ELSEVIER

Magnetic Resonance Imaging 21 (2003) 1251–1261

MAGNETIC RESONANCE IMAGING

Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping

Rainer Goebel[a,*], Alard Roebroeck[a], Dae-Shik Kim[b], Elia Formisano[a]

[a]*Department of Cognitive Neuroscience, Faculty of Psychology, Maastricht University, Maastricht, The Netherlands*
[b]*Center for Magnetic Resonance Research, University of Minnesota Medical School, Minneapolis, MN, USA*

Received 15 August 2003; received in revised form 22 August 2003; accepted 23 August 2003

Abstract

We present a framework aimed to reveal directed interactions of activated brain areas using time-resolved fMRI and vector autoregressive (VAR) modeling in the context of Granger causality. After describing the underlying mathematical concepts, we present simulations helping to characterize the conditions under which VAR modeling and Granger causality can reveal directed interactions from fluctuations in BOLD-like signal time courses. We apply the proposed approach to a dynamic sensorimotor mapping paradigm. In an event-related fMRI experiment, subjects performed a visuomotor mapping task for which the mapping of two stimuli ("faces" vs "houses") to two responses ("left" or "right") alternated periodically between the two possible mappings. Besides expected activity in sensory and motor areas, a fronto-parietal network was found to be active during presentation of a cue indicating a change in the stimulus-response (S-R) mapping. The observed network includes the superior parietal lobule and premotor areas. These areas might be involved in setting up and maintaining stimulus-response associations. The Granger causality analysis revealed a directed influence exerted by the left lateral prefrontal cortex and premotor areas on the left posterior parietal cortex.

Keywords: Granger causality; Effective connectivity; Vector autoregressive models; Time-resolved fMRI

1. Introduction

Functional brain imaging has contributed substantial insights into the neural correlates of human information processing and cognitive operations. Yet limitation in temporal resolution has led researchers to focus on relevant information about *where* information is processed in the human brain (*functional segregation*). To gain a deeper understanding of *how* the brain processes information, more knowledge about the interaction of activated brain areas (*functional integration*) is needed.

Following the seminal work of several researchers [1–4], functional brain integration has been investigated during various cognitive or sensorimotor tasks using positron emission tomography (PET) and functional magnetic resonance imaging (fMRI). Significant methodological developments, such as the application of covariance structural equation modeling [4] and nonlinear system identification techniques [5] to neuroimaging data, have supported the idea that a statistical model of interacting neuronal systems can be obtained from metabolic/hemodynamic measurements of task-related neural activation. These models of the interaction between brain areas are often referred to as models of *effective connectivity*, defined as *the influence one neuronal system exerts over another* [1].

Here we present a general framework to investigate effective connectivity (or *directed* influences) between activated brain areas using vector autoregressive (VAR) modeling of time-resolved fMRI time-series in the context of Granger causality [6,7]. In this framework, time-resolved fMRI measurements provide topographical as well as temporal information about the brain areas subserving a cognitive task (see e.g., Refs. [8,9]). This is a very relevant aspect since temporal information of sufficient accuracy constitutes a prerequisite for applying vector autoregressive modeling (see below) or similar methods that aim to characterize not only instantaneous effects between coactivated brain regions but also "causal" (directed) effects acting over time. Such non-instantaneous effects occur if activity changes in area A affect activity changes in area B at a later point in

* Corresponding author. Tel.: +31-43-3884-014; fax: +31-43-3884-125.
E-mail address: r.goebel@psychology.unimaas.nl (R. Goebel).

0730-725X/03/$ – see front matter
doi:10.1016/j.mri.2003.08.026

time. VAR modeling of fMRI time-series and computation of Granger causality maps provide the mathematical framework for modeling effective connectivity.

2. Autoregressive modeling and Granger causality

Functional connectivity has been defined as "the temporal correlations between remote neurophysiological events" and effective connectivity as "the influence one neural system exerts over another" [2]. The latter definition is essentially a statement about causal relations between systems. In the following, we will describe how functional connectivity and effective connectivity are defined and modeled, respectively, within our proposed multivariate framework. We treat the sequence of fMRI measurements of selected regions of interest x_i (individual voxels or an average over multiple voxels) as the components of a discrete vector time-series $\mathbf{x}[n] = (x_1[n],\ldots, x_M[n])^T$, where n represents time, and "T" denotes matrix transposition. Without loss of generality it can be assumed that this vector is zero-mean, i.e., that $E(\mathbf{x}[n]) = (0,\ldots, 0)^T$, where "E" denotes the expectation operator.

2.1. Functional connectivity and effective connectivity

The linear functional connectivity between the components of $\mathbf{x}[n]$ is fully contained in its cross-covariance matrix:

$$\mathbf{R}_{\mathbf{xx}}[k] = \mathbf{E}(\mathbf{x}[n]\mathbf{x}^T[n+k]) = \begin{bmatrix} r_{11}[k] & r_{12}[k] & \ldots & r_{1M}[k] \\ r_{21}[k] & r_{22}[k] & \ldots & r_{2M}[k] \\ \vdots & \vdots & \ddots & \vdots \\ r_{M1}[k] & r_{M2}[k] & \ldots & r_{MM}[k] \end{bmatrix} \quad (1)$$

The off-diagonal element $r_{ij}[k]$, $i \neq j$, is the scalar cross-correlation function between $x_i[n]$ and $x_j[n]$ at lag k. The diagonal element of $r_{ii}[k]$ is the scalar autocorrelation of x_i at lag k. It holds that: $\mathbf{R}_{\mathbf{xx}}[-k]=\mathbf{R}_{\mathbf{xx}}[k]^T$. Implicit in our definitions of the mean and autocorrelation function is the assumption that they do *not* depend on n, i.e., we are assuming that the signals we are dealing with are wide sense stationary (WSS). Any entry $\gamma_{ij}[k]$ of the cross-covariance matrix may be interpreted as a non-parametric measure of *linear* association between component *i* and component *j* at lag *k*. Thus, the autocorrelation function is a model-free characterization of statistical association between time-series, without any regard for the underlying dependence-structure between the components x_i (and possible external components z_j), which generated such association.

To make inferences about such underlying structure, i.e., about effective connectivity, extra assumptions are needed and have to be incorporated into a multivariate process-model of the vector time-series $\mathbf{x}[n]$. Such assumptions can be of two sorts. First, structural assumptions determine *which* components x_i can depend directly on or be directly influenced by which other components x_j (or exogenous components z_j). In the method of Covariance Structural Equation Modeling (CSEM) as applied to neuroimaging data [4] these structural assumptions would constitute the so-called *anatomic model.* Eventually, DTI tractography could provide a more data-driven way of forming these assumptions (see Le Bihan et al., this issue; Kim et al., this issue). Second, functional dependence assumptions, determine *how,* mathematically, the value $x_i[n]$ can statistically depend on (or is a function of) values $x_j[n-k,\ldots, n]$ of other components (or values $z_j[n-k,\ldots, n]$ of exogenous components). Such functional dependence assumptions are essentially contained in the specific process-model one chooses to employ. For instance, in CSEM, the functional dependence assumptions are that the value $x_i[n]$ can only be a function of the instantaneous values of other components, as in $x_1[n] = a_2x_2[n] + a_3x_3[n] + b_1z_1[n] + e[n]$, where $e[n]$ denotes unexplained noise.

We propose to treat $\mathbf{x}[n]$ as a vector autoregressive (VAR) process, and thus to use vector autoregressive models to make inferences on effective connectivity. We choose autoregressive modeling to assess the degree of dependence between components for several reasons. First, VAR models are dynamical models that can capture the temporal structure in the variations of individual components and in the interdependence between them. Second, the parameters of autoregressive models are relatively easy to estimate by solving a *linear* regression problem. Third, many random processes can be very well approximated by a sufficiently high order AR model. Finally, as we will show, VAR models form a natural context in which measures of directed influence based on the concept of Granger causality can be defined.

2.2. Granger causality

Taking the temporal structure of signal time-courses into account is related to our commonsense concept of causality: causes always precede effects. Something in the future can not cause something in the past or present. All events taking place at a certain point in time must have had their cause at an earlier stage. These considerations have led the econometrist Clive Granger to propose a definition of causality for temporally structured data, i.e., time-series [6,7]. Conceptually, it amounts to the following: if a time-series *y* causes (or has an influence on) *x*, then knowledge of *y* should help predict future values of *x*. Thus, causality (or influence) is framed in terms of predictability. More in detail, given two discrete time-series *x* and y, we say that *y Granger causes x* if we can predict the current value of *x*, $x[n]$ using past values of *x and y* (i.e., the information set $D = \{ y\text{-}, x\text{-}\} = \{ y[n-1], y[n-2],\ldots, x[n-1], x[n-2],\ldots\}$) better than we can when using past values of *x* alone (i.e., the set $D - y\text{-} = x\text{-}$). In this way the temporal structure in the dependency between *x* and *y* is used to decide on the *direction* of

possible influences between them. If there are other possible causal influences of relevance, then we add these to the information set *D*. For instance, if in fact there exists a third time series *z* that has an influence on both *x* and *y*, then we should add the values *z*- = {*z*[*n*−1], *z*[*n*−2],. . .} to the set *D*. Otherwise, using just the set *D* as above, spurious causality between *x* and *y* could arise. Furthermore, Granger proposed to speak of *instantaneous causality* between *x* and *y* when we can predict *x*[n] better from *D* + { *y*[*n*]} than we can from *D* alone. So defined, instantaneous causality can be seen to have no direction.

To make his definitions operational, Granger proposed to use linear auto-regressive models to produce predictions (or forecasts) of the value of *x*[*n*] from values in *D*. The sum of squared errors in the forecasts that such an auto-regressive model makes can then be used as a measure of how good it can predict *x*[*n*]. Moreover, we can compare models that use a different set of past values in their prediction of *x*[*n*] by comparing their respective error variances. Granger's method has even been generalized to a measure of directed linear influence between two groups of time-series (possibly conditional on a third group) using vector auto-regressive models [10,11], which we will use in this investigation. Several variants of these techniques have recently been applied to neurophysiological data to gain insight in the direction of influences between neural systems [12–14].

2.3. Vector autoregressive models

The vector time-series **x**[n] can be modeled as an AR process as:

$$\mathbf{x}[n] = -\sum_{i=1}^{p} \mathbf{A}[i]\mathbf{x}[n-i] + \mathbf{u}[n] \tag{2}$$

where **u**[*n*] is (multivariate) white noise, with cross-covariance matrix var(**u**[*n*]) = **Σ,** if *k* = 0, otherwise var(**u**[*n*]) = 0. The matrices **A**[*i*] are called the autoregression (AR) coefficients because they regress **x**[*n*] onto its own past. We call *p* the order of the auto-regression and will refer to the above model, with adjustable parameters **A**[*i*] and **Σ** to be estimated, as a VAR(*p*) model. There are two important interpretations of the above VAR model. First, it can be considered to model **x**[*n*] as the output of a multivariate linear filter driven by the white noise input **u**[*n*]. This filter has a rational transfer function containing the **A**[*i*] in the denominator matrix-polynomial. This interpretation makes clear that the model really captures the temporal structure of **x**[*n*], since, because the input **u**[*n*] has no (linear) temporal structure *by definition,* all temporal structure present in **x**[*n*] must be contained in the **A**[*i*]. Second, the VAR model can be thought of as a linear prediction model, that predicts the current value of **x**[*n*] based on a linear combination of the most recent *p* past values. Consequently, the current value of a component $x_i[n]$ is predicted based on a linear combination of its own past values *and* the past values of the other components. The second interpretation of the VAR model shows its value in quantifying Granger causality between (groups of) components.

2.4. Effective connectivity: directed influence

Geweke [10] has proposed a measure of linear influence (or feedback, as he calls it) $\mathbf{F}_{x,y}$ between the time-series **x**[*n*] and **y**[*n*] which can be regarded as an implementation of the concept of Granger causality in terms of vector autoregressive models. The influence measure $\mathbf{F}_{x,y}$ is the sum of three components: the linear influence from **x** to **y** ($\mathbf{F}_{x\to y}$), the linear influence from **y** to **x** ($\mathbf{F}_{y\to x}$), and the instantaneous influence between **x** and **y** ($\mathbf{F}_{x\cdot y}$). The measure can be defined using the residual cross-covariance matrices of the following three VAR models involving the *K*-dimensional series **x**[*n*] and *L*-dimensional series **y**[*n*]:

$$\begin{aligned}
\mathbf{x}[n] &= -\sum_{i=1}^{p} \mathbf{A}_x[i]\mathbf{x}[n-i] + \mathbf{u}[n] \qquad \mathrm{var}(\mathbf{u}[n]) = \Sigma_1 \\
\mathbf{y}[n] &= -\sum_{i=1}^{p} \mathbf{A}_y[i]\mathbf{y}[n-i] + \mathbf{v}[n] \qquad \mathrm{var}(\mathbf{v}[n]) = \mathbf{T}_1 \\
\text{and with } \mathbf{q}[n] &= \begin{bmatrix} \mathbf{x}[n] \\ \mathbf{y}[n] \end{bmatrix}: \\
\mathbf{q}[n] &= -\sum_{i=1}^{p} \mathbf{A}_q[i]\mathbf{q}[n-i] + \mathbf{w}[n] \qquad \mathrm{var}(\mathbf{w}[n]) = \mathbf{Y} = \begin{bmatrix} \Sigma_2 & \mathbf{C} \\ \mathbf{C}^T & \mathbf{T}_2 \end{bmatrix}
\end{aligned} \tag{3}$$

where **q**[*n*] is *O*-dimensional (with *O* = *K* + *L*), Σ_1 and Σ_2 are *K* by *K*, $\mathbf{T}_1$ and $\mathbf{T}_2$ are *L* by *L*, and **Y** is *O* by *O*. It is these residual cross-covariance matrices Σ_1, Σ_2, and **Y,** that are currently of interest to us, because they quantify how well we are able (using linear AR models) to predict current values of **x** and **y** from their past values. The measures of total linear dependence between **x** and **y,** linear influence from **x** to **y,** linear influence from **y** to **x,** and instantaneous

influence between **x** and **y** are defined to be, respectively [10]:

$$\begin{aligned}\mathbf{F}_{\mathbf{x},\mathbf{y}} &= \ln(|\Sigma_1| \cdot |\mathbf{T}_1|/|\mathbf{Y}|)\\ \mathbf{F}_{\mathbf{x}\to\mathbf{y}} &= \ln(|\mathbf{T}_1|/|\mathbf{T}_2|)\\ \mathbf{F}_{\mathbf{y}\to\mathbf{x}} &= \ln(|\Sigma_1|/|\Sigma_2|)\\ \mathbf{F}_{\mathbf{x}\cdot\mathbf{y}} &= \ln(|\Sigma_2| \cdot |\mathbf{T}_2|/|\mathbf{Y}|)\end{aligned} \tag{4}$$

where " Σ " denotes the determinant of Σ. From these definitions, it can be seen that it holds that:

$$\mathbf{F}_{\mathbf{x},\mathbf{y}} = \mathbf{F}_{\mathbf{x}\to\mathbf{y}} + \mathbf{F}_{\mathbf{y}\to\mathbf{x}} + \mathbf{F}_{\mathbf{x}\cdot\mathbf{y}} \tag{5}$$

Here we are assuming that the finite order AR-models are valid descriptions of the time-series **x**[*n*], **y**[*n*], and **q**[*n*], which also implies the assumption that **q**[*n*] is WSS and thus that **x**[*n*] and **y**[*n*] are jointly WSS. In this case, we can give the following interpretations of the measures. The four measures take their values in the interval [0, ∞], i.e., they are by construction nonnegative. $\mathbf{F}_{x,y}$ is a measure of the total linear dependence between the series **x** and **y.** If nothing of the value at a given instant of one can be explained by a linear model containing all the values (past, present, and future) of the other, $\mathbf{F}_{x,y}$ will evaluate to zero. $\mathbf{F}_{x\to y}$ is a measure of linear directed influence from **x** to **y.** If past values of **x** will not improve the best linear prediction of the current value of **y** over the prediction obtained from using past values of **y** alone, then $\mathbf{T}_1 = \mathbf{T}_2$ and $\mathbf{F}_{x\to y}$ will be zero. Conversely, if past values of **x** *do* improve the prediction of the current value of y, then $\mathbf{T}_2 < \mathbf{T}_1$ and $\mathbf{F}_{x\to y} > 0$. Since it holds that $\mathbf{T}_2 \leq \mathbf{T}_1$, $\mathbf{F}_{x\to y}$ will always be nonnegative. As we can interpret the determinant of a correlation or covariance matrix as a measure of generalized variance, $\mathbf{T}_1$ is the generalized variance of the mean squared error in predicting **y**[*n*] by a linear projection on its own past values {**y**[*n*−1], **y**[*n*−2],. . .}. Therefore, $\mathbf{F}_{x\to y}$ quantifies the reduction in this generalized variance obtained by adding past values of **x** to the projection set. A similar interpretation holds, of course, for $\mathbf{F}_{y\to x}$.

Thus, the two directed components, $\mathbf{F}_{x\to y}$ and $\mathbf{F}_{y\to x}$, use the arrow of time to decide on the direction of influence. However, the total linear dependence between **x** and **y** does not often consist fully of these directed components. Much of the total linear dependence can be contained in the undirected instantaneous influence $\mathbf{F}_{x\cdot y}$ between them. Essentially, $\mathbf{F}_{x\cdot y}$ quantifies the improvement in the prediction of the current value of **x** (or **y**) by including the current value of **y** (or **x**) in a linear model already containing the past values of **x** *and* **y.** From this symmetry it can be seen that $\mathbf{F}_{x\cdot y}$ indeed contains no directional information at all. In practice, nonzero values of $\mathbf{F}_{x\cdot y}$ can be caused by directed influence between **x** and **y** at a finer time-scale than that at which **x** and **y** are observed. Thus, poor temporal sampling of the processes of interest (at a frequency lower than that required to detect relevant interactions) can obscure the true directed linear influence between them. True directional influence (as computed with the above measures) might then either not be detected or might 'leak' into the instantaneous component, hiding the direction of influence from us.

Besides the problem associated with poor temporal sampling (which is prominent in fMRI measurements) there is an additional issue that troubles the interpretation of the influence measures as detecting true causality between the observed processes. This is the problem of spurious causality [7] that can appear between two processes when both are influenced by other external sources that are not taken into account. The influence measures defined above are only valid when **x** is the sole source of influence on **y** and vice versa. Any additional external source of influence will confound the inferences made from these measures. Thus, we should take *any* source of influence on either process into account. Mathematically, this amounts to including any such external process $z_i[n]$ in our analysis as a confound. Together these confounders form the vector process **z**[*n*] that we have to condition on to remove its effects from the analysis.

Acknowledging this, the measures of linear dependence defined above can be extended to measures of *conditional* linear dependence by including the *M*-dimensional process **z**[*n*] in each of the VAR models [11]. The interpretation of the conditional measures is quite similar to that of the unconditional measures. The main difference is, of course, that we now control for confounding external sources of influence. From the above definitions it is clear that all we need to estimate these measures of influence (conditional or unconditional) for observed real world signals (like fMRI measurements) are estimates of the residual correlation matrices of certain VAR models.

3. Simulation example

As remarked above, an important problem in inferring interactions at a neuronal population level from fMRI data are our indirect access to the signals of interest (Fig. 1). The fMRI signal can be considered a filtered and sampled version of the Local Field Potential (LFP) signal that is itself a measure of the activity fluctuations of local population of neurons [15]. Previous studies [12,13] have shown that techniques based on VAR-modeling and Granger causality can capture the dynamic structure of the LFP signal and can infer neuronal interactions from it. The simulation example reported here investigates the effect of 1) Hemodynamics (i.e., filtering) and 2) temporal sampling on our ability to extract from the fMRI signal, population interactions that supposedly take place at the level of LFP signals. More extensive simulations varying the assumed properties of the hemodynamic response function and temporal sampling will be reported elsewhere (Roebroeck et al., in preparation).

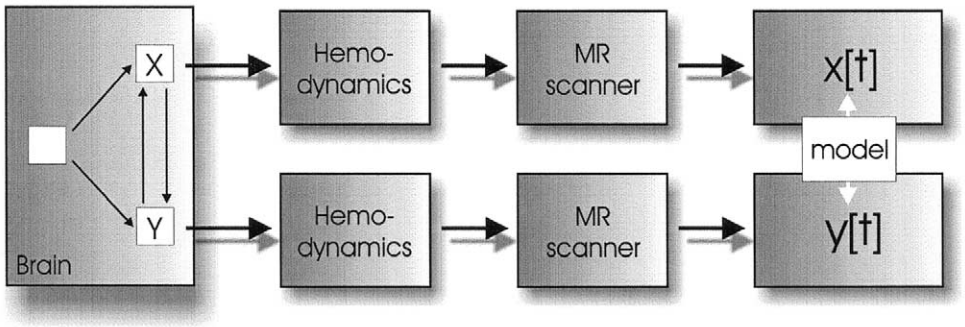

Fig. 1. An illustration of the indirect access to interacting brain regions with fMRI. *Hemodynamics* and the *MR scanner* contribute unwanted artifacts to the signals of interest and might confound modeling efforts. Confounding is especially deleterious when the unwanted contributions are different for the brain regions under investigation (e.g., different hemodynamic responses in different regions).

The LFP signals $x[n]$ and $y[n]$ of two interacting neuronal populations were generated as a realization of a bi-dimensional first-order VAR process with:

$$\mathbf{A}[1] = \begin{bmatrix} -0.8454 & 0 \\ -0.5 & -0.8454 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.2853 & 0 \\ 0 & 0.2853 \end{bmatrix} \quad (6)$$

The timestep of the simulation was taken to be 100 ms, and the process was constructed to have spectral content in the lower frequency ranges (<1Hz). Furthermore, by construction there is a directed influence from x to y, but no influence from y to x, as can be seen from the off-diagonal elements of $\mathbf{A}$ [1]. There is no real instantaneous influence between the channels, since the off-diagonal elements of Σ are zero. The LFP signals simulated as a realization (1000 timepoints = 100 sec) of this process are plotted in the top part of Fig. 2.

The order p, used to compute influence measures $\mathbf{F}_{x\to y}$, $\mathbf{F}_{y\to x}$, and $\mathbf{F}_{x\cdot y}$ for the simulated signals was set to that which minimized several order selection criteria (Akaike

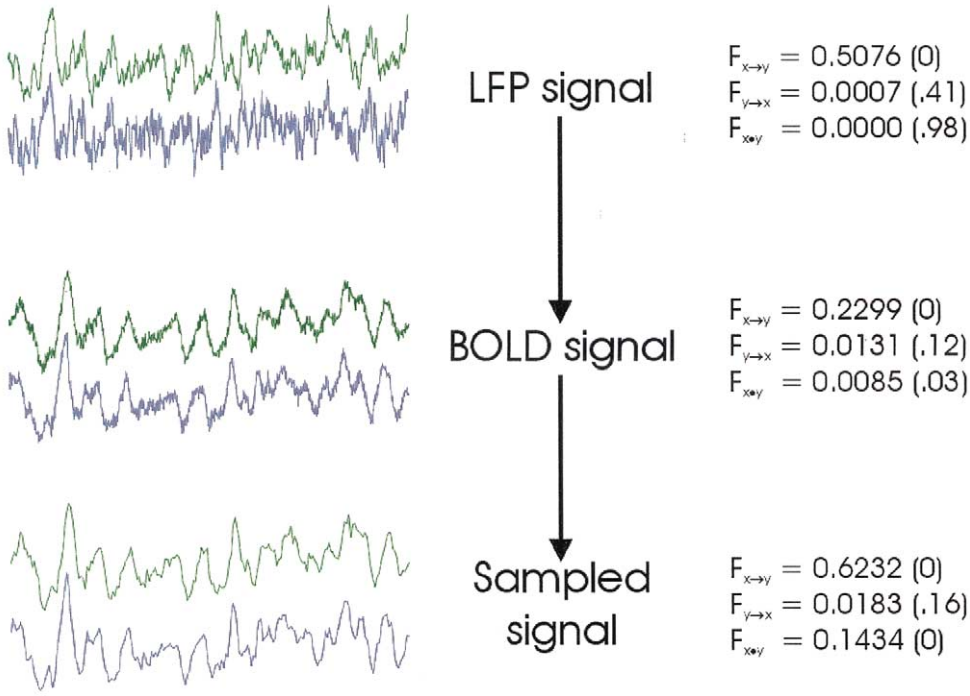

Fig. 2. Simulation of unidirectional influence at the LFP level (upper part), the effects of hemodynamic convolution (middle part) and temporal subsampling (lower part). The simulation shows: 1) that the modeled influence from X to Y is easily detectable at all three levels; 2) that there is a clear leakage of the directed influence into the instantaneous term at the BOLD level and even more in the subsampled signal; and 3) that there is an increase in significance of the non-modeled influence from Y to X in the BOLD and subsampled signal. However, the values of the two directed influences remain clearly separated at all levels. Note that because of temporal subsampling, the absolute values of the influence measures at the final level can not be compared to the values at the other levels.

Information criterion, Bayesian Information Criterion, and the Hannan-Quinn criterion). These criteria are designed to trade-off the reduction in error-variance against the increase in the number of parameters and are thus able to give an indication of the model order needed to capture the dynamics of a given signal. For the simulated LFP signals, the criteria (correctly) indicated that an order $p = 1$ was appropriate. Furthermore, to assign significance levels to the computed measures a bootstrapping procedure was used (see e.g., 11, 12). In short, a bootstrapped distribution under the null-hypothesis ($\mathbf{F}_i = 0$) for the influence measures was obtained by independently simulating the models estimated for x and y individually and re-computing the measures for these independent realizations. The achieved significance level reported here is the proportion of values in this null-distribution that are larger than the value computed for the original signal. To construct null-distributions of $\mathbf{F}_{x\to y}$, $\mathbf{F}_{y\to x}$, and $\mathbf{F}_{x\cdot y}$, 200 realizations were simulated.

The computed values of the influence measures and their bootstrapped significance levels for the simulated LFP signals are reported in the top-left of Fig. 2. The influence from x to y over time is correctly reflected in a high value of $\mathbf{F}_{x\to y}$ and an achieved significance level of 0. The absence of any influence from y to x and instantaneous influence is also correctly reflected in very low non-significant values for $\mathbf{F}_{y\to x}$, and $\mathbf{F}_{x\cdot y}$, respectively.

BOLD signals were simulated by filtering the LFP signals through the hemodynamic impulse-response function (HRF) modeled as a gamma function [16]. The parameters in this model were set to values corresponding to short (0.5 sec) stimulus durations [17]. Particularly, the tau parameter, controlling the width of the HRF, was set to 0.5. After filtering, 20% of white gaussian noise was added reflecting physiological noise in the hemodynamics. The resulting signals are plotted in the middle part of Fig. 2. The autoregressive order used for the computation of the influence measured was set to 5, as suggested by the order selection criteria. The computed influence values show that the influence from x to y is still correctly identified, as shown by the large highly significant value of $\mathbf{F}_{x\to y}$. Nevertheless, a marked decrease in magnitude of $\mathbf{F}_{x\to y}$ as compared to that computed for the original LFP is obvious. This seems to suggest that the loss of some temporal structure in the signals in the filtering somewhat decreases our ability to extract directed interactions. The slightly inflated values for $\mathbf{F}_{y\to x}$, and $\mathbf{F}_{x\cdot y}$ also seem to support this. However, the overall pattern of interactions present in the original LFP signal (influence from x to y but *not* from y to x) is still clearly reflected in the computed influence measures, and thus the effect of hemodynamics on our ability to detect temporally directed influences seems to be far from destructive.

The final sampled BOLD signal as obtained from the scanner in an fMRI experiment was generated by sampling the BOLD signal every 5 points, comparable to obtaining T2* weighted images with a volume TR of 500 ms. After sampling 10% of white gaussian noise was added reflecting measurement error. The sampled BOLD signal is plotted in the bottom part of Fig. 2. The autoregressive order used for the computation of the influence measured was set to 2, as suggested by the order selection criteria. Once again, the influence from x to y is still correctly identified, as shown by the large highly significant value of $\mathbf{F}_{x\to y}$. It should be noted that the values of the influence measures cannot be compared across time-series with a different time-unit. Thus the increase in absolute magnitude of the terms (particularly the $\mathbf{F}_{x\to y}$ term) has no interpretable meaning. Absence of influence from y to x is once again correctly reflected in a small (especially as compared to $\mathbf{F}_{x\to y}$) non-significant value for $\mathbf{F}_{y\to x}$. As expected, and discussed above, the instantaneous term $\mathbf{F}_{x\cdot y}$ has become inflated to a large significant value, due to the effect of filtering and sampling. However, such leakage of true directed influence into the undirected instantaneous term induced by smoothing and sampling the original series is not at all destructive to our inferences. The example simulation clearly shows that Granger causality analysis on fMRI-data can pick up certain kinds of temporally lagged influences between interacting neuronal populations. More extensive simulations, varying e.g., the order and spectral content of the original VAR process, the assumed HRF, and the sampling interval, must be performed to examine the range of interactions which can be reliably detected by this analysis (Roebroeck et al., in preparation).

4. An application: dynamic sensorimotor mapping

We applied the described method to a dynamic sensorimotor mapping paradigm. Sensorimotor coordination is dynamic in nature and involves the selection and execution of appropriate actions based on the perceived, changing environment and previous experience. Here we focus on a particularly interesting aspect of sensorimotor coordination, namely how sensory-motor associations are established and dynamically changed over time. In particular, the task we used requires sudden remappings of established stimulus-response couplings. It is expected that such a remapping process takes at least several hundreds of milliseconds until it is "implemented" in cortical systems mediating the coupling of sensory to motor areas and therefore can be investigated using our framework. We chose an event-related design to also study transient responses to individual events like a presentation of a stimulus or execution of a response.

There are several behavioral studies on related phenomena most notably in the context of task switching paradigms [18]. Although previous imaging studies have provided important knowledge of which areas are involved in sensorimotor coordination and task switching in the human brain [19–23] the direction of interactions between these areas are largely unknown (but see Ref. [24]).

It is important to notice that the computed influence measures for a given region of interest (ROI) were mapped

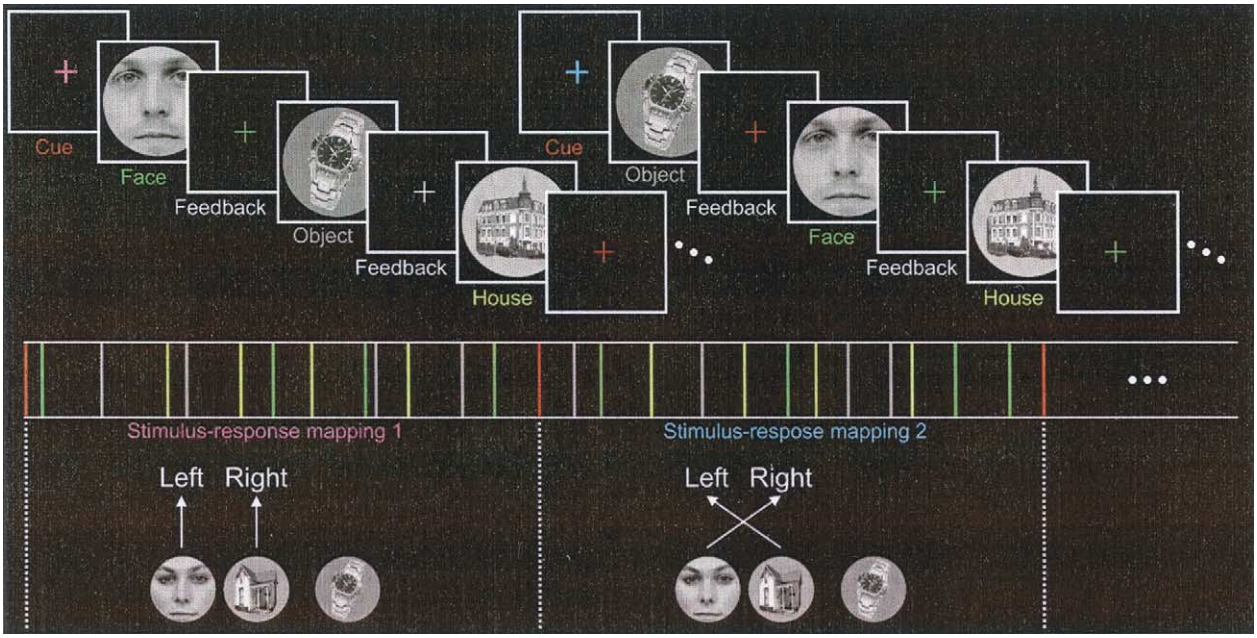

Fig. 3. The stimuli (top), temporal organization of events (middle), and required responses (bottom) in the visuomotor mapping task. White dots in the event sequence imply continuation of the task regime. Refer to the text for further explanation. Note that the number of trials in between two remapping cues can vary within and between runs, which determines the "switching load" for the subject (see text for further explanation).

over all voxels in the scanning volume. In this way, regions that interact with a given ROI are identified by the maps, instead of prespecified by the experimenter in the form of an anatomic model. This represents a more data-driven, exploratory method of investigation than used in other methods (e.g., CSEM) using as little a priori knowledge or expectations as possible.

5. Materials and methods

Previously, we reported on directed influences between brain regions in a sensorimotor task switching paradigm [25]. Here, we focus on how these influences change with differences in task switching load for a second subject.

5.1. Cognitive task

The subject performed three conditions of a visuomotor mapping task in which two stimulus categories had to be mapped to two responses ("left" or "right"). In a "fast-switching" condition and a "slow-switching" condition the mapping of the two stimulus categories ("houses" and "faces") to the responses alternated periodically between the two possible mappings (see Fig. 3). A remapping cue (500 ms change in the color of the fixation cross) indicated a change in the required stimulus-response mapping (S-R mapping) for the following trials. In addition to the face-stimuli and house-stimuli, pictures of objects appeared that required no response. Stimuli were shown for 120 ms with a stimulus onset asynchrony (SOA) of 2-6 s. Feedback on the correctness of responses was given at every trial (green fixation cross for a correct response; red fixation cross for an incorrect response). In the fast-switching condition, the S-R mapping changed every 2 to 6 trials, while in the slow-switching condition, the S-R mapping changed every 15 trials. The third condition performed by the subject was a "no-switching" condition, in which two stimulus categories ("animals" and "fruit") had to be mapped consistently to right and left responses without changes in the mapping. As in the other conditions, "no-go" trials in the form of object images were also presented. The subject performed 2 runs of each of the three conditions.

5.2. MRI scanning and experimental setup

Images were acquired using a 3 Tesla scanner ("Trio," Siemens, Erlangen, Germany). Functional images were acquired T2* weighted echo planar sequence (echo time (TE) 28 ms, volume repetition time (TR) 1000 ms, field of view 224 mm $\times$ 224 mm, 64 $\times$ 64 matrix, giving 3.5 mm x 3.5 mm in-plane resolution). The images consisted of 18 oblique transverse slices (interleaved acquisition), 5 mm thick with a 1 mm inter slice gap. For the slow-switching runs 540 volumes were acquired; for the fast-switching and no-switching runs 500 volumes were scanned. Structural images were acquired using a T1 MPRAGE sequence (echo time 4 ms, 256 $\times$ 256 $\times$ 192 matrix, 1 $\times$ 1 $\times$ 1 mm^3 voxels). Stimulus presentation, response registration, and synchronization to the scanner acquisition were performed using the software program *Presentation* (Neurobehavioral systems, San Francisco, CA).

5.3. Data analysis

Imaging data were analyzed using BrainVoyager 2000 (Brain Innovation, Maastricht, The Netherlands). The anatomic volume was transformed to the Talairach coordinate system [26]. The cortical surface was reconstructed [27] and inflated for visualization of results. The time courses of activation of individual voxels were constructed from the functional images and corrected for the temporal difference in acquisition of different slices (slice scan time correction) using sinc interpolation. Subsequently, linear trends and low frequency components (up to and including four cycles in the time course) were removed prior to any analysis. Voxel time courses were then coregistered to the structural volume and transformed into Talairach space with a resolution of 3 $\times$ 3 $\times$ 3 mm using trilinear interpolation. No spatial or temporal smoothing was applied to the functional time courses.

5.4. Conventional statistical analysis (general linear model)

Regional activations were analyzed using a GLM, testing for the (differential) contribution of several predictor functions to the explanation of variation in individual voxel time-courses. Six predictor functions reflecting the main stimuli and cues in the task were constructed as box-car functions (value one at the single scan where the relevant event took place, value zero otherwise) filtered through a linear model of the BOLD response [16]. Predictors were created for the mapping cue, the control stimulus, left hand responses and right hand responses. Individual voxel time courses were regressed onto a model containing these predictors and an additional constant-level predictor to correct for the signal level.

5.5. Granger causality mapping

The network of interacting regions subserving performance of the visuomotor mapping task was investigated by mapping the influence measures discussed above over the whole brain, giving what we call Granger Causality Maps (GCMs). Each of the computed GCMs centers on a single region of interest (reference region) that is considered *both* as the source of influence to voxels in the rest of the brain *and* as the target of influence from voxels in the rest of the brain. The reference regions were chosen as the activated regions found with the GLM analysis. In the notation used above, the averaged BOLD response time-course of voxels in a specified reference region was considered as the time-course $x[n]$. Subsequently, the BOLD response time-course of each single voxel in the functional volume was taken as the time-course $y[n]$ and the influence measures $\mathbf{F}_{x\to y}$, $\mathbf{F}_{y\to x}$, and $\mathbf{F}_{x\cdot y}$ were computed. For each reference region this procedure resulted in three GCMs: 1) the map of $\mathbf{F}_{x\to y}$ (Reference to Voxel map) showing voxels which are influenced by the activity in the reference region; 2) the map of $\mathbf{F}_{y\to x}$ (Voxel to Reference map) showing voxels whose activity influence the activation in the reference ROI; and 3) the map of $\mathbf{F}_{x\cdot y}$ (Instantaneous influence map) showing voxels whose activation shows an instantaneous dependency relation with activation of the reference ROI without any clear direction in time. The GCMs were computed both for the fast-switching runs and for the no-switching runs. The resulting maps were then subtracted (fast-switching –no-switching) to show the increases in influence values from the no-switching condition to the fast-switching condition. The influence maps were computed using an autoregressive model order of 1. Thus, Granger causality between brain regions was considered looking *one* TR (i.e., 1 s) into the past. Autoregressive models were estimated using an orthogonalization procedure [28] allowing us to compute pooled estimates over the two runs within the conditions performed by the subject.

6. Results

6.1. Activated regions and deconvolved event-related time courses

Maps for regional activation and deconvolved event-related BOLD responses, computed over all runs of multiple subjects revealed a widespread network comprising frontal, parietal, and ventral visual regions (see Fig. 2B in Ref. [25]). Activation for the remapping cue was observed prominently in posterior parietal areas (somewhat lateralized to the left) and premotor areas, both on the medial (supplementary motor area and presupplementary motor area) and lateral cortical surface (dorsal and ventral lateral premotor cortex). Furthermore, visual and prefrontal areas are also activated at the mapping cue. The deconvolved event-related responses for the mapping cue for these frontoparietal areas showed a transient, but temporally extended (over a few seconds) rise of activity when the stimulus-response mapping changed. Therefore, we investigated the interaction between these regions further by choosing a highly activated left parietal area as a reference region for the GCM analysis.

6.2. Granger causality mapping and influence analysis

The difference maps (fast-switching–no switching) for a reference region in the left posterior parietal cortex (PPC) are shown in Fig. 4. Two important points have to be made about the interpretation of these maps. First, all inferences about influences between regions of the cortex can only be interpreted with respect to the reference region. Strictly, for a given set of influence maps we *cannot* talk about interactions between two regions which are both *not* the reference region. Second, direct influence can only be inferred in as far as the reference region and a given voxel are each

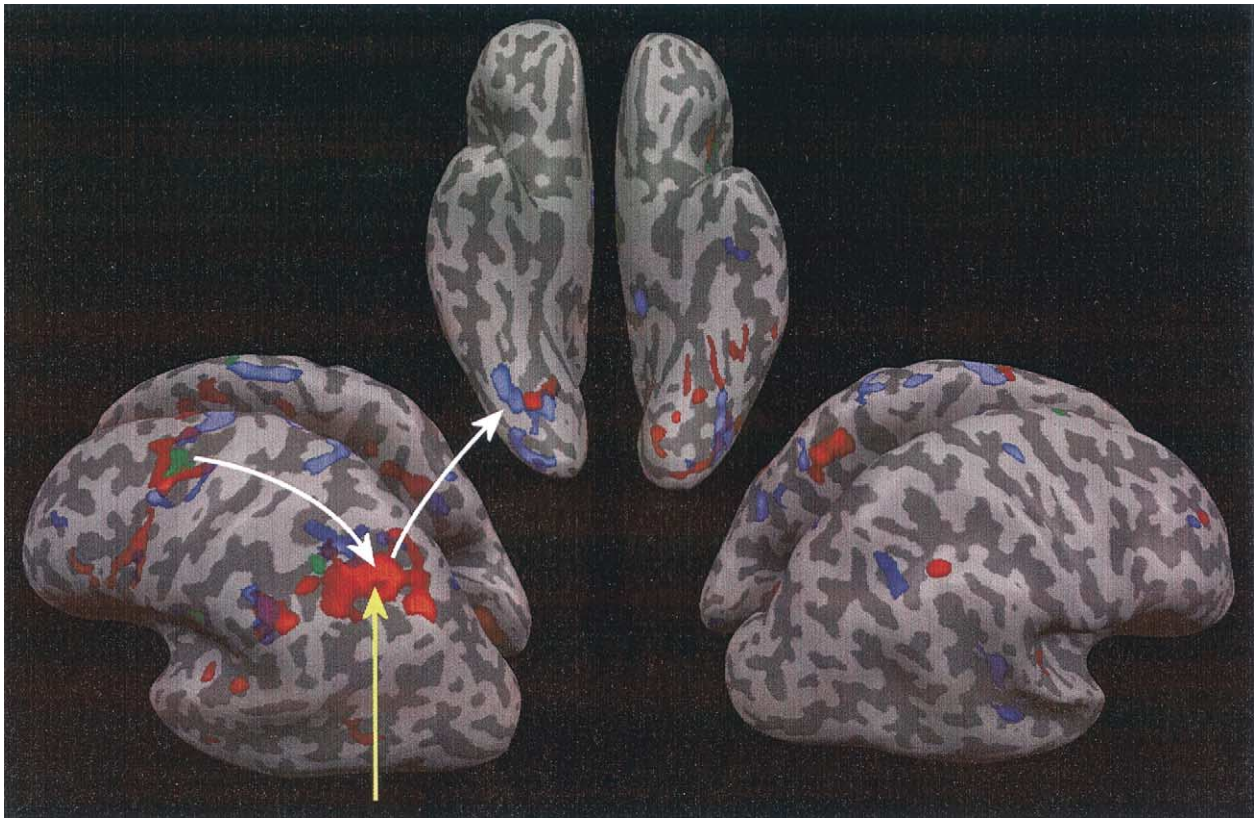

Fig. 4. Difference map showing the increases in influence values from the no-switching condition to the fast-switching condition for a reference region in the left posterior parietal cortex (yellow arrow). White arrows highlight some of the main directed influences discussed in the text. To obtain this difference map, linear Granger Causality Maps (GCMs) were computed separately for the fast and no-switching conditions and then subtracted (fast-switching –no-switching). The difference map for instantaneous influence between the reference area and other parts of the brain (instantaneous map) is shown in red. The difference map for influence from the reference region to other parts in the brain (reference-to-voxel map) is shown in blue. The difference map for influence from other parts of the brain to the reference region (voxel-to-reference map) is shown in green.

other's only source of influence. Indirect influence (e.g., mediated by a third region) will wrongly appear as direct influence unless additional mediating sources of influence are taken into account (by adding them to the model).

The difference-GCMs for the left posterior parietal region show an increase of instantaneous dependence in the fast-switching runs with other cortical regions, notably left lateral premotor and prefrontal regions and visual areas. More interestingly, the voxel-to-reference map for this region also clearly shows an increased influence exerted by left premotor regions in fast-switching runs. Furthermore, the reference-to-voxel map shows an increase in influence exerted on higher order visual areas in the inferotemporal cortex. Both findings are in accordance with earlier results for a different subject [25].

7. Discussion

We have presented a framework for modeling directed cortical interactions based on time-resolved fMRI and a mathematical realization of the concept of Granger causality.

The proposed framework rests on the assumption that time-resolved fMRI contains enough temporal information to determine directed influences based solely on temporal precedence. Recent work suggests that time-resolved fMRI has a temporal resolution in the sub-second range. While such a temporal resolution can not reveal neuronal interactions, it might be sufficient to reveal temporal dependencies among cognitive components of complex cognitive tasks [9].

The simulation presented here suggests that low frequency influences in the LFP signal are detectable in the hemodynamically convolved and temporally sampled fMRI-signal. Other simulations have suggested that modeled influences in higher frequency ranges are not detectable in the fMRI signal if a standard model for hemodynamic convolution is used. This would suggest that the directed influences reported here are based on low frequency fluctuations in the BOLD signal, temporally lagged between regions. This observation puts this method in relation to fMRI mental chronometry, which is also based on lagged temporal information in trial-related signal fluctuations. While fMRI mental chronometry primarily exploits event-related onset latency differences, GCM analysis can exploit

such lagged dependencies over extended temporal periods in the ongoing signal fluctuations.

Due to physiological noise, scanner noise, and limited temporal resolution, we expect that transient short-lagged interactions between brain areas cannot be resolved with our approach. Nonetheless, these directed influences could appear in our analysis in the (non-directional) instantaneous map. From this consideration it also follows that on the basis of the instantaneous map it cannot be inferred that no directed effects among areas exist, but only that the temporal resolution of the data in not sufficient to detect them. In our task, we could observe directed effects between frontal and parietal areas indicating that the frontal areas drive or "Granger cause" activity fluctuations in specific subregions of the activated parietal areas. We interpret this directed influence as neural correlates of executive control and working memory. More specifically, we assume that frontal areas are involved in generating and maintaining an appropriate self-instruction for the current sensory-motor mapping and that the corresponding neural representations act upon parietal areas to implement a respective motor program. Our finding and its interpretation can only be tentative at present since we have performed the influence analysis only in two subjects. We are currently analyzing the data of other subjects to evaluate whether the directed influence from frontal to parietal areas can be generalized to the population level.

7.1. Future improvements

Since exploitation of temporal information in fMRI time courses is critical for the usefulness of our approach, one should optimize all scanning-related aspects influencing temporal resolution [9].

One of the attractive properties of the chosen cognitive task is that the same visual stimuli require a different response depending on the current mapping context. We would thus expect that influence patterns between brain areas also change with respect to the currently valid mapping rule. These potential mapping-specific influence maps could not be revealed in the current analysis because the maps were computed over the whole time course. As a future improvement, we want to compute influence maps for each mapping type separately to potentially reveal dynamic remapping effects. Such a "windowed" extension to our approach might also be useful for other paradigms in which aspects of the task are changing over time within or across functional runs.

Another improvement would be to include nonlinear terms in the multivariate vector autoregresseive modeling approach. This would allow, for example, to reveal modulatory (multiplicative) effects in which one area influences the coupling strength between two other areas. In our task, for example, we would expect that the sensory-motor coupling (mediated probably by parietal regions) changes with respect to the current mapping rule because in one mapping context, houses (faces) require a left (right) response and in the other, houses (faces) require a right (left) response. The necessary sensory-motor associations could be implemented in the parietal lobe as discussed above. It could be, however, also established by modulating the flow of information from the sensory areas (i.e., FFA, PPA) to parietal and motor areas. We would expect such a modulatory sensory-motor effect especially in the first trials after presentation of a (re-) mapping cue.

Ideally, individual information on anatomic connectivity from DTI (or other similar non-invasive technique) could be used as structural constraint for the VAR modeling of functional time-series. Verifying the feasibility of this idea and its methodological implementation will require further research.

We think the Granger causality mapping approach can form a valuable complement to other connectivity methods, and we hope that will stimulate further investigations of how brain areas communicate with each other.

Acknowledgments

This work was supported by the Human Frontiers Science Program.

References

[1] Buchel C, Friston K. Assessing interactions among neuronal systems using functional neuroimaging. Neural Netw 2000;13:871–82.
[2] Friston KJ, Ungerleider LG, Jezzard P, Turner R. Characterizing modulatory interactions between V1 and V2 in human cortex with fMRI. Hum Brain Mapp 1995;2:211–24.
[3] Horwitz B. Functional interactions in the brain: use of correlations between regional metabolic rates. J Cereb Blood Flow Metab 1991; 11:A114–20.
[4] McIntosh AR, Gonzales-Lima F. Structural equation modelling and its application to network analysis in functional brain imaging. Hum Brain Mapp 1994;2:2–22.
[5] Friston KJ, Buchel C. Attentional modulation of effective connectivity from V2 to V5/MT in humans. Proc Natl Acad Sci USA 2000; 97:7591–6.
[6] Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 1969;37:424–38.
[7] Granger CWJ. Testing for causality: a personal viewpoint. J Econ Dynamics Control 1980;2:329–52.
[8] Formisano E, Linden DE, Di Salle F, Trojano L, Esposito F, Sack AT, Grossi D, Zanella FE, Goebel R. Tracking the mind's image in the brain I: time-resolved fMRI during visuospatial mental imagery. Neuron 2002;35:185–94.
[9] Formisano E, Goebel R. Tracking cognitive processes using fMRI mental chronometry. Curr Opin Neurobiol 2003;13:174–81.
[10] Geweke J. Measurement of linear dependence and feedback between multiple time series. J Am Stat Assoc 1982;77:304–13.
[11] Geweke J. Measures of conditional linear dependence and feedback. J Am Stat Assoc 1984;79:907–15.
[12] Bernasconi C, Konig P. On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. Biol Cybern 1999;81:199–210.
[13] Freiwald WA, Valdes P, Bosch J, Biscay R, Jimenez JC, Rodriguez LM, Rodriguez V, Kreiter A K, Singer W. Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. J Neurosci Meth 1999;94:105–19.

[14] Kaminski M, Ding M, Truccolo WA, Bressler SL. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biol Cybern 2001;85: 145–57.
[15] Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. Neurophysiological investigation of the basis of the fMRI signal. Nature 2001;412:150–7.
[16] Boynton GM, Engel SA, Glover GH, Heeger DJ. Linear systems analysis of functional magnetic resonance imaging in human V1. J Neurosci 1996;16:4207–21.
[17] Liu H, Gao J. An investigation of the impulse functions for the nonlinear BOLD response in functional MRI. Magn Reson Imaging 2000;18:931–8.
[18] Allport DA, Styles EA, Hsieh S. Shifting intentional set: exploring the dynamic control of tasks. In Moscovitch CUM, editor. Attention and Performance XV. Cambridge, MA: MIT Press, 1994. p. 421–52.
[19] Dove A, Pollmann S, Schubert T, Wiggins CJ, von Cramon DY. Prefrontal cortex activation in task switching: an event-related fMRI study. Brain Res Cogn Brain Res 2000;9:103–9.
[20] Rushworth MF, Hadland KA, Paus T, Sipila PK. Role of the human medial frontal cortex in task switching: a combined fMRI and TMS study. J Neurophysiol 2002;87:2577–92.
[21] Sohn MH, Ursu S, Anderson JR, Stenger VA, Carter CS. Inaugural article: the role of prefrontal cortex and posterior parietal cortex in task switching. Proc Natl Acad Sci USA 2000;97:13448–53.
[22] Toni I, Ramnani N, Josephs O, Ashburner J, Passingham RE. Learning arbitrary visuomotor associations: temporal dynamic of brain activity. Neuroimage 2001;14:1048–57.
[23] Toni I, Rushworth MF, Passingham RE. Neural correlates of visuomotor associations. Spatial rules compared with arbitrary rules. Exp Brain Res 2001;141:359–69.
[24] Toni I, Rowe J, Stephan KE, Passingham RE. Changes of corticostriatal effective connectivity during visuomotor learning. Cereb Cortex 2002;12:1040–7.
[25] Goebel R, Roebroeck AF, Kim D-S, Formisano E. A framework for the investigation of directed cortical interactions: theoretical background and application to dynamic sensorimotor mapping. In: Kanwisher N, Duncan J, editors. Functional neuroimaging of visual cognition. Attention and Performance XX. Cambridge: Oxford University Press. In press.
[26] Talairach J, Tournoux P. Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging. Stuttgart: Thieme, 1988.
[27] Kriegeskorte N, Goebel R. An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical MR volumes. Neuroimage 2001;14:329–46.
[28] Bagarinao E, Sato S. Algorithm for vector autoregressive model parameter estimation using an orthogonalization procedure. Ann Biomed Eng 2002;30:260–71.