Comparing Dynamic Causal Models

W.D. Penny*, K.E. Stephan, A. Mechelli and K.J. Friston,
Wellcome Department of Imaging Neuroscience,
University College London.

October 31, 2003

Abstract

This article describes the use of Bayes factors for comparing Dynamic Causal Models (DCMs). DCMs are used to make inferences about effective connectivity from functional Magnetic Resonance Imaging (fMRI) data. These inferences, however, are contingent upon assumptions about model structure, that is, the connectivity pattern between the regions included in the model. Given the current lack of detailed knowledge on anatomical connectivity in the human brain, there are often considerable degrees of freedom when defining the connectional structure of DCMs. In addition, many plausible scientific hypotheses may exist about which connections are changed by experimental manipulation, and a formal procedure for directly comparing these competing hypotheses is highly desirable. In this article, we show how Bayes factors can be used to guide choices about model structure, both with regard to the intrinsic connectivity pattern and the contextual modulation of individual connections. The combined use of Bayes factors and DCM thus allows one to evaluate competing scientific theories about the architecture of large-scale neural networks and the neuronal interactions that mediate perception and cognition.

*Corresponding author: w.penny@fil.ion.ucl.ac.uk

1 Introduction

Human brain mapping has been used extensively to provide functional maps showing which regions are specialised for which functions [14]. A classic example is the study by Zeki et al. (1991) [48] who identified V4 and V5 as being specialised for the processing of colour and motion, respectively. More recently, these analyses have been augmented by functional integration studies, which describe how functionally specialised areas interact and how these interactions depend on changes of context. These studies make use of the concept of effective connectivity defined as the influence one region exerts over another as instantiated in a statistical model. A classic example is the study by Buchel and Friston [8] who used Structural Equation Modelling (SEM) to show that attention to motion modulates connectivity in the dorsal stream of the visual system.

In a recent paper [18] we have proposed the use of Dynamic Causal Models (DCMs) for the analysis of effective connectivity. DCM posits a causal model whereby neuronal activity in a given region causes changes in neuronal activity in other regions, via inter-regional connections, and in its own activity, via self-connections. Additionally, any of these connections can be modulated by contextual variables like cognitive set or attention. The resulting neurodynamics of the modeled system then give rise to fMRI time series via local hemodynamics which are characterised by an extended Balloon model [16, 10].

A DCM is fitted to data by tuning the neurodynamic and hemodynamic parameters so as to minimise the discrepancy between predicted and observed fMRI time series. Importantly, however, the parameters are constrained to agree with a-priori specifications of what range the parameters are likely to lie within. These constraints, which take the form of a prior distribution, are then combined with data via a likelihood distribution to form a posterior distribution according to Bayes' rule. Changes in effective connectivity can then be inferred using Bayesian inference based on the posterior densities.

In this paper we apply Bayesian inference not just to the parameters of DCMs, as in [18], but to the models themselves. This allows us to make inferences about model structure, that

is, which of several alternative models is optimal given the data. Such decisions are of great practical relevance because we still lack detailed knowledge about the anatomical connectivity of the human brain [35]. Decisions about the intrinsic connectivity of DCMs are therefore usually based on inferring connections from supposedly equivalent areas in the Macaque brain for which the anatomical connectivity is well known [43]. This procedure has many pitfalls, however, including a multitude of incompatible parcellation schemes and frequent uncertainties about the homology and functional equivalence of areas in the brains of man and monkey. This problem may be less severe in sensory systems, but is of particular importance for areas involved in higher cognitive processes like language [1]. Thus, there are often considerable degrees of freedom when defining the connectional structure of DCMs of the human brain. We show how Bayes factors can be used to guide the modeller in making such choices. A second question concerning model structure is which of the connections included in the model are modulated by experimentally controlled contextual variables (e.g. attention). This choice reflects the modeller's hypothesis about where context-dependent changes of effective connectivity occur in the modeled system. We demonstrate how Bayesian model selection can be used to distinguish between competing models that represent the many plausible hypotheses.

The paper is structured as follows. In section 2 we introduce briefly the neurobiological context in which DCM is usually applied. We focus particularly on hierarchical models and the distinction between anatomical and functional characterisations. In section 3 we review Dynamic Causal Modelling from a theoretical perspective by defining the neurodynamic and hemodynamic models. In section 4 we describe Bayesian estimation and the Bayes factors that are used to weigh evidence for and against competing scientific hypotheses. Results on simulated and experimental data are presented in section 5.

1.1 Notation

We use upper-case letters to denote matrices and lower-case to denote vectors. $\mathsf{N}(m, \Sigma)$ denotes a uni/multivariate Gaussian with mean m and variance/covariance Σ. I_K denotes the $K \times K$ identity matrix, 1_K is a $1 \times K$ vector of 1s, 0_K is a $1 \times K$ vector of zeros, if X is a matrix,

X_{ij} denotes the i,jth element, X^T denotes the matrix transpose and $\mathsf{vec}(X)$ returns a column vector comprising its columns, $\mathsf{diag}(x)$ returns a diagonal matrix with leading diagonal elements given by the vector x, $\otimes$ denotes the Kronecker product (see Appendix) and $\log x$ denotes the natural logarithm.

2 Neurobiological issues

Many applications of DCM, both in this article and in previous work [18, 32], refer to "bottom-up" and "top-down" processes, and we envisage that a large number of future applications of DCM will rest on this distinction. Some of the possible DCM architectures for modeling these processes may, at first glance, seem at odds with traditional cognitive theories that relate bottom-up processes to so-called "forward" connections and top-down processes to "backward" connections [46]. Here we try to clarify this relationship, using some simple examples from the visual system, and emphasize the need for precise terminology when distinguishing between the levels of anatomical connectivity (forward vs. backward connections) and cognitive processes (bottom-up vs. top-down).

Classical theories of visual information processing posit a hierarchy of cortical areas, each performing a specialized analysis and feeding the results of its computations to the next (i.e. higher) level [30]. The anatomical basis for information transfer from lower to higher areas in this bottom-up model are so-called "forward" (or "feedforward") connections that terminate in the granular layer (i.e. layer IV) of the higher area and originate in both supra- and infragranular layers of the source area [13]. Stimulus-dependent bottom-up processes are not sufficient, however, to explain the effects of contextual factors (e.g. cognitive set, expectation, or attention) that can induce substantial changes in information processing. These modulatory processes are often referred to as top-down processes and are mediated anatomically by so-called "backward" (or "feedback") connections from higher to lower areas which both originate and terminate in infra- and supragranular layers.

The neurophysiological mediation of top-down processing is complex and not well understood, but comprises at least two different mechanisms (see below). Although differential lam-

inar patterns cannot currently be represented in DCMs, one can model simple hierarchies of areas in DCM, and in these hierarchies connections can be classified as forward or backward based on the relative position of the areas in the hierarchy (see Fig. 24 in [18]). It may appear natural to assume that, in DCM, bottom-up effects should always be modeled by a modulation of forward connections, and top-down effects should be modeled by a modulation of backward connections. However, this is not the case.

Consider a very simple example of a DCM that consists of the two reciprocally connected visual areas V1 and V5, with V1 receiving visual input (VIS STIM) (Fig. 3A). Let us imagine that some visual stimuli are moving, whereas others are stationary. It is well established that V5 is particularly sensitive to motion information, i.e. V5 shows increased responsiveness to V1 inputs whenever the stimulus is moving [4] (Fig. 3B). In DCM, this bottom-up process would be modeled by modulating the V1-V5 forward connection by a factor that indicates stimulus motion (MOT, Fig. 3A). However, top-down processes can also be expressed through a modulation of forward connections. For example, imagine that (i) stimuli are always moving, and (ii) attention is sometimes directed to the motion of the stimuli and sometimes to some other stimulus property (e.g. colour). Previous studies have demonstrated that V5 responses to V1 inputs are enhanced whenever motion is attended [8, 17, 33, 11, 45]. This attentional top-down effect conforms to a "gain control" mechanism and is mediated neurophysiologically by backward connections from higher areas (represented by "X" in Fig. 3D). These influence those neurons in V5 which receive inputs from V1 via forward connections [5, 24], to enhance their responsiveness to V1 inputs, possibly through interactions between dendritic and somatic postsynaptic potentials [42] (see Fig. 3D) or voltage-dependent NMDA receptors. Although this level of detail cannot currently be modeled in DCMs, we can model precisely the same mechanism, at a coarser level, by allowing the V1-V5 forward connection to be modulated by attention (Fig. 3C). This approach has been applied to primate single cell data [38].

The behaviour of this model then corresponds to the observed neurophysiology: the magnitude of stimulus-dependent responses in V5 (i.e. the V5 responses to V1 inputs) is augmented whenever motion is attended. These examples show that modulation of forward connections can

represent a bottom-up process (if the contextual input refers to a stimulus property; Fig. 3A) as well as a top-down mechanism (if the contextual input represents cognitive set like attention, Fig. 3C).

In addition to stimulus-locked, multiplicative gain control mechanisms, attentional top-down modulation can be achieved by at least one more process. For example, during attention an enduring shift in the baseline responses of visual areas has even been observed in the absence of stimuli [28, 27, 11]. Neurophysiologically, this additive baseline shift is believed to be mediated by backward connections that do not, as in the case of the gain control mechanism, simply sensitize post-synaptic cells to inputs from lower areas, but exert a more direct, "driving" effect on neurons in the target area [28]. There are various ways of modeling this. If one does not know what area might represent the source of this attentional top-down effect, one can model the influence of attention to motion onto V5 as a direct, additive increase in V5 activity (ATT-MOTION, Fig. 4A). If, however, one has reason to believe that a particular area, e.g. the superior parietal cortex (SPC) in this example, mediates this effect, it can be included in the model as shown in Fig. 4B. Here, attention drives SPC whose backward connections activate V5. This models an increase in attentional effects in a purely additive way, but may be a sufficient explanation for the data.

Further plausible ways of modeling top-down mechanisms in DCMs exist, including modulation of self-connections (which would correspond to modeling a context-dependent change of intra-areal self-inhibition), but we will not go into further details here. The main message of this section is that, depending on the exact mechanism that one models and the nature of the modulatory input, top-down effects can be mediated both by modulation of forward and backward connections. To this end, it is useful to distinguish between the type of anatomical connections included in the model (forward vs. backward connections) and the cognitive processes modeled (bottom-up vs. top-down). We will return to some of these issues later because they provide a very nice example of alternative architectures for attention that can be disambiguated using Bayesian model selection.

3 Dynamic Causal Models

Dynamic Causal Models have been proposed recently [18] as a method for the analysis of functional integration. The first step, in such an analysis, is the identification of a set of $i = 1..L$ regions that comprise the system we wish to study. These can be found via results of previous imaging studies or from analyses of functional specialisation using standard General Linear Model (GLM) approaches [15]. The second step is the specification of a set of $j = 1..M$ experimental variables that act as inputs to the system. Each input can be of a driving nature, whereby activity in a given area is directly altered, or of a modulatory nature, whereby changes in activity occur indirectly via changes in connection strengths. In Buchel and Friston [8], for example, the driving input was the experimental variable describing when moving images were presented to a subject and the modulatory input was a variable describing when that subject was instructed to attend to possible velocity changes. The neurophysiological system comprised three regions in the visual pathway.

The effective connectivity in DCM is characterised by a set of 'intrinsic connections' that specify which regions are connected and whether these connections are unidirectional or bidirectional. We also define a set of input connections that specify which inputs are connected to which regions, and a set of modulatory connections that specify which intrinsic connections can be changed by which inputs. The overall specification of input, intrinsic and modulatory connectivity comprise our assumptions about model structure. This in turn represents a scientific hypothesis about the structure of the large-scale neuronal network mediating the underlying cognitive function.

Figure 1 shows an example of a DCM network. DCMs comprise a bilinear model for the neurodynamics and an extended Balloon model [16, 10] for the hemodynamics. The next two sub-sections cover each of these topics, specifying what the model parameters are and their prior distributions. Section 2.3 then describes the likelihood distribution for a DCM model and specifies how the neurodynamic and hemodynamic priors are combined into an overall DCM prior.

3.1 Neurodynamics

The neurodynamic parameters are the intrinsic, modulatory and input connectivity matrices that define the multivariate differential equation governing neuronal activity

$$\dot{z}_t = \left(A_u + \sum_{j=1}^{M} u_t(j) B_u^j \right) z_t + C u_t \tag{1}$$

where t indexes continuous time and the dot notation denotes a time derivative. This is known as a bilinear model because the dependent variable, $\dot{z}_t$, is linearly dependent on the product of z_t and u_t. That u_t and z_t combine in multiplicative fashion endows the model with 'nonlinear' dynamics that can be understood as a nonstationary linear system that changes according to u_t. Importantly, because u_t is known, parameter estimation is tractable.

The neuronal activity z_t is an $L \times 1$ vector comprising activity in each of the L regions and the input u_t is an $M \times 1$ vector comprising the scalar inputs $u_t(j)$ where $j = 1..M$. The intrinsic connectivity matrix A_u and modulatory connectivity matrix B_u^j are of dimension $L \times L$ and the input matrix C is of dimension $L \times M$. Here, the u subscripts in A_u and B_u^j denote that the matrix elements are 'unnormalised' as described below. Later we will describe 'normalised' A and B^j matrices.

In the intrinsic and modulatory matrices, an entry in row i and column k denotes a connection from region k to region i. If the regions form a hierarchy then entries in the lower diagonal therefore constitute forward connections and entries in the upper diagonal are backward connections. In the example network in figure 1 the matrices are

$$\begin{aligned} A_u &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \\ B_u^2 &= \begin{bmatrix} B_{11}^2 & 0 \\ 0 & B_{22}^2 \end{bmatrix} \\ C &= \begin{bmatrix} C_{11} & 0 \\ 0 & C \end{bmatrix} \end{aligned} \tag{2}$$

where, for example, A_{21} is the forward connection to the higher cortical region, region 2, from the lower region, region 1.

Following [18], the self-connections are enforced to take on the same value, σ, in all regions

by constraining the connectivity matrices as follows

$$\begin{aligned} A_u &= \sigma(A - I) \\ B_u^j &= \sigma B^j \end{aligned} \tag{3}$$

where the A and B^j matrices on the right are in 'normalised' form. The A matrix is defined to have zero entries on the diagonal resulting in diagonal entries in A_u that are identically equal to $-\sigma$. This enforces the intrinsic neuronal time constants to be the same in all regions. The normalised coupling parameters are more interpretable as they express the strength of a connection between regions relative to the strength of self-connections. Further, normalised connections are more robust to slice-timing errors and to regional variations in hemodynamic response [18].

In the abscence of coupling between areas the time-constant of neuronal transients (ie. the half-life) is given by $\tau = \log 2/\sigma$. A priori we know what range of time-constants are physiologically plausible and we can put this information into DCM via the prior distribution

$$p(\sigma) = \mathsf{N}(\eta_\sigma, C_\sigma) \tag{4}$$

where $\eta_\sigma = 1$ and C_σ is set so as to render the probability of obtaining negative σ's arbitrarily small. In this paper, as in [18], we use a value of $C_\sigma = 0.105$ which makes this probability 0.001. The distribution of time-constants is therefore given by $p(\tau) = p(\sigma)\partial\sigma/\partial\tau$. We note that the expected neuronal time constant, $< \tau >$, is therefore determined by both η_σ and C_σ (note $< \tau > \neq \frac{\log 2}{<\sigma>}$ as the transformation between σ and τ is nonlinear [34]). The nature of this prior distribution can be appreciated by drawing samples from it as shown in Figure 2. This shows that the expected neuronal time constant is about 900ms.

We can then define a vector of neurodynamic parameters as the neuronal time constant concatenated with vectorised connectivity matrices. That is

$$\theta^c = \begin{bmatrix} \sigma \\ \mathsf{vec}(A) \\ \mathsf{vec}(B) \\ \mathsf{vec}(C) \end{bmatrix} \tag{5}$$

Model structure is defined by specifying which entries in the above matrices are allowed to take

on non-zero values ie. which inputs and regions are connected. A given model, say model m, is then defined by its pattern of connectivity. Note that only connections which are allowed to be non-zero will appear in θ^c. For a network with N_a intrinsic, N_b modulatory and N_c input connections θ^c will have $N_\theta = N_a + N_b + N_c + 1$ entries.

Priors are placed on the A and B^j matrices so as to encourage parameter estimates that result in a stable dynamic system (see section 2.3.1 in [18] for a discussion). For each connection in A and B^j the prior is

$$p(A_{ik}) = \mathsf{N}(0, v_a) \tag{6}$$
$$p(B^j_{ik}) = \mathsf{N}(0, v_b)$$

where the prior variance v_a is set to ensure stability with high probability (see Appendix A.3 in [18] for a discussion of this issue). For each connection in C the prior is

$$p(C_{im}) = \mathsf{N}(0, v_c) \tag{7}$$

These priors are so-called 'shrinkage-priors' because the posterior estimates shrink towards the prior mean, which is zero. The size of the prior variance determines the amount of shrinkage. The above information can be concatenated into the overall prior

$$p(\theta^c) = \mathsf{N}(\theta^c_p, C^c_p) \tag{8}$$

where the p subscripts denote priors and

$$\begin{aligned} \theta^c_p &= [\eta_\sigma, 0_{N_\theta - 1}]^T \\ C^c_p &= \mathsf{diag}[C_\sigma, v_a 1_{N_a}, v_b 1_{N_b}, v_c 1_{N_c}] \end{aligned} \tag{9}$$

This completes our description of the prior distribution of neurodynamic parameters. For any given θ^c we can integrate equation 1 and obtain neuronal time series for each region. These are shown for our example DCM in the bottom panel of figure 1. We can then relate this neuronal activity to an fMRI time series via the hemodynamic process described in the following section.

3.2 Hemodynamics

In DCM the hemodynamics are described by the Balloon model first described by Buxton et al. [10] and developed further by Friston et al. [19, 16]. DCM uses a separate Balloon model for each region. For the ith region, neuronal activity z_i causes an increase in vasodilatory signal s_i that is subject to auto-regulatory feedback. Inflow f_i responds in proportion to this signal with resulting changes in blood volume v_i and deoxyhemoglobin content q_i

$$\begin{aligned} \dot{s}_i &= z_i - \kappa_i s_i - \gamma_i(f_i - 1) \\ \dot{f}_i &= s_i \\ \tau_i \dot{v}_i &= f_i - v_i^{1/\alpha_i} \\ \tau_i \dot{q}_i &= f_i \frac{1-(1-\rho_i)^{1/f_i}}{\rho_i} - v_i^{1/\alpha_i} \frac{q_i}{v_i} \end{aligned} \tag{10}$$

where in region i, κ_i is the rate of signal decay, γ_i is the rate of flow-dependent elimination, τ_i is the hemodynamic transit time, α_i is Grubb's exponent and ρ_i is the resting oxygen extraction fraction. The biophysical parameters can be concatenated into the vector $\theta^h = \{\kappa_i, \gamma_i, \tau_i, \alpha_i, \rho_i\}$, for $i = 1..L$. Priors are placed on the biophysical parameters to ensure biological plausibility

$$p(\theta^h) = \mathsf{N}(\theta_p^h, C_p^h) \tag{11}$$

where

$$\begin{aligned} \theta_p^h &= 1_L \otimes h_{mean} \\ C_p^h &= I_L \otimes H_{cov} \end{aligned} \tag{12}$$

and h_{mean}, H_{cov} are the prior means and covariances which are the same for each region and $\otimes$ denotes the Kronecker product (see Appendix). The prior means and variances are shown in table 1 in [18] and were computed from data collected during a word presentation fMRI experiment as follows. For each of 128 voxels, hemodynamic parameters were estimated using a nonlinear function minimization routine [19] and the means and variances of the parameter estimates over these 128 voxels were then used as our prior means and variances. The predicted

BOLD signal in region i is then related to blood volume and deoxyhemoglobin content as follows

$$\begin{aligned} h_i &= g(v_i, q_i) \\ &= 2(7\rho_i(1-q_i) + 2(1-\frac{q_i}{v_i}) + (2\rho_i - 0.2)(1-v_i)) \end{aligned} \tag{13}$$

For a particular setting of the biophysical parameters, θ^h, one can take the neuronal activity in a given region, z_i, integrate equation 10 and pass the resulting blood volume and deoxyhemoglobin content values, v_i and q_i, through the nonlinearity in equation 13. This then gives rise to an fMRI time series.

The nature of the prior distribution over hemodynamic parameters can be appreciated by plotting the hemodynamic response to neuronal transients for various values of θ_h sampled from $p(\theta_h)$, as shown in figure 2. The average hemodynamic response peaks at 4s which perhaps seems a little early. However, one must bear in mind that these are responses to transients from isolated regions. Connected regions result in more persistent neuronal dynamics which have the effect of delaying the peak hemodynamic response as shown for example in figure 10 of [18].

3.3 Overall prior and likelihood

We concatenate all neurodynamic and hemodynamic parameters into the overall p-dimensional parameter vector

$$\theta = \begin{bmatrix} \theta^c \\ \theta^h \end{bmatrix} \tag{14}$$

This vector contains all the parameters of a DCM model that we need to estimate. Consequently the prior mean and covariance are given by

$$\begin{aligned} \theta_p &= \begin{bmatrix} \theta_p^c \\ \theta_p^h \end{bmatrix} \\ C_p &= \begin{bmatrix} C_p^c & 0 \\ 0 & C_p^h \end{bmatrix} \end{aligned} \tag{15}$$

The neurodynamics and hemodynamics combine to produce a multivariate time series of observations as follows

$$\dot{x} = f(x, u, \theta) \tag{16}$$

$$h(\theta, u) \quad = \quad g(x)$$

with states $x = \{z, s, f, v, q\}$. For given input u, and DCM parameters θ, model predictions can be produced by integrating the state equation as described in [18, 16]. This integration is efficient because most fMRI experiments result in input vectors that are highly sparse. For a data set with N_s scans we can then create a $LN_s \times 1$ vector of model predictions $h(\theta, u)$ covering all time points and all areas (in the order all time points from region 1, region 2 etc.). The observed data y, also formatted as an $LN_s \times 1$ vector, is then modelled as

$$y = h(\theta, u) + X\beta + w \tag{17}$$

where w is an $LN_s \times 1$ vector of Gaussian prediction errors with mean zero and covariance matrix C_e, X contains effects of no interest and β is an unknown vector of parameters to be estimated. In [18] the error covariance described both autoregressive and white noise processes and simulations showed the estimation was robust to misspecification of the error process. It is sufficient therefore to characterise the prediction error in each region as a white noise process. That is, $C_e = I_{N_s} \otimes \Lambda$ where Λ is an $L \times L$ diagonal matrix with Λ_{ii} denoting error variance in the ith region.

4 Bayesian Estimation and Inference

This section consists of two parts that describe how Bayesian inference is used (i) to estimate the parameters a DCM model and (ii) to compare different models. These may be regarded as the first and second levels of Bayesian inference.

In the first part we describe how the DCM prior and likelihoods are combined via Bayes rule to form the posterior distribution. Section 3.1 sets out some notation and section 3.2 describes how the posterior is computed iteratively using an Expectation-Maximisation (EM) algorithm.

The second part, starting in section 3.3, describes how to compute the model evidence. This can be decomposed into two types of term: accuracy terms and complexity terms. The best model, or one with the highest evidence, strikes an optimal balance between the two. In section 3.4 we describe how Bayes factors, ratios of model evidences, are used to compare different

models and in section 3.5 suggest how Bayes factors be used to make decisions. We also present a coding perspective on Bayesian model comparison in section 3.6.

Readers not familiar with Bayesian modelling are referred to [20]. More specifically the Laplace approximations, model evidences and Bayes factors that we shall encounter are described in [25, 26, 37].

4.1 Parameter Priors and Likelihoods

We now set out some notation that both summarises the definitions in section 2 and that will be used to derive further quantities, such as posterior distributions and model evidences. The parameter prior and likelihood are

$$\begin{aligned} p(\theta|m) &= \mathsf{N}(\theta_p, C_p) \\ p(y|\theta, m) &= \mathsf{N}(h(\theta, u), C_e) \end{aligned} \tag{18}$$

These can be expanded as

$$\begin{aligned} p(\theta|m) &= (2\pi)^{-p/2}|C_p|^{-1/2}\exp(-\frac{1}{2}e(\theta)^T C_p^{-1} e(\theta)) \\ p(y|\theta, m) &= (2\pi)^{-N_s/2}|C_e|^{-1/2}\exp(-\frac{1}{2}r(\theta)^T C_e^{-1} r(\theta)) \end{aligned} \tag{19}$$

where

$$\begin{aligned} e(\theta) &= \theta - \theta_p \\ r(\theta) &= y - h(\theta, u) - X\beta \end{aligned} \tag{20}$$

are the 'parameter errors' and 'prediction errors'.

4.2 Estimation of Parameter Posteriors

From Bayes' rule the posterior distribution is equal to the likelihood times the prior divided by the evidence [20]

$$p(\theta|y, m) = \frac{p(y|\theta, m)p(\theta|m)}{p(y|m)} \tag{21}$$

Taking logs gives

$$\log p(\theta|y,m) = \log p(y|\theta,m) + \log p(\theta|m) - \log p(y|m) \tag{22}$$

The parameters that maximise this posterior probability, the Maximum Posterior (MP) solution, can then be found using a Gauss-Newton optimisation scheme whereby parameter estimates are updated in the direction of the gradient of the log-posterior by an amount proportional to its curvature (see e.g. [36]). The model parameters are initialised to the mean of the prior density.

If the proportion of data points to model parameters is sufficiently large, as is the case with DCM models of fMRI time series, then the posterior is well approximated with a Gaussian. The aim of optimisation is then to estimate the mean and covariance of this density which can be achieved using an Expectation-Maximisation (EM) algorithm described in section 3.1 of [16]. In the E-step, the posterior mean, $\hat{\theta}$, and the posterior covariance, $\hat{\Sigma}$, are updated using a Gauss-Newton step and in the M-step the hyper-parameters of the noise covariance matrix, C_e, are updated. These steps are iterated until the posterior distribution

$$p(\theta|y,m) = \mathsf{N}(\theta_{MP}, \Sigma_{MP}) \tag{23}$$

is reached. The posterior density can be used to make inferences about the size of connections as shown, for example, in Figure 12.

In statistics, approximation of a posterior density by a Gaussian centred on the maximum posterior solution is known as a Laplace approximation [25]. The parameters of no interest, β, can also be estimated by forming an augmented parameter vector that includes θ and β and an augmented observation model, as described in Equation 7 of [18].

4.3 Model Evidence, BIC and AIC

The structure of a DCM model is defined by specifying which regions are connected to each other, via the intrinsic connectivity matrix, and which inputs can alter which connections, via the modulatory matrix. A given model, say model m is then defined by this pattern of connectivity. Different models can be compared using the evidence for each model and this can

be thought of as a second-level of Bayesian inference. The model evidence is computed from

$$p(y|m) \quad = \quad \int p(y|\theta, m)p(\theta|m)d\theta \tag{24}$$

Note that the model evidence is simply the normalisation term from the first level of Bayesian inference, given in equation 21. In the appendix we show that, using the Laplace approximation, this leads to an expression for the log model evidence consisting of an accuracy and complexity term defined as follows

$$\log p(y|m)_L = Accuracy(m) - Complexity(m) \tag{25}$$

where

$$Accuracy(m) \quad = \quad -\frac{1}{2}\log|C_e| - \frac{1}{2}r(\theta_{MP})^T C_e^{-1} r(\theta_{MP}) \tag{26}$$

$$Complexity(m) = \frac{1}{2}\log|C_p| - \frac{1}{2}\log|\Sigma_{MP}| + \frac{1}{2}e(\theta_{MP})^T C_p^{-1} e(\theta_{MP}) \tag{27}$$

Use of base-e or base-2 logarithms leads to the log-evidence being measured in 'nats' or 'bits' respectively. The first term in $Accuracy(m)$ can be expressed as the product of the noise variances Λ_{ii} over all regions and the second term will be close to unity as the Λ_{ii} are estimated based on the observed errors $r(\theta_{MP})$. The complexity terms will be discussed further in section 4.6.

The evidence embodies the two conflicting requirements of a good model, that it fit the data and be as simple as possible. The requirement that the model be simple is intuitively appealing and concurs with notions such as Occam's Razor - that one should accept the simplest explanation that fits the data. But is there a mathematical reason for preferring simple models ? Figure 5, presents an argument from Mackay [29], which shows that indeed there is. In brief, although complex models can explain more data, they are suboptimal for any given data.

Computation of $\log p(y|m)_L$ requires inversion of the prior covariance matrix (ie. C_p^{-1}). To compute this quantity it is therefore recommended to use a full-rank prior over the hemodynamic parameters. Alternatively, one can use a lower-rank prior (as in [18]) and compute $\log p(y|m)_L$ by first projecting the hemodynamic parameters onto the relevant subspace.

A drawback of the Laplace approximation, to the model evidence, is its dependence on parameters of the prior density e.g. the prior variance on intrinsic connections v_a. This dependence is particularly acute in the context of DCM where v_a is chosen to ensure (with high probability) that the optimisation algorithm converges to a stable solution. This means it is difficult to compare models with different numbers of connections.

We therefore do not employ the Laplace approximation in this paper but make use of alternative approximations. The first, the Bayesian Information Criterion [41], is a special case of the Laplace approximation which drops all terms that don't scale with the number of data points. In the appendix we show that for a DCM it is given by

$$BIC \quad = \quad Accuracy(m) - \frac{p}{2}\log N_s \tag{28}$$

where p is the number of parameters in the model. The second criterion we use is Akaike's Information Criterion (AIC) [3]. AIC is maximised when the approximating likelihood of a novel data point is closest to the true likelihood, as measured by the Kullback-Liebler divergence (this is shown in [39]). For DCM, AIC is given by

$$AIC \quad = \quad Accuracy(m) - p \tag{29}$$

Though not originally motivated from a Bayesian perspective, model comparisons based on AIC are asymptotically equivalent to those based on Bayes factors [2], ie. AIC approximates the model evidence.

Empirically, BIC is observed to be biased towards simple models and AIC to complex models [25]. Indeed, inspection of Equations 28 and 29 shows that for values of p and N_s typical for DCM, BIC pays a heavier parameter penalty than AIC.

4.4 Bayes factors

Given models $m = i$ and $m = j$ the Bayes factor comparing model i to model j is defined as [25, 26]

$$B_{ij} = \frac{p(y|m=i)}{p(y|m=j)} \tag{30}$$

where $p(y|m=j)$ is the evidence for model j found by exponentiating the approximations to the log-evidence in equations 25, 28 or 29. When $B_{ij} > 1$, the data favour model i over model j, and when $B_{ij} < 1$ the data favour model j.

The Bayes factor is a summary of the evidence provided by the data in favour of one scientific theory, represented by a statistical model, as opposed to another. Just as a culture has developed around the use of p-values in classical statistics (eg. $p < 0.05$), so one has developed around the use of Bayes factors. Raftery [37], for example, presents an interpretation of Bayes factors as shown in Table 1. Jefferys [23] presents a similar grading for the comparison of scientific theories. These partitionings are somewhat arbitrary but do provide rough descriptive statements.

Bayes factors can also be directly interpreted as odds ratios where $B_{ij} = 100$, for example, corresponds to odds of 100-to-1. Bayes factors can be used to convert a prior odds ratio into a posterior odds ratio. For equal prior odds the posterior odds is equal to the Bayes factor. From this we can compute the equivalent posterior probability of hypothesis i as shown, for example, in Table 1.

Bayes factors in Bayesian statistics play a similar role to p-values in classical statistics. In [37], however, Raftery argues that p-values can give misleading results, especially in large samples. The background to this assertion is that Fisher originally suggested the use of significance levels (the p-values beyond which a result is deemed significant) $\alpha = 0.05$ or 0.01 based on his experience with small agricultural experiments having between 30 and 200 data points. Subsequent advice, notably from Neyman and Pearson, was that power and significance should be balanced when choosing α. This essentially corresponds to reducing α for large samples (but they did'nt say *how* α should be reduced). Bayes factors provide a principled way to do this.

The relation between p-values and Bayes factors is well illustrated by the following example due to Raftery [37]. For linear regression models one can use Bayes factors or p-values to decide whether to include an extra regressor. For a sample size of $N_s = 50$, positive evidence in favour of inclusion (say, $B_{12} = 3$) corresponds to a p-value of 0.019. For $N_s = 100$ and 1000 the corresponding p-values reduce to 0.01 and 0.003. If one wishes to decide whether to include multiple extra regressors the corresponding p-values drop more quickly.

Importantly, unlike p-values, Bayes factors can be used to compare non-nested models. They also allow one to quantify evidence in favour of a null hypothesis. Raftery shows [37] how Bayes factors can be computed for linear and logistic regression, generalized linear models and Structural Equation Models. Examples of using Bayes factors for assessing forensic evidence and in probabilistic models in general are given in Mackay [29], and Raftery [37] gives applications in sociology. For example, Raftery compared the two hypotheses about social mobility (how the occupations of fathers and sons are related) in industrialised countries; Hypothesis 1, that social mobility patterns were different in different countries and Hypothesis 2, that frequencies of transition between occupations were similarly symmetric across countries, the data supporting the second hypothesis with $B_{21} > 150$.

A possible disadvantage of Bayes factors is their dependence on parameters of the prior distributions. For this reason we have decided to use AIC and BIC approximations to the model evidence, as described in the previous section.

4.5 Making decisions

If one wishes to make decisions based on Bayes factors then some cut-off value is required. In Bayesian decision theory the choice of cut-off is guided by a 'loss function' or 'utility' which captures the costs of making false positive and false negative decisions [6].

In this paper we suggest a conservative strategy which is to compute Bayes factors based on AIC and BIC and to make a decision *only* if both factors are in agreement. In particular, if both AIC and BIC provide Bayes factors of at least e (the natural exponent 2.7183) we regard this as 'consistent' evidence. Further, we regard consistent evidence as the basis for decision-making, for example the decision to fit new models or the decision to regard one of a number of hypotheses as a 'working hypothesis'.

The reason for this cut-off is as follows. For a simpler model to be favoured over a complex one, the limiting factor is due to AIC. If the simpler model has δ_p fewer parameters and both models are equally accurate then the change in log evidence is $-\delta_p$ nats. The smallest value $\delta_p = 1$ gives a Bayes factor of e.

For a more complex model to be favoured over a simpler one the limiting factor is due to BIC. In this case we can work out the number of scans required to achieve a Bayes factor of e by noting that the change in log-evidence is

$$\Delta BIC = \frac{N_s}{2} \log \left(1 + \frac{\delta_s}{100}\right) + \frac{\delta_p}{2} \log N_s \tag{31}$$

where δ_s is the percentage increase in signal variance. Figure 6, for example, shows that for $\delta_s = 2$ which is typical of the fMRI model comparisons in this paper, about 400 data points are required. Generally, for smaller δ_p and δ_s it is harder to tell models apart. Overall, we 'accept' one model over another if there is a 'nats difference' between them.

For the case of comparing a simpler model to a more complex one with $\delta_p = 1$ this cut-off results in a very conservative test. This is because even if the two models are truly equally accurate, on any given finite data set one model will appear more accurate than the other. Because this will be the simpler model for half of such data sets the sensitivity of the test is 50%. This test does, however, have a high specificity as no decision is made if the cut-off is not exceeded. As δ_p increases so does the sensitivity.

Finally, we note that a Bayes factor of e corresponds to a posterior probability of 73%, ie. there is a 27% probability that our decision is incorrect ! This may seem extraordinarily high but, as indicated in the previous section, our experience with p-values does not translate in a straightforward way to posterior probabilities.

If we assume that quantities governing statistical inference, such as the variance of parameter estimates, scale in DCM as they do in linear regression then, given typical fMRI sample sizes of 200-400 scans, a Bayes factor of e would correspond to a p-value of less than 0.01 (see linear regression example in section 4.4). This seems quite reasonable.

4.6 Coding Perspective

In this section we consider Bayesian model comparison from an information theoretic or 'coding' perspective. Imagine one wished to transmit a data set over a communication channel. This could be done by simply digitizing the data and transmitting it. It would occupy a certain number of bits of the channel. Alternatively, one could fit a model to the data and then send

the model parameters and the prediction errors, the total number of bits required being the sum of the parameter bits and the error bits. Better models require fewer bits to be transmitted and for data containing discernible patterns model-based coding is superior. This is the rationale behind the Minimum Description Length (MDL) model comparison criterion [47]. In fact, a version of MDL [40] is equal to the negative of the BIC, ie. MDL=-BIC. The link with Bayesian inference is that the sender and receiver must agree on the transmission protocol so that they know how to encode and decode the messages. The choice of coding scheme for the parameters corresponds to the choice of prior and the choice of coding scheme for the errors corresponds to the likelihood.

In information theory [12] the 'information content' of an event, x, is related to its probability by

$$S(x) = \log \frac{1}{p(x)} = -\log p(x) \tag{32}$$

More precisely, Shannons coding theorem implies that x can be communicated at a 'cost' that is bounded below by $-\log p(x)$. Use of base-e or base-2 logarithms leads to this cost being measured in 'nats' or 'bits' respectively. In what follows we refer to $S(x)$ as the cost of communicating x.

By looking at the appropriate terms in the log-evidence one can read off the cost of coding the prediction errors region by region and the cost of coding each type of parameter. For the Laplace approximation we can equate

$$-\log p(y|m) = \sum_i S_e(i) + \sum_k S_p(k) + S_d \tag{33}$$

with equation 25 where $S_e(i)$ is the cost of prediction errors in the ith region, $S_p(k)$ is the cost of the kth parameter and S_d is the cost of the dependency between parameters (captured in the posterior covariance matrix). We see that

$$\begin{aligned} S_e(i) &= 0.5 \log \Lambda_{ii} + 0.5 \frac{1}{\Lambda_{ii}} r_i(\theta_{MP})^T r_i(\theta_{MP}) \\ S_p(k) &= 0.5 \log \frac{\sigma^2_{prior}(k)}{\sigma^2_{posterior}(k)} + 0.5 \frac{1}{\sigma^2_{prior}(k)} e_k(\theta_{MP})^T e_k(\theta_{MP}) \end{aligned} \tag{34}$$

where Λ_{ii} denotes the error variance in the ith region (defined in section 2.3), $\sigma^2_{posterior}(k)$ is the posterior variance of the kth parameter taken from the relevant diagonal in the posterior covariance matrix Σ_{MP} and $\sigma^2_{prior}(k)$ is the prior variance of the kth parameter and is taken from the appropriate diagonal entry in C_p. For example, if k refers to an intrinisic connection $\sigma^2_{prior}(k) = v_a$. Equation 34 shows that the costs of archetypal intrinsic, modulatory and input connections are determined by v_a,v_b and v_c. This again highlights the dependence of the Laplace approximation on these quantities. In contrast, the AIC and BIC criteria assume that the cost of coding a parameter is the same regardless of which parameter it is. For AIC this cost is 1 nat and for BIC it is $0.5 \log N_s$ nats.

For a given fitted DCM we can decompose the model evidence into the costs of coding prediction errors, region by region, and the cost of coding the parameters. It is also possible to decompose Bayes factors into prediction error and parameter terms and this will give an indication as to why one model is favoured over another.

5 Applications

In this section we describe fitting DCMs to fMRI data from an Attention to Motion experiment and a Visual Object Categorisation experiment. We also describe fitting models to simulated data to demonstrate the face validity of the model comparison approach. These data were generated so as to have similar Signal to Noise Ratios (SNRs) to the fMRI data, where SNR is defined as the ratio of signal amplitude to noise amplitude [34]. For regions receiving driving input the SNRs were approximately 2 for the Attention data and 0.5 for the Visual Object data. These SNRs were computed by dividing the standard deviation of the DCM predictions by the estimated observation noise standard deviation. We typically chose the SNR of the simulated data to be unity.

5.1 Comparing Intrinsic Connectivity

In this section, we use Bayes factors to compare DCMs with different intrinsic connectivity patterns. The ability to determine the most likely intrinsic connectivity pattern of a model given the observed functional data is highly relevant in practice because there still is very little

detailed knowledge about anatomical connections in the human brain [35]. Definitions of human brain models therefore usually rely on inferring connections from supposedly equivalent areas in the Macaque brain where the connectivity pattern is known at a great level of detail [43]. The difficulties associated with this approach have been described in the Introduction. Additional uncertainty is due to the problem that, even if one knew all anatomical connections between a given set of areas, the question would remain whether all of these connections are functionally relevant within a given functional context.

To demonstrate how Bayes factors can help in cases of uncertainty about the intrinsic connectivity, we investigated the example of two simple models with hierarchically arranged regions. These two models differed in their connectional structure by the presence or absence of reciprocal connections. Specifically, we used DCMs comprising three regions and three input variables and generated 360 data points from the two models shown in Figure 7. Model 1 had a unilateral forward structure and model 2 a reciprocal architecture. We used the connectivity parameters shown in the Figure, hemodynamic parameters set to the prior expectation and an interval between scans of $TR = 2s$. The inputs u_1, u_2 and u_3 are the boxcar functions shown in Figure 8. These inputs are identical to the input variables from the Attention to Visual Motion analysis described in a later section. The simulated time series were created by integrating the state equations (Equation 16). We then added observation noise to achieve an SNR of unity in the regions receiving driving input and repeated this procedure to generate ten data sets from each model structure.

For each data set we then fitted two models, one having forward connections and the other reciprocal connections. We then computed Bayes factors using the AIC and BIC approximations to the model evidence. The results, in Table 2, provide consistent evidence (in the sense defined in section 4.5) in favour of the correct model in all cases. The results in this table show *average* Bayes factors where averaging took place in log-space (eg. $< \log B_{12} >$).

Table 3 shows a breakdown of the Bayes factor for a typical run on simulated data from the model with feedforward connectivity. The 'cost', S, column gives the cost in bits of coding each prediction or parameter error, and the overall cost is given by the sum of the individual

costs. The Bayes factor column shows the corresponding components of the Bayes factor, given by 2^{-S}, with the overall value given by the *product* of individual components. Any apparent discrepancy between individual entries and overall values is due to the fact that entries are only displayed to two decimal places. Bayes factor components larger than 1 favour model 1. In the remainder of this paper, there are several tables showing a partitioning of Bayes factors that use this format.

Table 3 shows that the forward model is favoured because the number of bits required to code the errors is about the same, but fewer bits are required to code the parameters. That is, the forward model is equally accurate but more parsimonious.

Table 4 shows a breakdown of the Bayes factor for a typical run on simulated data from the model with reciprocal connectivity. Here, the reciprocal model is favoured as it is more accurate, especially in regions R1 and R2, that is, in the regions which receive direct feedback.

Overall, these results demonstrate that Bayes factors can indeed be used to compare models with different intrinsic connectivities.

5.2 Comparing Modulatory Connectivity

In this section, we use simulated data and a simple model of hemispheric specialization (lateralization) to demonstrate the practical relevance of Bayes factors for comparing models with different modulatory connectivity. Traditionally, lateralization has often been envisaged to reflect differences in the local computational properties of homotopic areas in the two hemispheres. Recent studies have indicated, however, that asymmetries in the intra-hemispheric functional couplings may be an equally important determinant of hemispheric specialization [31, 44]. This section demonstrates the ability of DCM to correctly identify asymmetries of modulatory intra-hemispheric connectivity despite the presence of reciprocal inter-hemispheric connections between homotopic regions.

We generated 256 data points ($TR = 2s$) from model 1 shown in the top panel of Figure 9, where modulation of connectivity takes place in the left hemisphere. We used the connectivity parameters shown in the figure and hemodynamic parameters were set equal to their prior

expectation. The driving input u_1 consisted of delta functions with interstimulus intervals drawn from a uniform distribution with minimum and maximum values of 2 and 8s. The modulatory input u_2 consisted of a boxcar function with period 40s. The simulated time series were created by integrating the state equations (Equation 16). We then added observation noise so as to achieve an SNR of unity in the region receiving driving input. This procedure was repeated to generate ten data sets.

For each data set we then fitted two DCM models, model 1 assuming that connectivity is modulated in the left hemisphere and model 2 assuming that it is modulated in the right. Deciding which is the best model is not a trivial task as information can pass between hemispheres via the lateral connections. Informally, however, one should be able to infer which model generated the data for the following reason. Both models predict that L2 and R2 activity will be modulated indirectly by the contextual input u_2. For data generated from the left-hemisphere model, L2 will be modulated more than R2 (vice-versa for the right-hemisphere model). Thus if model 2 does a reasonable job of predicting R2 activity it will necessarily do a poor job of predicting L2 activity (and vice-versa). Formally, the hypotheses embodied in the networks can be evaluated by fitting the models and computing the Bayes factor, B_{12}. For our ten data sets both AIC and BIC gave, on average, $B_{12} = 17$ providing consistent evidence in favour of the correct hypothesis. This same value would have resulted (but this time for B_{21}) had we fitted the models to right hemisphere data. This is because model 2 is equivalent to model 1 after a relabelling of regions.

Why did the Bayes factors favour the left-hemisphere model ? The short answer is that it is the correct model. A more detailed answer can be provided by showing the breakdown of the Bayes factor in table 5 which was computed for a typical run. This breakdown shows clearly that the main reason model 1 is favoured is because it predicts activity in L2 more accurately. Model 2 does a good job of predicting activity in R2 but a poor job in L2. The AIC and BIC criteria produce the same Bayes factor because both networks have the same number of connections. The models are therefore compared solely on the basis of accuracy.

We also considered a model, model 3, with both left and right modulatory connections.

This was fitted to simulated data generated from model 1. For our ten data sets AIC gave, on average, a Bayes factor B_{13} of 1.78 and BIC gave 10.50. Thus the Bayes factors tell us that, overall, we can't be confident that the data came from model 1. On exactly 5 data sets, however we obtained $B_{13} > e$, so in these 5 cases we would correctly conclude that the data came from model 1. On the other 5 data sets we would draw no conclusion. This gives an indication as to the conservativeness of the 'consistent' evidence rule.

We then compared Bayes factors for data generated from model 3, where the modulatory effect was the same on both sides. For ten data sets AIC gave, on average a Bayes factor B_{31} of 2.39 and BIC gave 0.40. Thus the Bayes factors tell us we can't be confident that the data came from model 3. The reason for this uncertainty is that we are asking quite a subtle question - the increase in percentage of signal variance explained by model 3 over model 1 simply isn't large enough to produce consistent Bayes factors.

Finally, to show the unambiguous nature of model selection in the context of discriminable models, we generated data from a fourth model where the modulatory effect on the left side was as before, an increase in connection strength between L1 and L2 from 0.3 to 0.9 (mediated with an intrinsic connection of 0.3 and a modulatory connection of 0.6), but the modulation on the right side was a *decrease* in connection strength from 0.9 to 0.3 (mediated via an intrinsic connection of 0.9 and a modulatory connection of -0.6). Over ten data sets AIC and BIC gave Bayes factors B_{41}, on average, of 2330 and 396 indicating strong and consistent evidence in favour of the correct hypothesis.

Overall, these simulations show that Bayes factors can be used to make inferences about modulatory connections. As predicted by theory (see section 3.5) the sensitivity of the model comparison test increases with larger differences in the number of model parameters or increasing differential signal strength. Put simply, models with greater structural or predictive differences are easier to discriminate.

5.3 Attention to Visual Motion

In previous work we have established that attention modulates connectivity in a distributed system of cortical regions mediating visual motion processing [8, 17]. These findings were based on data acquired using the following experimental paradigm. Subjects viewed a computer screen which displayed either a fixation point, stationary dots or dots moving radially outward at a fixed velocity. For the purpose of our analysis we can consider three experimental variables. The 'photic stimulation' variable indicates when dots were on the screen, the 'motion' variable indicates that the dots were moving and the 'attention' variable indicates that the subject was attending to possible velocity changes. These are the three input variables that we use in our DCM analyses and are shown in Figure 8.

In this paper we model the activity in three regions V1, V5 and superior parietal cortex (SPC). The original 360-scan time series were extracted from the data set of a single subject using a local eigendecomposition and are shown in Figure 10.

We initially set up three DCMs each embodying different assumptions about how attention modulates connectivity between V1 and V5. Model 1 assumes that attention modulates the forward connection from V1 to V5, model 2 assumes that attention modulates the backward connection from SPC to V5 and model 3 assumes attention modulates both connections. These models are shown in Figure 11. Each model assumes that the effect of motion is to modulate the connection from V1 to V5 and uses the same reciprocal hierarchical intrinsic connectivity. Later we will consider models with different intrinsic connections.

We fitted the models and the Bayes factors are shown in Table 6. These show that the data provide consistent evidence in favour of the hypothesis embodied in model 1, that attention modulates solely the forward connection from V1 to V5.

We now look more closely at the comparison of model 1 to model 2. The estimated connection strengths of the attentional modulation were 0.23 for the forward connection in model 1 and 0.55 for the backward connection in model 2. The posterior distribution of this first connection is shown in Figure 12. The posterior probabilities of these connections being greater than the

threshold $\gamma = (\log 2)/4$ (ie. the probabilities that the modulatory effects occur within 4 seconds) are 0.78 and 0.97.

A breakdown of the Bayes factor B_{12} in table 7 shows that the reason model 1 is favoured over model 2 is because it is more accurate. In particular, it predicts SPC activity much more accurately. Thus, although model 2 does show a significant modulation of the SPC-V5 connection, the required change in its prediction of SPC activity is sufficient to compromise the overall fit of the model. If we assume models 1 and 2 are equally likely apriori then our posterior belief in model 1 is 0.78.

This example makes an important point. Two models can only be compared by computing the evidence for each model. It is not sufficient to compare values of single connections. This is because changing a single connection changes overall network dynamics and each hypothesis is assessed (in part) by how well it predicts the data, and the relevant data are the activities in a distributed network.

We now focus on model 3 that has *both* modulation of forward and backward connections. Firstly, we make a statistical inference to see if, within model 3, modulation of the forward connection is larger than modulation of the backward connection. For this data the posterior distribution of estimated parameters tells us that this is the case with probability 0.75. This is a different sort of inference to that made above. Instead of inferring which is more likely, modulation of a forward or backward connection, we are making an inference about which effect is stronger when both are assumed present.

However, this inference is contingent on the assumption that model 3 is a good model. The Bayes factors in Table 6, however, show that the data provide consistent evidence in favour of the hypothesis embodied in model 1, that attention modulates *only* the forward connection. Table 8 shows a breakdown of B_{13}. Here the dominant contribution to the Bayes factor is the increased parameter cost for model 3.

So far, our models have all assumed a reciprocal intrinsic connectivity. We examine the validity of this assumption by also fitting a model with purely forward connections (model 4) and a model having a full intrinsic connectivity (model 5). These models are otherwise identical

to model 1. Table 9 shows Bayes factors of the fitted models that provide consistent evidence favouring model 1 over model 4. But between models 1 and 5 there is no consistent evidence either way. We can therefore be confident that our assumption of reciprocally and hierarchically organised intrinsic connectivity is a reasonable one.

5.4 Visual Object Categories

Functional imaging studies have reported the existence of discrete cortical regions in occipito-temporal cortex that respond preferentially to different categories of visual object such as faces, buildings and letters. In previous work [32] we have used DCM to explore whether such category-specificity is the result of modulation of backward connections from parietal areas or modulation of forward connections from primary visual areas.

In this section we focus on a single area in Mid-Occipital (MO) cortex which responded preferentially to images of faces. We set up DCMs comprising three regions, V3, MO and superior-parietal cortex (SPC). Full descriptions of the experimental design, imaging acquisition and extraction of regional time series are available in [32]. Our analyses used the data from 'subject 1' and our regions are those used in the DCM analysis in figure 1 of [32]. The time series consist of 1092 scans. The original data files can be obtained from the National fMRI Data Center (http://www.fmridc.org) and their acquisition is described in Ishai et al. [22].

The aim of our analyses was to find out if the specificity of the face-responsive area could be better attributed to increased connectivity from V3 or from SPC. To this end we fitted three models to the data which are shown in Figure 13. These models postulate modulation of the forward connection to MO (model 1), modulation of the backward connection to MO (model 2) and modulation of both connections (model 3). All three models assume a reciprocal and hierarchically organised intrinsic connectivity. Later we will look at models with different intrinsic connectivity. We fitted the models and the Bayes factors are shown in Table 10. These provide evidence in favour of the hypothesis embodied in model 1, that the processing of faces modulates only the forward connection from V3 to MO.

We now turn to the assumption of reciprocal and hierarchically organised intrinsic connectiv-

ity and test its validity by fitting a model with purely feedforward connections (model 4) and full intrinsic connectivity (model 5). These models are otherwise identical to model 1. The Bayes factors of the fitted models are shown in Table 11 and provide consisten evidence in favour of model 1 over model 4. Between models 1 and 5, however, there is no consistent evidence either way. We can therefore be content that our assumption of reciprocal connectivity is sufficient.

We now look in more detail at two of the pairwise model comparisons. Table 12 provides a breakdown of the Bayes factor for model 1 versus model 2. This shows that the largest contributions to the Bayes factors are the better model fits in V3 and MO. Because both models have the same number of connections the relative BIC and AIC parameter costs are zero. The models are therefore compared solely on the basis of which is more accurate. Table 13 shows a breakdown of the Bayes factor for model 1 versus model 4, indicating that the increased accuracy of the model with reciprocal intrinsic connectivity more than compensates for its lack of parsimony, with respect to the model with purely forward connections.

6 Discussion

We have described Bayesian inference procedures in the context of Dynamic Causal Models. DCMs are used in the analysis of effective connectivity and posterior distributions can be used, for example, to assess changes in effective connectivity caused by experimental manipulation. These inferences, however, are contingent on assumptions about the intrinsic and modulatory architecture of the model ie. which regions are connected to which other regions and which inputs can modulate which connections.

To date, the specification of intrinsic connectivity has been based on our knowledge, for example, of anatomical connectivity in the Macaque. Whilst this approach may be tenable for sensory sytems it is more problematic for higher cognitive systems. Moreover, even if we knew the anatomical connectivity the question would remain as to whether these connections were functionally relevant in a given functional context. The use of Bayes factors to guide the choice of intrinsic connectivity is therefore of great practical relevance. In this paper we have shown how they can be used, for example, to decide between feedforward, reciprocal and fully

connected structures. We have also shown how Bayes factors can be used to compare models with different modulatory connectivity. This is important as it is the changes in connectivity that are usually of primary scientific interest.

The use of Bayes factors for model comparison is somewhat analagous to the use of F-tests in the General Linear Model. Whereas t-tests are used to assess individual effects, F-tests allow one to assess the significance of a set of effects. Bayes factors play a similar role but additionally allow inferences to be constrained by prior knowledge. Moreover, it is possible to simultaneously entertain a number of hypotheses and compare them using Bayesian evidence. Importantly, these hypotheses are not constrained to be nested.

In this paper we have used AIC and BIC approximations to the model evidence and defined a criterion of 'consistent' evidence on which decisions can be based. This was motivated by the fact that the AIC approximation is known to be biased towards complex models and BIC to simpler models. In future we envisage improved approximations, perhaps based on Laplace approximations where the prior variances are inferred using Empirical Bayes. We are also aware of a number of improvements to the AIC criterion [7].

Model comparison of effectivity connectivity models has previously been explored in the context of SEM by Bullmore et al. [9]. This work has established the usefulness of such approaches for comparing nested structural equation models which are most suitable for the analysis of PET data. In our work, we compare DCM models which are currently most suited for the analysis of fMRI data. Moreover, the model comparison approaches we have explored employ a Bayesian perspective enabling the comparison of non-nested models.

Currently, we are using Bayesian model comparison over a limited set of models defined by the modeller. This allows the user to compare a handful of working hypotheses about the large-scale organisation of their neurocognitive system of interest. Future work may develop automatic model search procedures. These would embody standard Bayesian procedures whereby model search proceeds by considering only those models in an 'Occam window' [37]. Similar model search procedures have been previously explored in the context of SEM by Bullmore et al. [9]. Another future research avenue is to use Bayesian model averaging [21], where instead of

choosing the 'best' model, models are combined using the evidence as a weighting factor.

The combined use of Bayes factors and DCM provides us with a formal method for evaluating competing scientific theories about the forms of large-scale neural networks and the changes in them that mediate perception and cognition.

References

[1] F. Aboitiz and V. Garcia. The evolutionary origin of the language areas in the human brain. a neuroanatomical perspective. *Brain Research Review*, 25:381–396, 1997.

[2] H. Akaike. Information measures and model selection. *Bulletin of the International Statistical Institute*, 50:277–290, 1973.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrox and F. Caski, editors, *Second international symposium on information theory*, page 267, 1973. Budapest: Akademiai Kiado.

[4] T.D. Albright and G.R. Stoner. Visual motion perception. *Proceedings of the National Academy of Sciences USA*, 92:2433–2440, 1995.

[5] A. Angelucci, J.B. Levitt, E.J. Waltonand J.M. Hupe, J. Bullier, and J.S. Lund. Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22:8633–8646, 2002.

[6] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 2000.

[7] H. Bozdogan. Model selection and Akaike's information criterion: the general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[8] C. Buchel and K.J. Friston. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, 7:768–778, 1997.

[9] E. Bullmore, B. Horwitz, G. Honey, M. Brammer, S. Williams, and T. Sharma. How good is good enough in path analysis of fMRI data ? *NeuroImage*, 11:289–301, 2000.

[10] R.B. Buxton, E.C. Wong, and L.R. Frank. Dynamics of blood flow and oxygenation changes during brain activation: The Balloon Model. *Magnetic Resonance in Medicine*, 39:855–864, 1998.

[11] D. Chawla, G. Rees, and K.J. Friston. The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neuroscience*, 2(7):671–676, 1999.

[12] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[13] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[14] R.S.J. Frackowiak, K.J. Friston, C.D. Frith, R.J. Dolan, and J.C. Mazziotta, editors. *Human Brain Function*. Academic Press USA, 1997.

[15] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[16] K.J. Friston. Bayesian Estimation of Dynamical Systems: An Application to fMRI. *NeuroImage*, 16:513–530, 2002.

[17] K.J. Friston and C. Buchel. Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 97(13):7591–7596, 2000.

[18] K.J. Friston, L. Harrison, and W. Penny. Dynamic Causal Modelling. *NeuroImage*, 19(4):1273–1302, 2003.

[19] K.J. Friston, A. Mechelli, R. Turner, and C.J. Price. Nonlinear responses in fMRI: The Balloon model, Volterra kernels and other hemodynamics. *NeuroImage*, 12:466–477, 2000.

[20] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1995.

[21] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417, 1999.

[22] A. Ishai, L.G. Ungerleider, A. Martin, J.L. Schouten, and J.V. Haxby. The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience*, 12:35–51, 2000. Supplement 2.

[23] H. Jefferys. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, 31:203–222, 1935.

[24] R.R. Johnson and A. Burkhalter. A polysynaptic feedback circuit in rat visual cortex. *Journal of Neuroscience*, 17:7129–7140, 1997.

[25] R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. Technical Report 254, University of Washington, 1993. http://www.stat.washington.edu/tech.reports/tr254.ps.

[26] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[27] S. Kastner, M.A. Pinsk, P. De Weerd, R. Desimone, and L.G. Ungerleider. Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22:751–761, 1999.

[28] S.J. Luck, L. Chelazzi, S.A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of Neurophysiology*, 77:24–42, 1997.

[29] D.J.C Mackay. *Information Theory, Inference and Learning Algorithms*. Springer, 2000.

[30] D. Marr. *Vision*. Freeman, New York, 1982.

[31] A.R. McIntosh, C.L. Grady, L.G. Ungerleider, J.V. Haxby, S.L. Rapoport, and B. Horwitz. Network analysis of cortical visual pathways mapped with pet. *Journal of Neuroscience*, 14:655–666, 1994.

[32] A. Mechelli, C.J. Price, U. Noppeney, and K.J. Friston. A dynamic causal modelling study of category effects: Bottom-up or top-down mediation ? Journal of Cognitive Neuroscience, In Press, 2003.

[33] K.M. O'Craven, B.R. Rosen, K.K. Kwong, A.Treisman, and R.L. Savoy. Voluntary attention modulates fmri activity in human mt-mst. *Neuron*, 18:591–598, 1997.

[34] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[35] R.E. Passingham, K.E. Stephan, and R. Kotter. The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience*, 3:606–616, 2002.

[36] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C*. Cambridge, 1992.

[37] A.E. Raftery. Bayesian model selection in social research. In P.V. Marsden, editor, *Sociological Methodology*. Cambridge, Mass., 1995.

[38] J.H. Reynolds, L. Chelazzi, and R. Desimone. Competitive mechanisms subserve attention in macaque areas v2 and v4. *Journal of Neuroscience*, pages 1736–1753, 1999.

[39] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1995.

[40] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing, 1989. Singapore.

[41] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[42] M. Siegel, K.P. Kording, and P. Konig. Integrating top-down and bottom-up sensory processing by somato-dendritic interactions. *Journal of Computational Neuroscience*, 8:161–173, 2000.

[43] K.E. Stephan, L. Kamper, A. Bozkurt, G.A. Burns, M.P. Young, and R. Kotter. Advanced database methodology for the collation of connectivity data on the macaque brain (cocomac). *Philosophical Transactions of the Royal Society London B Biological Sciences*, 356:1159–1186, 2001.

[44] K.E. Stephan, J.C. Marshall, K.J. Friston, J.B. Rowe, A. Ritzl, K. Zilles, and G.R. Fink. Lateralized cognitive processes and lateralized task control in the human brain. *Science*, 301:394–386, 2003.

[45] S. Treue and J.H. Maunsell. Attentional modulation of visual motion processing in cortical areas mt and mst. *Nature*, 382:539–541, 1996.

[46] L.G. Ungerleider, S.M. Courtney, and J.V. Haxby. A neural system for human visual working memory. *Proceedings of the National Academy of Sciences USA*, 95:883–890, 1998.

[47] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computing Journal*, 11:185–195, 1968.

[48] S. Zeki, J.D. Watson, C.J Lueck, K.J Friston, C. Kennard, and R.S. Frackowiak. A direct demonstration of functional specialization in human visual cortex. *Journal of Neuroscience*, 11:641–649, 1991.

A The Kronecker Product

If A is an $m_1 \times m_2$ matrix and B is an $n_1 \times n_2$ matrix, then the Kronecker product of A and B is the $(m_1 n_1) \times (m_2 n_2)$ matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1m_2}B \\ \dots & & \\ a_{m_1 1}B & & a_{m_1 m_2}B \end{bmatrix} \tag{35}$$

B Approximating the model evidence

B.1 Laplace approximation

The model evidence is given by

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta \tag{36}$$

This can be approximated using Laplace's method

$$\begin{aligned} p(y|m)_L &\approx p(y|m) \\ &= (2\pi)^{-p/2}|C_p|^{-1/2}(2\pi)^{-N_s/2}|C_e|^{-1/2}I(\theta) \end{aligned} \tag{37}$$

where

$$I(\theta) = \int \exp(-\frac{1}{2}r(\theta)^T C_e^{-1} r(\theta) - \frac{1}{2}e(\theta)^T C_p^{-1} e(\theta))d\theta \tag{38}$$

Substituting $e(\theta) = (\theta - \theta_{MP}) + (\theta_{MP} - \theta_p)$ and $r(\theta) = (y - h(\theta_{MP})) + (h(\theta_{MP}) - h(\theta))$ into the above expression, and removing terms not dependent on θ from the integral, then gives

$$I(\theta) = \left[\int \exp(-\frac{1}{2}(\theta - \theta_{MP})^T \Sigma_{MP}^{-1}(\theta - \theta_{MP}))d\theta\right] \tag{39}$$

$$. \quad \left[\exp(-\frac{1}{2}r(\theta_{MP})^T C_e^{-1} r(\theta_{MP}) - \frac{1}{2}e(\theta_{MP})^T C_p^{-1} e(\theta_{MP}))\right] \tag{40}$$

where the first factor is the normalising term of the multivariate Gaussian density. Hence

$$\begin{aligned} I(\theta) &= (2\pi)^{p/2}|\Sigma_{MP}|^{1/2} \exp(-\frac{1}{2}r(\theta_{MP})^T C_e^{-1} r(\theta_{MP}) \\ &- \frac{1}{2}e(\theta_{MP})^T C_p^{-1} e(\theta_{MP})) \end{aligned} \tag{41}$$

Substituting this expression into Eq 37 and taking logs gives

$$\begin{aligned} \log p(y|m)_L &= -\frac{N_s}{2}\log 2\pi - \frac{1}{2}\log|C_e| - \frac{1}{2}\log|C_p| + \frac{1}{2}\log|\Sigma_{MP}| \\ &- \frac{1}{2}r(\theta_{MP})^T C_e^{-1} r(\theta_{MP}) - \frac{1}{2}e(\theta_{MP})^T C_p^{-1} e(\theta_{MP}) \end{aligned} \tag{42}$$

When comparing the evidence for different models we can ignore the first term as it will be the same for all models. Dropping the first term and rearranging gives

$$\log p(y|m)_L = Accuracy(m) - \frac{1}{2}\log|C_p| + \frac{1}{2}\log|\Sigma_{MP}| - \frac{1}{2}e(\theta_{MP})^T C_p^{-1} e(\theta_{MP}) \tag{43}$$

where

$$Accuracy(m) = -\frac{1}{2}\log|C_e| - \frac{1}{2}r(\theta_{MP})^T C_e^{-1} r(\theta_{MP}) \tag{44}$$

is the *accuracy* of model m.

B.2 Bayesian Information Criterion

Substituting Eq. 41 into Eq. 37 gives

$$p(y|m)_L = p(y|\theta_{MP}, m)p(\theta_{MP}|m)(2\pi)^{p/2}|\Sigma_{MP}|^{1/2} \tag{45}$$

Taking logs gives

$$p(y|m)_L = \log p(y|\theta_{MP}, m) + \log p(\theta_{MP}|m) + \frac{p}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_{MP}| \tag{46}$$

The dependence of the first three terms on the number of scans is $O(N_s)$, $O(1)$ and $O(1)$. For the 4th term entries in the posterior covariance scale linearly with N_s^{-1}

$$\begin{aligned} \lim_{N_s \to \infty} \frac{1}{2} \log |\Sigma_{MP}| &= \frac{1}{2} \log |\frac{\Sigma_{MP}(0)}{N_s}| \\ &= -\frac{p}{2} \log N_s + \frac{1}{2} \log |\Sigma_{MP}(0)| \end{aligned} \tag{47}$$

where $\Sigma_{MP}(0)$ is the posterior covariance based on $N_s = 0$ scans. This last term therefore scales as $O(1)$. Schwarz [41] notes that in the limit of large N_s equation 46 therefore reduces to

$$\begin{aligned} BIC &= \lim_{N_s \to \infty} \log p(y|m)_L \\ &= \log p(y|\theta_{MP}, m) - \frac{p}{2} \log N_s \end{aligned} \tag{48}$$

This can be re-written as

$$BIC = Accuracy(m) - \frac{p}{2} \log N_s \tag{49}$$

Table 1: **Interpretation of Bayes factors**. Bayes factors can be interpreted as follows. Given candidate hypotheses i and j a Bayes factor of 20 corresponds to a belief of 95% in the statement 'hypothesis i is true'. This corresponds to strong evidence in favour of i.

B_{ij}	$p(m = i\|y)(\%)$	Evidence in favour of model i
1 to 3	50-75	Weak
3 to 20	75-95	Positive
20 to 150	95-99	Strong
≥ 150	≥ 99	Very Strong

Table 2: **Comparing Intrinsic Connectivity** The table shows the Bayes factor B_{12} averaged over 10 runs of feedforward data (from model 1), and B_{21} averaged over 10 runs of reciprocal data (from model 2). AIC and BIC consistently provide between positive and very strong evidence in favour of the correct model.

	B_{12}	B_{21}
AIC	4.7	2×10^8
BIC	230	4×10^6

Table 3: **Comparing Intrinsic Connectivity** The table shows the contributions to the Bayes factor B_{12} for a typical feedforward data set. The largest single contribution is the cost of coding the parameters. The overall Bayes factors provide positive (AIC) and very strong (BIC) evidence in favour of the true model.

Source	Model 1 vs. Model 2 Relative Cost (bits)	Bayes Factor B_{12}
R1 error	0.03	0.98
R2 error	-0.12	1.09
R3 error	0.20	0.87
Parameters (AIC)	-2.89	7.39
Parameters (BIC)	-8.49	360
Overall (AIC)	-2.77	6.84
Overall (BIC)	-8.38	330

Table 4: **Comparing Intrinsic Connectivity** The table shows contributions to the Bayes factor B_{21} for a typical reciprocal data set ie. model 2 is true. The largest single contribution to the Bayes factor is the cost of coding the prediction errors. The overall Bayes factors provide very strong evidence in favour of the true model.

Source	Model 2 vs. Model 1 Relative Cost (bits)	Bayes Factor B_{21}
R1 error	-24.8	2×10^7
R2 error	-6.94	123
R3 error	-0.81	1.75
Parameters (AIC)	2.89	0.14
Parameters (BIC)	8.49	0.003
Overall (AIC)	-29.66	8×10^8
Overall (BIC)	-24.06	2×10^6

Table 5: **Comparing Modulatory Connectivity** Breakdown of contributions to the Bayes factor for model 1 with 'left-hemisphere' modulation versus model 2 having 'right-hemisphere' modulation for a typical left-hemisphere data set. The largest single contribution to the Bayes factor is the increased model accuracy in region L2, where 2.97 fewer bits are required to code the prediction errors. The overall Bayes factor of 8.74 provides positive evidence in favour of the left-hemisphere hypothesis. Because both network structures have the same number of connections the relative cost of parameters under both AIC and BIC is zero.

Source	Model 1 vs. Model 2 Relative Cost (bits)	Bayes Factor B_{12}
L1 error	-0.47	1.38
L2 error	-2.97	7.82
R1 error	-0.19	1.14
R2 error	0.49	0.71
Parameters (AIC)	0.00	1.00
Parameters (BIC)	0.00	1.00
Overall (AIC)	-3.13	8.74
Overall (BIC)	-3.13	8.74

Table 6: **Attention Data - comparing modulatory connectivities** Bayes factors provide consistent evidence in favour of the hypothesis embodied in model 1, that attention modulates (solely) the bottom-up connection from V1 to V5. Model 1 is preferred to models 2 and 3.

	B_{12}	B_{13}	B_{32}
AIC	3.56	2.81	1.27
BIC	3.56	19.62	0.18

Table 7: **Attention Data:** Breakdown of contributions to the Bayes factor for model 1 versus model 2. The largest single contribution to the Bayes factor is the increased model accuracy in region SPC, where 5.64 fewer bits are required to code the prediction errors. The overall Bayes factor B_{12} of 78 provides strong evidence in favour of model 1.

Source	Model 1 vs. Model 2 Relative Cost (bits)	Bayes Factor B_{12}
V1 error	7.32	0.01
V5 error	-0.77	1.70
SPC error	-8.38	333.36
Parameters (AIC)	0.00	1.00
Parameters (BIC)	0.00	1.00
Overall (AIC)	-1.83	3.56
Overall (BIC)	-1.83	3.56

Table 8: **Attention Data:** Breakdown of contributions to the Bayes factor for model 1 versus model 3. The largest single contribution to the Bayes factor is the cost of coding the parameters. The table indicates that both models are similarly accurate but model 1 is more parsimonious. The overall Bayes factor B_{13} provides consistent evidence in favour of the (solely) bottom-up model.

Source	Model 1 vs. Model 3 Relative Cost (bits)	Bayes Factor B_{13}
V1 error	-0.01	1.01
V5 error	0.02	0.99
SPC error	-0.05	1.04
Parameters (AIC)	-1.44	2.72
Parameters (BIC)	-4.25	18.97
Overall (AIC)	-1.49	2.81
Overall (BIC)	-4.29	19.62

Table 9: **Attention Data - comparing intrinsic connectivities** There is consistent evidence in favour of model 1 over model 4, but, between models 1 and 5, there is no consistent evidence either way.

	B_{14}	B_{15}
AIC	1×10^{20}	0.06
BIC	1×10^{19}	3.13

Table 10: **Visual Object Data - comparing modulatory connectivity.** Bayes factors provide evidence in favour of the hypothesis embodied in model 1, that the processing of faces modulates (solely) the bottom-up connection from V3 to M0. Model 1 is preferred to models 2 and 3, and model 3 is preferred to model 2.

	B_{12}	B_{13}	B_{32}
AIC	7950	2.75	2890
BIC	7950	33.47	237

Table 11: **Visual Object Data - comparing intrinsic connectivities** Bayes factors provide evidence in favour of model 1 over model 4, but between models 1 and 5 there is no consistent evidence either way.

	B_{14}	B_{15}
AIC	2280	0.01
BIC	15.4	2.00

Table 12: **Visual Object Category Data** Breakdown of contributions to the Bayes factor for model 1 versus model 2. The largest contributions to the Bayes factor are the better model fits in V3 and MO. The overall Bayes factor B_{12} of 7950 provides very strong evidence in favour of model 1.

Source	Model 1 vs. Model 2 Relative Cost (bits)	Bayes Factor B_{12}
V3 error	-10.59	1545
MO error	-6.01	64.6
SPC error	3.65	0.08
Parameters (AIC)	0.00	1.00
Parameters (BIC)	0.00	1.00
Overall (AIC)	-12.96	7950
Overall (BIC)	-12.96	7950

Table 13: **Visual Object Category Data** Breakdown of contributions to the Bayes factor for the DCM with reciprocal and hierarchically organised intrinsic connectivity (model 1) versus the DCM with feedforward intrinsic connectivity (model 4). The increased accuracy of model 1 more than compensates for its lack of parsimony.

Source	Model 1 vs. Model 4 Relative Cost (bits)	Bayes Factor B_{14}
V3 error	-8.10	274
MO error	-2.97	7.83
SPC error	-2.97	7.83
Parameters (AIC)	2.89	0.14
Parameters (BIC)	10.09	0.0009
Overall (AIC)	-11.15	2280
Overall (BIC)	-3.95	15.4

Figure 1: **DCM Neurodynamics.** The top panel shows a Dynamic Causal Model comprising $L = 2$ regions and $M = 2$ inputs. The input variable u_1 drives neuronal activity z_1. Informally, neuronal activity in this region then excites neuronal activity z_2 which then re-activates activity in region 1. Formally, these interactions take place instantaneously according to equation 1. The time constants are determined by the values of the intrinsic connections A_{11}, A_{12}, A_{21} and A_{22}. Input 2, typically a contextual input such as instructional set, then acts to change the intrinsic dynamics via the modulatory connections B^2_{11} and B^2_{22}. In this example, the effect is to reduce neuronal time-constants in each region as can be seen in the neuronal time series in the bottom panel.

Figure 2: **Samples from priors.** These distributions characterise our expectations about what the neuronal transients and hemodynamic responses should look like. For each value of σ, sampled from $p(\sigma)$, we generated the neuronal response to a unit impulse, this response being a neuronal transient. Then, for each neuronal transient we drew a sample θ_h from $p(\theta_h)$ and generated a hemodynamic response. The figures show samples of (a) neuronal transients, (b) hemodynamic responses, (c) a histogram of time constants of neuronal transients (mean=880ms), and (d) a histogram of peak hemodynamic response times (mean=4.1s). The histograms in (c) and (d) are made up from 10,000 samples and the plots in (a) and (b) consist of the first 100 samples.

Figure 3: **DCM models of modulatory processes.** **A**: A simple DCM that includes visual areas V1 and V5. Visual stimuli drive activity in V1 that is reciprocally connected to V5. The strength of the forward connection V1-V5 depends on whether stimuli are moving or stationary, i.e. V1-V5 is modulated by a vector MOT indicating the presence of motion in the visual input. **B**: The bottom-up process modeled by A is shown schematically at a synaptic level. The strength of the input from the V1 neuron to the dendritic tree of the V5 neuron is enhanced for moving stimuli. The strength of the synaptic transmission (green circle) simply follows the strength of the input from V1. **C**: Same DCM as in A, except that this model allows for modulation of the V1-V5 forward connection by attention to motion (ATT). **D**: Same schema as in B, but showing the top-down gain control process modeled by C at a synaptic level. Here, the strength of the synaptic response of the V5 neuron to inputs from the V1 neuron (green circle) is modulated by simultaneous inputs from a higher attention-related area X to the same V5 neuron (red circle). These inputs change the biophysical properties of the dendritic tree of the V5 neuron, rendering it more susceptible to inputs from V1 neurons. Various potential mechanisms for this modulation exist, e.g. see [42].

Figure 4: **DCM models of additive processes.** **A**: Same basic DCM as in Fig. 3, but without a modulation of either connection. Instead, attention to motion leads to a direct (additive) increase of V5 activity, independent of the presence and nature of visual input. This represents a simple model of top-down baseline shift processes without specifying which areas represent the physiological source of the top-down influence. **B**: In addition to A, this DCM includes the superior parietal cortex (SPC) as a putative source of attentional top-down influences onto visual areas.

Figure 5: **Why simple models are preferable.** The figure plots the evidence for model 1, $p(y|m_1)$, and the evidence for model 2, $p(y|m_2)$, against y, the space of all possible data sets. Here, a data set y_i would be fMRI time series from regions of interest. The complex model, model 2, can 'explain' more data sets than the simple model, model 1. If one observes y_3, a data set that both models can explain, then by virtue of the densities $p(y|m)$ having to integrate to unity, $p(y_3|m_1)$ will be larger than $p(y_3|m_2)$. Thus, the simple model is preferred. This figure is adapted from Mackay [29].

Figure 6: **Dependence of BIC on number of samples.** The figure plots the Bayes factor B_{12} computed from BIC versus the number of scans N_s where model 1 has one more parameter than model 2 and the relative increase in signal variance is (a) 1% and (b) 2%, the latter being typical of fMRI data used in this paper. The horizontal line shows a Bayes factor of e.

Figure 7: **Comparing intrinsic connectivity structures.** Synthetic DCM models comprising the three regions R1, R2 and R3. Model 1 (left panel) has only forward connections and model 2 (right panel) has a reciprocal connectivity. In both networks activity is driven by input u_1 and forward connections are modulated by inputs u_2 and u_3. These inputs are shown in Figure 8.

Figure 8: **Comparing intrinsic connectivity: inputs.** The plots bottom to top show the driving input u_1 and modulatory inputs u_2 and u_3. These inputs were used together with the network structures in Figure 7 to produce simulated data. These inputs are also identical to the 'Photic', 'Motion' and 'Attention' variables used in the analysis of the Attention to Visual Motion data (see Figures 10 and 11).

Figure 9: **Comparing modulatory connectivity.** Synthetic DCM models comprising four regions: L1 and L2 in the 'left-hemisphere' and R1 and R2 in the 'right hemisphere'. The networks have driving input entering the 'lower' areas L1 and R1, and an intrinsic connectivity comprising within-hemisphere feedforward connections and reciprocal lateral connections between hemispheres. In model 1 (top panel), feedforward connectivity is modulated in the left hemisphere and in model 2 (bottom panel) feedforward connectivity is modulated in the right hemisphere.

Figure 10: **Attention data.** fMRI time series (rough solid lines) from regions V1, V5 and SPC and the corresponding estimates from DCM model 1 (smooth solid lines).

Figure 11: **Attention models**. In all models photic stimulation enters V1 and the motion variable modulates the connection from V1 to V5. Models 1, 2 and 3 have reciprocal and hierarchically organised intrinsic connectitivty. They differ in how attention modulates the connectivity to V5, with model 1 assuming modulation of the forward connection, model 2 assuming modulation of the backward connection and model 3 assumes both. Models 4 and 5 assume modulation of the forward connection, but have a purely feedforward intrinsic connectivity (model 4) or a fully connected intrinsic architecture (model 5).

Figure 12: **Attention model - posterior distribution.** The plot shows the posterior probability distribution of the parameter B^1_{21}. This is the connection from region 1 (V1) to region 2 (V5) that is modulated by attention (the 3rd input). The mean value of this distribution is 0.23. This is also shown in Figure 11. We can use this distribution to compute our belief that this connection is larger than some threshold γ. If we choose eg. $\gamma = (\log 2)/4 = 0.17$ then this corresponds to computing the probability that this modulatory effect occurs within 4 seconds. In DCM faster effects are mediated by stronger connections (see eg. Equation 1). For our data we have $p(B^3_{21} > \gamma) = 0.78$.

Figure 13: **DCM models of category-specificity.** Models 1, 2 and 3 have reciprocal and hierarchically organised intrinsic connectitivty. Model 1 has modulation of the forward connection to MO, model 2 has modulation of the backward connection to MO, model 3 has both.

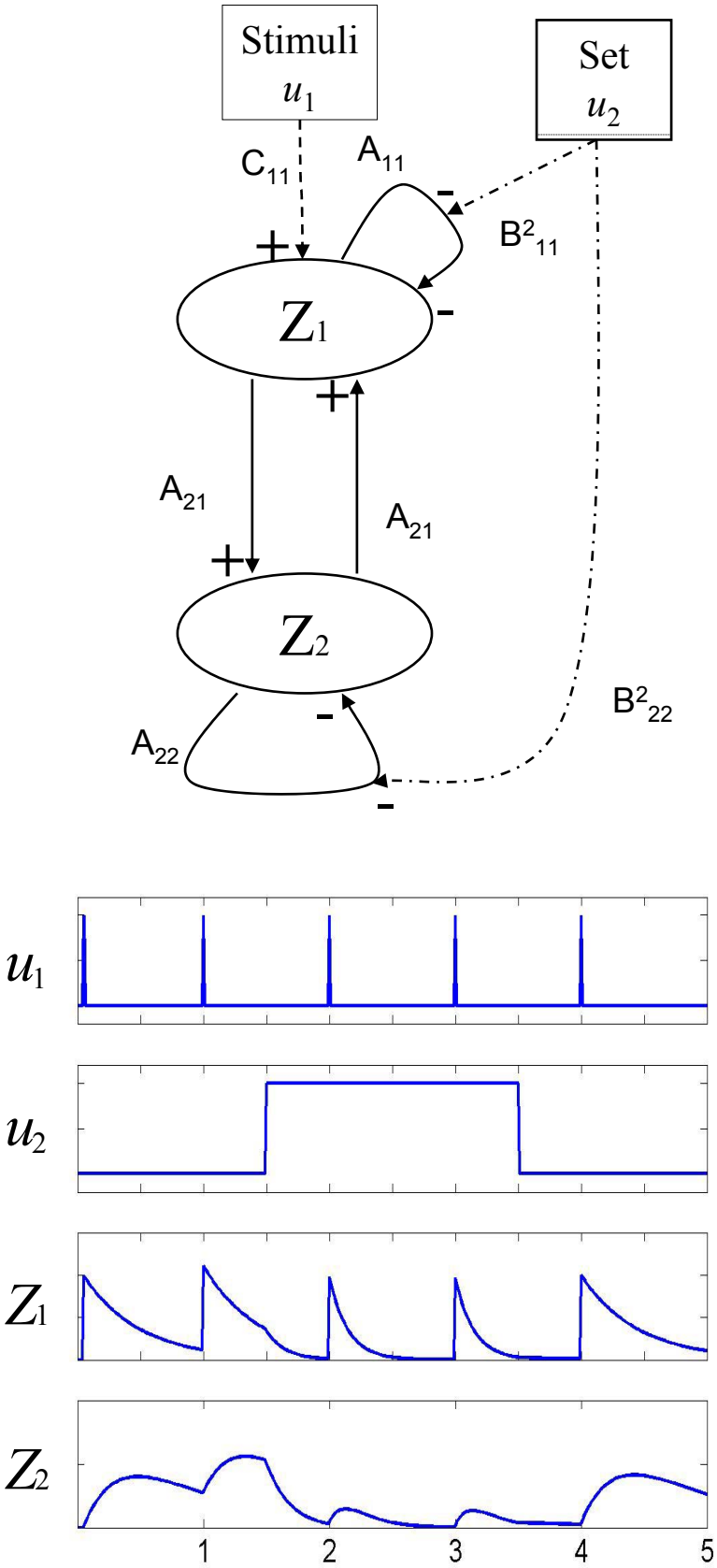
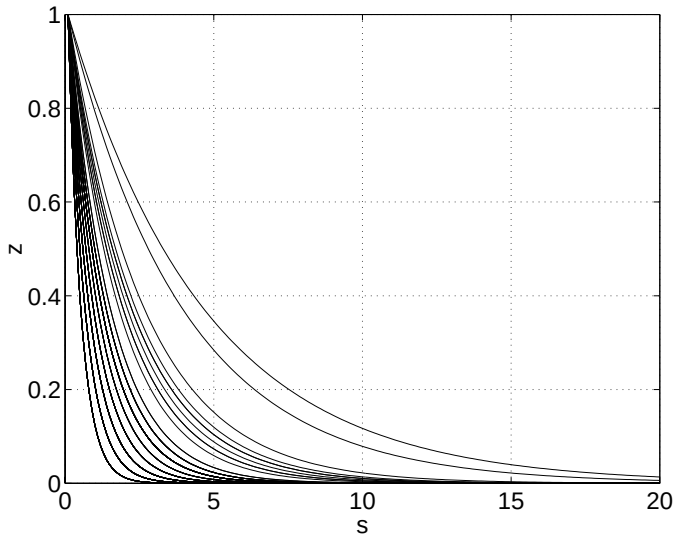
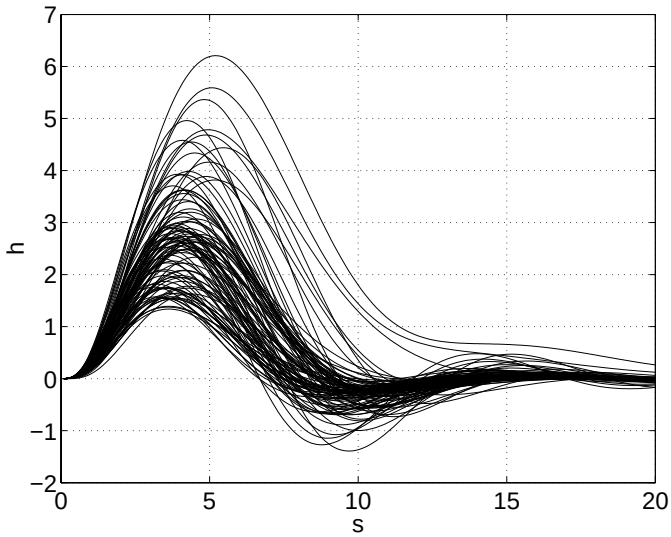

Stimuli
u_1
Set
u_2
C_{11}
A_{11}
B^2_{11}
Z_1
A_{21}
A_{21}
Z_2
B^2_{22}
A_{22}
u_1
u_2
Z_1
Z_2
1
2
3
4
5

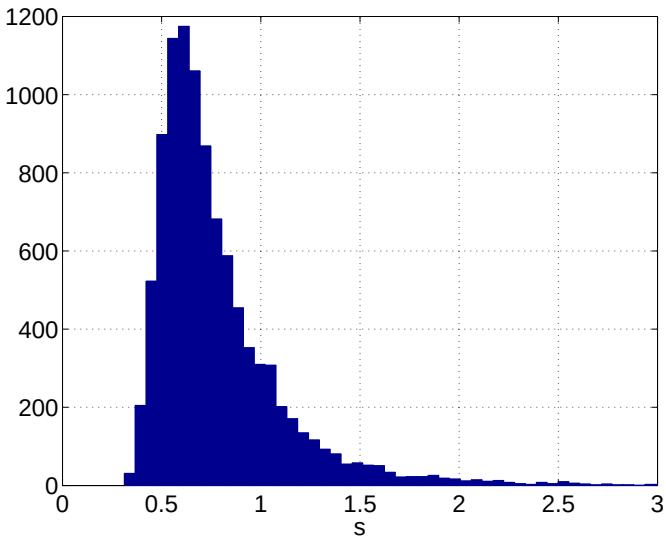
Figure 1



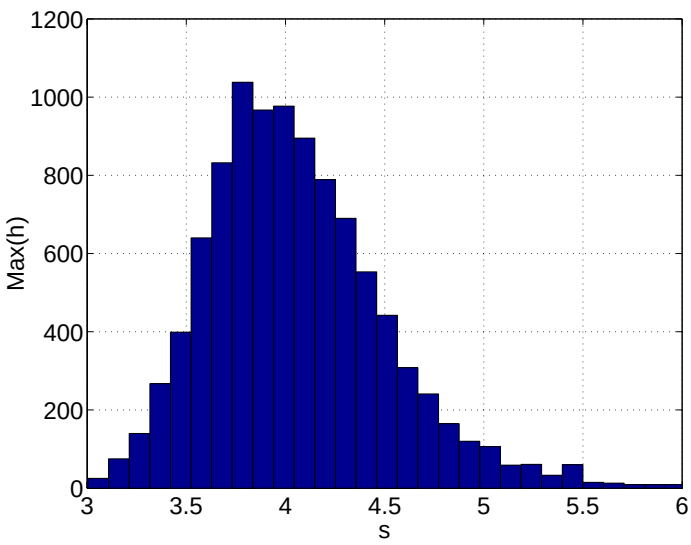
(a)



(b)



(c)



(d)

Figure 2

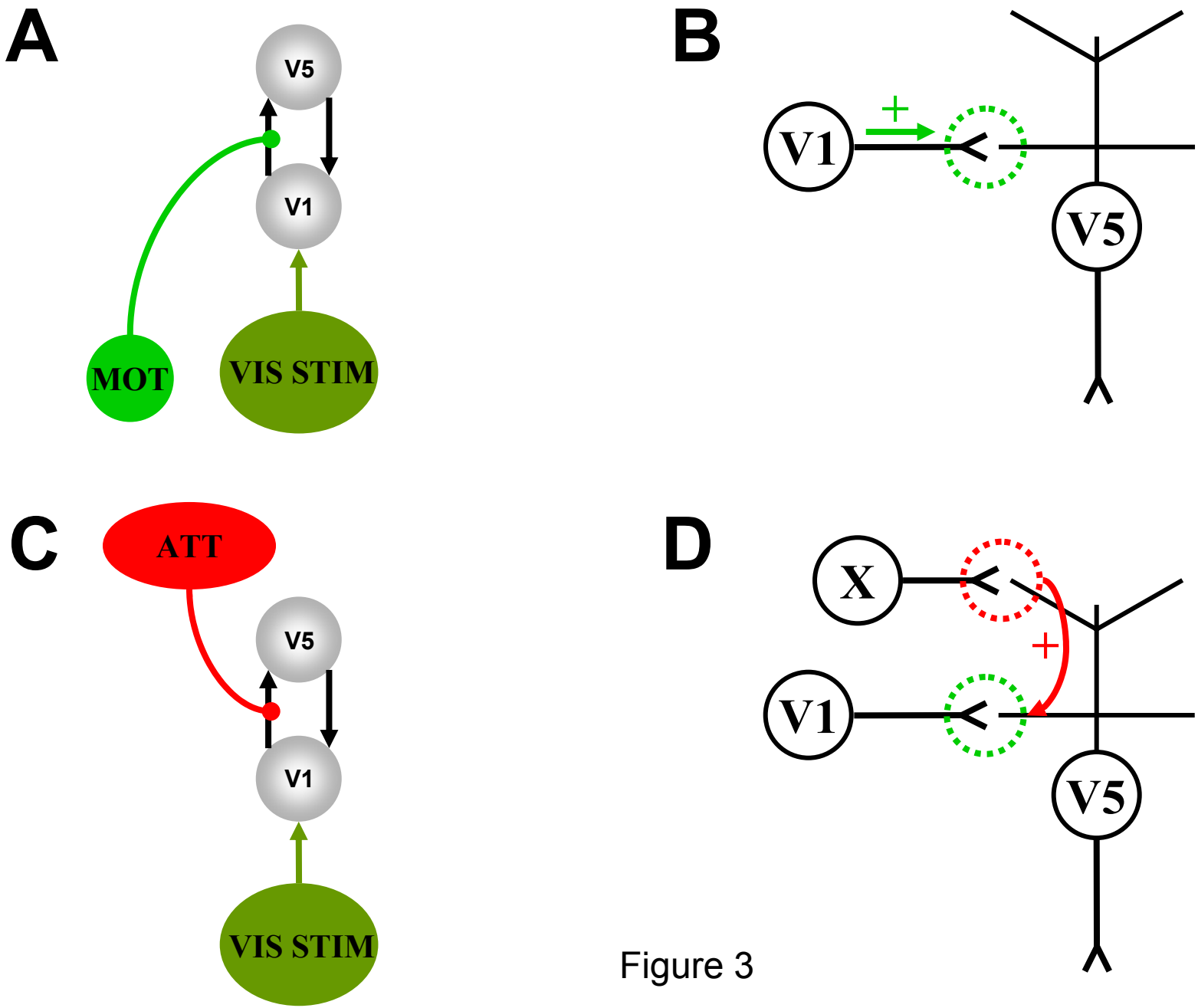


Figure 3

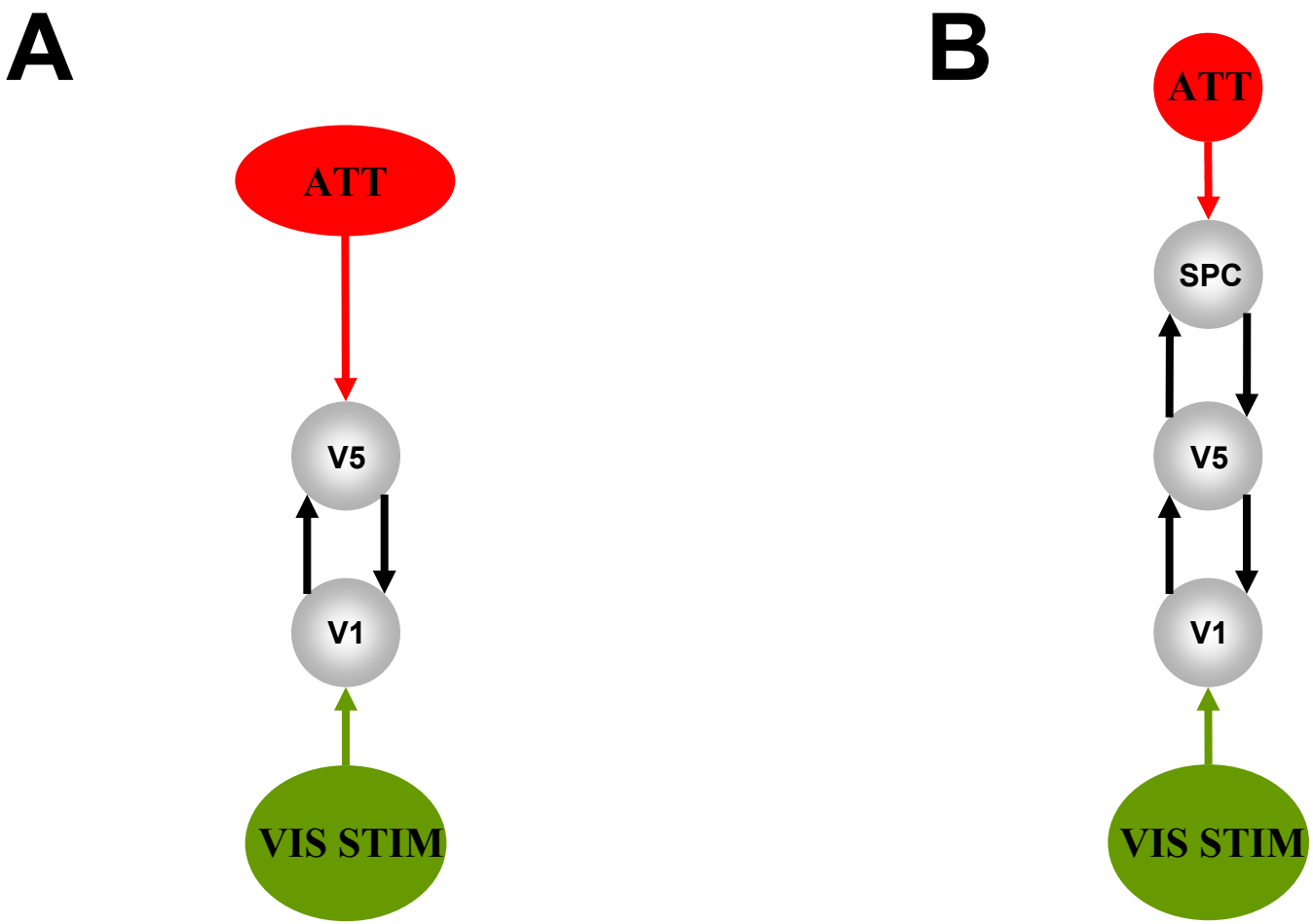


Figure 4

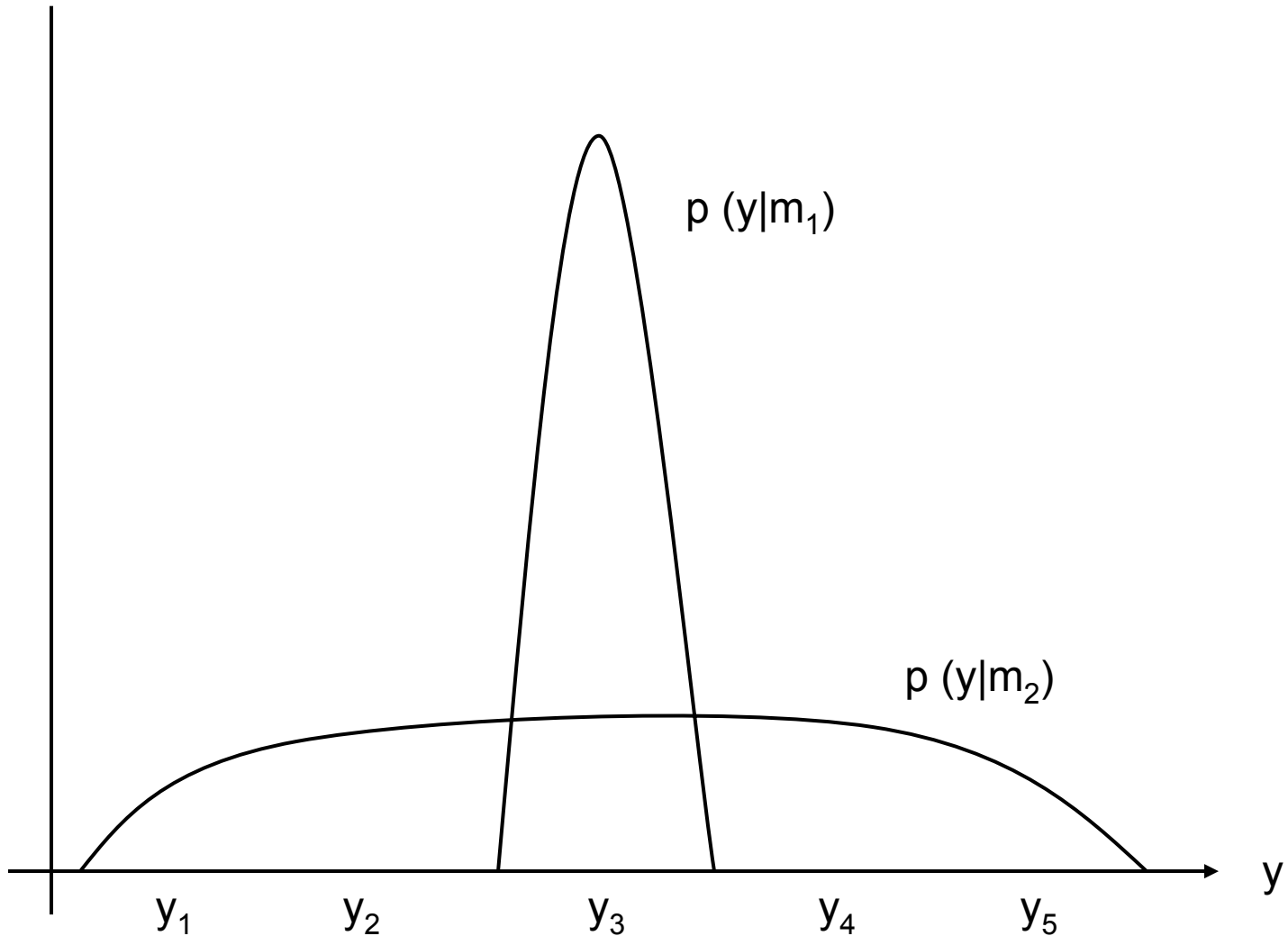
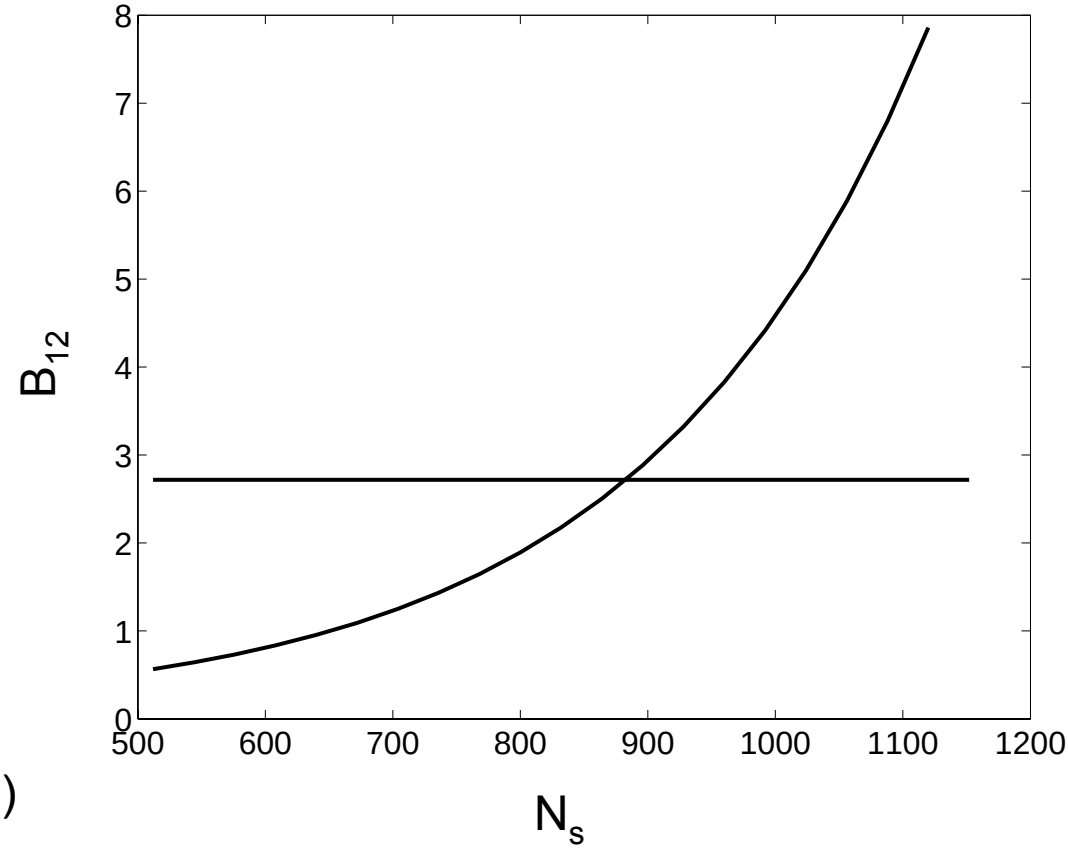
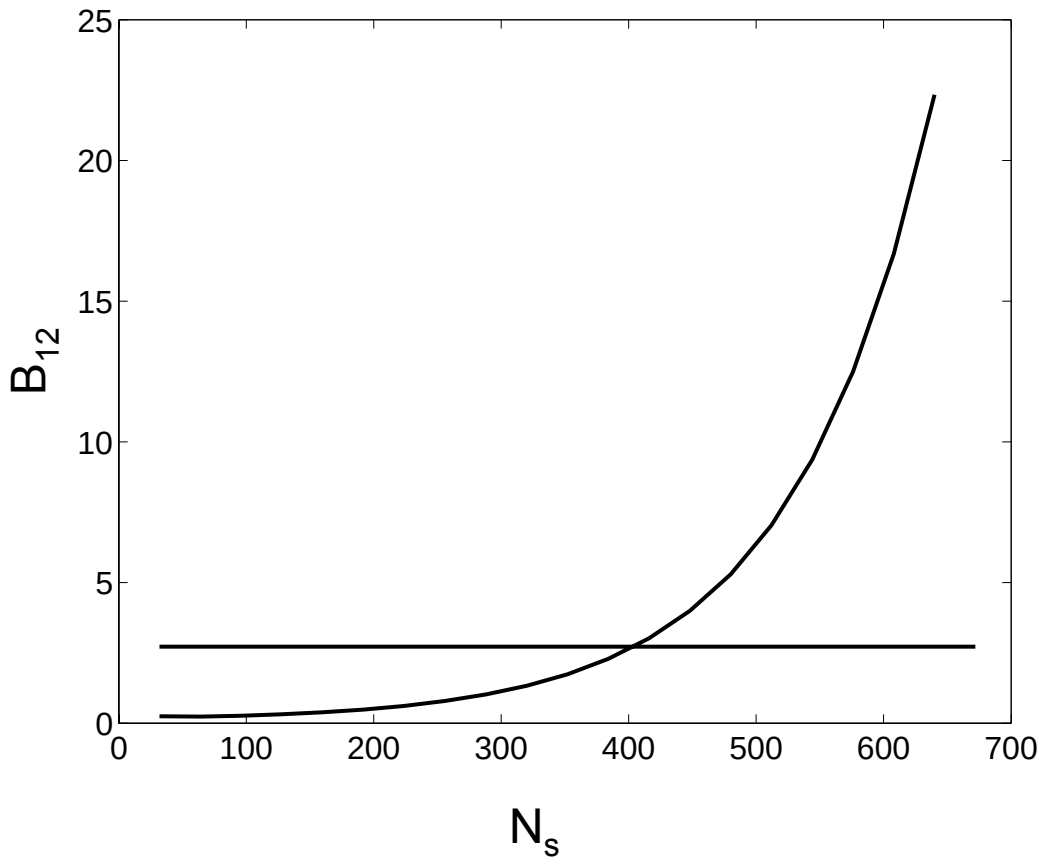


Figure 5


8
7
6
5
4
3
2
1
0
B_{12}
500
600
700
800
900
1000
1100
1200
N_s

(a)


25
20
15
10
5
0
B_{12}
0
100
200
300
400
500
600
700
N_s

(b)

Figure 6

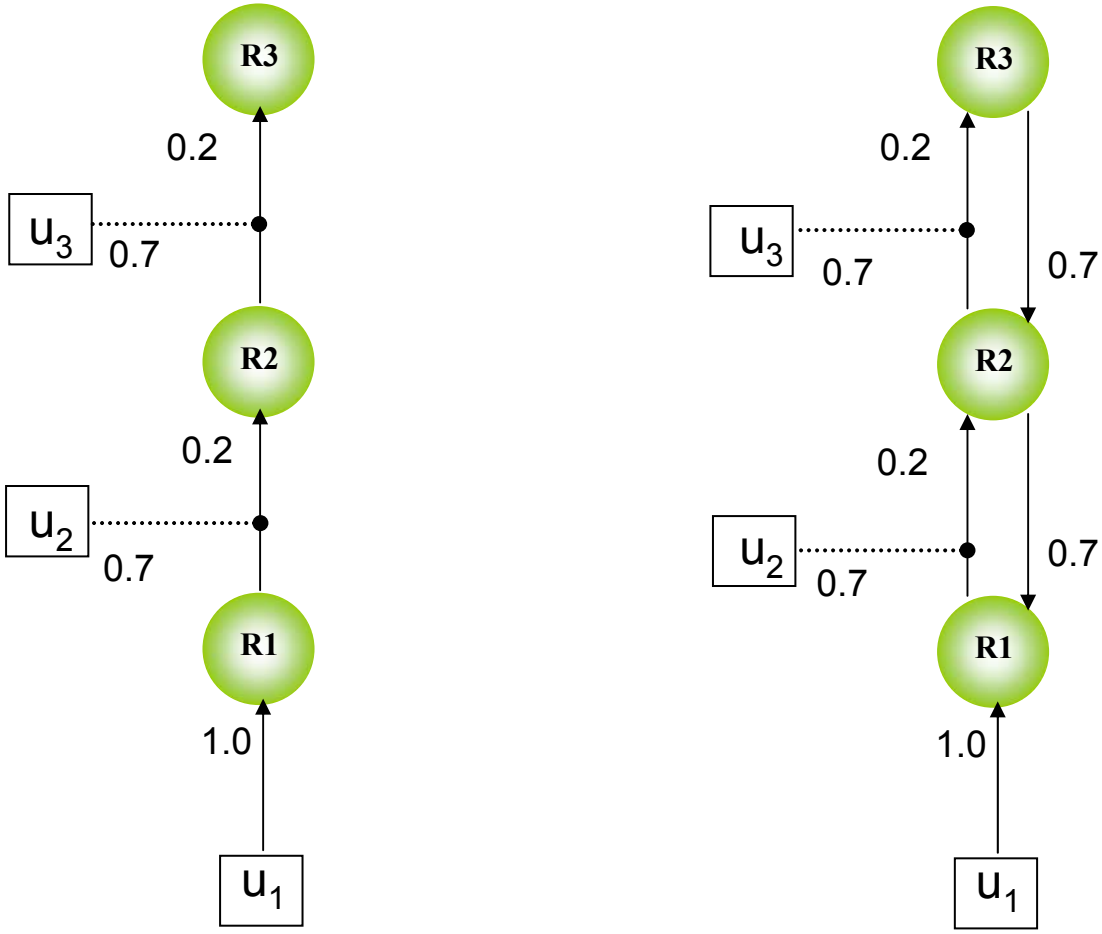


Figure 7

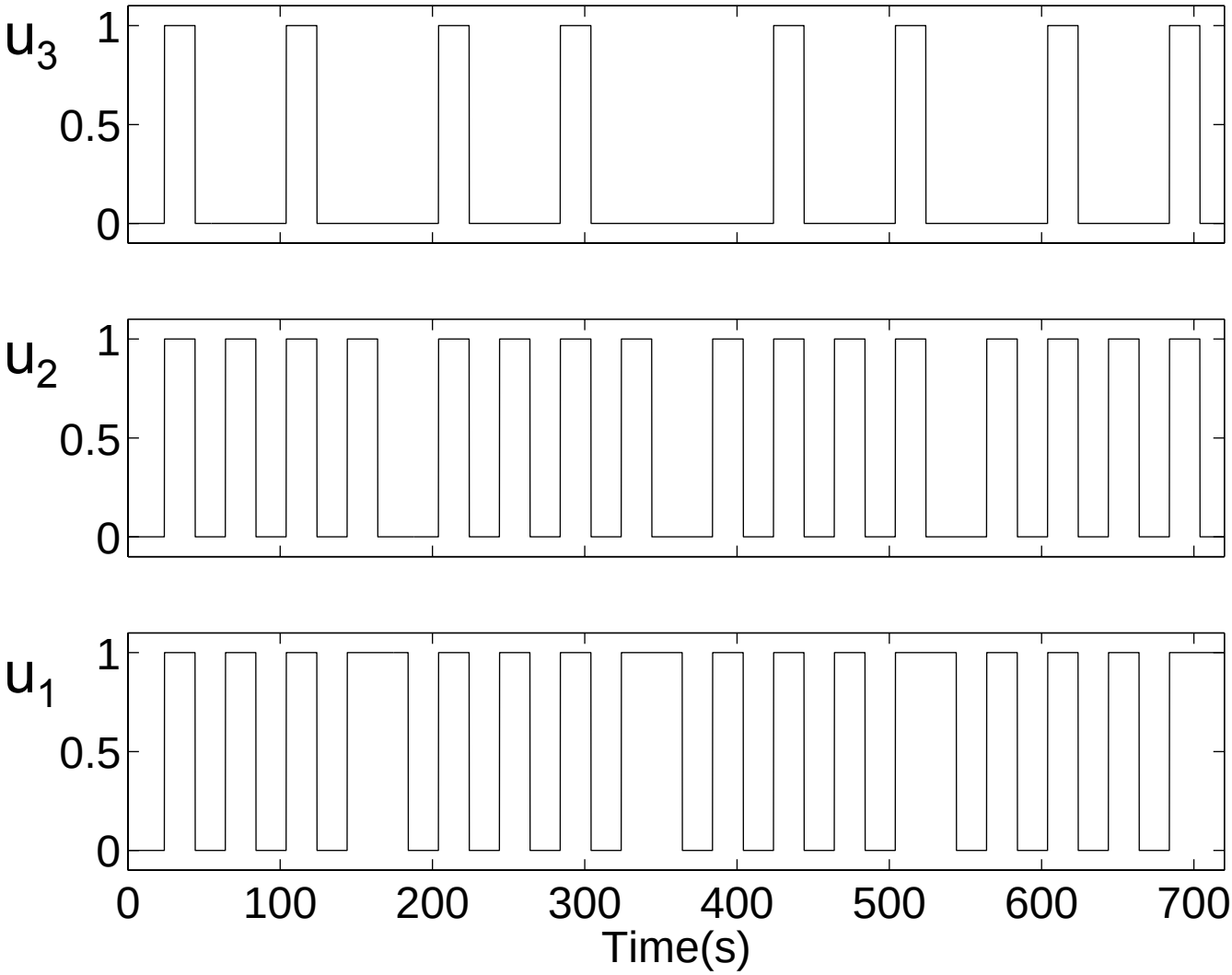
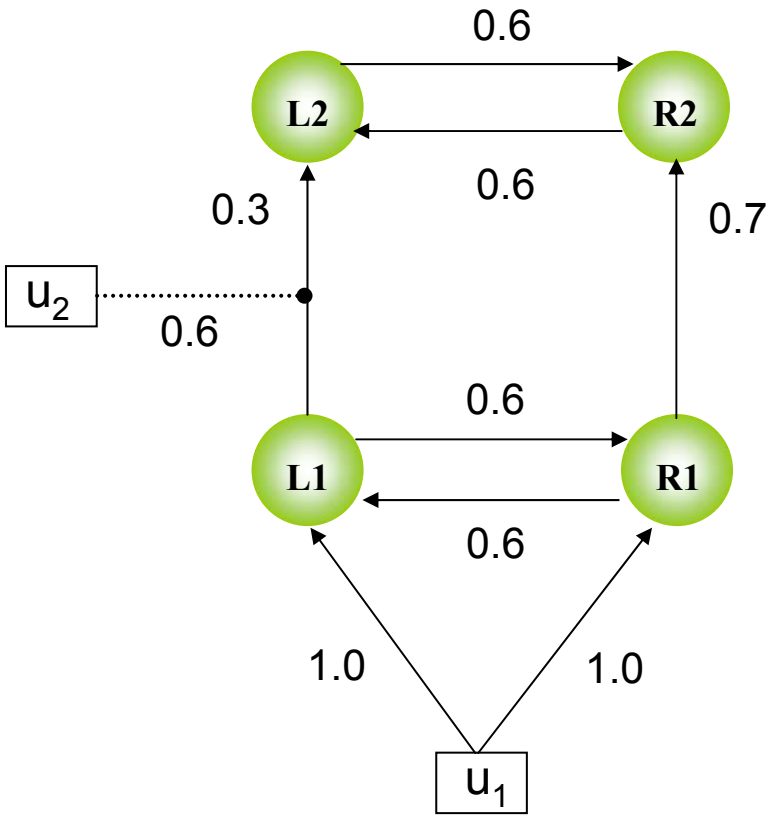


Figure 8



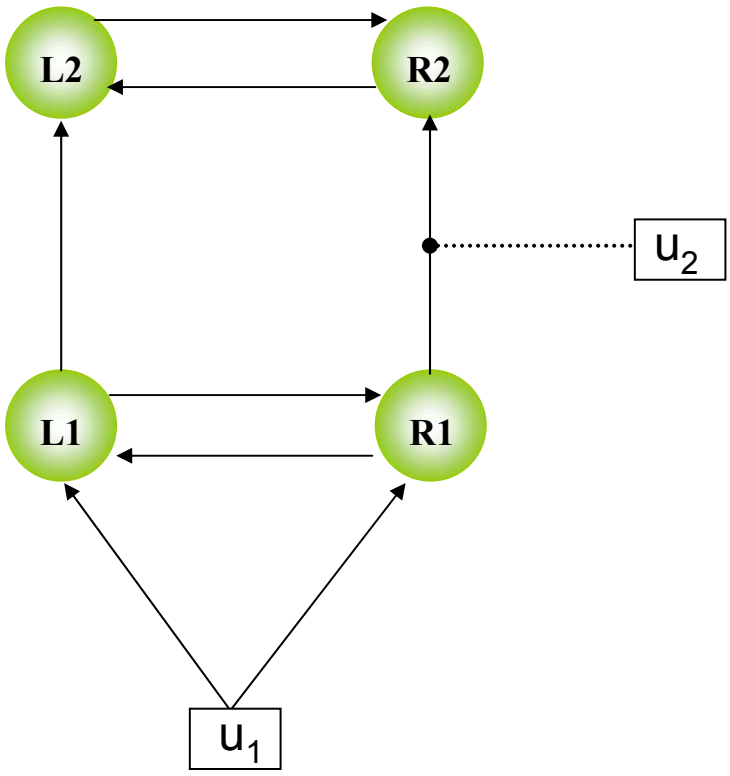


Figure 9

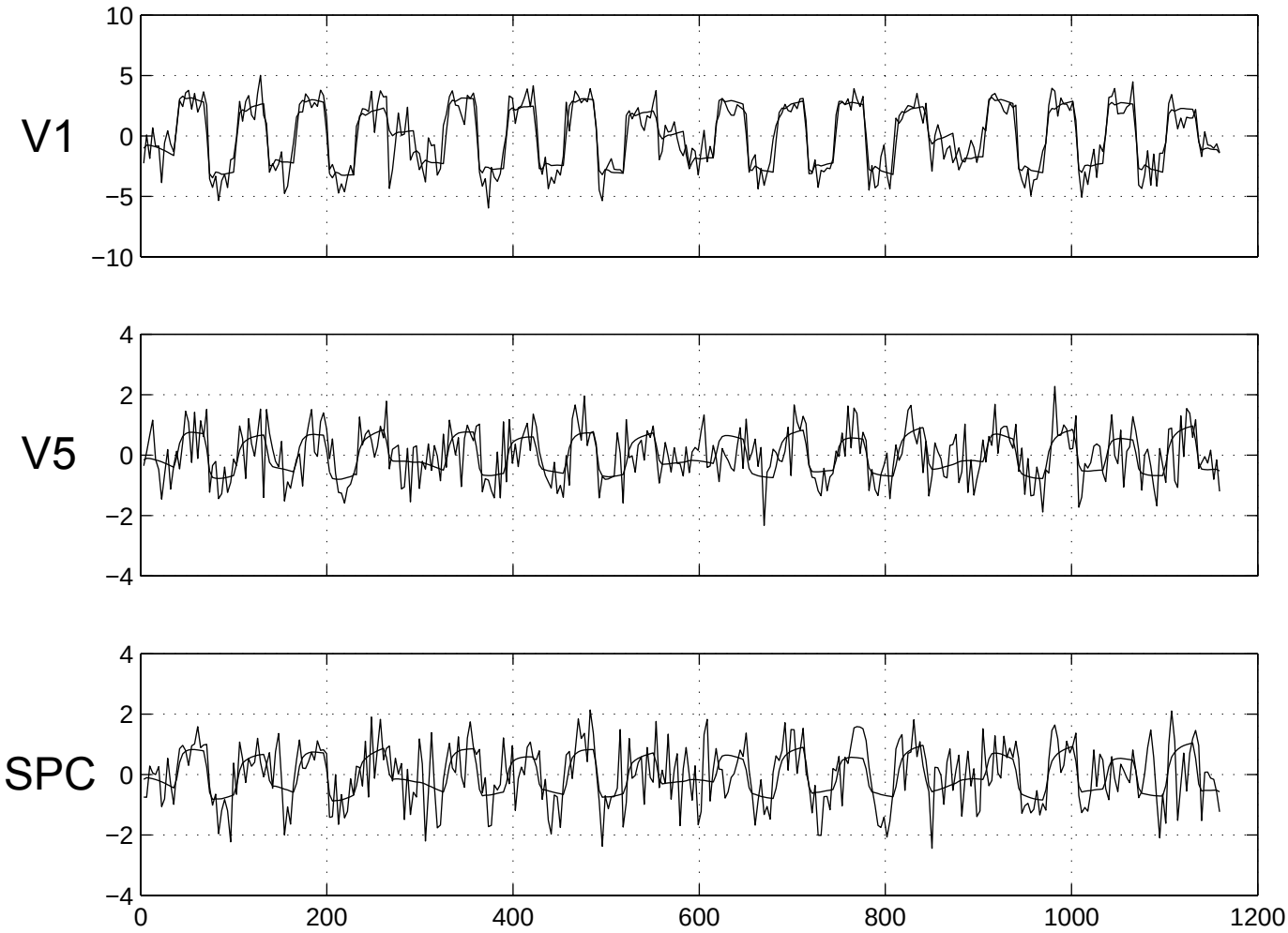
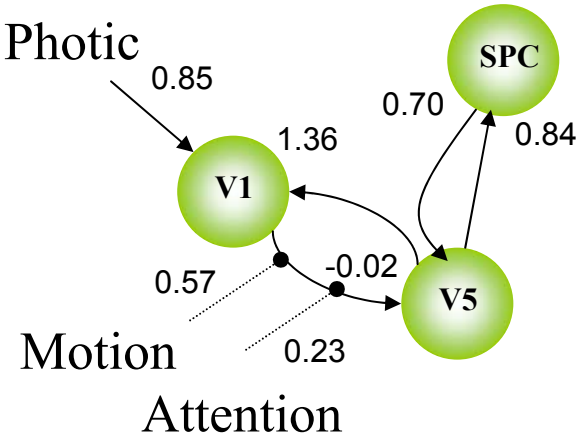
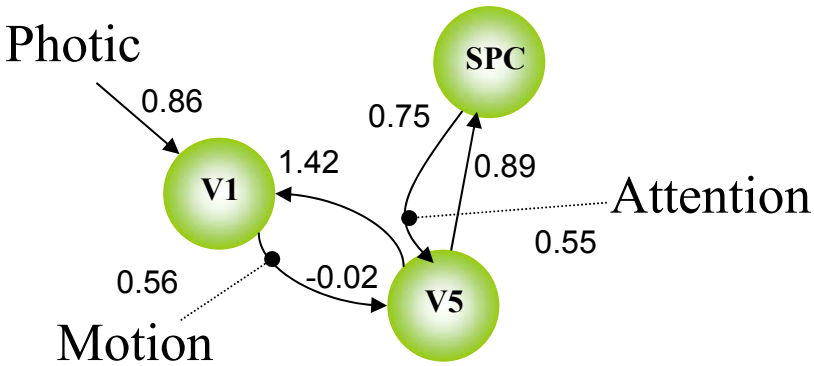


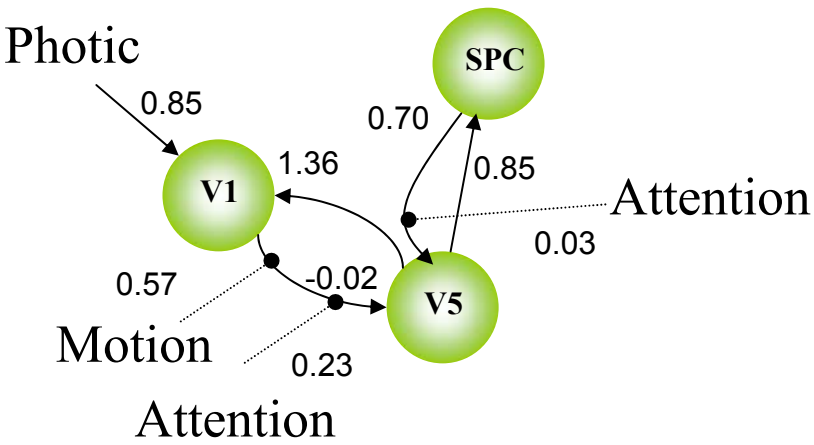
Figure 10



Model 1

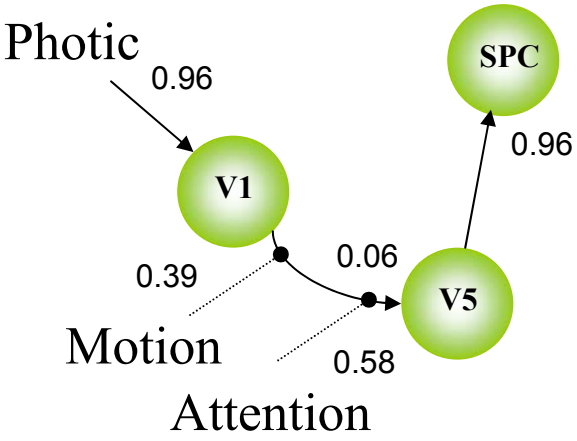


Model 2

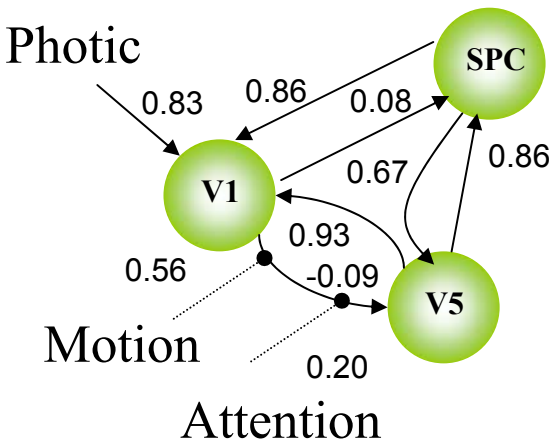


Model 3

Figure 11



Model 4



Model 5

Figure 11 (cont'd)

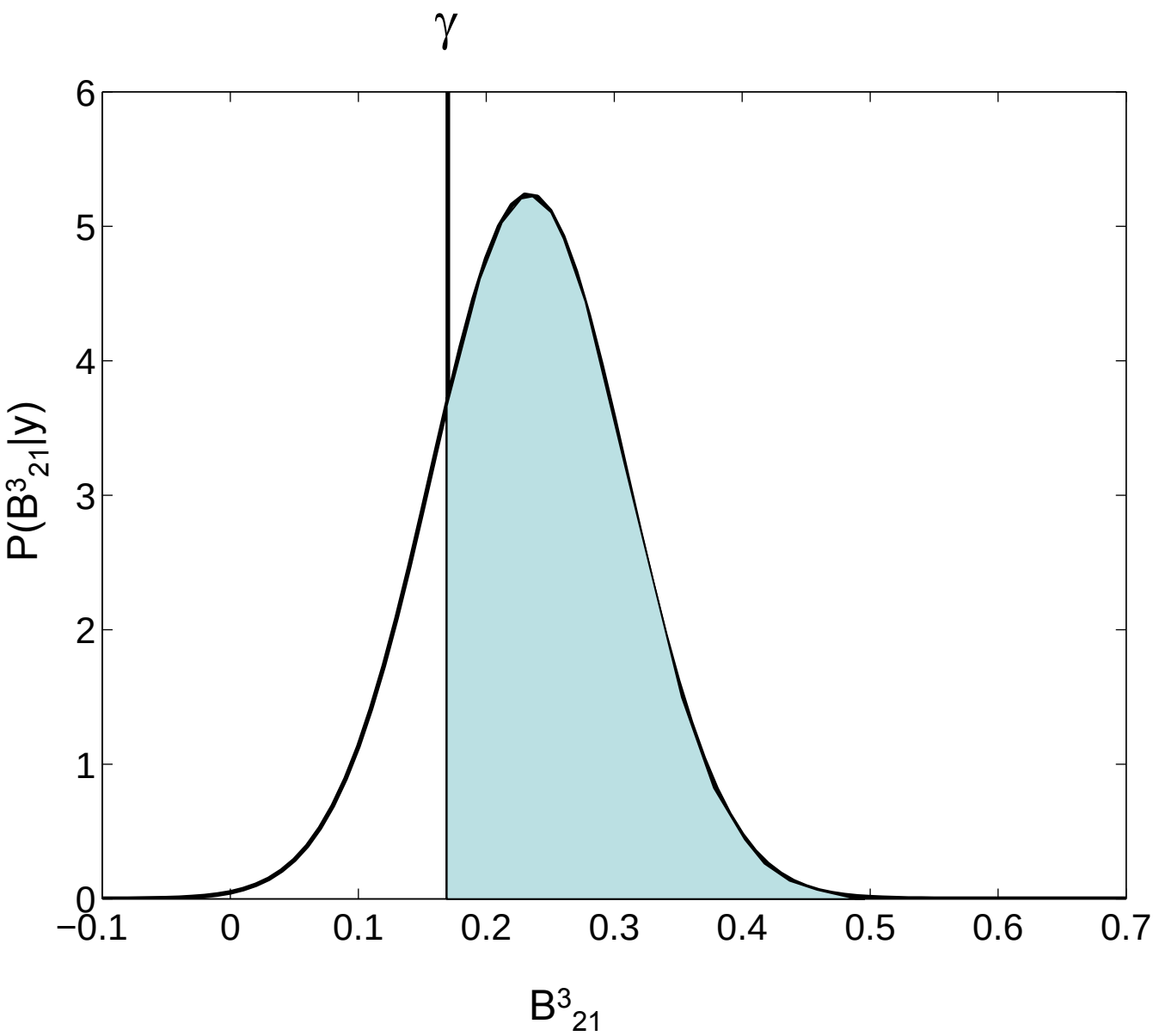


Figure 12

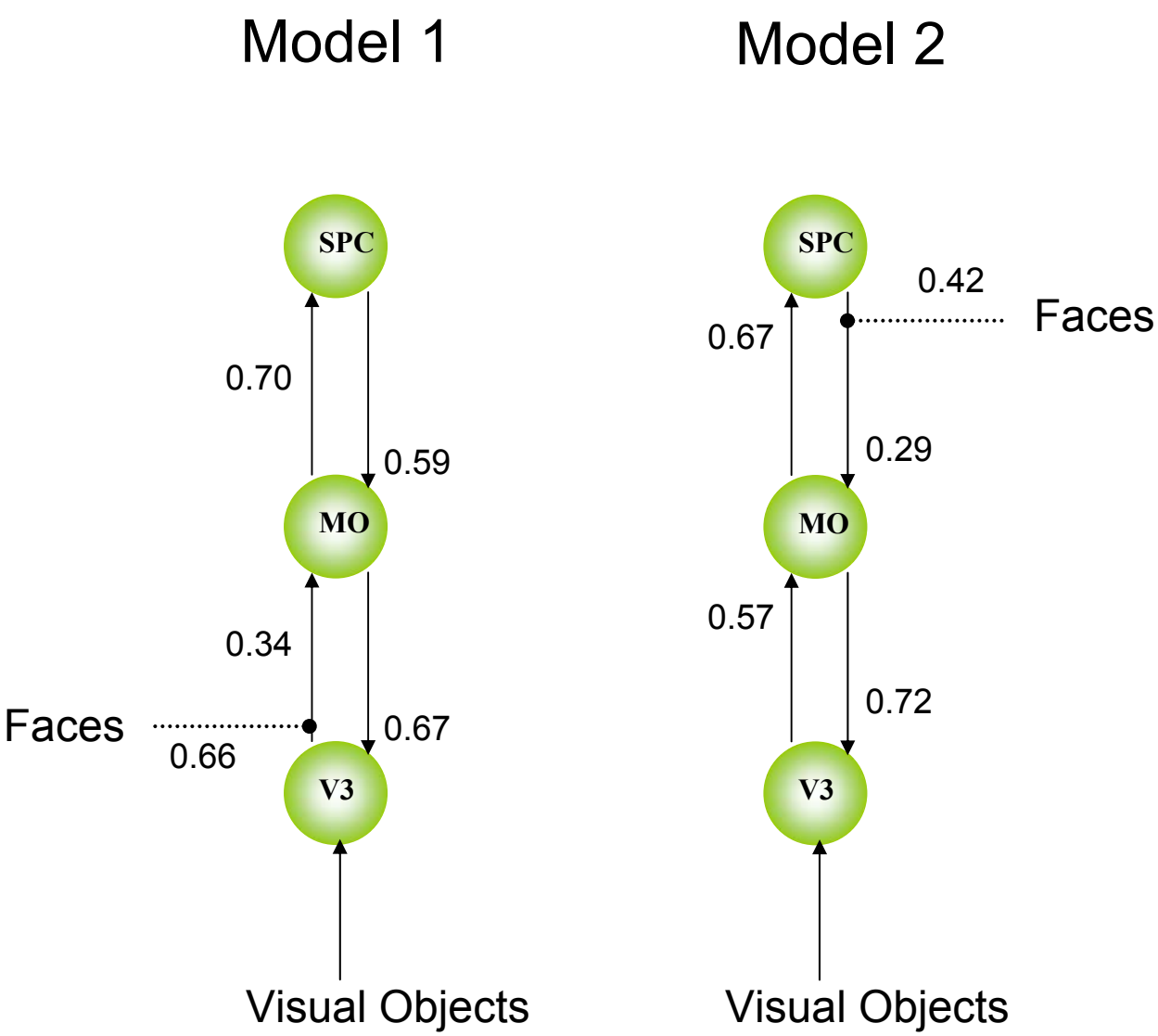

Model 1
SPC
0.70
0.59
MO
0.34
Faces
0.66
0.67
V3
Visual Objects
Model 2
SPC
0.42
Faces
0.67
0.29
MO
0.57
0.72
V3
Visual Objects

Figure 13

Model 3

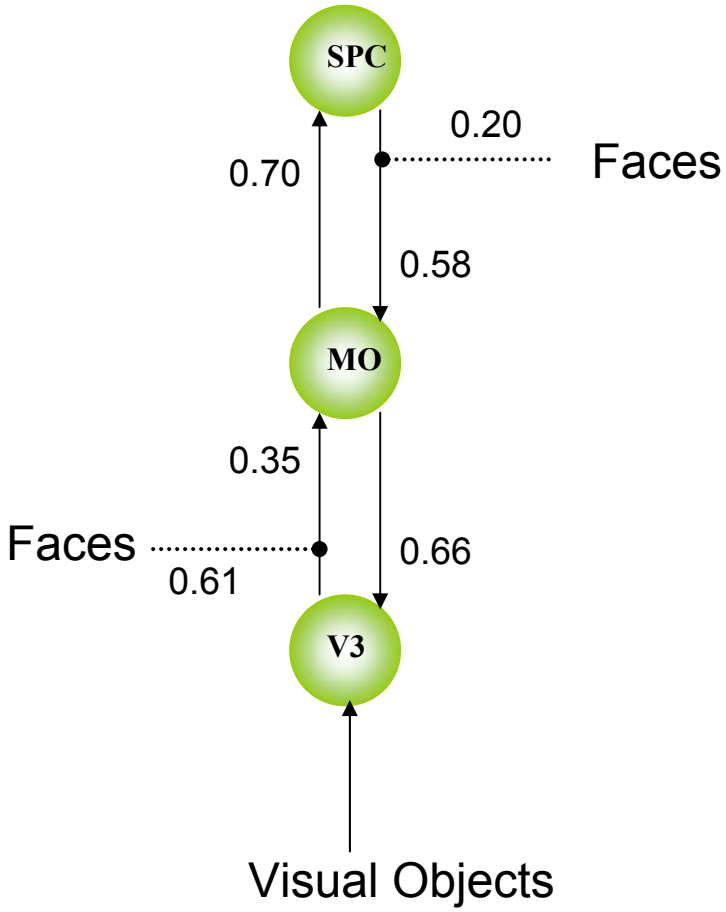


Figure 13 (cont'd)

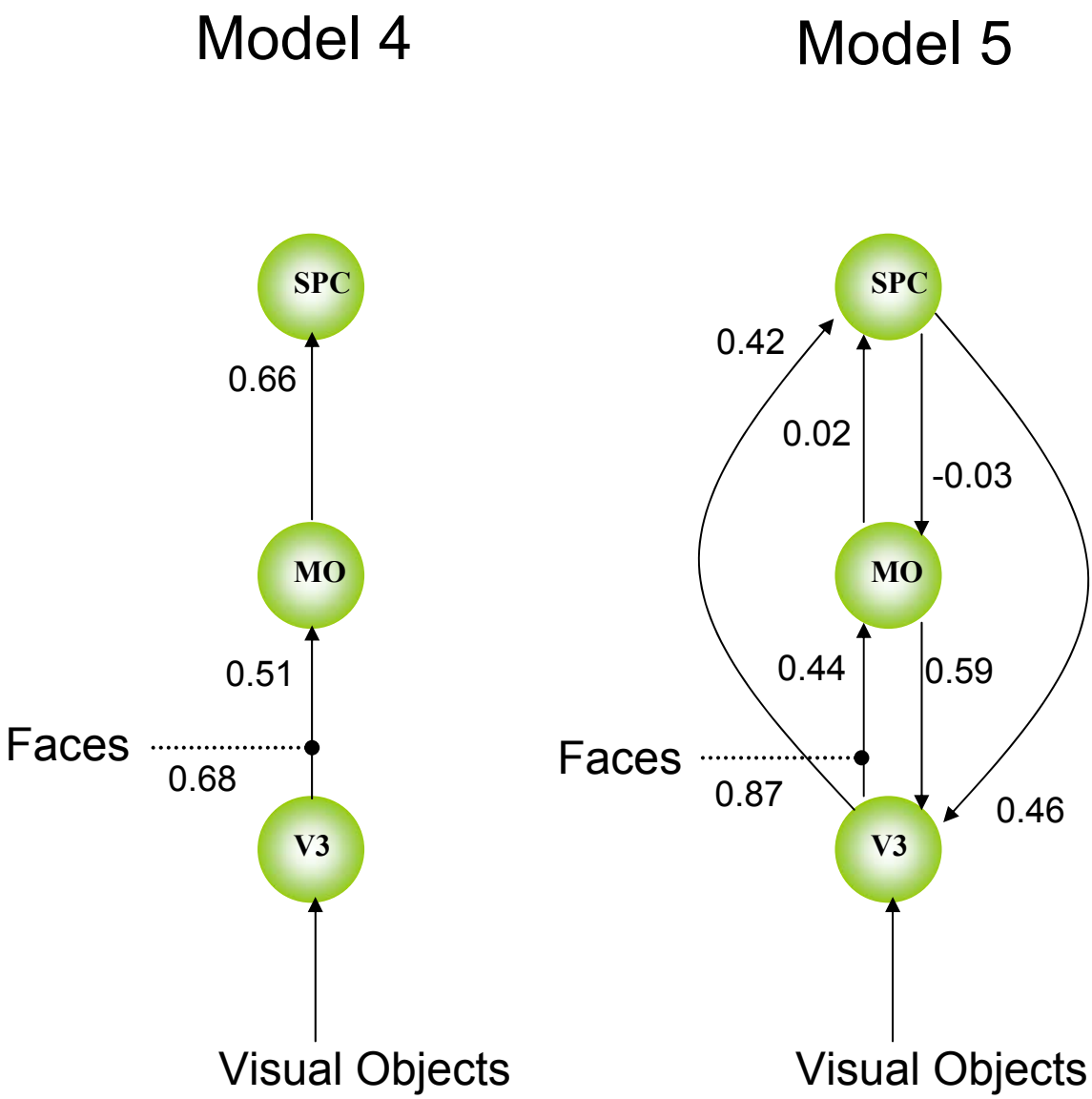


Figure 13 (cont'd)