A Generalized and Drifting Time Corrected Approach using Wiener-Granger Causality and VAR(p) Process for Detecting Directed Functional Communication between Brain Regions and Behavior Prediction

Oliver Y. Chén

Abstract

Brain talks. It communicates with innner and outer world by means of perception, learning and memory, decision making, sensory and motor control, etc. Here we demonstrate that there are directional information flows of various magnitudes between regions within the brain. Electroencephalography (EEG) records neuronal electricity activity at millisecond-range over a period of time; and functional magnetic resonance imaging (fMRI) measures changes in blood flow and oxygenation in the whole brain in response to neuronal activation. We show that the neural communication between different regions is either unidirectional or bidirectional; and each information flow is of different magnitudes, by using EEG and fMRI data. The major paths of information flows between different subjects overlap overall, during both resting and task states, with individual-specific paths identified. Furthermore, based upon the information flow paths, neuronal activity observed in one region can predict neural patterns of another region; and can together predict behavior. The study provides anatomical and functional connectivity insights between brain regions and can assist diagnosis and treatment of brain diseases caused by pathway obstruction.

Keywords behavior prediction, brain connectivity, VAR, Granger causality, EEG, fMRI

1 Introduction

Brain talks. It talks in a way that it is able to perceive information from the outer world, process and store it, and use it for prediction, task and decision making. There has been a tremendous amount of effort to understand the means by which it communicates within itself, from anotomical, clinical, cognitive, and behavioral approaches, to imaging. Despite the pervasive research in anatomical, functional, and effective connnectivity to understand the information exchange between brain regions and its effect on behavior, it remains open to address the following three questions together and compactly. While given a treatment, how does the treatment affects brain activities (information) of different regions and consequently how does information flows from one part of the brain to another (who talks first); how strong is each information flow (who talks more); and how do these information flows predict behavior (what does the conversation say)?

In general, the study of brain connectivity can be mainly divided into the following categories: experimentally, resting state and task state; connectivity types: anatomical, functional, and effective; methodologically: correlation based, synchrony based, and dynamic system (e.g. PDE) based; data modalities: fMRA EEG, MEG, ECoG, etc.; and to study the brain influence on behavior, classification and prediction.

A popular approach to investigate the functional brain connectivity is via the correlation approach. In the resting state, it delineates the potential resting functional network; and across tasks, if the absolute magnitudes or signs of correlations change, there is a potential quantitive or qualatitive functional influence between regions. Functional connectivity, however, does not rest on any statistical dependencies models among observed responses and therefore does not provide inference about the coupling between two regions, i.e., effective connectivity (Friston (1994), Friston (2011)). There are works using functional connectivity for subject classification (e.g. Craddock:2009). These approaches, increasingly appreciated, however, do not provide

hypothesis of differences in coupling because they are based upon a mapping from imaging data (consequences) to the categorical cause (Friston, 2011). A more proper way for classification using imaging data is via the effective connectivity analysis, where models of coupling among hidden brain states are considered (Friston, 2011). An important and interesting area of studying effective connectivity is the studying of the directionality and magnitude of information flows between brain regions. There are a few approaches. For example, lag based approach (e.g. Granger causality, Granger (1969)), conditional independence (e.g. Bayesian network, Jensen (1996), Nielsen and Jensen (2009), Charniak (1991)), and higher order statistics (e.g. pairwise conditional probability, Patel et al. (2006)). There are studies particularly studying the directed information exchange between brain regions (see Hinrichs et al (2008), Liu and Aviyente (2009), Liu and Aviyente (2012) Chen et al (2014)). The dynamic causal models (DCMs, Mgenerative et al) and its extensions (e.g. generative embedding, Brodersen, which combines DCMs and support vector machines (SVMs)) allows one to study the biophysical dynamic system at the neural level, and classify patients by using fMRI data; however, it has a few complicatons: the sophysticated model buiding makes parameter estimation relatively complicated; there could be identifiable issues (which can be alleviated by including prior information; but strong priors would increase bias and decreas reliability); and it requires knowledge of physiological quantities such as synaptic connection strengths, which are not always available. The dynamic directional model (DDM, Zhang et al. (2015)) alleviates the above issues with DCMs and uses intracranial electrocorticographic (ECoG) time series to study brain effective connectivity. Their work is exceptional, but their main focuse is on the dynamic connectivity between brain regions and is not on behavior prediction; and due to the limitation of the ECoG data (e.g. subject dependent and varying electrode placement locations), only selected regions can be studied.

Here we introduce a new approach that adapts the vector autoregressive (VAR) models,

granger causality models (GCMs), and structural equation models (SEMs) into a more rigorous framwork to explore the directed influence between brain regions and behavior prediction, with a potential causal interpretation. The VAR models under the granger framework enables us to study the temporal relations between recorded neural time series, and relate the neural activity to behavior, with minimal physiological information; the readily available electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance (fMRI) data provide extensive potential to investigate information flows between brain regions noninvasively; the model accounts for the variability by means of stochastic processes; and the framwork allows behavior prediction using neural activity measurements.

Vector autoregressive (VAR) models have been pervasive in economics. See (Sims (1980), Lütkepohl (1991), Watson (1994), Lütkepohl (1999), Waggoner and Zha (1999)); further see Hamilton (1994), Campbell, Lo and MacKinlay (1997), Cuthbertson (1996), Mills (1999) and Tsay (2001) for VAR models in application to financial data. Recently, there has been a growing popularity of implementing VAR models in the brain activity data to study the temporal neural activity and its interaction between different brain regions; and along with the Granger-causality models, in studying the directed influcence between different brain regions.

Since its formal introduction in 1969, Granger model (Wiener (1956), Granger (1969)) has been extensively used in modeling economitric data; and more recently, in modeling neural data (e.g., for fMRI data, see Roebroeck et al. (2005), (Bressler et al., 2008), Sridharan et al. (2008), Zhou et al. (2009), Barrett et al. (2010); for EEG and MEG data, see Astolfi et al. (2007), Gow et al. (2008), Brovelli et al. (2004), Bressler et al. (2007); and see Bressler and Seth (2011) for an overview). Granger causation uses one time series to forcast another. We say a random variable X "granger causes" another variable Y, if Y is more predictive using information of its own past and the past of X, than using only the past information of Y. Recent years have seen a great increase of using Granger models in the field of neuroscience. The model is used

to investigate the connectivity and information flow between different regions, their effect size, and how they change over time. Nevertheless, the question of true "causality" is philosophical, and because of the *post hoc ergo propter hoc* fallacy, econometricians and neuroscientists assert that the Granger test finds only "predictive causality". While we do not claim that intensive neural activity preceding another can be used as a proof of causation, we claim that it could be used as a useful indication of potential causality, and shall discuss this in the discussion section.

It is of neurobiological interest and importance to noninvasively detect the directed influence between brain regions, quantify the magnitude of the directing influence, and understand how the directed influence between task and resting states, it between different levels of the same task, and it between different tasks, change. Finally, it is interesting to predict behavior based upon measured brain activity and the estimated underlying connectivity network.

The rest of the paper is organized as follows: In Section 2 we discuss the motivating thermal pain data set. In Section 3 we introduce the model for behavior prediciton using neural activity measurements. We also provide a brief overview of the VAR process and its estimation in Section 8. In Section 3.3 we provide a procedure to adjust time lags in fMRI data to make causal claims. In Section 4, we introduce an estimation procedure; in Section 4.5, we include a Wald test for performing inference and some asymptotic properties of the obtained estimates; and in Section 4.6, we propose a solution when the time series data is high-dimensional. In Sections 5 and 6, the efficacy of the approach is illustrated through simulations and an application to fMRI data. Finally, in Section 7, we conclude the paper with a discussion.

2 Data Structure

Consider any two brain regions M and N (potentially high dimensional) under a particular stimulus X. Let the number of voxels in each region be as V_m and V_n, respectively. From neurobiological point of view, within each region the neurons that are wired together, fired

together. For every voxel in the region, a time series can be measured at every $t \in \mathcal{T}$; then for the entire region, a vector of time series can be recorded.

Therefore, for brain regions M and N, the data matrices are $\mathbf{M}$ and $\mathbf{N}$, where $\mathbf{M}$ is a $V_m \times T$ and $\mathbf{N}$ is a $V_n \times T$, and each row represent a time series vector with length T.

Suppose the effect of treatment on the voxels in each region has time lag p. The process in each region can be delineated by two seperate vector autoregressive (VAR) processes of time lag p. Because a model with too small a p under-represent the data and fail to capture the spectral features that require higher order (Mitra and Pesaran, 1999) and a model with too large a p complicate parameter estimation, one need to select a suitable p to ensure a model that explains certain amount of variance and is reasonably easy to estimate. Common approaches include Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz et al., 1978) .

The questions of direct scientific interest are: (1) how the treatment affect these paired VAR processes; (2) how information flows between regions M and N: $M \to N$, $N \to M$, or $M \leftrightarrow N$; (3) how strong is each information flow; and (4) how VAR processes together affect the outcome Y.

3 The model

Assume the time series measured from each node is wide-sense stationary (WSS). Models with WSS ensure the estimates are tractible. To check WSS, one may verify that the autocorrelation function declines sharply (Hamilton, 1994) or via formal testing (e.g. the unit root test (Dickey and Fuller, 1979), (Phillips and Perron, 1988), the KPSS test (Kwiatkowski et al., 1992)). For nonstationary data, one may consider differencing the data, or use approaches that deal with time varying causal relationship (e.g., locally stationary approach by Ding et al. (2000), adaptive recursive least-square model by Hesse et al. (2003), and nonparemetric approach such as

spectral factorization by Dhamala et al. (2008)).

3.1 Temporal Treatment effect

Before the network is defined, we determine the magnitute and direction of information flow between each paired regions using the linear feedback from N to M ($F_{N\to M}$), from M to N ($F_{M\to N}$), and instantaneous linear feedback ($F_{M\cdot N}$) (Geweke, 1982). For a brief overview of the VAR process and the "Geweke test" for linear feedbacks, see Section 8.

Let us begin by assuming $F_{M\to N} \gg F_{N\to M}$. Then the information flow can be described in 1a.

Specifically,

$$\mathbf{m}(t) = \boldsymbol{\alpha}_1(t)T + \sum_{i=1}^{p} \mathbf{E}_1(i)\mathbf{m}(t-i) + \boldsymbol{\eta}_1(t) \quad Var\big(\boldsymbol{\eta}_1(t)\big) = \boldsymbol{\Sigma}_3; \tag{1}$$

$$\mathbf{n}(t) = \boldsymbol{\beta}_1(t)T + \sum_{i=1}^{p} \mathbf{F}_1(i)\mathbf{n}(t-i) + \sum_{i=1}^{p} \mathbf{G}_1(i)\mathbf{m}(t-i) + \boldsymbol{\epsilon}_1(t) \quad Var\big(\boldsymbol{\epsilon}_1(t)\big) = \mathbf{T}_3; \tag{2}$$

Assume there is a bidirectional information flow between M and N. Then the information flow can be described in 1b.

Specifically,

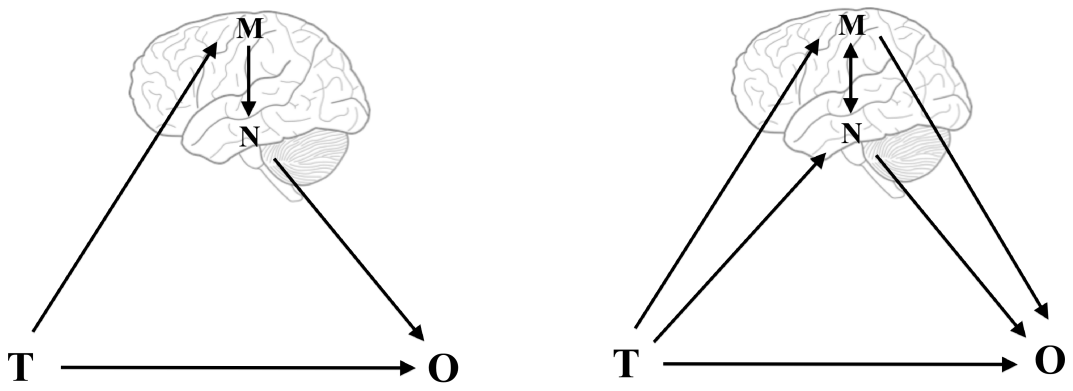
$$\begin{aligned} \mathbf{m}(t) &= \mathbf{a}_1(t)T + \sum_{i=1}^{p} \mathbf{E}_2(i)\mathbf{m}(t-i) + \sum_{i=1}^{p} \mathbf{E}_3(i)\mathbf{n}(t-i) + \boldsymbol{\eta}_2(t) \quad Var\big(\boldsymbol{\eta}_2(t)\big) = \boldsymbol{\Sigma}_4; \\ \mathbf{n}(t) &= \mathbf{b}_1 T + \sum_{i=1}^{p} \mathbf{F}_2(i)\mathbf{n}(t-i) + \sum_{i=1}^{p} \mathbf{G}_2(i)\mathbf{m}(t-i) + \boldsymbol{\epsilon}_2(t) \quad Var\big(\boldsymbol{\epsilon}_2(t)\big) = \mathbf{T}_4; \end{aligned} \tag{3}$$

$$\boldsymbol{\Gamma} = Var\begin{pmatrix} \boldsymbol{\eta}_2(t) \\ \boldsymbol{\epsilon}_2(t) \end{pmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_4 & \mathbf{C}_2 \\ \mathbf{C}_2^{\intercal} & \mathbf{T}_4 \end{bmatrix}.$$

3.2 Behavior Prediction

Prediction in unidirectional path

Let us assume $F_{M\to N} \gg F_{N\to M}$. Then the information flow can be described in 1a.



(a) Unidirectional information flow

(b) Bidirectional information flow

Figure 1: Graphs of information flows

Specifically, consider (1), (2), and

$$O = \gamma_0 + \gamma_1 T + \sum_{t \in \mathcal{T}} \left\{ \mathbf{H}_1(t)\mathbf{n}(t) \right\} + \boldsymbol{\xi}_1. \tag{4}$$

Note that in (3) we assume if the information flow between regions M and N is only unidirectional $M \to N$, then there is no effect from $m(t)$ on Y.

Prediction in bidirectional path Assume there is a bidirectional information flow between M and N. Then the information flow can be described in the following graph. Then the information flow can be described in 1b.

Specifically, consider (3) and

$$O = r_0 + r_1 T + \sum_{t \in \mathcal{T}} \left\{ \mathbf{H}_2(t)\mathbf{n}(t) + \mathbf{H}_3(t)\mathbf{m}(t) \right\} + \boldsymbol{\xi}_2. \tag{5}$$

The overal prediction of neural activities on Y can be written as:

$$O = R_0 + R_1 T + \sum_{t \in \mathcal{T}} \left\{ \sum_{\mathbf{n}} \mathbf{H}_2(t)\mathbf{n}(t) + \sum_{\mathbf{m}} \mathbf{H}_3(t)\mathbf{m}(t) + \right\} + \boldsymbol{\xi}_3. \tag{6}$$

3.3 Causal Inference and Granger Causality with fMRI Data

There are some problems relating granger causality models (see Friston (1994) for an overview). A major issue for making causal claims (predictive causality) using fMRI data is the time delays due to the indirect measurement of blood flow.

Let us begin by consider V voxels. Let $M_1(t), \cdots, M_V(t)$ denote the true unknown temporal neural activity intensity for each voxel. Let $\tilde{M}_1(t), \cdots, \tilde{M}_V(t)$ be the corresponding temporal fMRI measurements. For every voxel $v \in \{1, \cdots, V\}$,

$$\tilde{M}_v(s) = f_v\big(M_v(s + \delta_v), s\big) + \epsilon_v, \tag{7}$$

where δ_v is the time delay at voxel v between the true neural activity intensity and the measured fMRI intensity, f_v is some known function, and ϵ_v is measures the noise.

In reality, the time delay for each voxel is unknown and varies across the brain. To estimate it, we consider put a prior on it (for simplicity we further assume i.i.d.), such that:

$$\delta_v \overset{i.i.d.}{\sim} \mathcal{P}(\boldsymbol{\theta}).$$

A reasonable prior is Beta function, because δ_v is positive and bounded. To estimate $\boldsymbol{\theta}$, we introduce the following procedure.

First, for simplicity, assume f_v is linear, $\epsilon_v = 0$:

$$\tilde{M}_v(s) = M_v(s + \delta_v).$$

Denote $\mathcal{C} = \big\{(i,j) : M_i \to M_j\big\}$, the knowledge of all connectivity; and $\mathcal{C}^k = \big\{(i,j) : M_i \to M_j\big\}$, the knowledge of connectivity we have. $\mathcal{C}^k \subset \mathcal{C}$. Hence:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \mathbb{E}\Big\{ \sum_{(i,j)\in\mathcal{C}^k} \Big| \hat{F}_{M_i \to M_j}(\boldsymbol{\theta}) - F_{M_i \to M_j} \Big| \Big\} \right\}, \tag{8}$$

where $F_{M_i \to M_j}$ denotes the known directionality between M_i and M_j and $\hat{F}_{M_i \to M_j}(\boldsymbol{\theta})$ denotes the time adjusted directionality between M_i and M_j, given $\boldsymbol{\theta}$.

4 Estimation

4.1 Estimation of Resting State within-region Information Flow

Equation (15) can be written as:

$$\mathbf{Y}^m = \mathbf{X}^m \boldsymbol{\beta}^m + \boldsymbol{\epsilon}^m, \tag{9}$$

where $\mathbf{Y}^m = \left(\mathbf{m}_1^\intercal, \cdots, \mathbf{m}_{V_m}^\intercal\right)^\intercal$, $\mathbf{X}^m = \left(\mathbf{I}_{V_m} \otimes \mathbf{Z}^m\right)$,

$\mathbf{Z}^m = \begin{pmatrix} 1, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & \mathbf{m}^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{m}^\intercal(t-1), & \mathbf{m}^\intercal(t-2), & \cdots, & \mathbf{m}^\intercal(t-p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{m}^\intercal(T-1), & \mathbf{m}^\intercal(T-2), & \cdots, & \mathbf{m}^\intercal(T-p) \end{pmatrix}$ is $T \times k$, $k = Vp + 1$, and

$\mathbf{m}_i = \left(m_i(1), \cdots, m_i(T)\right)^\intercal, \forall i = 1, \cdots, V.$

Use the two-step approach in 8.4, we have:

$$\{\hat{\boldsymbol{\beta}}^m\}^{\text{glm}} = \left[(\mathbf{X}^m)^\intercal \hat{\boldsymbol{\Sigma}}^m \mathbf{X}^m\right]^{-1} (\mathbf{X}^m)^\intercal (\hat{\boldsymbol{\Sigma}}^m)^{-1} \mathbf{Y}^m,$$

where $\hat{\boldsymbol{\Sigma}}^m = \dfrac{\hat{\boldsymbol{\epsilon}}^m (\hat{\boldsymbol{\epsilon}}^m)^\intercal}{T-k}$, $\hat{\boldsymbol{\epsilon}}^m = \mathbf{Y}^m - \mathbf{X}^m \{\hat{\boldsymbol{\beta}}^m\}^{\text{ols}}$, and $\{\hat{\boldsymbol{\beta}}^m\}^{\text{ols}} = \left[(\mathbf{X}^m)^\intercal \mathbf{X}^m\right]^{-1} (\mathbf{X}^m)^\intercal \mathbf{Y}^m$.

Similarly, $\{\hat{\boldsymbol{\beta}}^n\}^{\text{glm}}$ for (15) can be derived.

4.2 Estimation of Resting State between-region Information Flow

With some modification from (19), (17) can be written as:

$$\mathbf{Y}^{m,n} = \mathbf{X}^{m,n} \boldsymbol{\beta}^{m,n} + \boldsymbol{\epsilon}^{m,n}, \tag{10}$$

where $\mathbf{Y}^{m,n} = \begin{pmatrix} \mathbf{M} \\ \mathbf{N} \end{pmatrix}$, $\mathbf{X}^{m,n} = \begin{pmatrix} \mathbf{I}_{V_m} \otimes \mathbf{Z}^m, & \mathbf{I}_{V_m} \otimes \mathbf{Z}^n, & \mathbf{0}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{0}, & \mathbf{I}_{V_n} \otimes \mathbf{Z}^n, & \mathbf{I}_{V_n} \otimes \mathbf{Z}^m \end{pmatrix}$, $\boldsymbol{\beta}^{m,n} = \begin{pmatrix} \boldsymbol{\beta}^{m \to m} \\ \boldsymbol{\beta}^{n \to m} \\ \boldsymbol{\beta}^{n \to n} \\ \boldsymbol{\beta}^{m \to n} \end{pmatrix}$,

$$\mathbf{Z}^m = \begin{pmatrix} 1, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & \mathbf{m}^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{m}^\intercal(t-1), & \mathbf{m}^\intercal(t-2), & \cdots, & \mathbf{m}^\intercal(t-p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{m}^\intercal(T-1), & \mathbf{m}^\intercal(T-2), & \cdots, & \mathbf{m}^\intercal(T-p) \end{pmatrix} \text{ is } T \times k,\ k = Vp+1, \text{ and } \mathbf{m}_i =$$

$$\big(m_i(1), \cdots, m_i(T)\big)^\intercal, \forall i = 1, \cdots, V, \text{ and } \mathbf{Z}^n = \begin{pmatrix} 1, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & \mathbf{n}^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{n}^\intercal(t-1), & \mathbf{n}^\intercal(t-2), & \cdots, & \mathbf{n}^\intercal(t-p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{n}^\intercal(T-1), & \mathbf{n}^\intercal(T-2), & \cdots, & \mathbf{n}^\intercal(T-p) \end{pmatrix}$$

is $T \times k$, $k = V_n p + 1$, and $\mathbf{n}_i = \big(n_i(1), \cdots, n_i(T)\big)^\intercal$, $\forall i = 1, \cdots, \mu$.

Use the two-step approach in 8.4, we have:

$$\{\hat{\boldsymbol{\beta}}^{m,n}\}^{\text{glm}} = \big[(\mathbf{X}^{m,n})^\intercal \hat{\boldsymbol{\Sigma}}^{m,n} \mathbf{X}^{m,n}\big]^{-1} (\mathbf{X}^{m,n})^\intercal (\hat{\boldsymbol{\Sigma}}^{m,n})^{-1} \mathbf{Y}^{m,n},$$

where $\hat{\boldsymbol{\Sigma}}^{m,n} = \dfrac{\hat{\boldsymbol{\epsilon}}^{m,n}(\hat{\boldsymbol{\epsilon}}^{m,n})^\intercal}{2T - k_1 - k_2}$, $\hat{\boldsymbol{\epsilon}}^{m,n} = \mathbf{Y}^{m,n} - \mathbf{X}^{m,n}\{\hat{\boldsymbol{\beta}}^{m,n}\}^{\text{ols}}$, and $\{\hat{\boldsymbol{\beta}}^{m,n}\}^{\text{ols}} = \big[(\mathbf{X}^{m,n})^\intercal \mathbf{X}^{m,n}\big]^{-1} (\mathbf{X}^{m,n})^\intercal \mathbf{Y}^{m,n}$.

4.3 Estimation of Temporal Treatment effect

(1) can be written as:

$$\mathbf{Y}^m = \mathbf{X}^{T,m} \boldsymbol{\beta}^{T,m} + \boldsymbol{\eta}_1, \tag{11}$$

where $\mathbf{Y}^m$ is the same as (9), $\mathbf{X}^{T,m} = \big(\mathbf{I}_{V_m} \otimes \mathbf{Z}^{T,m}\big)$,

$$\mathbf{Z}^{T,m} = \begin{pmatrix} 1, & T, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & T, & \mathbf{m}^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & T, & \mathbf{m}^\intercal(t-1), & \mathbf{m}^\intercal(t-2), & \cdots, & \mathbf{m}^\intercal(t-p) \\ \vdots, & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & T, & \mathbf{m}^\intercal(T-1), & \mathbf{m}^\intercal(T-2), & \cdots, & \mathbf{m}^\intercal(T-p) \end{pmatrix} \text{ is } T \times k,\ k = Vp+2,$$

and $\mathbf{m}_i = \big(m_i(1), \cdots, m_i(T)\big)^\intercal$, $\forall i = 1, \cdots, V$.

(2) can be written as:

$$\mathbf{Y}^n = \mathbf{X}^{n\to n,m\to n}\boldsymbol{\beta}^{n\to n,m\to n} + \boldsymbol{\epsilon}_1,$$

where $\mathbf{X}^{n\to n,m\to n} = \begin{pmatrix} \mathbf{I}_{V_n} \otimes \mathbf{Z}^n, & \mathbf{I}_{V_n} \otimes \mathbf{Z}^m \end{pmatrix}$, and $\mathbf{Y}^n, \mathbf{Z}^n, \mathbf{Z}^m$ follow from the notations in (10).

Finally, (3) can be written as:

$$\mathbf{Y}^{m,n} = \mathbf{X}^{T,m,n}\boldsymbol{\beta}^{T,m,n} + \boldsymbol{\Gamma},$$

where $\mathbf{Y}^{m,n}$ is given in (10), $\mathbf{X}^{T,m,n} = \begin{pmatrix} \mathbf{I}_{V_m} \otimes \mathbf{Z}^{T,m}, & \mathbf{I}_{V_m} \otimes \mathbf{Z}^{T,n}, & \mathbf{0}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{0}, & \mathbf{I}_{V_n} \otimes \mathbf{Z}^{T,n}, & \mathbf{I}_{V_n} \otimes \mathbf{Z}^{T,n} \end{pmatrix}$, $\boldsymbol{\beta}^{T,m,n} = \begin{pmatrix} \boldsymbol{\beta}^{T,m\to m} \\ \boldsymbol{\beta}^{T,n\to m} \\ \boldsymbol{\beta}^{T,n\to n} \\ \boldsymbol{\beta}^{T,m\to n} \end{pmatrix}$,

$\mathbf{Z}^{T,m}$ given in (11), and $\mathbf{Z}^{T,n} = \begin{pmatrix} 1, & ,T & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & ,T & \mathbf{n}^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots, & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & ,T & \mathbf{n}^\intercal(t-1), & \mathbf{n}^\intercal(t-2), & \cdots, & \mathbf{n}^\intercal(t-p) \\ \vdots, & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & ,T & \mathbf{n}^\intercal(T-1), & \mathbf{n}^\intercal(T-2), & \cdots, & \mathbf{n}^\intercal(T-p) \end{pmatrix}$ is $T \times k$, $k = V_n p + 2$, and $\mathbf{n}_i = \big(n_i(1), \cdots, n_i(T)\big)^\intercal, \forall i = 1, \cdots, \mu$.

$\boldsymbol{\beta}^{T,m}$, $\boldsymbol{\beta}^{n\to n,m\to n}$, and $\boldsymbol{\beta}^{T,m,n}$ coefficients above can be estimated using the two-step procedure in 8.4 straightforwardly.

4.4 Estimation in Behavior Prediction

(1) can be written as:

$$\mathbf{Y}^O = \mathbf{X}^{T,m,n}\boldsymbol{\beta}^{T,m,n},$$

where $\mathbf{Y}^O = \{Y_1, \cdots, Y_J\}$, $\mathbf{X}^{T,m,n} = \begin{pmatrix} 1 & T_1 & vec^\intercal(\mathbf{Z}_1^n) & vec^\intercal(\mathbf{Z}_1^m) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & T_J & vec^\intercal(\mathbf{Z}_J^n) & vec^\intercal(\mathbf{Z}_J^m) \end{pmatrix}$, $\mathbf{Z}_1^n$ are j^{th} treatment specific matrix, and in particular, for the j^{th} treatment:

$$\mathbf{Z}_j^m = \begin{pmatrix} 1, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & \mathbf{m}_j^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{m}_j^\intercal(t-1), & \mathbf{m}_j^\intercal(t-2), & \cdots, & \mathbf{m}_j^\intercal(t-p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{m}_j^\intercal(T-1), & \mathbf{m}_j^\intercal(T-2), & \cdots, & \mathbf{m}_j^\intercal(T-p) \end{pmatrix}$$ is $T \times k_{V_m}$, $k_{V_m} = V_m p + 1$,

$\mathbf{Z}_1^n$, $\mathbf{m}_j^\intercal(t-s) = \big(m_{i,j}(t-s), \cdots, m_{i,j}(t-s)\big)^\intercal$, for $1 \leq s \leq t$, and $\forall i = 1, \cdots, V$ nodes in M, and $\mathbf{Z}_j^n = \begin{pmatrix} 1, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & \mathbf{n}_j^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{n}_j^\intercal(t-1), & \mathbf{n}_j^\intercal(t-2), & \cdots, & \mathbf{n}_j^\intercal(t-p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{n}_j^\intercal(T-1), & \mathbf{n}_j^\intercal(T-2), & \cdots, & \mathbf{n}_j^\intercal(T-p) \end{pmatrix}$ is $T \times k_{V_n}$, $k_{V_n} = V_n p + 1$,

and $\mathbf{n}_j^\intercal(t-s) = \big(n_{i,j}(t-s), \cdots, n_{i,j}(t-s)\big)^\intercal$, for $1 \leq s \leq t$, and $\forall i = 1, \cdots, V_n$ nodes in N.

Furthermore, $\boldsymbol{\beta}^{T,m,n} = \big(r_0, r_1, (\boldsymbol{\beta}^n)^\intercal, (\boldsymbol{\beta}^m)^\intercal\big)^\intercal$. Notice, vec$(\mathbf{Z}_j^n)^\intercal$ is $1 \times Tk_V$ and vec$(\mathbf{Z}_j^m)^\intercal$ is $1 \times Tk_{V_n}$.

4.5 Inference

Wald statistics

To test

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

from (20) we can form the Wald statistic:

$$Wald = \big(\mathbf{R}\hat{\boldsymbol{\beta}}^{gls} - \mathbf{r}\big)^\intercal \Big\{ \mathbf{R}[\hat{\boldsymbol{\Sigma}} \otimes (\mathbf{Z}^\intercal\mathbf{Z})^{-1}]\mathbf{R}^\intercal \Big\}^{-1} \big(\mathbf{R}\hat{\boldsymbol{\beta}}^{gls} - \mathbf{r}\big). \tag{12}$$

4.6 Dimension Reduction

The number of parameters in a model grows at $\mathcal{O}(m^2)$ for m-dimensional linear vector time series. For nonlinear vector time series, the growth rate is even faster. Therefore, dimension reduction approaches are necessary to extract the majority of the information and increase the computing speed. There are two major categories of dimension reduction approaches: time domain approaches and frequency approaches. For the frequency approches (see Brillinger (1981) and Shumway and Stoffer (2000). The time domain approaches can be further divided in to procipal component approach; the Box and Tiao canonical analysis (Box and Tiao, 1977); reduced rank models ((Robinson, 1973), (Ahn and Reinsel, 1990), (Reinsel and Ahn, 1992), and (Reinsel and Velu, 1998); scalr component models (Tiao and Tsay, 1989), and state space models (Akaike, 1974), (Aoki, 1987), (Hannan and Deistler, 1988), (Harvey, 1989), and (Durbin and Koopman, 2001). (Peña and Poncela, 2006) gives a good overview over dimension reduction in multivariate time series.

In this paper, we use the reduced rank models (Velu et al, 1986) to conduct dimension reduction for VAR(p) processes.

Consider a VAR(p) process in the matrix form:

$$\mathbf{y}(t) = \mathbf{A}\mathbf{y}^*(t) + \boldsymbol{\mu}(t); \tag{13}$$

where $\mathbf{A} = \big(\mathbf{A}(1), \cdots, \mathbf{A}(p)\big)$ and $\mathbf{y}^*(t) = \big(\mathbf{y}(t-1)^\intercal, \cdots, \mathbf{y}^\intercal(t-p)\big)^\intercal$. If $\mathbf{A}$, $m \times m$ has reduced rank, then $\mathbf{A}$ can be written as:

$$\mathbf{A} = \mathbf{B}_r\mathbf{C}_r$$

where $\mathbf{B}_r$ and $\mathbf{C}_r$ are $m \times r$ and $r \times m$, respectively. Then (13) can be written as:

$$\mathbf{y}(t) = \mathbf{B}_r\mathbf{z}(t) + \boldsymbol{\mu}(t),$$

where $\mathbf{z}(t) = \mathbf{C}_r\mathbf{y}^*(t)$.

5 Simulation Study

6 Application to Data from a Thermal Pain Study

"The study of neural interactions will be important for relating cognitive theories with brain operations. It is highly unlikely that the functional organization of the brain follows the independent modular organization of psychological constructs. Therefore, it is also unlikely that a single brain region has only one cognitive function. Instead, functionally specialized anatomical networks within the brain may be more easily related to cognitive constructs. There may not be a single brain area that represents "attentio" for instance, but there are more likely numerous brain areas whose interactions represent attention operations. The important point is that it may be possible for parts of the same anatomical network to be involved in another function when the interactions change. " "Indeed, these views are consistent with the adaptive characteristics of the central nervous system, in which functional organiza tion is viewed as dynamically related to the particular environmental demands rather than a static property [John and Schwartz, 1978; McIntosh and Gonzalez- Lima, 1993, 1994; Merzenich and Sameshima, 1993; Pascaul-Leone et al., 1994; Recanzone et al., 1992; Scheich et al., 1992; Wolpaw and Lee, 1989; Zohary et al., 19941. Brain imaging techniques will make an important contribution to the understanding of this property, given their ability to asses5 activity across many neural regions. Covariance analyses like structural equation modeling will contribute to this understanding by determining the functional relations among these areas and showing how they are modified depending on the behavioral/cognitive requirements."

"In the central nervous system there are numerous ways one area can have a functional impact on another. Many neural systems have a parallel anatomical organization by which connections between areas can be both direct (hierarchical) and indirect. Effects decomposition of anatomically based structural equation models allows for the evaluation of whether the in-

fluence of one region on another is through a direct effect or is mediated through one or more indirect routes."

SEM problems: "The example presented above shows that the anatomical foundation is the key feature for neural structural equation models. An immediate concern with the use of the anatomical constraint is the degree to which the anatomical model reflects reality. Any system of equations in which there are unknowns to be solved benefit from constraints to possible solutions. Using the connective anatomy of the system helps to constrain solutions. However, if all major and minor paths were included, most models would contain reciprocal loops at nearly every level, with some interconnections between levels, both feedforward and feedback. When all possible anatomical connections are included, it is likely that an underdetermined system of equations would result, in which there are either the same number of known and unknown elements or more unknown elements. In either case, unique solutions are not obtainable. A hierarchical model building strategy could be used whereby a smaller part of the model is estimated first, the parameters fixed at those estimates, and then more paths and regions added to complete the model [McIntosh and Gonzalez-Lima, 1992al. In most cases, however, some compromise between anatomical accuracy and interpretability may be needed. Any modeling effort, whether based on simulations or data-fitting, is necessarily a simplifications and represents an approximation of reality. It is the degree of simplification that determines the utility of the model."

"Brain imaging researchers will often discuss the results in terms of âĂIJfunctional networksâĂİ without specific references to how these networks are formed. By requiring that the networks be expressed in a covariance- based model, the assumptions of the researcher as to the organization of the network are more easily seen."

7 Discussion

In this paper, the model only considers linear features of VAR processes. The true neural connectivity and underlying functional information flows may be, however, nonlinear. For nonlinear functions, one may extend our approach by applying the kernal approach (Ancona et al. (2004), Marinazzo et al. (2008)) for short and noisy data, or use a transfer entrophy. It is worthwhile mentioning that, although in neurosicence applications some systems are nonlinear, (1) nonlinear systems can usually be well approximated by linear functions (such as cases with large-scale interactions, see Mclntosh and Gonzalez-Lima (1994)), and stationary Gaussian process is necessarily linear (Barnett et al., 2009).

8 Appendix

8.1 Vector Autoregressive Process

Here we review the VAR(p) process in the context of brain regions.

Consider n voxels each of which is measured at equally spaced time points. For a particular $t \in \mathcal{T}$, the neural activity measured from these n voxels can be written as:

$$\mathbf{y}(t) = \big(y_1(t), y_2(t), \cdots, y_n(t)\big)^{\mathsf{T}},$$

where $\mathbf{y}(t)$ is a column vector of length n, and for $i \in \{1, \cdots, n\}$, $y_i(t)$ is the neural activity of the i^{th} voxel measured at time $t \in \mathcal{T}$.

Suppose the neural activity of the i^{th} voxel measured at time t depends on its own past p time points and the nerual activity of the rest $n-1$ voxels during the past p time points, the p-lag vector autoregressive (VAR(p)) model can be written as:

$$\mathbf{y}(t) = \mathbf{c} + \mathbf{\Pi}_1 \mathbf{y}(t-1) + \mathbf{\Pi}_2 \mathbf{y}(t-2) + \cdots + \mathbf{\Pi}_p \mathbf{y}(t-p) + \boldsymbol{\epsilon}(t), \tag{14}$$

where $\mathbf{\Pi}_1$ is an $n \times n$ coefficient matrices and $\boldsymbol{\epsilon}(t)$ is an $n \times 1$ zero mean serieally uncorrelated or independent white noise vector process with time invariant covariance $Var(\boldsymbol{\epsilon}(t)) = \mathbf{\Sigma}$.

For example, for $n = 2$ and $p = 2$:

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} y_1(t-1) \\ y_2(t-1) \end{pmatrix} + \begin{pmatrix} \pi_{11}^2 & \pi_{12}^2 \\ \pi_{21}^2 & \pi_{22}^2 \end{pmatrix} \begin{pmatrix} y_1(t-2) \\ y_2(t-2) \end{pmatrix} + \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \end{pmatrix},$$

or equivalently

$$y_1(t) = c_1 + \pi_{11}^1 y_1(t-1) + \pi_{12}^1 y_2(t-1) + \pi_{11}^2 y_1(t-2) + \pi_{12}^2 y_2(t-2) + \epsilon_1(t);$$

$$y_2(t) = c_2 + \pi_{21}^1 y_1(t-1) + \pi_{22}^1 y_2(t-1) + \pi_{21}^2 y_1(t-2) + \pi_{22}^2 y_2(t-2) + \epsilon_2(t),$$

where $cov(\epsilon_1(s), \epsilon_2(t)) = \sigma_{12}$ for $s = t$, and 0, otherwise.

The mean-adjusted form of (14) can be written as:

$$\mathbf{y}(t) - \boldsymbol{\mu} = \mathbf{\Pi}_1\big(\mathbf{y}(t-1) - \boldsymbol{\mu}\big) + \mathbf{\Pi}_2\big(\mathbf{y}(t-2) - \boldsymbol{\mu}\big) + \cdots + \mathbf{\Pi}_p\big(\mathbf{y}(t-p) - \boldsymbol{\mu}\big) + \boldsymbol{\epsilon}(t),$$

where the unconditional mean $\boldsymbol{\mu} = \big(\mathbf{I}_n - \mathbf{\Pi}_1 - \cdots - \mathbf{\Pi}_p\big)^{-1}\mathbf{c}$.

A more general form of VAR(p) can be written as:

$$\mathbf{y}(t) = \mathbf{\Pi}_1\mathbf{y}(t-1) + \mathbf{\Pi}_2\mathbf{y}(t-2) + \cdots + \mathbf{\Pi}_p\mathbf{y}(t-p) + \mathbf{\Pi}^Z\mathbf{Z}(t) + \mathbf{\Pi}^X\mathbf{X}(t) + \boldsymbol{\epsilon}(t),$$

where $\mathbf{Z}(t)$ accouts for deterministic terms such as a linear trend or seasonal dummy variable, and $\mathbf{X}(t)$ accounts for exogeneous variable. In this paper we assume that $\mathbf{Z}(t)$ and $\mathbf{X}(t)$ are absorbed in the error term $\boldsymbol{\epsilon}(t)$.

Notice the number of parameters in a model grows at $\mathcal{O}(m^2)$ for m-dimensional linear vector time series.

8.2 Resting State Information Flow

Within region information flow:

Assume there is no information flow between regions M and N. Then with in each region of M and N, $\mathbf{m}(t)$ and $\mathbf{m}(t)$ are two independent VAR(p) processes.

$$\mathbf{m}(t) = \mathbf{c}_1(t) + \sum_{i=1}^{p} \mathbf{A}_1(i)\mathbf{m}(t-i) + \mathbf{u}_1(t), \quad Var\big(\mathbf{u}_1(t)\big) = \mathbf{\Sigma}_1; \tag{15}$$

$$\mathbf{n}(t) = \mathbf{c}_2(t) + \sum_{i=1}^{p} \mathbf{B}_1(i)\mathbf{n}(t-i) + \mathbf{v}_1(t), \quad Var\big(\mathbf{v}_1(t)\big) = \mathbf{T}_1. \tag{16}$$

Total information flow:

Assume there is both information flow within M and N, and between M and N. Then $\mathbf{m}(t)$ and $\mathbf{m}(t)$ can be written as:

$$\begin{aligned} \mathbf{m}(t) &= \mathbf{c}_3(t) + \sum_{i=1}^{p} \mathbf{A}_2(i)\mathbf{m}(t-i) + \sum_{i=1}^{p} \mathbf{C}_2(i)\mathbf{n}(t-i) + \mathbf{u}_2(t), \quad Var\big(\mathbf{u}_2(t)\big) = \mathbf{\Sigma}_2; \\ \mathbf{n}(t) &= \mathbf{c}_4(t) + \sum_{i=1}^{p} \mathbf{B}_2(i)\mathbf{n}(t-i) + \sum_{i=1}^{p} \mathbf{D}_2(i)\mathbf{m}(t-i) + \mathbf{v}_2(t), \quad Var\big(\mathbf{v}_2(t)\big) = \mathbf{T}_2, \end{aligned} \tag{17}$$

where $\mathbf{\Sigma} = Var\begin{pmatrix} \mathbf{u}_2(t) \\ \mathbf{v}_2(t) \end{pmatrix} = \begin{bmatrix} \mathbf{\Sigma}_2 & \mathbf{C} \\ \mathbf{C}^\intercal & \mathbf{T}_2 \end{bmatrix}$.

8.3 Test of Influence Flow Direction

The measure of linear feedback from N to M ($F_{N\to M}$), from M to N ($F_{M\to N}$), and it of instantaneous linear feedback ($F_{M\cdot N}$) are defined (Geweke, 1982) as follows:

$$F_{N\to M} = \ln\left(\frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_2|}\right);$$
$$F_{M\to N} = \ln\left(\frac{|\mathbf{T}_1|}{|\mathbf{T}_2|}\right);$$
$$F_{M\cdot N} = \ln\left(\frac{|\mathbf{T}_2|\,|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}|}\right).$$

The summation of these terms is defined as the measure of linear dependence (Gel'fand and Yaglom, 1959), or $F_{M,N} = \ln\left(\frac{|\mathbf{T}_2|\,|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}|}\right)$, which gives the "the measure of information". Specifically:

$$F_{M,N} = F_{N\to M} + F_{M\to N} + F_{M\cdot N}.$$

8.4 Estimation

Assume that the VAR(p) model is covariance stationary, and there are no restrictions on the parameters of the model. In SUR (Seemingly Unrelated Regression) notation, each equation in the VAR(p) model in (14) may be written as follows.

For $i = 1, \cdots, n$, where i indicates the i^{th} node in n nodes in a brain region:

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\pi}_i + \boldsymbol{\epsilon}_i, \tag{18}$$

where $\mathbf{y}_i = \big(y_i(1), \cdots, y_i(T)\big)^\intercal$ is $T \times 1$, $\mathbf{Z} = \begin{pmatrix} 1, & \mathbf{0}, & \mathbf{0}, & \cdots, & \mathbf{0} \\ 1, & \mathbf{y}^\intercal(1), & \mathbf{0}, & \cdots, & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{y}^\intercal(t-1), & \mathbf{y}^\intercal(t-2), & \cdots, & \mathbf{y}^\intercal(t-p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1, & \mathbf{y}^\intercal(T-1), & \mathbf{y}^\intercal(T-2), & \cdots, & \mathbf{y}^\intercal(T-p) \end{pmatrix}$ is $T \times k$, $k = np + 1$; and $\boldsymbol{\pi}_i$ is $k \times 1$; and in particular: for $1 \leq s \leq p$, $\mathbf{y}^\intercal(t-s) = \big(y_1(t-s), \cdots, y_n(t-s)\big)$.

Let $\boldsymbol{\Pi} = \big(\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_n\big)$ be the $k \times n$ coefficient matrix. Let vec$(\cdot)$ be the operator stacking inputs vertically, such that

$$\text{vec}\big(\boldsymbol{\Pi}\big) = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \vdots \\ \boldsymbol{\pi}_n \end{pmatrix}$$

is the $nk \times 1$.

Then the full model including n nodes is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{19}$$

where $\mathbf{Y} = \big(\mathbf{y}_1^\intercal, \cdots, \mathbf{y}_n^\intercal\big)^\intercal$, $\mathbf{X} = \big(\mathbf{I}_n \otimes \mathbf{Z}\big)$, and $\boldsymbol{\beta} = \text{vec}\big(\boldsymbol{\Pi}\big)$.

We use a two-stepe procedure for model estimation.

For the first step, we use the ordinary least square approach.

1. We obtain the OLS estimate of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y};$$

2. Compute residuals:

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{ols}};$$

3. Obtain the estimated covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \frac{\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^\intercal}{T-k} = \frac{\sum_{t=1}^{T} \hat{\boldsymbol{\epsilon}}_t \hat{\boldsymbol{\epsilon}}_t^\intercal}{T-k}.$$

For the second step, we use the generalized least square approach.

1. we update $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}}^{\text{gls}} = \left(\mathbf{X}^\intercal \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\intercal \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{Y};$$

2. the estimate $\hat{\boldsymbol{\beta}}^{\text{gls}}$ is consistent and asymptotically normal:

$$\begin{aligned} \sqrt{T-k}\left(\hat{\boldsymbol{\beta}}^{\text{gls}} - \boldsymbol{\beta}_0\right) &\sim MVN\left(\mathbf{0}, \left(\mathbf{X}^\intercal \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}\right)^{-1}\right) \\ &\sim MVN\left(\mathbf{0}, \left[(\mathbf{I}_n \otimes \mathbf{Z})^\intercal \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{I}_n \otimes \mathbf{Z})\right]^{-1}\right) \\ &\sim MVN\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}} \otimes (\mathbf{Z}^\intercal \mathbf{Z})^{-1}\right). \end{aligned} \tag{20}$$

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*(6), 716–723.

Ancona, N., D. Marinazzo, and S. Stramaglia (2004). Radial basis function approach to nonlinear granger causality of time series. *Physical Review E 70*(5), 056221.

Astolfi, L., F. Cincotti, D. Mattia, M. G. Marciani, L. A. Baccala, F. de Vico Fallani, S. Salinari, M. Ursino, M. Zavaglia, L. Ding, et al. (2007). Comparison of different cortical connectivity estimators for high-resolution eeg recordings. *Human brain mapping 28*(2), 143–157.

Barnett, L., A. B. Barrett, and A. K. Seth (2009). Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters 103*(23), 238701.

Barrett, A. B., L. Barnett, and A. K. Seth (2010). Multivariate granger causality and generalized variance. *Physical Review E 81*(4), 041907.

Bressler, S. L., C. G. Richter, Y. Chen, and M. Ding (2007). Cortical functional network organization from autoregressive modeling of local field potential oscillations. *Statistics in medicine 26*(21), 3875–3885.

Bressler, S. L. and A. K. Seth (2011). Wiener–granger causality: a well established methodology. *Neuroimage 58*(2), 323–329.

Bressler, S. L., W. Tang, C. M. Sylvester, G. L. Shulman, and M. Corbetta (2008). Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. *The Journal of neuroscience 28*(40), 10056–10061.

Brovelli, A., M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences of the United States of America 101*(26), 9849–9854.

Charniak, E. (1991). Bayesian networks without tears. *AI magazine 12*(4), 50.

Dhamala, M., G. Rangarajan, and M. Ding (2008). Analyzing information flow in brain networks with nonparametric granger causality. *NeuroImage 41*(2), 354–362.

Dickey, D. A. and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association 74*(366a), 427–431.

Ding, M., S. L. Bressler, W. Yang, and H. Liang (2000). Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biological cybernetics 83*(1), 35–45.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping 2*(1-2), 56–78.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity 1*(1), 13–36.

Gel'fand, I. M. and A. Yaglom (1959). *Calculation of the amount of information about a random function contained in another such function*. American Mathematical Society Providence.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association 77*(378), 304–313.

Gow, D. W., J. A. Segawa, S. P. Ahlfors, and F.-H. Lin (2008). Lexical influences on speech perception: a granger causality analysis of meg and eeg source estimates. *Neuroimage 43*(3), 614–623.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Hamilton, J. D. (1994). *Time series analysis*, Volume 2. Princeton university press Princeton.

Hesse, W., E. Möller, M. Arnold, and B. Schack (2003). The use of time-variant eeg granger causality for inspecting directed interdependencies of neural assemblies. *Journal of neuroscience methods 124*(1), 27–44.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, Volume 210. UCL press London.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics 54*(1), 159–178.

Marinazzo, D., M. Pellicoro, and S. Stramaglia (2008). Kernel method for nonlinear granger causality. *Physical Review Letters 100*(14), 144103.

Mclntosh, A. and F. Gonzalez-Lima (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping 2*(1-2), 2–22.

Mitra, P. P. and B. Pesaran (1999). Analysis of dynamic brain imaging data. *Biophysical journal 76*(2), 691–708.

Nielsen, T. D. and F. V. Jensen (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Patel, R. S., F. D. Bowman, and J. K. Rilling (2006). A bayesian approach to determining connectivity of the human brain. *Human brain mapping 27*(3), 267–276.

Phillips, P. C. and P. Perron (1988). Testing for a unit root in time series regression. *Biometrika 75*(2), 335–346.

Roebroeck, A., E. Formisano, and R. Goebel (2005). Mapping directed influence over the brain using granger causality and fmri. *Neuroimage 25*(1), 230–242.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Sridharan, D., D. J. Levitin, and V. Menon (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences 105*(34), 12569–12574.

Wiener, N. (1956). The theory of prediction modern mathematics for engineers vol 1, ed ef beckenbach.

Zhang, T., J. Wu, F. Li, B. Caffo, and D. Boatman-Reich (2015). A dynamic directional model for effective brain connectivity using electrocorticographic (ecog) time series. *Journal of the American Statistical Association 110*(509), 93–106.

Zhou, Z., M. Ding, Y. Chen, P. Wright, Z. Lu, and Y. Liu (2009). Detecting directional influence in fmri connectivity analysis using pca based granger causality. *Brain research 1289*, 22–29.