NeuroImage 117 (2015) 202–221

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Dynamic causal modelling of brain–behaviour relationships

L. Rigoux [a], J. Daunizeau [a,b,*]

[a] Brain and Spine Institute, Paris, France
[b] ETH, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 24 November 2014
Accepted 15 May 2015
Available online 22 May 2015

Keywords:
fMRI
Brain connectivity
DCM
Mediation analysis
Encoding
Decoding
Brain lesion
Deficit
Functional degeneracy
Functional recovery

ABSTRACT

In this work, we expose a mathematical treatment of brain–behaviour relationships, which we coin behavioural Dynamic Causal Modelling or *bDCM*. This approach aims at decomposing the brain's transformation of stimuli into behavioural outcomes, in terms of the relative contribution of brain regions and their connections. In brief, *bDCM* places the brain at the interplay between stimulus and behaviour: behavioural outcomes arise from coordinated activity in (hidden) neural networks, whose dynamics are driven by experimental inputs. Estimating neural parameters that control network connectivity and plasticity effectively performs a neurobiologically-constrained approximation to the brain's input–outcome transform. In other words, neuroimaging data essentially serves to enforce the realism of *bDCM*'s decomposition of input–output relationships. In addition, post-hoc artificial lesions analyses allow us to predict induced behavioural deficits and quantify the importance of network features for *funnelling* input–output relationships. This is important, because this enables one to bridge the gap with neuropsychological studies of brain-damaged patients. We demonstrate the face validity of the approach using Monte-Carlo simulations, and its predictive validity using empirical fMRI/behavioural data from an inhibitory control task. Lastly, we discuss promising applications of this work, including the assessment of functional degeneracy (in the healthy brain) and the prediction of functional recovery after lesions (in neurological patients).

Introduction

Humans are largely unaware of the mechanisms that determine the way they process information and how they respond to it. These mechanisms can be described at the psychological level (e.g., in terms of perceptions, emotions and/or intentions) and at the neurobiological level (e.g. in terms of the specific involvement of brain regions and/or neuromodulatory systems). Bridging these two levels of description is the hallmark of cognitive neuroscience. But can we use neuroimaging data to understand how brain networks *funnel* the impact of (incoming) relevant information onto the production of (overt) behaviour? This work exposes a mathematical approach that allows us to decompose the specific contribution of brain regions and their interactions to the behavioural response, given whole-brain neuroimaging time series.

The gold-standard of modern neuropsychological investigations of brain–behaviour relationships relies upon identifying the behavioural deficits of patients exhibiting focal brain lesions or atrophies (Godefroy et al., 1998). Although this approach provides invaluable evidence for causal brain–behaviour relationships, it has two severe drawbacks: (i) it limits the scope of neuroscientific investigations to those behavioural processes that are specifically impaired in accessible brain-damaged patients, and (ii) its interpretation can be partially confounded by functional recovery induced by brain plasticity mechanisms (see e.g., Duffau et al., 2003). The latter issue is a consequence of the brain's *functional degeneracy* (Price and Friston, 2002; Friston and Price, 2003), i.e. the ability of its (structurally different) elements to perform the same function or yield the same output (Edelman and Gally, 2001). Taken together, these concerns make functional neuroimaging in healthy subjects a necessary complement to brain-damaged patient studies.

Functional degeneracy parallels the notion of *functional integration* within brain networks (Price and Friston, 2002), which suggests that the functional role of cerebral components (brain regions, neural ensembles, neurons, …) is largely determined by the influence they exert onto each other (Zeki and Shipp, 1988; Tononi et al., 1994). In this context, the past decade has witnessed a paradigm switch in human brain mapping research. In addition to localizing brain regions that encode specific sensory, motor or cognitive processes, neuroimaging data is nowadays further exploited to understand how information is transmitted through brain networks (Sporns, 2007). The ambition here is to ask questions such as: "what is the nature of the information that region A passes on to region B?". Such analysis of brain imaging data relies on biophysical models of how the brain is wired and how it reacts in different situations (Valdes-Sosa et al., 2011), a seminal example of which is Dynamic Causal Modelling or DCM (Friston et al., 2003). DCM has become a standard tool for identifying the connectivity structure and plasticity of functional brain networks from neuroimaging data

* Corresponding author at: Motivation, Brain and Behaviour Group, Brain and Spine Institute, 47, bvd de l'Hopital, 75013 Paris, France.
E-mail address: jean.daunizeau@gmail.com (J. Daunizeau).

http://dx.doi.org/10.1016/j.neuroimage.2015.05.041
1053-8119/

(Daunizeau et al., 2011a). It has proven useful in disclosing neurobiological mechanisms underlying, for example, associative learning (Den Ouden et al., 2010), speech comprehension (Leff et al., 2008), action observation (Hillebrandt et al., 2014) or motivational processes (Schmidt et al., 2011). At present, DCM is the most suitable framework within which to address the problem of comprehending how information enters, propagates, and reverberates through brain networks (Smith et al., 2011). This is because it is based upon a generative model that describes how stimuli and/or task instructions induce changes in (hidden) coupled neuronal states that cause variability in the observed (local) neuroimaging time series. DCM does not, however, explain how distributed brain responses are causally involved in the production of functional outcomes (e.g., perceptual content, memory retrieval, decisions, etc.).

This is the issue we propose to address in this work. In doing so, our objective is less about picturing the connectivity of brain networks than about explaining how it eventually controls trial-by-trial functional outcomes. More precisely, we aim at performing a neurocognitive decomposition of the *transfer function* from experimental inputs (stimuli and/or task instructions) to functional outcomes (behavioural responses), through activity in the underlying large-scale brain network dynamics. Critically, this decomposition is performed under the constraint that its intermediary (neural) states are maximally similar to brain measurements. This requires (i) augmenting DCM with a mapping from hidden neuronal states to behavioural outputs, (ii) extending current probabilistic system identification techniques to deal with concomitant empirical recordings of neuroimaging and behavioural data time series, and (iii) developing post-hoc artificial lesions' analyses that can flag those network features that are critical for a given input–output transfer function. As will become clearer later on, we will borrow inspiration from classical "decoding" schemes (Haynes and Rees, 2006; Serences and Saproo, 2012) to construct our mathematical mapping from brain activity to functional outcomes. In this view, segregation and integration within brain networks determine how incoming information selectively flows through (large-scale) brain networks, whose elements cooperate to produce functional outcomes. Here, neuroimaging data serves to identify critical parameters (e.g., synaptic weights and their modulation by experimental manipulations) that control the input–state–output transfer function. In turn, this endows our decomposition of input–output relationships with neurobiological realism. The model also allows us to simulate the behavioural deficits that would follow from anatomical lesions on either brain regions or their connections. This is important, because this enables us to quantify how important those large-scale network features are for funnelling the input–output relationship. It also means that we are in a position to quantitatively relate functional neuroimaging studies in healthy subjects with behavioural studies in brain-damaged patients.

This paper is organized as follows.

We first expose the relevant mathematical details of our approach, which we coin *behavioural DCM* or *bDCM* (cf. Fig. 1). In particular, we describe the rationale behind the parametric form of its three constituent mappings, namely: (i) the mapping from experimental inputs to hidden neural state dynamics, (ii) the mapping from neural states to neuroimaging time series, and (iii) the mapping from neural states to behavioural outcomes (which is inspired from decoding schemes). We also summarize the probabilistic (Bayesian) inference machinery, as well as our strategy for predicting the impact of network lesions. We then assess the face validity of the approach using Monte-Carlo simulations and demonstrate it in the context of empirical fMRI data from a healthy subject performing a (Go/No-Go) inhibitory control task. Finally, we discuss the limitations and promising extensions of our work.

Model and methods

In this section, we expose the mathematical details of *bDCM*. We first recall the generative model of DCM for fMRI data, which consists of a dynamical mapping from experimental inputs to evoked dynamics of hidden neural states, which drive spatio-temporal variations in the BOLD signal. We then show how to augment this model with a mapping from hidden neural states to overt behavioural outcomes, borrowing intuitions from decoding approaches. Finally, we summarize the ensuing probabilistic (Bayesian) model inversion scheme, as well as our strategy for predicting behavioural deficits that result from (simulated) lesions on the model. Importantly, we also highlight how to derive quantitative indices of the behavioural relevance of network connections, and how these differ from standard DCM estimates of effective connectivity.

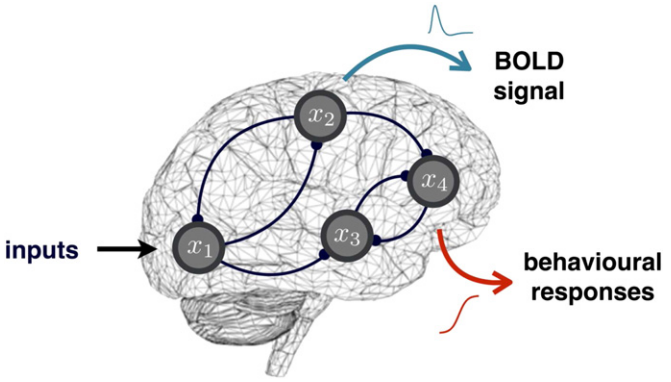

Fig. 1. Schematic representation of *bDCM*. Experimental stimuli enter the system as inputs, which are propagated and reverberated through the network, whose activity x evolves according to models borrowed from dynamical systems' theory. This evoked activity both generates a hemodynamic response visible in the fMRI BOLD signal and eventually produces a behavioural response (e.g. an observable decision such as a button press).

DCM for fMRI data: predicting distributed BOLD responses

DCM is based upon a generative model, i.e. a quantitative description of the mechanisms by which observed data are generated. In particular, large-scale brain networks are described in terms of segregated brain regions that influence each other through reciprocal connections, whose strength can be specifically modulated by the context (e.g. task instructions). A key property of this realistic scenario is that any input entering the network quickly reverberates through recurrent (feedback) connections, which compromises qualitative predictions of its impact on the network (Daunizeau et al., 2011b). In turn, modelling how experimental manipulations (u) drive the dynamics of distributed hidden neural states (x) has to rely upon dynamical system theory. Let us assume that distributed neural dynamics obey the following ordinary differential equation:

$$\dot{x} = f(x,u) \tag{1}$$

where $\dot{x} \equiv dx/dt$ is the rate of change of the system's neural states and f captures the biophysical mechanisms that determine the impact of u onto the temporal evolution of x. The parametric form of the neural states' evolution function that is used in DCM for fMRI derives from a second-order Taylor expansion of the unknown function f, as follows (Friston et al., 2003; Stephan et al., 2008):

$$\begin{aligned} f(x,u) &= \underbrace{\left.\frac{\partial f}{\partial x}\right|_0}_{A} x + \underbrace{\left.\frac{\partial f}{\partial u}\right|_0}_{C} u + \sum_j u_j \underbrace{\left.\frac{\partial^2 f}{\partial x \partial u_j}\right|_0}_{B^{(j)}} x + \sum_i x_i \underbrace{\left.\frac{\partial^2 f}{\partial x \partial x_i}\right|_0}_{D^{(i)}} x + \ldots \\ &\Downarrow \\ f(x,u,\theta) &\triangleq Ax + \sum_j u_j B^{(j)} x + Cu + \sum_i x_i D^{(i)} x \end{aligned} \tag{2}$$

where the gradients of f are evaluated at the system's steady state in the absence of inputs ($x = 0, u = 0$). The second line of Eq. (2) shows DCM's