less strongly than in the previous section, that pattern categorisation occurs in terms of lower order features. This can be achieved on going from one net to another, by the set of labels for the higher category net being a reduction of those for the lower one by not containing all of the features of the lower net. Thus, if the higher module has labels μ and lower one labels μ and v, the labels v can be considered as features of members of the categories labeled by μ. More generally, the weight matrix is assumed to be, extending (10),

$$w_{\alpha i \beta j} = \sum_{\mu, v} w^{\mu v}_{\alpha\beta} u^{\mu}_{\alpha i} u^{v}_{\beta j} \tag{14}$$

where the pattern labels μ and v each belong to sets appropriate for the past experience of their specific module. Such an approach leads to an extension of the CSEM Eq. (13), in the form

$$v^{\mu}_{\alpha} = \sum_{\beta v} w^{\mu v}_{\alpha\beta} v^{v}_{\beta} + I^{\mu}_{\alpha} \tag{15}$$

When the module on the left-hand side of (15) is a categorising one (so only having the labels μ), summation on the right will be over extra labels attached to the weights w corresponding to extra features. This corresponds solely to a categorisation process. Lateral connections between modules at the same level of the processing hierarchy can be included by assuming that the original Eq. (13) is valid in that case. Feedback from a higher to a lower level of categorisation may occur, so as to make precise the categorisation at the higher level by support from lower level activity. Evidence indicates that feedback has a modulatory effect [Friston et al., 1995], which would be included by nonlinearity on the right-hand side of (15) (along the lines of "sigma-pi" networks, which we will consider toward the end of the paper). That does not change the basic feature of (15), which is that different pattern sets are separately involved in different tasks.

So far, we have assumed that the projections v^{μ}_{α} are directly measurable as summed neural activity. That can be best justified by assuming there are lateral connections internally in each module that cause relaxation to the relevant pattern projection. It will then be the case that the projections are the neural activations being directly measured as creating blood flow.

The above has assumed that processing occurs in a hierarchical manner. This may be roughly valid for vision and somatosensation. In the frontal lobes, it is also valid for motor actions, with motor cortex being at the lowest level of the cortical processing hierarchy. However, the levels of various sites in prefrontal cortex are difficult to assess. If we follow Pandya and colleagues [Pandya and Yetarian, 1990; Petrides and Pandya, 1994], then we may still obtain a rough hierarchy of processing modules, with coupling from posterior to anterior cortical sites at comparable levels in the hierarchy.

The above also assumes that only projections of neural activity onto the original patterns experienced in the past by a module occurs. That will be true if there is suitably fast relaxation to the relevant pattern, as in an attractor net; this may occur by aid of feedback from higher modules as well as by lateral connections in the module itself. We assume that such feedback occurs to achieve relaxation fast in comparison to the haemodynamic time constants.

At the same time, new learning occurs, as is clear in episodic memory tasks such as in [Krause et al., 1999]. There is also semantic and procedural memory updating. These have not been explicitly included here, although this can be done by the addition of learning laws for the neural connection weights, with resulting learning laws filtering through into the CSEM system of equations; these features are beyond the confines of this paper, moving the CSEM system of equations into the time domain.

To summarise, a set of bridging rules has thereby been obtained to give CSEMs of form (15). The weight matrix between a set of active modules has been identified as the sum, over the distinct sets of patterns, of the connection strengths obtained by use of the CSEMs (15) for the different tasks involving these patterns; the label μ, v are to be regarded as a complex of variables denoting categories and features of the members of those categories, to which incoming activity has relaxed. The basic problem to be faced in using this set of CSEMs will be to determine the sets of pattern labels relevant for a given task across the active modules. This will have to be achieved by careful breakdown of the psychological paradigm involved into its components and resulting identification of the relevant patterns sets at different processing levels. Variations on these must then be tried, using significance level testing, described briefly, to select the most appropriate pattern set hierarchy to fit the data.

Conclusions

At this point, it is useful to give partial conclusions, so as to help indicate where we need to go next, as well as consider alternative derivations. In the previous subsections, we faced up to the problem of deriving CSEMs from the underlying neural Eqs. (7). We found that it was possible to derive them under the assumption of random coupling between blocks of the modules all involved in the same type of processing. The neurons in each of the blocks were then assumed to be coupled randomly to others in the same blocks in other modules (or to themselves). This led to CSEMs for the particular type of processing in which parts of each of the modules are connected to other similar components. The assumption of modules being decomposed into blocks dedicated to the same task may not be too dangerous if the components are considered as different cortical layers. However, in the case, for example, of vision, where in V1 there is some separation of processing to the ventral and dorsal streams, the components of the modules are not in separate parts of the brain, and will lead to summed, and therefore confusable, blood flow signals. To get around this problem, we extended the pattern storage approach to allow for different pattern sets stored in the relevant modules; to achieve activation of the relevant projections, we assumed that relaxation occurred to the patterns already stored in each module.

The results we have derived in this section appear to be specific to the coding scheme used. However, we have derived the CSEMs under a range of such schemes: random connections (suitably scaled), block connections across modules, pattern separation, either with or without hierarchies. All of these different schemes were arrived at by different assumptions and led to differing CSEMs according to the nature of the dependence of the connections strengths on the task. Thus, all of these different CSEMs can be used to determine the best fit model to a set of data. Because the pattern separation model has the best biological basis, it is expected to be most appropriate, but it should not necessarily bias the data analysis.

So far, we have only obtained CSEM-like equations for the underlying averaged neural activity. We next analyse blood flow and its relation to neural activity in more detail, to see whether the preliminary CSEMs preserve their appearance.

STRUCTURAL MODELS: THE BLOOD FLOW APPROACH

The blood flow signal: Only excitatories

In the preceding sections, we developed an approach to deriving covariance structural equation models of neural activity and the associated bridging laws. This was based on a simplified system of suitably coupled neurons, with output described by the neurons' mean firing rate. For modules coupled in a random fashion, with the elements of the weight matrix between the modules each chosen from a suitably scaled random distribution, the neurons in each module tended with time to respond in an identical manner. This led to too simple a version of a structural model, in which there was no distinction in given modules as to the various tasks they have been genetically or adaptively designed for. An extension of the weight matrices to incorporate such task-related distinctions for modules was then developed. This led in the previous subsection to hierarchical models that were shown to lead to suitable structural equation systems.

We now extend the preliminary CSEMs so as to incorporate blood flow, which does not notice the labels of neurons. At this juncture, Eqs. (13) or (15) could thereby better apply to MEG data than to PET and fMRI data. For MEG, it is the neurons themselves that are the source of the signal, as has been used so far in our analysis (although we still need to include the direction of apical dendrites to give a first approximation to the direction of the underlying local current dipole source current to obtain magnetic fields).

The signal for the blood flow is a localised spatial and temporal average of some function F of the membrane potentials $u_\alpha(\mathbf{r}, t)$, where we are now labelling the neurons by their position r on the cortical sheet as well as the module to which they belong. The signal being detected by PET or fMRI in module α at position $\mathbf{r}$ at time t is therefore of the form

$$s_\alpha(\mathbf{r}, t) = \sum_{r' \in N(r), t'} F[\mathbf{u}_\alpha(\mathbf{r}', t'), \mathbf{r} - \mathbf{r}', t - t'] \quad (16)$$

where the spatial dependence of F delineates the neighbourhood $N(\mathbf{r})$ of the position $\mathbf{r}$ over which $\mathbf{r}'$ is summed, and will depend on the amount of spatial smoothing caused by the signal spread [for example, by nitric oxide, as discussed in Krekelberg and Taylor, 1996] from neuronal activity calling for enhanced blood flow, and also that performed by the measuring device, as well as that intrinsic in the distribution of capillaries and arteries. It also depends on the anatomical choice of region. That there is a close relationship between blood flow and neuronal activity has been demonstrated recently for the rat whisker barrel cortex in [Yang et al., 1997]. Thus, the neighbourhood determined by the spatial spread in the function *F* is expected to be local.

The temporal averaging contained in (16) is also determined by the temporal form of *F*. Various forms have been suggested, such as a Gaussian, a Poisson distribution [Friston et al., 1995], or a triangular form, as in the balloon model of [Buxton et al., 1998]. This has been extended by Glover [1999] to be of trapezoidal shape to obtain a better fit to fMRI data on finger tapping at a variety of speeds. Moreover, in this latter work, nonlinearities were uncovered, as they were in the study of Friston et al. [1998]. These various forms for *F* will have to be used to obtain the optimal fit to the haemodynamic response to fast repetition.

A simple approximation to *F*, for slow repetition of stimuli (allowing linearity of *F* in *u*) is Glover [1999]:

$$F(u, \mathbf{r}, t) = \left[\sum_j c_j t^{n(j)} \exp(-t/t_j)\right] u \qquad (16a)$$

for suitable constants c_j and t_f and integers $n(j)$, with values as, say, suggested in Glover [1999]. We next take account of the spatial averaging occurring on the left-hand side of (16), using (16a), to replace $u(r')$ by its averaged value u_α in the module for the particular pattern set α. This label will then be handed on to the blood flow, leading to the observable s_α as the following function of the neuronal activity:

$$s_\alpha = [\sum_j c_j t^{n(j)} \exp(-t/t_j)] u_\alpha \qquad (16b)$$

Excitatory and inhibitory neuronal populations

So far, we have not tried to distinguish between excitatory and inhibitory neurons. The only manner in which inhibition could enter Eqs. (7) would be by negative values for the relevant weights from a given inhibitory neuron; in a similar fashion, excitatory neurons would only have positive weights to all other neurons. However, this in reality divides up the neurons into two populations; from now on, we will do that more explicitly. This will allow us to include more specific features of the local inhibitory neuron population, and at the same time attempt to be more realistic in the manner in which inhibition is achieved in reducing neuronal activity (although not necessarily the resulting blood flow). In this dual population approach, the rCBF/BOLD signal is related, at the linear level, to the weighted sum of the synaptic activity arriving at a neuron from both the excitatory and inhibitory populations. It has been suggested that inhibition is treated in the sum with the same (positive) weight as excitation [Arbib et al., 1995; Jueptner and Weiller, 1995; Tagamets and Horwitz, 1998], although few experiments have been performed to test this; the few good ones have supported this view [see Jueptner and Weiller, 1995; Horwitz and Sporns, 1994 for reviews]. Given this uncertainty, the excitatory and inhibitory neurons should more correctly be handled as hidden variables, with unknown measurement functions. If *u* and *v* denote their respective membrane potentials, then (16) must be extended to

$$s_\alpha(\mathbf{r}, t) = \sum_{r',t'} F[u_\alpha(\mathbf{r'}, t'), v_\alpha(\mathbf{r'}, t'), \mathbf{r} - \mathbf{r'}, t - t'] \qquad (17)$$

where the measurement function *F* is of unknown form. The quantities $s_\alpha(\mathbf{r}, t)$ on the left-hand side of Eq. (17) are now the observables, whereas the membrane potentials $u_\alpha(\mathbf{r}', t')$, $v_\alpha(\mathbf{r}', t')$, are the hidden variables. The summation in (17), similar to that in (16), is over a suitable neighbourhood around the measurement centre, determined by the dynamics of blood flow and of the smoothing caused by the measuring system. This can also be replaced by a simpler expression, corresponding to (16a), by adding a similar expression to the left-hand side of (16a) with *u* replaced by *v*:

$$F(u, \mathbf{r}, t) = [\sum_j c_j t^{m(j)} \exp(-t/t_j)] u + \left[\sum_j d_j t^{n(j)} \exp(-t/t'_j)\right] v \qquad (17a)$$

It is the expressions (17), (17a) which should be used to determine the observable effects of underlying neuronal activity.

CSEMs for populations of excitatory and inhibitory neurons

Having introduced the populations of separate excitatory and inhibitory neurons and their related observable blood flow effects, we must now develop the appropriate neuron dynamics for them to be able to deduce CSEMs for them. The dynamical equations involve the underlying set of neural variables $u_{\alpha i}(r', t')$, $v_{\alpha i}(r', t')$, satisfying the extension of Eq. (7) (assuming only linear output response for the inhibitory neurons) to

$$\tau_{\alpha i} du_{\alpha i}/dt = -u_{\alpha i} + \sum w_{\alpha i,,\beta j}\, f(u_{\beta j}) - \sum w'_{\alpha i,,\alpha j} v_{\alpha j} + I_{\alpha}(t) \tag{18a}$$

$$\tau_{\alpha i} dv_{\alpha i}/dt = -v_{\alpha i} + \sum w''_{\alpha i,,\beta j}\, f(u_{\beta j}) - \sum w''_{\alpha i,,\alpha j} v_{\alpha j} + I_{\alpha}(t) \tag{18b}$$

where the connection coefficients are all positive. Dropping the time constant terms as contributing only a small amount (as in the excitatory case previously), and assuming that only local connections arise from the inhibitory neurons, these equations reduce thereby to

$$u_{\alpha i} = \sum w_{\alpha i,,\beta j} f(u_{\beta j}) - \sum w'_{\alpha i,\alpha i} v_{\alpha i} + I_{\alpha}(t) \tag{19a}$$

$$v_{\alpha i} = \sum w''_{\alpha i,,\beta j} f(u_{\beta j}) - \sum w'''_{\alpha i,,\alpha i} v_{\alpha i} + I_{\alpha}(t) \tag{19b}$$

Under similar assumptions to those made in the previous section, Eqs. (19) can be reduced, assuming linearity of the output functions *f*(*u*) of the excitatory neurons, to relations involving only neuronal activities associated with a given pattern class, and lead to the coupled CSEMs

$$u_{\alpha} = \sum w_{\alpha,\beta} u_{\beta} - w'_{\alpha,\alpha} v_{\alpha} + I_{\alpha}(t) \tag{20a}$$

$$v_{\alpha} = \sum w''_{\alpha,,\beta} u_{\beta} - w'''_{\alpha,,\alpha} v_{\alpha} + I_{\alpha}(t) \tag{20b}$$

where the appropriate extra pattern labels on the averaged membrane potentials have been dropped for simplicity. To these must be added the projection of (17a) as relating the observables and the underlying neural activity:

$$s_{\alpha} = [\sum_j c_j t^{m(j)} \exp(-t/t_j)] u_{\alpha} + [\sum_j d_j t^{n(j)} \exp(-t/t'_j)] v_{\alpha} \tag{17b}$$

Equations (17b) and (20) now form the more complete set of CSEMs, with latent variables equal to the averaged separate excitatory and inhibitory neuron membrane potentials.

It is this form which is importantly different from the basic CSEMs referred to previously and used so far in analysing brain imaging. In this new form, the averaged neuronal membrane potentials are completely hidden; only the blood flow variables s_{α} are observable. This is as it should be; we cannot assume that the neuronal activities are directly observable, and indeed they clearly are not, as seen in Eq. (17a). Only a suitable linear combination of them, averaged with a kernel function over past time, is observable. Thus, the past results obtained by applying the simple versions of CSEMs to brain imaging, with no hidden variables, need to be redone using the present set of CSEMs. The answer to question 3 of the Introduction is thus a definite yes in the extreme. There are no directly observable neuronal variables in the CSEMs. Our search for bridging laws has thus led to the important result that the land on the CSEM side of the bridge needs to be altered completely.

It is possible to eliminate the inhibitory neuron variables in (20), but the resulting CSEMs for the excitatory averaged membrane potentials are then highly nonlinear in the underlying weight matrices, so that the forms (20) appears to be more useful. It is from them that the measured connection strengths can be immediately interpreted as the elements of the weight matrices of the original dynamical neuronal Eqs. (18). It is to be noted that, for a total of *M* modules, the number of free parameters in (20) is $M(M + 1)$, as compared to only $(1/2)M(M - 1)$ in (13). However, the increase in number of parameters must be accepted to give a proper account of the inhibitory neurons.

In addition, we point out that the synaptic weight matrices that are allowed to enter Eqs. (20) must have only nonnegative entries, an important extra constraint beyond that usually entering CSEM approaches to brain imaging. It has usually been considered that negative connection strengths arise from the inhibitory effects of one module on another; these are seen in several of the CSEM models produced so far as long-range effects, such as in Horwitz et al. [1992a]. However, there is no evidence, except for the basal ganglia, that inhibitory neurons have long-range axons. The only way to interpret such cortical inhibition is by means of long-range excitation onto inhibitory interneurons. However, because these interneurons are only locally connected, their contribution would not be properly included in the CSEMs used so far. Only through their indirect effects on the excitatory population, as described in Eqs. (20), and their direct effect on the blood flow as specified in Eq. (17b), can these opposing characteristics (that of inhibition both reducing neural activity yet increasing blood flow) be properly disentangled.

Nonlinear modulation effects can be included in the above discussion by extending the neuronal equations to quadratic or even more nonlinear neurons jointly in a number of inputs, as described earlier as sigma-pi neurons, by replacing (6) by:

$$u_{\alpha i} = \sum w_{\alpha i,,\beta j} u_{\beta j} + \sum w_{\alpha i,,\beta j \gamma k} u_{\beta j} u_{\gamma k} + \ldots + I_{\alpha i}(t) \quad (21)$$

where the dots indicate even higher-order terms than quadratic in neural activities. Such nonlinear effects can be carried over to structural equations by assuming an extended form of (10) in the case of pattern storage networks, with resulting reduction of the system of equations to the relevant mean pattern activities over each module. We may combine the sigma-pi CSEMs resulting from (21) with the blood flow observables (16), or extend the former to include separate inhibitory and excitatory populations and use (17b) to determine the blood flow. In all of these nonlinear extensions of CSEMs the neuron variables are always hidden, and the blood flow values are always the observed variables. As before, this gives a completely different form of CSEMs to those in present use in analysis of brain imaging data.

Recapitulation

Let us recapitulate the assumptions that have been used to arrive at these structural equations: (a) blood flow is proportional to some (in general nonlinear function) of the strength of the neuronal membrane potential averaged over a region about the central point **r,** and also averaged over the time from onset of the neural activity, in a time-translation invariant manner, (b) the dependence of blood flow on the underlying membrane potential is linear, and (c) the connection strength between neurons is roughly constant over the neighbourhoods of the neurons being summed over in the blood flow averaging process (16).

The first assumption has experimental support from the work of Glover just cited; that of (b) also has support provided the stimulus repetition rate is not too high. Assumption (c) is satisfactory if the neighbourhood is chosen small enough, such as a column in V1 or a single whisker barrel in layer 4 of somatosensory cortex [Yang et al., 1997]. It may even be valid for several columns in any region of the cortex if they are coding for very similar features. Thus, we expect this neighbourhood to have a diameter of about 300 μm or thereabouts. For a high enough applied magnetic field, such as at 7 T [Yang et al., 1997], it turns out that, using a suitably high t or z threshold, only such a small region was observed as activated; when the threshold was lowered, the neighbourhood observable contained several whisker barrels, and thus assumption (c) would not have been satisfied. In that case, it is expected that the connection strength arising from the CSEM analysis would correspond to an average of different underlying neural connection weights.

In conclusion, we see that assumptions (a) to (c) are satisfied under certain conditions and define the path coefficients by Eqs. (15) for a fused excitatory/inhibitory population [or (17b) and (20) in the case of separate excitatory and inhibitory neuronal populations] from the underlying synaptic weights. In other conditions, the CSEM connection strengths are only equal to an average over the underlying synaptic weights.

EXTENSIONS TO THE TEMPORAL DOMAIN

The main result we have reached are Eqs. (15) or (20) for the solely excitatory or for the joint inhibitory and excitatory populations, respectively. These are the respective equations of covariance structural equation modelling, the latter in particular with the neuronal activity generating hidden variables. To these must be added the observation equations; for fMRI and PET there are the various forms of (16), as, for example, (16a) and (17), as in (17b), respectively. Moreover, this reduction from the underlying neural network equations was only possible under some stringent conditions, as noted in previously.

Given that the CSEM equations are used to analyse brain imaging data in a regime where they satisfy the above conditions, what can be done with the resulting connection strengths to relate to the faster temporal dynamics of MEG and EEG? It would appear on the surface to be a relatively easy task to put time back into the CSEM Eqs. (15) or (20) [Taylor et al., 1997]. This could be done by two main steps: (1) Reintroducing time delays between the modules, because of the finite speed of axonal transmission. This corresponds to taking the time delays $t_{\alpha\beta}$ in the appropriate term on the right-hand side of those equations:

$$u_\alpha(t) = \sum_\beta M_{\alpha\beta} u_\beta(t - t_{\alpha\beta}) \quad (22)$$

where only the original form of CSEM is considered with a single population of neurons, for simplicity, with only one potential u for each module, and M denoting the CSEM connection strengths between the modules. (2) Reintroducing the neuronal time decay of activity in each module, so giving

$$\tau_\alpha du_\alpha(t)/dt = -u_\alpha(t) + \sum_\beta M_{\alpha\beta}u_\beta(t - t_{\alpha\beta}) + I_\alpha(t) \tag{23}$$

which is now appropriate to analyse in terms of MEG data.

To this must be added the resulting observable equations. These will depend on the detailed geometry of each of the modules, with orientation of the cortical surface or more generally geometric distribution of the apical dendrites in a module. If a neuron at position **r** in module α has apical dendrite along the vector **n**(**r**), then the current distribution leading to resulting magnetic or electric fields outside the head or on the scalp due to the activity u_α in module α will be

$$\mathbf{j}(\mathbf{r}, t) = u_\alpha(t)\mathbf{n}(\mathbf{r}) \tag{24}$$

It is then straightforward to arrive at the measurement value d_m at the mth coil of an MEG machine, with lead field $\phi_m(\mathbf{r})$, as

$$d_m(t) = \int \mathbf{j}(\mathbf{r}, t)\phi_m(\mathbf{r})\, d\mathbf{r} \tag{25}$$

in the standard manner [Taylor et al., 1999]. A similar result holds for EEG measurements replacing the magnetic lead field by the associated electric lead field. In the case of separate excitatory and inhibitory populations, the quantity to be used on the right-hand side of (24) in place of $u_\alpha(t)$ is solely the total membrane potential of the relevant excitatory population. We claim that the underlying neuronal Eqs. (7), or (18) in the case of coupled populations, together with the measurement Eq. (25) [with (24)], leads to a similar set of CSEMs of form of (23), or the related coupled excitatory and inbibitory ones, now with more sensitivity in the time domain than in the blood flow cases. In other words, this gives an extension of the CSEMs using fMRI/PET values of the connection strengths into the EEG/MEG domain.

Further temporal extensions can also be achieved by building into Eq. (23) a number of the possibilities noted previously and left out of the analysis so far (because of insensitivity of the heamodynamic response function to time courses somewhat less than a second): temporal features of ionic channels and synaptic responses, neuromodulatory effects, temporal features associated with complex cell geometry. These can explicitly and straighforwardly be inserted into the temporal CSEM (23), using the same connection strengths as deduced from the CSEM (18) (although we will not do that here explicitly).

There are, however, some temporal features noted that are more difficult to incorporate into the temporal CSEM (23), in particular, that of oscillatory coupling between neurons. This has been observed as important across modules in various tasks, both in awake and sleeping subjects [see, for a review of recent work, Ritz and Sejnowski, 1997]. Oscillations may be intrinsic to a single neuron or arise from coupling in neural populations, as in hippocampus [Kocsis et al., 1999; Traub et al., 1999]. MEG analyses have been developed which explicitly try to separate such oscillations [Tass et al., 1998]. Coupled oscillations across modules have also been observed by EEG, as demonstrated in Desmedt and Tomberg [1994].

The question is thus: how do we carry the CSEM connections strengths arising from fMRI/PET measurements into the domain of oscillatory MEG/EEG effects? In particular, are these former connection strengths the same as those to be used in models of oscillatory neuronal population dynamics?

The answer is unknown in terms of the possible processing complexities at the neuronal level that may be present, as spelled out previously, but the following is a possible simplified approach. It may be shown that, under suitable conditions on the relevant temporal features of delays, refractoriness and the timing patterns of excitatory and inhibitory inputs to a neuron, phase locking of oscillations will occur [Gerstner et al., 1996]. However, their discussion was concerned with spiking neurons. To be able to incorporate the greater sensitivity to time given to a network by spiking neurons, using the same CSEM, the neuronal response $f(u)$ must now be given by a suitable step function of the membrane potential u, with value

$$f[u, t] = 1\ (t_1 < t < t_1 + \delta, t_1 = \text{first } t \ni u(t) > \theta) \tag{26}$$

where δ is the width of the spike and θ the threshold of the response. There are numerous more complex spiking extensions of the mean firing rate model than to the leaky integrate-and-fire model of (26), but this nonlinearity is sufficient to give the sensitivity to the time of response of the neuron. Thus, an immediate extension of the temporal CSEM (23) to take account of spiking neurons would be to replace the weighted input $u_\beta(t)$ from the module β by the function $f[u_\beta, t]$. For the dual populations of excitatory and inhibitory neurons, such a system of coupled spiking neurons may satisfy the conditions for the production of synchronisation. The temporally extended CSEM model is to be constructed with the same connection strengths that have been measured by fMRI/PET, assuming that all the active modules were coupled by oscillations.

A more complete answer to this question can only be given by also showing that it is possible, using these arguments, or their extensions, to obtain the same temporal CSEMs starting from the underlying population dynamics of (18) (for the dual population case) but with the output response functions *f*() for the excitatory neurons in (18) given by (26). The crucial feature used in these pattern-reduction processes was that of linear dependence of the outputs of all of the neurons on their membrane values. Equation (26) is decidedly nonlinear. How can we proceed in that case?

Spatial averaging over ensembles of neurons, in which there is a spread of time constants and delay times (so giving a corresponding temporal averaging), leads to the responses of coupled spiking neurons by their mean firing rates [Amari, 1972; Bressloff and Taylor, 1991; Sejnowski, 1977]. But that is what was used earlier, so all that apparatus can be brought into play, and the CSEM story developed for MEG/EEG to lead to the temporally extended version with linearised responses, of Eq. (23) (and its extension to the separate excitatory and inhibitory population case). Thus, the CSEM Eqs. (17b) and (20) can be obtained in this case, using the spatial averaging involved in the measurement process.

In summary, the bridging laws between underlying neuronal population dynamics and imaging measurement extend from those for fMRI/PET to MEG/EEG by inserting suitable temporal properties into the module activation variables. These latter have been deduced, under the assumptions spelled out earlier, from the projection of the neuronal activities onto the set of patterns used as part of the overall filter/relaxation processing of each of the modules. Provided the spatial ensemble averaging by EEG/MEG measurement also produces a temporal averaging then the reduction from spiking neurons (synchronised or not) will lead to mean firing rate neurons. The bridging laws developed in the mean firing rate case thus apply to the more realistic spiking neuronal populations.

SOURCES OF VARIABILITY

What are the sources of variability in the structural models we have arrived at? This is an important question which has been considered in some detail by a number of workers, as in Horwitz et al. [1992a] or in Tagamets and Horwitz [1998], and from a simulation point of view in [Horwitz et al., 1999]. From the present viewpoint, there are the overall noise sources which arise from the neglected external inputs introduced in Eqs. (7) and (7a). There are also internal sources of noise in the response of each neuron, as especially arising from synaptic noise because of quantal release probability. That we could (but will not here) model in detail [Taylor, 1972]. These are some of the sources of noise giving intrasubject variability during an experiment. Others arise from modulatory influences, such as from a varied level of attention or from emotional sources during an experiment. In particular, the modification of attention was shown to be an important determiner of connection strengths in the simulation [Horwitz et al., 1999]. Intersubject variability arises from differences between values of the synaptic strengths $w_{\alpha i \beta j}$ or $w_{\alpha\beta}(r, s)$. This is another important source of variation usually considered in structural analysis, and from the present point of view can have important effects in the brain imaging situation. This was found to be the most important determiner of connection strengths in the simulation mentioned earlier [Horwitz et al., 1999]. There can also be variability arising from the coupling between blood flow/BOLD and neural activity both between and across subjects as well as of the haemodynamic parameters (latency, spread). Similarly, when we turn to the temporal aspects of Eqs. (7), or the temporal version (23), the intersubject variation of the time constants τ_α will be important for related MEG and EEG analysis.

To summarize, all of the parameters entering the covariance structural equation models (9) (and its later extension) can vary both intra- and intersubject. Which of these variations is the most important is to be determined by further analysis along the lines we have started here in defining the meaning of the path coefficients in terms of the underlying neural network parameters.

SUMMARY AND DISCUSSION

We have developed a set of assumptions to be able to deduce from the underlying neural network dynamics of the brain a suitable set of equations to allow understanding of the connected networks of brain areas now being investigated using structural models. These assumptions are of three sorts, those for fMRI/PET and further ones for EEG/MEG, and (a) involve the nature of coding for object categories used in tasks being imaged, in terms of patterns learned in the past as filters for feedforward processing or by relaxation by recurrence, (b) involve the random character of the coding between modules and of the scaling of weights in terms of inputs, or incorporate "pattern separation" by means of neural weights depending on previously seen patterns which decompose into orthogonal sets with a hierarchical structure in the pattern labels, and (c) require suitable linearisation of single neuron responses (although extension to sigma-pi modulatory networks is possible under further conditions).

These assumptions have been shown to lead to the structural Eqs. (15) that connect a certain set of cortical modules under a task condition and involving categories of stimuli. The module activations had then to be related to observations of blood flow, and this was done by the discussions related to Eqs. (16). Under a further assumption of linearity of the relation between blood flow and neural activity, it was shown possible to deduce observables directly expressible in terms of the same reduced, averaged quantities as arose in the CSEMs from the underlying neural activity.

At the next stage of the analysis, which involved a more detailed discussion of the relation between blood flow and neural activity producing it, the need to split the neurons into two populations was made explicit, with the resulting neuronal Eqs. (20) and blood flow dependence on neuronal activity (17b) being made explicit. Various assumptions needed to reduce these equations to CSEM form were then considered, leading to the equations extending the CSEM (15) to the two-population case. Moreover, the resulting CSEMs are able to support the various pattern separation assumptions. However, surprisingly, the neuronal activation variables always remained hidden under these various assumptions, and would always need to do so; only the blood flow (or BOLD signal) is observable.

We thus obtained answers to the three questions asked in the Introduction. The first question (how to justify CSEMs in brain imaging) was that this is possible, although not in the present form that they are being used, except by (1) fusing neuron populations together, (2) neglecting pattern separation features, and (3) assuming that blood flow is linearly proportional to neuronal activity. The first of these assumptions (neglect of separate excitatory and inhibitory populations) prevents a proper understanding of the effects of inhibition. The second of the assumptions (neglect of pattern separation) means that the resulting connection strengths cannot be properly related across paradigms. The third assumption can only be made assuming fusing of excitatory and inhibitory populations; if these are separate, then a linear sum arises, leading to the neural activities still being hidden. This again prevents the interpretation of inhibitory and excitatory effects realistically. To avoid these various defects, it is to be deemed correct to use the more general two-population form (17b) and (20) of the CSEMs, using the pattern separation ideas illustrated earlier.

These arguments imply that two populations of separate excitatory and inhibitory neurons should be used, their activities consisting of completely hidden variables. The resulting synaptic weight matrix for the neural interactivity can then be directly extracted from the connection strengths of the corresponding CSEMs, provided the condition of nonnegativity of the neuronal weight matrix has been extended to the connection strengths of the CSEMs. In this case, question two (the interpretation of the connection strengths) can be achieved directly by equating the connection strengths from the CSEMs and the synaptic weights between the underlying neurons.

Finally, question three of the Introduction has been answered as part of this previous discussion: all neuronal activities are hidden and the blood flow/BOLD measurements are the only observables. In general, these will not even be linear functions of the underlying neural variables, as experimental data now shows [Friston et al., 1998; Glover, 1999]. Neglecting these nonlinearities, the resulting CSEMs are still quite distinct from the forms of CSEMs in present use. The message coming from the above derivation is strong: use the new forms (17a,) (20) so as to be able to relate the connection strengths to underlying neural synaptic weights; otherwise, the meaning of the connection strengths are unclear.

The extension of the above ideas from the "static" regime of fMRI and PET to the highly dynamic one of EEG/MEG was achieved by additional assumptions concerning the possibility of performing temporal averaging as arising naturally from spatial averaging, as well as developing synchronised activity between coupled modules from the extended temporal equations. The spatial averaging along with a certain amount of time averaging, avoided the incorporation of spiking characteristics of neurons. However, the oscillatory nature of neural activity across cortical modules is unclear in this approach; detailed modelling will have to be performed of coupled two-population modules (with suitably temporal ionic currents) to determine what natural coupling will arise in the time domain from modules coupled with the connection strengths obtained from the above CSEMs.

The main conclusion above all others that we have reached is that the most appropriate CSEM model to use, if some relation between the connection strengths and the weight matrices is to be achieved, involves separate populations of excitatory and inhibitory neurons. It is this result which is quite surprising, given the nature of CSEMs used in brain imaging so far. But the "new look" on the CSEM approach to such analysis that has arisen from our neural network approach indicates that more can be put into the detailed form of the equations at the same time as valuable insights can be extracted. One of these insights is that there is a justification of the CSEM approach to brain imaging from the underlying neural network dynamics.

This is an important result: it justifies the extraction of connections strengths as involved with the causal flow of activity between modules. This causal flow of activity may be lost following the various approximations of relaxation and the long-time limit have been made. But it is there underpinning and justifying (modulo the approximations) the CSEMs, a feature absent from some of the large range of applications of structural models discussed, for example, in Bollen [1989] and Loehlin [1992]. It allows the description of the resulting CSEMs in terms of a flow of information or "activity" between the various modules with nonzero path strengths. Such a "flow" cannot usually be justified for structural models; here, the neural underpinning gives us the right to use such a term. Thus, we can talk about weak or strong "information flow" from one area to another according to the strength of the path coefficient connecting the first to the second module.

The range of CSEMs having been deduced earlier depends on the nature of the neural coding assumed to be instantiated in each of the modules. This plethora of CSEM versions allows for the deduction from experimental data, by means of significance testing, of the proper coding scheme that is being used in a particular paradigm. Further, the various approximations being made are vulnerable to further analysis, and the corresponding CSEMs to development to more complex forms.

The other important result is that all the neural activity is hidden; only the blood flow (or the corresponding magnetic or electric field) are observable. For the latter cases, the observation equations are well defined from the geometry of the head and the measuring devices. For the former, it is still necessary to determine the nature of the functions *F* described previously.

The bottom line from this paper is the new framework for CSEMs in terms of Eqs. (17b) and (20); these make all neuronal activity hidden, and introduce local inhibitory activity to allow inhibition to be taken into account. The new framework should then allow a better attack be made on transforming from the slow temporality of fMRI/PET to the fast one of MEG/EEG, both with spatial averaging, but the latter with controlled temporal additions to the ionic currents and possible emergence of oscillatory coupling between modules.

ACKNOWLEDGMENTS

One of us (JGT) would like to acknowledge hospitality at the Institute for Medicine, Juelich where this work was mainly performed.

REFERENCES

Abbott LF, Varela JA, Sen K, Nelson SB. 1997. Synaptic depression and cortical gain control. Science 275:220–224.

Amari S. 1972. Characteristics of random nets of analog neurons-like elements. IEEE Trans Syst, Man Cybernet 2:643–657.

Arbib MA, Bischoff A, Fagg AH, Grafton SA. 1995. Synthetic PET: Analyzing large-scale properties of neural networks. Hum Brain Mapp 2:225–233.

Basilevsky A. 1994. Statistical factor analysis and related methods, New York: Wiley.

Bollen KA. 1989. Structural equations with latent variables, New York: Wiley.

Bressloff PC, Taylor JG. 1991. Discrete time leaky integrator with synaptic noise. Neural Networks 4:789–801.

Buxton RB, Wong EC, Frank RL. 1998. Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. Magn Res Med 39:855–864.

Cessac B, Doyon B, Quoy M, Samuleides M. 1994. Mean-field equations, bifurcation map and route to chaos in discrete type neural networks. Physica D 74:24–44.

Churchland PS. 1986. Neurophilosophy: Towards a unified science of the mind-brain, Cambridge, MA: MIT Press.

Desmedt JE, Tomberg C. 1994. Transient phase-locking of 40 Hz electrical oscillations in prefrontal and parietal human cortex reflects the processing of conscious somatic perception. Neurosci Lett 168:126–129.

Friston KJ, Ungerleider LG, Jezzard P, Turner R. 1995. Characterizing modulatory interactions between areas V1 and V2 in human cortex: A new treatment of functional MRI data. Hum Brain Mapp 2:211–224.

Friston KJ, Josephs O, Rees G, Turner R. 1998. Non-linear event-related responses in fMRI. Magn Res Med 39:41–52.

Geman S. 1982. Almost sure stable oscillations in a large system of randomly coupled equations. SIAM J Appl Math 44:80–95.

Gerstner W, van Hemmen JL, Cowan JD. 1996. What matters in neuronal locking? Neural Comput 8:1653–1676.

Glover GH. 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9:416–429.

Hebb DO. 1949. The organization of behaviour: A neuropsychological theory. New York: Wiley.

Horwitz B. 1990. Simulating functional interactions in the brain: A model for examining correlations between regional cerebral metabolic rates. Int J Biomed Comput 26:149–170.

Horwitz B. 1991. Functional Interactions in the brain: Use of correlations between regional metabolic rates. J Cereb Blood Flow Met 11:A114–A120.

Horwitz B, Sporns O. 1994. Neural modeling and functional neuroimaging. Hum Brain Mapp 1:269–283.

Horwitz B, Soncrant TT, Haxby JV. 1992a. Covariance analysis of functional interactions in the brain using metabolic and blood flow data. In Advances In: Gonzalex-Lima F, Finkenstaedt T, Scheich H, editors. Metabolic mapping techniques for brain imaging of behavioural and learning functions (NATO Advanced Research Workshop). Dordrecht, The Netherlands Kluwer Academic Publishers, p 189–217.

Horwitz B, Grady CL, Haxby JV, Ungerleider LG, Schapiro MB, Mishkin M, Rapoport SI. 1992b. Functional associations among human posterior extrastriate brain regions during object and spatial vision. J Cogn Neurosci 4:311–322.

Horwitz B, Long TW, Tagametz MA. 1999. The neurobiological substrate of PET-fMRI functional connectivity. 5th International Conference on Functional Mapping of the Human Brain, Duesseldorf, Germany. NeuroImage 9:S392.

Jueptner M, Weiller C. 1995. Review: Does measurement of regional cerebral blood flow reflect synaptic activity? Implications for PET and fMRI. NeuroImage 2:148–156.

Kocsis B, Bragin A, Buzsaki G. 1999. Interdependence of multiple theta generators in the hippocampus: A partial coherence analysis. J Neurosci 19:6200–6212.

Krause BJ, Horwitz B, Taylor JG, Schmidt D, Mottaghy FM, Herzog H, Halsband U, Mueller-Gaertner H-W. 1999. Network analysis in episodic encoding and retrieval of word-pair associates. Eur J Neurosci 11:3293–3301.

Krekelberg B, Taylor JG. 1996. Nitric oxide in cortical map formation. J Chem Neuroanat 10:191–196.

Loehlin JC. 1992. Latent variable models. Hillsdale, NJ: Lawrence Erlbaum Associates.

Markram H, Tsodyks MV. 1996. Redistribution of synaptic efficiency between pyramidal neurons. Nature 382:807–810.

McClelland JL, Rumelhart DE, The PDP Research Group. 1986. Parallel distributed processing. Cambridge, MA: MIT Press.

McIntosh AR, Gonzalez-Lima F. 1994. Structural equation modeling and its application in functional brain imaging. Hum Brain Mapp 2:2–22.

McIntosh AR, Grady CL, Ungerleider LG, Haxby JV, Rapoport SI, Horwitz B. 1994. Network analysis of cortical visual pathways mapped with PET. J Neurosci 14:656–666.

Pandya DN, Yeterian EH. 1990. Architecture and connections of cerebral cortex: Implications for brain evolution and function. In: Scheibel AB, Wechsler AF, editors. Neurobiology of higher cognitive function. New York: Guilford Press, p 53–83.

Petrides M, Pandya DN. 1994. Comparative architectonic analysis of the human and the macaque frontal cortex. In: Boller F, Grafman J, editors. Handbook of neuropsychology. P Amsterdam: Elsevier Science B-V. 17–58.

Ritz R, Sejnowski TJ. 1997. Synchronous oscillatory activity in sensory systems: New vistas on mechanisms. Curr Opin Neurobiol 7:536–546.

Sejnowski TJ. 1977. Storing covariance with non-linearly interacting neurons. J Math Biol 4:301–321.

Tagamets MA, Horwitz B. 1998. Integrating electrophysiological and anatomical data to create a large-scale model that simulates a delayed match-to-sample human brain imaging study. Cereb Cortex 8:310–320.

Tass P, Rosenblum MG, Weule J, Kurths J, Pikovsky A, Volmann J, Schnitzler A, Freund H-J. 1998. Detection of n:m phase locking from noisy data: Application to magnetoencephalography. Phys Rev Lett 81:3291–3293.

Taylor JG. 1972. Spontaneous behaviour in neural networks. J Theor Biol 36:513–528.

Taylor JG, Ioannides AAS, Krause B, Mueller-Gaertner H-W. 1997. Structural equation modeling in time: An across-instruments approach. NeuroImage 5:S448.

Taylor JG, Ioannides AA, Mueller-Gaertner W-M. 1999. Mathematical analysis of lead field expansions. IEEE Trans Med Imaging 20:1–15.

Tononi G, Sporns O, Edelman GM. 1992. Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. Cereb Cortex 2:310–335.

Traub RD, Jeffreys JGR, Whittington. 1999. Fast oscillations in cortical circuits. Cambridge, MA: MIT Press.

Tulving E, Kapur S, Craik FIM, Markowitsch HJ, Houle S. 1994. Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography findings. Proc Natl Acad Sci USA 91:2016–2020.

Wennekers T, Paseman F. 1996. Synchronous chaos in high-dimensional modular neural networks. Int J Bifurc Chaos 6:2055–2067.

Yang X, Hyder F, Shulman RG. 1997. Functional MRI BOLD signal coincides with electrical activity in the rat whisker barrels. Magn Res Med 38:874–877.