A Unified Causal Model of Dynamic Causal Model and Granger Causal Model

Tian Ge[†] Jianfeng Feng[†,‡]

[†] Centre for Computational Systems Biology, Fudan University, PR China

[‡]Centre for Scientific Computing, University of Warwick, Coventry CV4 7AL, UK

May 12, 2009

Abstract

Two prominent approaches in exploring causal relationship from temporal biological data are Dynamic Causal Model (DCM) and Granger Causal Model (GCM) which have been intensively applied to fMRI data and are thought to be different radically. These two approaches are applicable to other types of temporal data such as microarray gene data, dynamical protein data, metabolic data etc. Here we present a novel approach which unifies the two approaches: unified causal model (UCM). The UCM is first intensively tested in toy models and then applied to biological data. Furthermore, the frequency domain decomposition of UCM is also presented.

1 Introduction

In order to realize the full potential of high-throughput data in biology we have to be able to convert them into biological knowledge. This process is best conceptualized as building a network: the experiment step on gene, protein, metabolite, brain area etc. gives us the part (node) and the next computational step is to discover the interactions (directed edges) connecting them. Causality analysis based upon experimental data has become one of the most powerful and valuable tools in discovering connections[23, 22]. Despite of the successful applications of the causality analysis in various areas in systems or computational biology, most developments are patchy and there is not a unified way to tackle the issue. In clinical science, for example, the emphasis of the development of causality is on how to clearly define causality [20]. A typical example is the so called Simpson paradox which demonstrates the ambiguity in defining the causal relationship. However, this issue disappears if we incorporate time into the definition of causality as Granger did [21]. Taking gene data (microarray data) as an instance, there are three approaches to deal with temporal data: a simple dynamical system approach, a dynamical Bayesian network approach and a Granger causality approach which is a generalization the dynamical system approach. In [9], we have discussed in details on the pros and cons of the latter two approaches. In fMRI data, two techniques have been introduced to address temporal dependencies and directed causal influences. They are Dynamic Causal Model(DCM) and Granger Causal Model(GCM). These two models have always been considered to differ radically from each other [1, 2]. DCM has state variables to cause the observed data and is believed to be causal models in a true sense while GCM is a phenomenological model which just tests

the statistical dependencies among the observations themselves and is irrespective of how the data are caused[1, 2, 3]. The significance of the two approaches in fMRI data alone is confirmed as in [1]: in 2008 there are around 450 papers devoted to both approaches. Here we have not counted published papers related to gene, protein and other types of biological data.

Hence the key question we want to address is the following. Can we present a unified approach of the various approaches mentioned above, in particular, the two causal models: DCM and GCM? The significance of such an approach is obvious and we would expect that the application of our approach could open many new issues in biology to be further investigated in the current tide of omic data and will play an important role in the 'next computational step'.

In this paper, we will show that DCM and GCM can be unified to the same framework. By setting up a conventional VAR model with additional deterministic inputs and also observation variables, we will have a more general model: unified causal model (UCM), to include both DCM and GCM as special cases and thus present a new and unified way to understand connectivity.

The UCM is first tested in toy models. With both state and observable variables, the interactions between nodes are successfully recovered with extended Kalman filter approach and extended partial Granger causality in both time and frequency domain. As expected, the GCM approach is not tailored for biological experiments. For example, in biological experiments, we often face the case that the data is recorded with and without a stimulus. The time gap between two adjacent stimuli is very short and we would expect the network structure remains unchanged during the whole experiment. However, the intensity of the

input is unknown to us, but it is demonstrated that our approach can also successfully reveal the network structure. This scenario is also the case for the data considered in [23, **?**], where the authors have treated the two cases separately, although there is a common and true structure.

Finally we applied UCM to experimental data recorded in the sheep inferotemporal cortex of the right and left hemisphere when it performs a discrimination and learning task [24].

2 Results

In order to evaluate the performance for the estimation of the state variables as well as the prediction of the parameters, we first apply the method to two toy models.

2.1 Toy Models

2.1.1 Toy Model 1

The toy model we use here is a triditional VAR model which has been extensively applied in tests of Granger causality [6]. Here we modify the model by adding two deterministic inputs $u(t)$ and $v(t)$. $u(t)$ is assumed to be a constant stimulation, i.e. $u(t) \equiv 0.1$ while $v(t)$ is assumed to a harmonic oscillators and has the form of sinusoidal function since biological rhythms are common phenomena of life [7]. Observation variables are also included in the toy model which are assumed to be functions of the state variables.Thus, we generate the

time series according to the following equations

$$\begin{aligned}
x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + 0.1x_1(t-1) \\
&\quad +0.1\cos[\tfrac{2\pi}{50}(t-1)] + \varepsilon_1(t) \\
x_2(t) &= -0.5x_1(t-1) + 0.5x_3(t-2) - 0.8 \times 0.1x_2(t-1) + \varepsilon_2(t) \\
x_3(t) &= -0.5x_2(t-1) + 0.5x_3(t-1) + 1.2 \times 0.1x_3(t-1) \\
&\quad -0.1\cos[\tfrac{2\pi}{50}(t-1)] + \varepsilon_3(t)
\end{aligned}$$

where ε_i, $i = 1, 2, 3$ are zero mean uncorrelated Gaussian noise with variance 0.5, 0.8 and 0.6 respectively. Hence, in this toy model, we have:

$$\begin{aligned}
A_1 &= \begin{bmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{bmatrix} = \begin{bmatrix} 0.95\sqrt{2} & 0 & 0 \\ -0.5 & 0 & 0 \\ 0 & -0.5 & 0.5 \end{bmatrix} \\
A_2 &= \begin{bmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{bmatrix} = \begin{bmatrix} -0.9025 & 0 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix} \\
B_1 &= \begin{bmatrix} b_{11}^1 & b_{12}^1 & b_{13}^1 \\ b_{21}^1 & b_{22}^1 & b_{23}^1 \\ b_{31}^1 & b_{32}^1 & b_{33}^1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.8 & 0 \\ 0 & 0 & 1.2 \end{bmatrix}, \; B_2 = 0, \; \vec{c} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \\
\tilde{A}_1 &= \begin{bmatrix} \tilde{a}_{11}^1 & \tilde{a}_{12}^1 & \tilde{a}_{13}^1 \\ \tilde{a}_{21}^1 & \tilde{a}_{22}^1 & \tilde{a}_{23}^1 \\ \tilde{a}_{31}^1 & \tilde{a}_{32}^1 & \tilde{a}_{33}^1 \end{bmatrix} = A_1 + uB_1 = \begin{bmatrix} 1.4435 & 0 & 0 \\ -0.5 & -0.08 & 0 \\ 0 & -0.5 & 0.62 \end{bmatrix} \\
u(t) &\equiv 0.1, \; v(t) = 0.1\cos[\tfrac{2\pi}{50}(t-1)]
\end{aligned}$$

Inspection of the above equations reveals that $x_1(t)$ is a direct source to $x_2(t)$, $x_2(t)$ and

$x_3(t)$ share a feedback loop and there is no direct connection between the remaining pairs of the state variables. Fig. 5A is an example of the 2000 time-steps of the data. Fig. 5B shows the network structure.

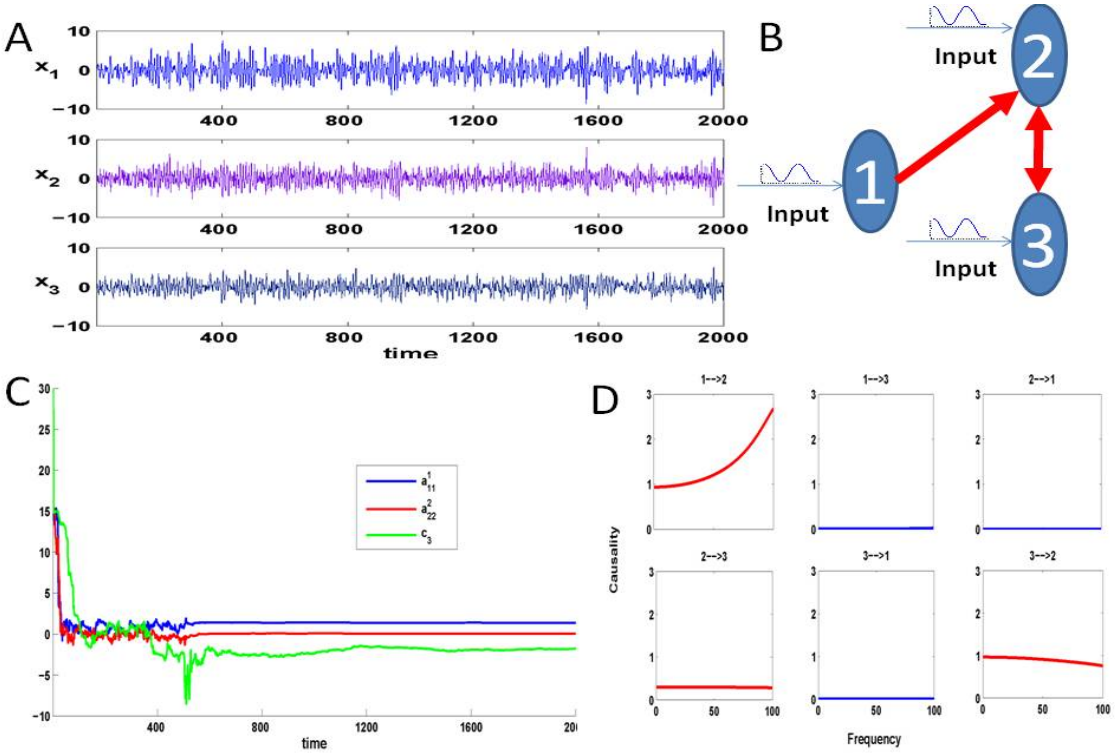

Figure 1: Results on Toy Model 1. **A**. Traces of the time series. **B**. The causal relationships considered in Toy Model 1 between the three state variables. **C**. The estimated parameters $\tilde{a}^1_{11}$,a^2_{22}, and c_3 for the simulated data in Toy Model 1. The initial values of the three parameters are all set to 30. The covariance matrix $Z(t)$ is first set to decay slowly to achieve faster convergence and then set to decay faster after five hundred time points to ensure a better accuracy. **D**. Frequency decomposition of all kinds of relationships between the state variables. Significant causal influences are marked by red.

The observation variables are

$$\begin{aligned} y_1(t) &= x_1(t) + \varepsilon_4(t) \\ y_2(t) &= x_2(t) + x_3(t) + \varepsilon_5(t) \\ y_3(t) &= x_1(t) - x_3(t) + \varepsilon_6(t) \end{aligned}$$

where ε_i, $i = 4, 5, 6$ are zero mean uncorrelated Gaussian noise with variance 0.1 and also uncorrelated with ε_i, $i = 1, 2, 3$.

Now, we can apply the method to this toy model, i.e., to estimate all the parameters $\tilde{A}_1, A_2, \vec{c}$ and state variables $x_i(t)$ from the deterministic input $u(t)$, $v(t)$ and noise observations. Simulations were performed for 2 seconds (2000 equally spaced time points). Fig. 5C shows that the parameters converge to their true values with only small fluctuations after several hundred data points even though the initial values of the parameters were assigned to be far away. It has already been pointed out that the covariance matrix $Z(t)$ of the noise in the parameter equation will affect the convergence rate and tracking performance [8]. In the situation here, a steep decay of the covariance matrix will lead to a better accuracy but the convergence is then slow. On the other hand, a slow decay will lead to a faster convergence but a larger fluctuation is observed. Hence, $Z(t)$ is carefully controlled to reduce to zero as the t increases (see Fig. 5C).

After the state variables have been recovered, we compute the partial Granger causality in both time and frequency domain (see Fig. 5D) and use the bootstrap approach to construct confidence intervals. Specifically, we simulate the fitted model to generate a data set of 1000 realizations of 2000 time points and use 3σ as the confidence interval. In this result, a causal connection illustrated as part of the network if and only if the lower bound of the 95% confidence interval of the causality is greater than zero. The results show that

our unified model can detect the causal relationship correctly in both time and frequency domain.

2.1.2 Toy model 2

When dealing with real data, it is quite common that we need to detect the causal influence between time series from several variables affected by some stimulus. The stimulus may be very complicated or hard to measure that it is impossible to formulate its form explicitly. However, if we ignore the influence of these input and use a traditional VAR model to detect the causality, it's quite probable that we will get a misleading structure even if we use a high-order VAR model.

We consider the following toy model which has exactly the same connection coefficients between the three state variables considered in Toy model 1 with an additional simple constant input function:

$$\begin{aligned} x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + p + \varepsilon_1(t) \\ x_2(t) &= -0.5x_1(t-1) + 0.5x_3(t-2) + p + \varepsilon_2(t) \\ x_3(t) &= -0.5x_2(t-1) + 0.5x_3(t-1) - p + \varepsilon_3(t) \end{aligned} \tag{1}$$

where ε_i, $i = 1, 2, 3$ are zero mean uncorrelated Gaussian noise with variance 0.5, 0.8 and 0.6 respectively.

Here, we assume that $p \equiv 0.5$ and the observation variables y_i, $i = 1, 2, 3$ are identical to the state variables with observation noise. The variance of the noise is 0.1. It is obviously that the network structure is the same one as shown in Fig. 5B. However, if we ignore the constant input and just use a VAR model to detect this structure, we will get the structure

showed in Fig. 2A with confidence intervals where two additional causal relationships (i.e. $1 \rightarrow 3, 3 \rightarrow 1$) are presented which shows that the real causal influence can no longer be correctly detected. Furthermore, when the input is not taken into consideration, the coefficients of the connection matrix will be meaningless and no longer provide us with the correct intuition of the strength of connection among the state variables. That is why we should include the stimulus into our model although sometimes we don't know the form or intensity of the stimulus.

With our unified model here, to some extend we can solve the issue above and detect the causal influence correctly among those state variables affected by some unknown stimulus intermittently, although our model is set up for deterministic input originally.

We now generate a time series of 10000 time points which is composed of 10 segments with equal length, a.e., $1-1000, 1001-2000, \cdots, 9001-10000$. Each segment takes the form

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = A_1 \begin{bmatrix} x_1(t-1) \\ x_2(t-1) \\ x_3(t-1) \end{bmatrix} + A_2 \begin{bmatrix} x_1(t-2) \\ x_2(t-2) \\ x_3(t-2) \end{bmatrix} + p \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1(t) \\ \varepsilon_2(t) \\ \varepsilon_3(t) \end{bmatrix} \tag{2}$$

where $A_1 = \begin{bmatrix} 0.95\sqrt{2} & 0 & 0 \\ -0.5 & 0 & 0 \\ 0 & -0.5 & 0.5 \end{bmatrix}$, $A_2 = \begin{bmatrix} -0.9025 & 0 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix}$ and ε_i, $i = 1, 2, 3$ are zero mean uncorrelated Gaussian noise with variance 0.5, 0.8 and 0.6 respectively.

The five segments $1-1000, 2001-3000, \cdots, 8001-9000$ are generated according to the above toy model without input, i.e., $p \equiv 0$, while the remaining five segments are assumed to include input of random intensity which will affect the state variables also

randomly. Specifically, within each segment, p is assigned a random value which is generated with the normal distribution $p \sim N(0, 1)$, and the same case with c_i: $c_i \sim N(0, 1)$, $i = 1, 2, 3$. Observation variables are still assumed to be identical to the state variables with observation noise whose variance is 0.1.

Hence, the network structure of the three state variables is still the same one shown in Fig. 5B while each state variable is affected by some input that we don't know the intensity. Fig. 2B shows the predicted network structure with confidence interval using our unified model. The results show that we can still detect the causal influence correctly in this situation.

2.2 Experimental data

Now, we apply our model to our experimental data. We consider the neuronal activity of the left and right inferior temporal cortex(IT) in a sheep's brain before and after a visual stimulus. The sheep is presented a blank screen with a central dot for one second and then a pair of faces for another second. 64 local field potentials in each each region are recorded with a sampling rate 1000Hz. The experiment is repeated 40 trials.

We first select 3 time series with the most significant variance from each region respectively. The visual stimulus in this experiment here is impossible to formulate at this moment. However, as we have showed above, we assume that the effect of the stimulus can be regarded as a constant input. The inclusion of the stimulus will make the model more resonable.

Another problem here is that if we detect the connection before and after the stimulus

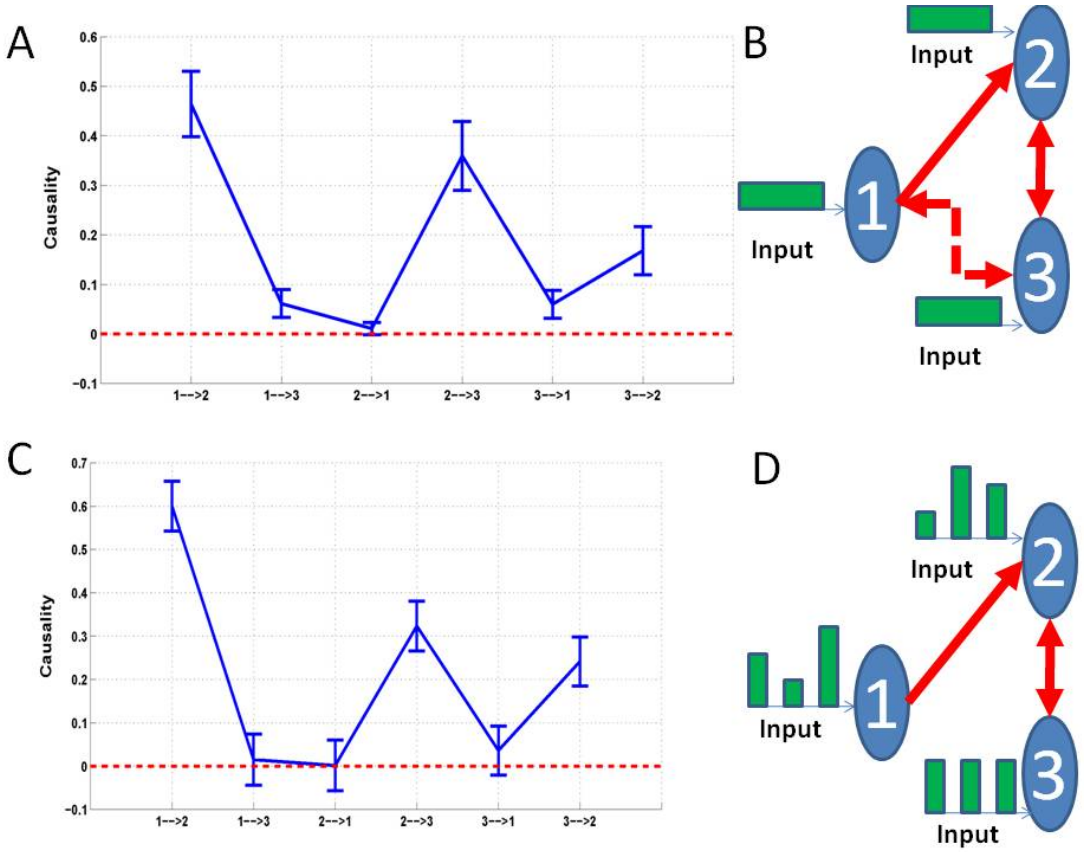

Figure 2: Results on Toy model 2. Network structures with and without stimulus. **A**. Confidence intervals of all links between units. The data is generated with Eq. (1), but we use $p = 0$ in our algorithms and a traditional VAR(10) model to detect the causal influence. **B**. Confidence intervals of all links between units. The data is generated with Eq. (2).

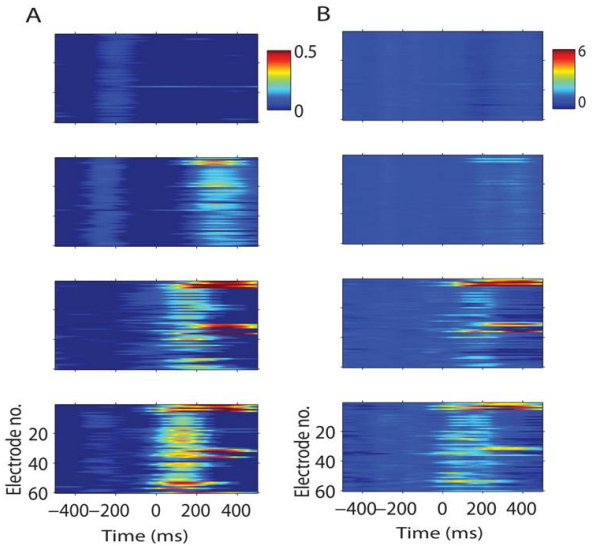

Figure 3: Rapid time course of learning effects on theta nested gamma. Pseudocolor panels show changes in: (a) Theta amplitude, and (b) Theta/gamma ratio in the right IT during the learning of one new face pair in Sheep B over sequential (top to bottom) blocks of 20-40 trials conducted over approximately 20 min (data plotted from 60 electrodes). Discrimination performance across the 4 blocks was 70%, 93%, 90% and 90% correct respectively (i.e. the learning criterion of $> 80\%$ was achieved in the second and subsequent blocks). The face pair stimulus occurs at time zero and the pseudocolor scale indicates normalised (by the maximum value during the stimulus) differences between pre and during stimulus.

respectively, we will get two totally different structures among the time series (not shown). However, it is widely believed that the connection will not change in such a short time. Hence, the structure we obtained in this way is still considered to be misleading.

In order to carry out a reasonable structure to illustrate the connection of the areas in the brain, we assume here that the connections among the time series don't change and we need to consider the time series before and after the stimulus as a whole and detect a unified structure. With the application of our model, we can include the intermittent stimulus and get a comparatively reliable structure. Fig. 4(Left) is the structure when we treat the data from the left and right hemisphere separately. The results are showed in Fig. 4 (Right) where we detect the influence of the time series from both regions respectively and also the interaction between two regions.

3 Methods

3.1 UCM Model

We consider the bilinear approximation to DCM taking the following form [1]:

$$\begin{cases} \dfrac{d\vec{x}(t)}{dt} &= [A + u(t)B]\,\vec{x}(t) + u(t)\vec{c} \\ \vec{y}(t) &= \tilde{g}(\vec{x}(t)) + \vec{\varepsilon}(t) \end{cases} \tag{3}$$

where $\vec{x}(t) = (x_1, \cdots, x_N)^T$ are state variables and $(\cdot)^T$ is the transpose of a vector, $u(t)$ is a known deterministic input,

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix}, B = \begin{bmatrix} b_{11} & \cdots & b_{1N} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{NN} \end{bmatrix}, \vec{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix}$$

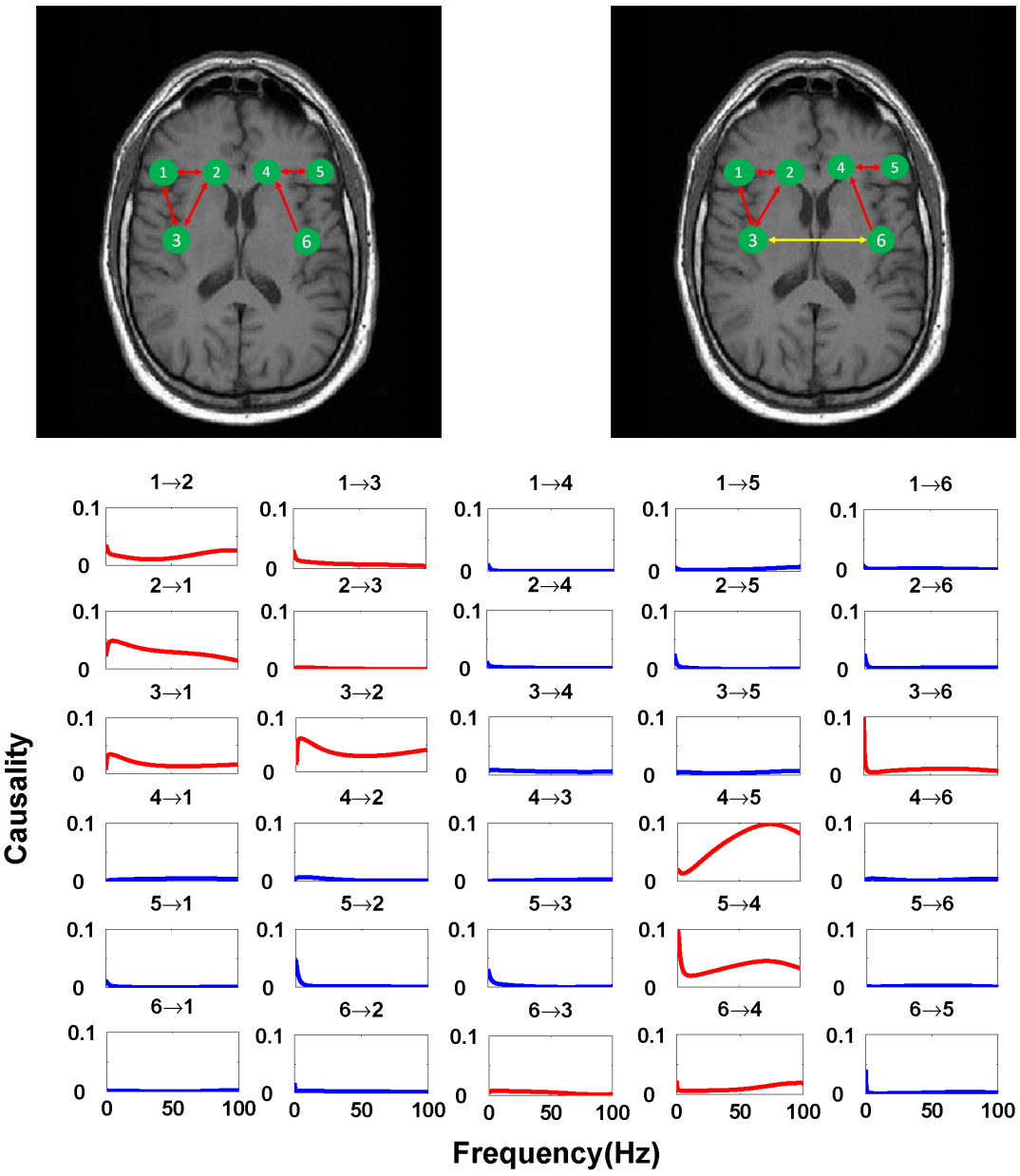

Figure 4: Network structure in the left and right hemisphere in inferotemporal cortex. Left. Data from two hemispheres are treated separately. Right. The network for both left and right hemisphere (a total of six sites) is detected simultaneously. Locations are not precise, only for illustrative purpose.

are parameters that mediate the intrinsic coupling among states, allow the inputs to modulate the coupling, and elicit the influence of extrinsic inputs on the states respectively. Then, the state variables $\vec{x}(t)$ enter a specific model to produce the outputs $\vec{y}(t)$ with the observation noise $\vec{\varepsilon}(t)$.

On the other hand, the GCM takes the form [1]:

$$\vec{x}(t) = A_1\vec{x}(t-1) + \cdots + A_p\,\vec{x}(t-p) + \vec{\varepsilon}(t) \tag{4}$$

where $A_i = \begin{bmatrix} a^i_{11} & \cdots & a^i_{1N} \\ \vdots & \ddots & \vdots \\ a^i_{N1} & \cdots & a^i_{NN} \end{bmatrix}$ and the model has a vector autoregressive representation with an order up to p.

The DCM is represented in terms of non-linear differential equations while the GCM is formulated in discrete time and the dependencies among state variables are approximated by a linear mapping over time-lags which seems to be quite different. However, if we first assume that there are no deterministic inputs $u(t)$ and the observation variables are identical to the state variables in DCM, we can apply a first-order difference to the DCM and obtain

$$[\vec{x}(t) - \vec{x}(t-1)] = A\vec{x}(t-1)$$

then

$$\vec{x}(t) = [A + I]\,\vec{x}(t-1) = A^*\vec{x}(t-1)$$

So, with identical observation variables and no inputs, the DCM can be regarded as a VAR(1) model which is a special case of the GCM represented by a VAR(p) model.

On the other hand, the GCM with autoregressive representation always takes the past

information into consideration while the DCM has no time-lags included in the differential equations. So, if we alter the DCM to the form:

$$\begin{aligned} \frac{d\vec{x}(t)}{dt} &= \int_0^t \left\{ [A + u(t-\tau)B]\, \vec{x}(t-\tau) + u(t-\tau)\vec{c} \right\} k(\tau) d\tau \\ \vec{y}(t) &= \tilde{g}(\vec{x}(t)) + \vec{\varepsilon}(t) \end{aligned}$$

where $k(\cdot)$ is a kernel function. Then the DCM also shares the feature of the GCM.

Hence, we can take the advantages of both DCM and GCM. In order to have a detailed comparison with the results in the literature, we borrow the Table 1 in [1] and is included in Discussion section. If we take the deterministic inputs intof consideration in the GCM model and also assume that the observations are a function of, and only of the state variables, the two models then can be unified as the following form:

$$\begin{cases} \vec{x}(t) &= [A_1 + u(t-1)B_1]\vec{x}(t-1) + \cdots \\ & \quad +[A_p + u(t-p)B_p]\vec{x}(t-p) + v(t-1)\vec{c} + \vec{\varepsilon}(t) \\ \vec{y}(t) &= \tilde{g}(\vec{x}(t)) + \vec{\varepsilon^*}(t) \end{cases} \tag{5}$$

where $u(t)$ and $v(t)$ are deterministic inputs, $\vec{\varepsilon}(t)$ and $\vec{\varepsilon^*}(t)$ are intrinsic and observation noise and are mutually independent. When $p = 1$ this is the DCM, and when $g(x) = x$ it is the GCM.

Now, if we can recover the state variables from the observation variables, all the problems can be considered in the frame of traditional Granger causality. The only difference is that here we have a deterministic input which is assumed to be known and will affect the connection of the state variables as well as the state of the variables directly.

3.2 UCM Algorithm

For the unified causal model (5), we now introduce an algorithm to estimate the state variables as well as all its parameters which will give us the first inspiration of the connection of the state variables.

Let $\vec{X}(t) = \begin{bmatrix} \vec{x}(t) \\ \vdots \\ \vec{x}(t-p+1) \end{bmatrix}, \vec{Y}(t) = \begin{bmatrix} \vec{y}(t) \\ \vdots \\ \vec{y}(t-p+1) \end{bmatrix}, \vec{U}(t) = \begin{bmatrix} u(t) \\ \vdots \\ u(t-p+1) \end{bmatrix}$

Then, the VAR(p) model can be reduced to a VAR(1) model which takes the form:

$$\begin{aligned} \vec{X}(t+1) &= \begin{bmatrix} A_1 + u(t)B_1 & A_2 + u(t-1)B_2 & \dots & A_p + u(t-p+1)B_p \\ I & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & I & 0 \end{bmatrix} \vec{X}(t) \\ &\quad + v(t) \begin{bmatrix} \vec{c} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \vec{w}(t) \\ &= A(\vec{\theta}, \vec{U}(t))\vec{X}(t) + C(\vec{\theta})v(t) + \vec{w}(t) = \vec{f}(\vec{X}(t), \vec{U}(t), v(t), \vec{\theta}) + \vec{w}(t) \\ \vec{Y}(t) &= \vec{h}(\vec{X}(t)) + \vec{s}(t) \end{aligned}$$

where $\vec{\theta}$ is the parameter vector to be estimated. $\vec{w}(t)$ and $\vec{s}(t)$ are both zero-mean uncorrelated Gaussian noise with covariance matrix $Q(t)$ and $R(t)$ respectively.

In order to apply the model to real data, we have to estimate both the states and parameters of the model from input variables and noise observations. A widely used method for this dual estimation is extended Kalman filter (EKF)[4,5]. Here we recursively approx-

imate the nonlinear system by a linear model and use the traditional Kalman filter for the linearized model.

Let $\vec{\xi} = \left[\vec{X}^T, \vec{\theta}^T\right]^T$ then

$$\begin{aligned}
\vec{\xi}(t+1) &= \begin{bmatrix} \vec{X}(t+1) \\ \vec{\theta}(t+1) \end{bmatrix} = \begin{bmatrix} \vec{f}(\vec{X}(t), \vec{U}(t), v(t), \vec{\theta}) \\ \vec{\theta}(t) \end{bmatrix} + \begin{bmatrix} \vec{w}(t) \\ \vec{\eta}(t) \end{bmatrix} \\
&= \begin{bmatrix} A(\vec{\theta}, \vec{U}(t))\vec{X}(t) + C(\vec{\theta})v(t) \\ \vec{\theta}(t) \end{bmatrix} + \vec{\zeta}(t) \\
&= \vec{g}(\vec{\xi}(t), \vec{U}(t), v(t)) + \vec{\zeta}(t)
\end{aligned}$$

where $\vec{\eta}(t)$ is uncorrelated Gaussian noise with covariance matrix $Z(t)$. Define

$$\begin{aligned}
\hat{\xi}_{t|t} &= E[\xi(t)|\vec{Y}(t), \vec{U}(t), v(t)] \\
\hat{\xi}_{t+1|t} &= E[\xi(t+1)|\vec{Y}(t), \vec{U}(t), v(t)] \\
\Omega_{t|t} &= E[(\xi(t) - \hat{\xi}_{t|t})(\xi(t) - \hat{\xi}_{t|t})^T|\vec{Y}(t), \vec{U}(t), v(t)] \\
\Omega_{t+1|t} &= E[(\xi(t+1) - \hat{\xi}_{t+1|t})(\xi(t+1) - \hat{\xi}_{t+1|t})^T|\vec{Y}(t), \vec{U}(t), v(t)]
\end{aligned}$$

where $\hat{\xi}_{t|t} = \begin{bmatrix} \hat{X}_{t|t} \\ \hat{\theta}_{t|t} \end{bmatrix}$, $\hat{\xi}_{t+1|t} = \begin{bmatrix} \hat{X}_{t+1|t} \\ \hat{\theta}_{t+1|t} \end{bmatrix}$. Then, the EKF algorithm for dual estimation consists of two steps: prediction and updating.

Prediction: Given the estimated state $\hat{\xi}_{t|t}$, the observation $\vec{Y}(t)$ and input $\vec{U}(t)$ and $v(t)$, we predict the state variables and the covariance matrix of prediction error of the system at time $t+1$.

$$
\begin{aligned}
\hat{\xi}_{t+1|t} &= E[\xi(t+1)|\vec{Y}(t),\vec{U}(t),v(t)] = E[\vec{g}(\xi(t),\vec{U}(t),v(t)) + \vec{\zeta}(t)|\vec{Y}(t),\vec{U}(t),v(t)] \\
&\approx E[(\vec{g}(\hat{\xi}_{t|t},\vec{U}(t),v(t)) + \tfrac{\partial \vec{g}}{\partial \xi^T}(\xi(t)-\hat{\xi}_{t|t})|\vec{Y}(t),\vec{U}(t),v(t)] \\
&= \vec{g}(\hat{\xi}_{t|t},\vec{U}(t),v(t)) = \begin{bmatrix} A(\hat{\theta}_{t|t},\vec{U}(t)) & 0 \\ 0 & I \end{bmatrix} \hat{\xi}_{t|t} + \begin{bmatrix} C(\hat{\theta}_{t|t}) \\ 0 \end{bmatrix} v(t) \\
\Omega_{t+1|t} &= E[(\xi(t+1)-\hat{\xi}_{t+1|t})(\xi(t+1)-\hat{\xi}_{t+1|t})^T|\vec{Y}(t),\vec{U}(t),v(t)] \\
&= E[(\vec{g}(\xi(t),\vec{U}(t),v(t)) + \vec{\zeta}(t) - \hat{\xi}_{t+1|t})(\vec{g}(\xi(t),\vec{U}(t),v(t)) \\
&\qquad +\vec{\zeta}(t) - \hat{\xi}_{t+1|t})^T|\vec{Y}(t),\vec{U}(t),v(t)] \\
&\approx E[(\tfrac{\partial \vec{g}}{\partial \xi^T}(\xi(t)-\hat{\xi}_{t|t}) + \zeta(\vec{t}))(\tfrac{\partial \vec{g}}{\partial \xi^T}(\xi(t)-\hat{\xi}_{t|t}) + \zeta(\vec{t}))^T|\vec{Y}(t),\vec{U}(t),v(t)] \\
&= F_t \Omega_{t|t} F_t^T + \Psi(t)
\end{aligned}
$$

where $\Psi(t) = \begin{bmatrix} Q(t) & 0 \\ 0 & Z(t) \end{bmatrix}$ and

$$
F_t = \begin{bmatrix} \frac{\partial f}{\partial X^T} & \frac{\partial f}{\partial \theta^T} \\ 0 & I \end{bmatrix} = \begin{bmatrix} A(\vec{\theta},\vec{U}(t)) & \frac{\partial}{\partial \theta^T}[A(\vec{\theta},\vec{U}(t))\vec{X} + C(\vec{\theta})v(t)] \\ 0 & I \end{bmatrix}_{\theta=\hat{\theta}_{t|t}, X=\hat{X}_{t|t}}
$$

Updating: We use the new observation $\vec{Y}(t+1)$ at time $t+1$ to update the state of the system.

$$
\begin{aligned}
\hat{\xi}_{t+1|t+1} &= \hat{\xi}_{t+1|t} + G(t+1)[\vec{Y}(t+1) - \vec{h}(\hat{X}_{t+1|t})] \\
\Omega_{t+1|t+1} &= [I - G(t+1)H(t+1)]\Omega_{t+1|t}
\end{aligned}
$$

where

$$
\begin{aligned}
G(t+1) &= \Omega_{t+1|t} H^T(t+1)[H(t+1)\Omega_{t+1|t}H^T(t+1) + R(t+1)]^{-1} \\
H(t+1) &= \begin{bmatrix} \frac{\partial h}{\partial X^T} & 0 \end{bmatrix}_{\hat{X}_{t+1|t}}
\end{aligned}
$$

3.3 UCM Definition of Causality

After recovering the state variables using the UCM algorithm as above, we can define the causality with the idea proposed by Granger. The difference is that, in our UCM model, two deterministic inputs $u(t)$ and $v(t)$ are added to the normal autoregressive representation. Here, we provide the formulation of UCM causality in both time domain and frequency domain.

3.3.1 Causality in the Time Domain

For simplicity of notation, here we only formulate UCM for two time series X_t and Y_t. To generalize them to more general case of multi time series, we refer the reader to [11]. Assume that X_t and Y_t in our UCM have the following representation:

$$\begin{aligned} X_t &= \sum_{j=1}^{p}[a_{1j} + u(t-j)b_{1j}]X_{t-j} + C_{1x}v(t-1) + \epsilon_{1t} \\ Y_t &= \sum_{j=1}^{p}[d_{1j} + u(t-j)e_{1j}]Y_{t-j} + C_{1y}v(t-1) + \epsilon_{2t} \end{aligned} \tag{6}$$

A joint representation in our UCM that includes the past information of both processes X_t and Y_t can be written as:

$$\begin{aligned} X_t &= \sum_{j=1}^{p}[a_{2j} + u(t-j)b_{2j}]X_{t-j} + \sum_{j=1}^{p}[d_{2j} + u(t-j)e_{2j}]Y_{t-j} + C_{2x}v(t-1) + \epsilon_{3t} \\ Y_t &= \sum_{j=1}^{p}[f_{2j} + u(t-j)g_{2j}]X_{t-j} + \sum_{j=1}^{p}[h_{2j} + u(t-j)k_{2j}]Y_{t-j} + C_{2y}v(t-1) + \epsilon_{4t} \end{aligned} \tag{7}$$

where p is the maximum number of lagged observations included in the model. ϵ_{it}, $i = 1, 2, 3, 4$ are prediction errors with variance Σ_i and are uncorrelated over time. Then, according to the causality definition of Granger, if the prediction of one process can be im-

proved by incorporating the past information of the second process, then the second process causes the first process. So, in the unified model here, we define that if the variance of prediction error for the process X_t is reduced by the inclusion of the past information of the process Y_t, then, a causal relation from Y_t to X_t exists. This can be quantified as

$$F_{Y\to X} = \ln \frac{\Sigma_1}{\Sigma_3}$$

If $F_{Y\to X} = 0$, there is no causal influence from Y_t to X_t and if $F_{Y\to X} > 0$, there is. Similarly, we can define the causal influence from X_t to Y_t as

$$F_{X\to Y} = \ln \frac{\Sigma_2}{\Sigma_4}$$

3.3.2 Causality in the Frequency Domain

Our UCM also allows a frequency domain decomposition to detect the intrinsic causal influence which provides valuable information.

We define the lag operator L to be $LX_t = X_{t-1}$ and assume here that the input $u(t)$ is a constant, i.e. $u(t) \equiv u$. Then, the joint representation of both processes X_t and Y_t in equation (5) can be expressed as:

$$\begin{aligned} X_t &= \textstyle\sum_{j=1}^{p}[a_{2j} + ub_{2j}]X_{t-j} + \sum_{j=1}^{p}[d_{2j} + ue_{2j}]Y_{t-j} + C_{2x}v(t-1) + \epsilon_{3t} \\ &= \textstyle\sum_{j=1}^{p}\tilde{a}_{2j}X_{t-j} + \sum_{j=1}^{p}\tilde{b}_{2j}Y_{t-j} + C_{2x}v(t-1) + \epsilon_{3t} \\ Y_t &= \textstyle\sum_{j=1}^{p}[f_{2j} + ug_{2j}]X_{t-j} + \sum_{j=1}^{p}[h_{2j} + uk_{2j}]Y_{t-j} + C_{2y}v(t-1) + \epsilon_{4t} \\ &= \textstyle\sum_{j=1}^{p}\tilde{c}_{2j}X_{t-j} + \sum_{j=1}^{p}\tilde{d}_{2j}Y_{t-j} + C_{2y}v(t-1) + \epsilon_{4t} \end{aligned} \tag{8}$$

Rewrite equation (7) in terms of lag operator, we have:

$$\begin{bmatrix} \tilde{a}_2(L) & \tilde{b}_2(L) \\ \tilde{c}_2(L) & \tilde{d}_2(L) \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \epsilon_{3t} \\ \epsilon_{4t} \end{bmatrix} + v(t-1) \begin{bmatrix} C_{2x} \\ C_{2y} \end{bmatrix} \tag{9}$$

where $\tilde{a}_2(0) = 1$, $\tilde{b}_2(0) = 0$, $\tilde{c}_2(0) = 0$, $\tilde{d}_2(0) = 1$.

Since what we really care about is the causal relationship caused by the intrinsic connection of the state variables rather than the outside driving force, i.e. the input $v(t)$, after fitting the model (8) and getting the covariance matrix of the prediction error, we just go on with the following model:

$$\begin{bmatrix} \tilde{a}_2(L) & \tilde{b}_2(L) \\ \tilde{c}_2(L) & \tilde{d}_2(L) \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \epsilon_{3t} \\ \epsilon_{4t} \end{bmatrix} \tag{10}$$

which means that after fitting the UCM with input $v(t)$ to eliminate outside influence, we just focus on the intrinsic causal influence in the frequency domain.

After normalizing equation (9) using the transformation proposed by Geweke [16, 17] to eliminate the cross term in the spectra, we assume that we have the normalized equation in the form:

$$\begin{bmatrix} \bar{a}_2(L) & \bar{b}_2(L) \\ \bar{c}_2(L) & \bar{d}_2(L) \end{bmatrix} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} \bar{\epsilon_{3t}} \\ \bar{\epsilon_{4t}} \end{bmatrix} \tag{11}$$

Fourier transforming both sides of equation (10) leads to

$$\begin{bmatrix} \bar{a}_2(\omega) & \bar{b}_2(\omega) \\ \bar{c}_2(\omega) & \bar{d}_2(\omega) \end{bmatrix} \begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} \bar{E}_x(\omega) \\ \bar{E}_y(\omega) \end{bmatrix} \tag{12}$$

Recasting equation (11) into the transfer function format we obtain

$$\begin{bmatrix} X(\omega) \\ Y(\omega) \end{bmatrix} = \begin{bmatrix} H_{xx}(\omega) & H_{xy}(\omega) \\ H_{yx}(\omega) & H_{yy}(\omega) \end{bmatrix} \begin{bmatrix} \bar{E}_x(\omega) \\ \bar{E}_y(\omega) \end{bmatrix} \tag{13}$$

After proper ensemble averaging we have the spectral matrix

$$S(\omega) = H(\omega)\Sigma H^*(\omega) = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{bmatrix}$$

where * denotes the complex conjugate and matrix transpose and $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$ is the covariance matrix of the prediction errors in equation (10). Hence, we can define the causal influence from Y_t to X_t at frequency ω as

$$f_{Y \to X}(\omega) = \ln \frac{S_{xx}(\omega)}{H_{xx}(\omega)\Sigma_{xx}H^*_{xx}(\omega)}$$

Similarly, we can define the causal influence from X_t to Y_t at frequency ω as well.

Note that although here we just provide the definition of pairwise Granger causality for UCM, it's obvious that similar methods can be easily applied to the definition of conditional, partial or complex Granger causality in both time and frequency domain[7, 9, 11, 13, 15]. Since explicit meaning of the parameters in the UCM (i.e. the intrinsic coupling among state variables, the strength of the inputs to modulate the coupling and the influence of the inputs on the state variables directly), we can also get an idea of the connection of the state variables and how the inputs affect them from the fitted model before we translate it into a single number.

3.4 Experiment

Animals and visual discrimination training Three female sheep were used (Ovis aries, one Clun Forest and two Dorsets). All experiments were performed in strict accordance with the UK 1986 Animals Scientific Procedures Act. During the experiments the animals were housed inside in individual pens. They were trained initially over several months to perform operant-based face (sheep) or non-face (objects) discrimination tasks with the animal making a choice between two simultaneously presented pictures one of which was

associated with a food reward. During stimulus presentations animals stood in a holding trolley and indicated their choice of picture by pressing one of two touch panels located in the front of the trolley. The food reward was delivered automatically to a hopper between the two panels. The life-sized pictures were back projected onto a screen 0.5m in front of the animal using a computer data projector. A white fixation spot was presented constantly in between trials to maintain attention and experimenters waited until the animals viewed this spot before triggering presentation of the image pairs. The stimulus images remained in view until the animal made an operant response (generally around 1-3 s). In each case for successful learning of a face or object pair required that a performance criterion of $>80\%$ correct choice over 40 trials (i.e. 40 pairs) was achieved consistently. By the end of training animals were normally able to reach the $> 80\%$ correct criterion after 40-80 trials and maintain this performance. Some previously learned (for periods ranging from 10 days to 9 months) stimulus pairs were then presented during subsequent electrophysiological recording experiments although the animals were mainly presented with novel stimulus pairs as well so that effects of learning on neurophysiological parameters could be assessed. For each sheep successful recordings were made in response to up to 11 different face or non face object pairs (Sheep A: 6 novel face and 1 novel object pair; B 7 novel face pairs, 3 previously learned face pairs and one previously learned object pair; Sheep C: 2 novel face pairs and 3 previously learned face pairs. Learning effects were monitored over between 80-189 trials and data was collected over blocks of 20-40 trials. For the face pairs Sheep A and B were discriminating between the faces of different familiar or unfamiliar individuals (face identity discrimination) whereas for Sheep C discrimination was between calm and stressed face expressions in the same animal (n=3 pairs) or in different animals (n=2 pairs).

With this animal the calm face was the rewarded stimulus. Where novel face or object pairs were being learned during recordings then where the $> 80\%$ performance criterion was achieved this was after between 20 and 80 training trials. Examples of the face and object pairs used are shown in Fig S1.

Electrophysiological recordings and analysis. Following initial behavioral training sheep were implanted under general anesthesia (fluothane) and full aseptic conditions with either unilateral (one animal) or bilateral planar 64-electrode (for configuration see Fig 4A) arrays (epoxylite coated, etched, tungsten wires with 250 μm spacing - total array area 2mm x 2mm, electrode impedence 0.2μm) aimed at the IT. Holes (0.5cm diameter) were trephined in the skull and the dura beneath cut and reflected. The electrode bundles were introduced to a depth of 20-22 mm from the brain surface using a stereotaxic micromanipulator and fixed in place with dental acrylic and stainless-steel screws attached to the skull. Two of these screws acted as reference electrodes, one for each array. Electrode depths and placements were calculated with reference to X-rays, as previously described (S1). Electrodes were connected to 34 pin female plugs (2 per array) which were cemented in place on top of the skull (using dental acrylic). Starting 3 weeks later the electrodes were connected via male plugs and ribbon cables to a 128 channel electrophysiological recording system (Cerebus 128 Data Acquisition System - Cyberkinetics Neurotechnology Systems Inc., USA) and recordings made during performance of the different face and non-face pair operant discrimination tasks. This system allowed simultaneous recordings of both neuronal spike and local event-related (LFP) activity from each electrode. Typically, individual recording sessions lasted around 30 min and for 80-200 individual trials. There was at least a week between individual recording sessions in each animal.

The LFPs were sampled at 2kHz and MUA spikes at 30kHz and digitized for storage from 3 seconds prior to the stimulus onset to 3 seconds after the stimulus onset (stimulus durations were generally 1-3 s). Neural recordings from our data collection systems consist of two raw large files of LFP and MUA. We used customized written Spike 2 (Cambridge Electronic Design, Cambridge, UK) scripts to translate them into common text files arranged by trial or channel for convenience in further analysis.

For data analysis the stored signals LFPs or MUA data contaminated with noise such as from animal chewing food were excluded as were LFPs with unexpectedly high power. For LFPs, offline filtering was applied in the range of 1-200 Hz and trend was removed before spectral analysis. Any trials having more than 5 points outside the mean 5 standard deviation range were discarded before the analysis. Blocks of trials where no visual evoked potential could be discriminated (only occurred in Sheep C) were also excluded. The mean and the standard deviation values of different parameters were calculated across all trials for 500 ms before and 500 ms during the presentation of the visual discrimination pairs (i.e. prior to the performance of any operant response). The LFPs and MUA responses were all aligned to the onset of the visual stimuli. All analysis were carried out using custom written routines in Matlab (The Mathworks Inc, Natick, MA).

At the end of the experiments animals were enthanized with an intravenous injection of sodium pentobarbitone and the brains removed for subsequent histological confirmation of X-rays that array placements were within the IT cortex region.

Analysis of visual evoked potential (VEP) and MUA The VEP was extracted from the LFPs by trial-averaging after aligning the data to the onset of stimulus. Two major peaks were identified from the VEP in the initial 500ms of stimulus presentation: positive peak at

100ms (P100) and a negative peak at 300ms (N300). We calculated the latency for these two peaks by finding the time corresponding to the maximum and minimum peak value respectively. The amplitudes of these two peaks were calculated as their peak values after subtracting the average baseline in the 100ms before stimulus onset.

4 Discussion

Advantage of unified approach. Different from all previous methods in estimating Granger causality in the literature where essentially a regression method is employed, in UCM we incorporate noise observation variables and apply the extended Kalman filter to recover the state variables. Additional inputs are also included in UCM on the basis of a autoregressive model. The advantage of such an approach over the previous methods is obvious. The UCM is more reasonable when we are faced with experimental data affected by some stimulus and applicable to the case that we can not track the state variables respectively but just a function of them or the observation noise is considerable. Furthermore, all the previous methods in estimating Granger causality are batch learning: before estimation, it requires to collect all data. The extended Kalman filter, on the other hand, is an on-line learing: we can now update Granger causality instaneously. One may argue that this is a common feature of online learning vs. batch learning. However, it is novel in the context of Granger causality. When we face biological data, this feature becomes more significant. As we know, adaption, or learning in animal in particular, is very unique and it makes it difficult to analyze as well since adaption introduces nonstationary into the system. The classical way of estimating Granger causality can cope with this difficulty by introducing sliding

windows in analyzing data. Of course, to select an optimal window size is always an issue in such an approach. However, in Kalman filter, we can get the inspiration of the connection of the state variables from the connection matrix and such an issue is automatically solved.

In comparison with DCM, the advantage of UCM are the following. First, it naturally allows time delay in the model. Time delay is ubiquitous in a biological system, no matter it is in gene circuits, protein circuits, metabolic circuits, neuronal networks and brain areas. Secondly, using Granger causality, we are able to summarize the causal effect into a single number which is more transparent and easy to understand, in particular in a system with a time delay. Thirdly, it allows a frequency domain decomposition. As we all know, when we deal with a dynamical system, it is sometime much easier to view it in the frequency domain rather than in the time domain.

Other types of data. In the current paper, we have only applied UCM to LFP. It is out of the question that it can be applied to other types of data. For example, in microarray data, we are routinely collected several thousands genes in many hours. With the dynamical microarry data, we are in the position to ask some interesting biology questions such as the senescence of a plant leave, the response of the a cell to bioinfectious etc. In protein dynamics, the dynamical data of a few hundred proteins are available. The data allows us to assess how several proteins form a protein complex to function. Hence although we have restricted ourselves to LFP, the applicability of the UCM is certainly beyond what we reported here.

Learning effect. **This is what I want to do here actually**. Biological phenomena occur at different time scales. For example, when an animal respond to a dangerous stimulus

Commonalities and Differences between DCM, GCM and UCM	DCM	GCM	UCM
commonalities			
Multivariate analysis of time-series data	Yes	Yes	Yes
Models directed coupling	Yes	Yes	Yes
Inference on models	Yes	Yes	Yes
Differences			
Causality based on temporal precedence	No	Yes	GCM is more general
Causality based on control theory	Yes	No	Yes
Requires known inputs	Yes	No	In general yes, but see example 2
Requires orthogonal innovations	No	Yes	Not necessary
Requires stationary processes	No	Yes	Could use sliding window
Requires a specific biophysical model	Yes	No	Yes
Models non-linear coupling	Yes	No	Yes
Inference on model parameters	Yes	No	Yes
Frequency decomposition	No	Yes	Yes

Figure 5:

(a prey), it requires it to react instantaneously (within a few hundred msecs). However, learning should be at a longer time scale: taking hours or days to change, although SDTP is within 40 msecs. Here we have carried a detailed analysis on how the

Nonlinear. One might suggest that in UCM, the interaction between each unit is linear, which is certainly an over simplification of real situations. Nevertheless, as we have demonstrated before, we can easily include nonlinearity in our analysis.

$$\begin{cases} \vec{x}(t) &= [A_1 + u(t-1)B_1]K_1(\vec{x}(t-1)) + \cdots \\ & \quad +[A_p + u(t-p)B_p]K_P(\vec{x}(t-p)) + v(t-1)\vec{c} + \vec{\varepsilon}(t) \\ \vec{y}(t) &= \tilde{g}(\vec{x}(t)) + \vec{\varepsilon^*}(t) \end{cases} \tag{14}$$

where $K_i, i = 1, \cdots, p$ are kernal functions and the others are the same as in Eq. (5).

Why frequency

Acknowledgment. J.F. was partially supported by grants from UK EPSRC CARMAN and an EU grant BION.

References

[1] Friston K(2009) Causal modelling and brain connectivity in functional magnetic resonance imaging . Plos Biol 7(2): e1000033.doi:10.1371/journal.pbio.1000033

[2] David O, Guillemain I, Saillet S, Reyt S, Deransart C, et al. (2008) Identifying neural drivers with functional MRI: An electrophysiological validation. Plos Biol 6(12): e315. doi:10.1371/journal.pbio.0060315

[3] Friston KJ, Harrison L, Penny W(2003) Dynamic causal modelling. Neuroimage 19: 1273-1302

[4] Sun X, Jin L, Xiong M (2008) Extended Kalman Filter for Estimation of Parameters in Nonlinear State-Space Models of Biochemical Networks. PLoS ONE3(11): e3758. doi:10.1371/journal.pone.0003758

[5] Li P, Goodall R, Kadirkamanathan V (2004) Estimation of parameters in a linear state space model using a Rao-Blackwellised particle filter. IEE Proc Control Theory Appl 151: 727C738

[6] Boris Gourvitch, Rgine Le Bouquin-Jeanns, Grard Faucon (2006) Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. Biol Cybern 95:349C369 doi: 10.1007/s00422-006-0098-0

[7] Jianhua Wu, James L.Sinfield, Jianfeng Feng. Analysis of Biological Rhythms Using a Harmonic Causal Method. (Submitted)

[8] Nelson TA (2000) nonlinear estimation and modeling of noisy time-series by dual Kalman filter methods. PhD thesis, Oregon Graduate Institute of Science and Technology.

[9] Zou CL, Feng J.F. (2009) Granger causality vs. Dynamic Bayesian network inference: A Comparative Study *BMC Bioinformatics* 10:122 doi:10.1186/1471-2105-10-122.

[10] Feng J.F. Yi DY, Krishna R, Guo SX, Buchanan-Wollaston V.(2009) Listen to genes: dealing with microarray data in the frequency domain PLoS One 4(4): e5098. doi:10.1371/journal.pone.0005098.

[11] Guo SX, Wu J.F.H., Ding MZ, Feng J.F.(2008) Uncovering interactions in the frequence domain PLoS Comput Biol 4(5): e1000087. doi:10.1371/journal.pcbi.1000087

[12] ES Nikitin, DV Vavoulis,Feng J.F.,M OShea, PR Benjamin and G Kemenes(2008) Persistent sodium current is a non-synaptic substrate for long-term memory Current Biology vol. 18: 1221-1226

[13] Guo SX, A. Seth, K. Kendrick, Feng J.F. (2008) Partial Granger causality: eliminating exogenous inputs and latent variables, J.F. Neurosci. Methods , vol. 172: 79-83 (IF=1.8)

[14] Wu J.H., Liu XG., Feng J.F. (2007) Detecting causality between different frequencies, J. Neuroscience Methods 167:367-375.

[15] Chen YH, Rangarajan G., Feng J.F. and Ding M.Z. Analyzing multiple nonlinear time series with extended Granger causality, Phys. Lett. A, vol. 324, 26-35.

[16] Geweke J. Measurement of Linear Dependence and Feedback Between Multiple Time Series. Journal of the American Statistical Association, 1982;77: 304-13.

[17] Geweke J. Measures of Conditional Linear Dependence and Feedback Between Time Series. Journal of the American Statistical Association, 1984;79: 907-15.

[18] Christophe Ladroue, Jianfeng Feng. Beyond element-wise interactions: Defining group-to-group interactions for biological processes. (Submitted)

[19] Handbook of Time Series Analysis. Björn Schelter, Matthias Winterhalder, Jens Timmer. 2006 WILEY-VCH Verlag GmbH & Co.KGaA, Weinheim. ISBN:3-527-40632-9. Chapter 17. Granger Causality: Basic Theory and Application to Neuroscience. Mingzhou Ding, Yonghong Chen and Steven L. Bressler

[20] Simpson, Edward H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Ser. B* **13**: 238C241.

[21] J. Pearl (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press: Cambridge, UK.

[22] Camacho DM and Collins JJ (2009). Systems biology strikes gold. *Cell* **137**: 24-26.

[23] Cantone I, Marucci L, Iorio F, Ricci M, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D, and Cosma MP (2009), A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches *Cell*, **137**: 172-181.

[24] Kendrick KM, Y Zhan, H Fischer, A U Nicol, XJ Zhang, Feng J.F. (2009) Learning alters theta-nested gamma oscillations in inferotemporal cortex *Nature Precedings* hdl: 10101/ npre. 2009.3151.1.