Identification of Directed Influence: Granger Causality, Kullback-Leibler Divergence, and Complexity

Abd-Krim Seghouane
Abd-krim.seghouane@nicta.com.au
National ICT Australia, Canberra Research Laboratory, Australian National University, College of Engineering and Computer Science, Canberra 2601, Australia

Shun-ichi Amari
amari@brain.riken.jp
RIKEN Brain Science Institute, Mathematical Neuroscience Laboratory, Saitama 351-0198, Japan

Detecting and characterizing causal interdependencies and couplings between different activated brain areas from functional neuroimage time series measurements of their activity constitutes a significant step toward understanding the process of brain functions. In this letter, we make the simple point that all current statistics used to make inferences about directed influences in functional neuroimage time series are variants of the same underlying quantity. This includes directed transfer entropy, transinformation, Kullback-Leibler formulations, conditional mutual information, and Granger causality. Crucially, in the case of autoregressive modeling, the underlying quantity is the likelihood ratio that compares models with and without directed influences from the past when modeling the influence of one time series on another. This framework is also used to derive the relation between these measures of directed influence and the complexity or the order of directed influence. These results provide a framework for unifying the Kullback-Leibler divergence, Granger causality, and the complexity of directed influence.

1 Introduction

With the widespread acceptance of the network description onto brain processing functions, identification and quantification of directed interactions between brain structures that support the processing of specific brain functions (perceptual, cognitive or motor functions) are fundamental issues in neuroscience. For example, information about the functioning or dysfunctioning of the brain can be inferred from the network structure. The current theories of schizophrenia emphasize that the core aspects of the pathophysiology are due to the disconnection hypothesis (Friston, 1998) rather than deficits in specific brain areas or neurotransmitter systems. Neuroimage techniques, such as electroencephalography (EEG), magnetoencephalography (MEG), functional near infrared imaging (fNIR), and functional magnetic magnetic resonance imaging (fMRI) are commonly employed to generate time series measurements of brain structure activities to address questions of functional and effective connectivity that are fundamental for the description of these brain networks (Roebroeck, Formissamo, & Goebel, 2005; Hesse, Moller, Arnold, & Schack, 2003). The key distinction between functional and effective connectivity pertains to whether one is assessing simple statistical dependencies (functional connectivity) or trying to infer the parameters of an underlying model of connectivity (effective connectivity). The interesting issue, from our perspective, is that Granger causality and related analysis allow the in vivo study of effective connectivity through the statistical analysis of neuroimaging time series. Although we have been in motivated this work by the analysis of neuroimaging time series, all the arguments presented in this letter apply to any inference about directed causality in any time series.

Granger causality analysis (Granger, 1969; Geweke, 1982; Kim, Putrino, Ghosh, & Brown, 2011; Quinn, Coleman, & Kiyavash, 2011) and transfer entropy methods (Schreider, 2000; Massey, 1990; Kamitake, Harashima, & Miyakawa, 1984; Amblard & Michel, 2011; Quinn, Coleman, Kiyavash, & Hatsopoulos, 2011) have become increasingly used approaches to explore directed influences between brain structures using functional neuroimage data (Hinrichs, Heinze, & Schoenfeld, 2006; Seghouane, 2011). While both approaches target the same objective—measuring the effect that one brain structure has on another—the transfer entropy specifically uses the Kullback-Leibler divergence to measure the influence of extra information. The general idea of Granger causality is defined in terms of upgrading predictability. If a signal X causes a signal Y, the knowledge of the past of both X and Y should improve the prediction of the future of Y in comparison to the knowledge of the past of Y only, whereas the idea behind transfer entropy is defined in terms of influence on conditional probabilities as measured with the Kullback-Leibler divergence (Kullback & Leibler, 1951). If a signal Y does not cause a signal X, then the probability density describing the future of X conditioned on its past should not be different from the probability density describing the future of X conditioned on its past and the past of Y as measured by the Kullback-Leibler divergence.

In this letter, we make the simple point that all current statistics used to make inferences about directed influences in functional neuroimage time series are variants of the same underlying quantity. This includes directed transfer entropy, transinformation, Kullback-Leibler formulations, conditional mutual information, and Granger causality. In the autoregressive modeling framework, the relation between these information measures of directed influence, Granger causality, and the complexity of directed influence is derived in both the univariate and multivariate cases. The rest of this letter is organized as follows. After a brief formulation of the problem, the different measures of directed influence are introduced in section 2, and the different relations of these measures are described. The autoregressive modeling is introduced in section 3, and the link with Granger causality and the complexity of directed influence is established for the univariate case. The multivariate case is treated in section 4. A simulation example is described in section 5, and the conclusion is given in section 6.

2 Problem Formulation and Measures of Directed Influence

In a typical fMRI experiment, several regions of interest (ROIs) are a priori identified in the brain. Each ROI is represented in the fMRI data set by multiple voxels, where each voxel is a variable comprising a single time series reflecting changes in the underlying metabolic signal. A standard approach used to assess directed influence between two ROIs is to derive a single time series for each ROI either by averaging or extracting a principal component (Zhou, Chen, Ding, Lu, & Liu, 2009); alternatively repeated pairwise analysis can be performed on pairs of voxels.

Assume two stochastic processes of length N, $X = (x_k : k = 1, \ldots, N)$ and $Y = (y_k : k = 1, \ldots, N)$ corresponding to two ROIs between which some interaction exists. Let $x = (x_1, \ldots, x_N)$ and $y = (y_1, \ldots, y_N)$ be two scalar-valued time series of length N sampled from X and Y, respectively. At time k, the discrepancy between the probability densities $p(x_k|x_{k-1}, \ldots, x_{k-p}) = p(x_k|x_k^p)$ and $p(x_k|x_{k-1}, \ldots, x_{k-p}, y_{k-1}, \ldots, y_{k-q}) = p(x_k|x_k^p, y_k^q)$ can be used to test the directed influence $Y \to X$. In the absence of interaction, Y has no influence on X, and these two densities are equal (Schreider, 2000). The directed influence Y has on X can then be quantified by measuring the expectation of the discrepancy between $p(x_k|x_k^p)$ and $(p(x_k|x_k^p, y_k^q)$, which can be quantified using the Kullback-Leibler divergence (Kullback & Leibler, 1951) as follows:

$$E_{x_k^p, y_k^q}\left\{KL\left(P\big(x_k|x_k^p\big)||P\big(x_k|x_k^p, y_k^q\big)\right)\right\} = \sum_{x_k, x_k^p, y_k^q} p\big(x_k, x_k^p, y_k^q\big) \log \frac{p\big(x_k|x_k^p, y_k^q\big)}{p\big(x_k|x_k^p\big)}. \tag{2.1}$$

If the information set J_k available at time k is x_{k-j}, $1 \leq j \leq p$. Let J_n' define the expended information set J_n plus y_{k-j}, $1 \leq j \leq q$. Then, based on the definition of Granger causality (Granger & Newbold, 1986), y_{k-j} does not cause, x_k with respect to J_n' if $P(x_k|J_n) = P(x_k|J_n')$ for all $k > 0$, so that the extra information in J_n' does not affect the conditional distribution. Therefore, equation 2.1 is an extension of this definition where the Kullback-Leibler divergence is used to quantify the distance between the conditional distribution $P(x_k|J_n)$ and $P(x_k|J_n')$.

The above relation can then be used to derive the following measure of directional interaction between two brain areas or ROIs characterized by two stochastic processes of length N,

$$\begin{aligned} DI_a(Y \to X) &= \sum_{k=1}^{N} \sum_{x_k, x_k^p, y_k^q} p\left(x_k, x_k^p, y_k^q\right) \log \frac{p\left(x_k | x_k^p, y_k^q\right)}{p\left(x_k | x_k^p\right)} \\ &= \sum_{k=1}^{N} H\left(x_k | x_k^p\right) - H\left(x_k | x_k^p, y_k^q\right) \\ &= \sum_{k=1}^{N} I\left(y_k^q, x_k | x_k^p\right), \end{aligned} \tag{2.2}$$

where $H(.|.)$ and $I(.,.|.)$ represent the conditional entropy and the conditional mutual information, respectively. The measure described in equation 2.2 is linked to two other entropy-based measures introduced in the literature to describe the amount and the direction of interaction between two time series. For example, for $p = q = k - 1$,

$$\begin{aligned} DI_a(Y \to X) &= \sum_{k=1}^{N} I\left(y_k^q, x_k | x_k^p\right) \\ &= \sum_{k=1}^{N} I\left(y_k^{k-1}, x_k | x_k^{k-1}\right). \end{aligned} \tag{2.3}$$

Equation 2.2 resembles one of the first introduced directed interaction measure (Massey, 1990),

$$DI_b(Y \to X) = \sum_{k=1}^{N} I\left(y_k y_k^{k-1}, x_k | x_k^{k-1}\right), \tag{2.4}$$

where $y_k y_k^{k-1} = y_k, \ldots, y_1$. The only difference between equations 2.3 and 2.4 is the presence of the sample y_k in the conditional mutual information expression. Equation 2.4 corresponds to finding the mutual information between the time series Y up to k rather than $k - 1$ for equation 2.3 and the current sample of X, conditioned on the past $k - 1$ samples of X.

Kamitake et al. (1984), introduced another example of a measure of directional interaction, also known as transinformation,

$$DI_c(Y \to X) = \sum_{k=1}^{N} I\left(y_k, x^f | x_k x_k^p y_k^p\right), \tag{2.5}$$

where $x_k^p = x_{k-1}, \ldots, x_{k-p}$ and $y_k^p = y_{k-1}, \ldots, y_{k-p}$ are the past p samples of X and Y, respectively, and $x^f = x_{k+1}, \ldots, x_{k+q}$ are the q future samples of X with $p + q + 1 = N$. This measure was used in Hinrichs et al. (2006) to evaluate effective connectivity in a patient with homonymous hemianopsia due to a posterior cerebral artery stroke. Relation 2.5 corresponds to finding the mutual information between the current sample of Y and future q samples of X conditioned on the past p samples of X, the past p samples of Y, and the current sample of X. The proposed measure, equation 2.2, measures the influence of the past samples of Y on the current sample of X, while equation 2.5 measures the influence of the current sample of Y on the future samples of X. In equation 2.2, only the past samples of Y are considered rather than the past and current samples of Y because it is unlikely that the current measure of Y will instantaneously have direct influence on the current measure of X. This is also useful to make a link with transition probabilities and autoregressive modeling.

Proposition 1. *The equations that characterize the relationship between equations 2.4 and 2.5 are*

$$DI_c(Y \to X) = I(y^N, x^N) - DI_b(X \to Y) \tag{2.6}$$

and

$$DI_c(X \to Y) = I(x^N, y^N) - DI_b(Y \to X) \tag{2.7}$$

from which we deduce (Al-Khassaweneh & Aviyente, 2008)

$$DI_c(Y \to X) - DI_c(X \to Y) = DI_b(Y \to X) - DI_b(X \to Y), \tag{2.8}$$

where $I(y^N, x^N)$ is the mutual information between X and Y.

Relations 2.6 and 2.7 are similar to the conversation law (Massey & Massey, 2005). Their derivation is given in appendix A.

Note also that with $p = k - 1$, we have

$$\begin{aligned} DI_c(Y \to X) &= \sum_{k=1}^{N} I\left(y_k, x^f | x_k x_k^{k-1} y_k^{k-1}\right) \\ &= \sum_{k=1}^{N} I\left(y_k^{k-1}, x_k | x_k^{k-1}\right) \\ &= DI_a(Y \to X). \end{aligned} \tag{2.9}$$

Derivation 2.9 is based on the chain rule of Permuter, Kim, and Weissman (2011) and is also given in appendix A.

The computation of the above measure requires estimating multivariate probability density functions. Two different approaches can be used to estimate these multivariate entropies from experimental data: nonparametric or parametric. The nonparametric approach requires the choice of a kernel and the estimation of its bandwidth (Hinrichs et al., 2006; Schreider, 2000). The parametric approach requires the selection of the appropriate model and the estimation of its parameters. We adopt the parametric approach using autoregressive modeling in this letter for its relation with Granger causality (Granger, 1969; Geweke, 1982).

3 Parametric Estimation of the Directed Influence

For practical applications, processes X and Y can be modeled using AR models Kaminski et al. (2001),

$$x_k = a_1 x_{k-1} + a_2 x_{k-2} + \cdots + a_p x_{k-p} + \varepsilon_k, \quad k = p+1, \ldots, N, \tag{3.1}$$

where ε_k are identical and independently distributed (i.i.d.). $N(0, \sigma_p^2)$, independent of x_1, $a_1, \ldots, a_p \neq 0$, and $x_{p+1}, \ldots, x_N$ is a scalar value time series sampled from X. When this model is used, the conditional probability density,

$$p(x_k | x_k^p) = p(x_k - a_1 x_{k-1} - \cdots - a_p x_{k-p}) = p(\varepsilon_k). \tag{3.2}$$

On the other hand, the ARX model,

$$x_k = a_1 x_{k-1} + a_2 x_{k-2} + \cdots + a_p x_{k-p} + b_1 y_{k-1} + b_2 y_{k-2} + \cdots + b_q y_{k-q} + \varepsilon'_k,$$
$$k = \max(p, q) + 1, \ldots, N, \tag{3.3}$$

where ε_k are i.i.d. $N(0, \sigma_{p,q}^2)$, independent of $y_{\max(p,q)+1-q}$, $x_{\max(p,q)+1-p}$ and $a_1, \ldots, a_p, b_1, \ldots, b_q \neq 0$, which incorporates the q past samples of Y to improve the prediction of X at the current time, can be used to model the conditional probability density:

$$\begin{aligned} p(x_k | x_k^p, y_k^q) &= P(x_k - a_1 x_{k-1} - \cdots - a_p x_{k-p} - b_1 y_{k-1} - \cdots - b_q y_{k-q}) \\ &= p(\varepsilon'_k). \end{aligned} \tag{3.4}$$

Equation 2.2 is well adapted for use with models 3.2 and 3.4,

$$\begin{aligned} DI_a(Y \to X) = {} & \frac{N-p}{2} - \frac{N - \max(p,q)}{2} + (N-p) \ln \left(\sqrt{2\pi \sigma_p^2} \right) \\ & - (N - \max(p,q)) \ln \left(\sqrt{2\pi \sigma_{p,q}^2} \right), \end{aligned} \tag{3.5}$$

and in case $\max(p, q) = p$, equation 3.5 becomes

$$DI_a(Y \to X) = \frac{N-p}{2} \ln\left(\frac{\sigma_p^2}{\sigma_{p,q}^2}\right).$$

In practice, the candidate models 3.1 and 3.3 are estimated from the data. Substituting the maximum likelihood estimates $\hat{\sigma}_p^2$ and $\hat{\sigma}_{p,q}^2$ in the equation above for σ_p^2 and $\sigma_{p,q}^2$, we obtain

$$DI_a(Y \to X) = \frac{N-p}{2} \ln\left(\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p,q}^2}\right). \tag{3.6}$$

This represents the Granger causality defined in terms of the predictive error variances introduced in Granger (1969) and Geweke (1982) up to a multiplication constant. In this context, if the variability of the error ε'_k, $k = p + 1, \ldots, N$ of the ARX model, equation 3.3, as measured by its variance $\hat{\sigma}_{p,q}^2$, is smaller than the variability of the error ε_k of the AR model, equation 3.1, as measured by its variance $\hat{\sigma}_p^2$, then there is an improvement in the prediction of X due to Y. In contrast to Granger causality, the proposed directed influence measure is framed not in terms of prediction error but in terms of the discrepancy between conditional probabilities. In the framework of autoregressive modeling, the proposed directed influence measure is equivalent to Granger causality.

Proposition 2. *In asymptotic terms or large sample approximation, note that*

$$E_0\left\{DI_a(Y \to X)\right\} \simeq \frac{q}{2}, \tag{3.7}$$

where E_0 represents the expectation with respect to the true unknown generating probability density.

The derivation is given in appendix B.

Therefore, $DI_a(Y \to X)$ (and the Granger causality) measures the assistance of Y in predicting the future of X by asymptotically measuring the order or complexity by which Y influences X. As a result, there is an attractive complementarity between the notions' complexity, information, and prediction when it comes to measuring the interaction or influence between time series. The parametric estimation of the proposed measure, equation 2.2, using models 3.1 and 3.3, requires estimating the parameters of these models as well as the determination of their orders. The estimation of the parameters can be derived using the least squares method or the Yule-Walker equations (Brockwell & Davis, 1991). The estimation of the orders p and q can be obtained using a model selection criterion (Seghouane & Amari, 2007; Seghouane & Bekara, 2004; Seghouane, 2010).

4 The Multivariate Case

Instead of an approach that consists of deriving a single time series for each ROI, a possible alternative to assess the directed influence between two ROIs would be to consider multivariate random variables represented by vector time series characterizing the group of voxels comprising each ROI (Barrett, Seth, & Barnett, 2010). In this case, the autoregressive modeling requires the use of vector autoregressive models.

Assume that $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$ are realizations of two m-dimensional stochastic processes reflecting changes in the metabolic m-dimensional signals of two ROIs comprising m voxels. Then the vector autoregressive modeling equivalent to equation 3.1 is given by

$$X_k = A_1 X_{k-1} + A_2 X_{k-2} + \cdots + A_p X_{k-p} + \varepsilon_k, \quad k = p+1, \ldots, N, \tag{4.1}$$

where $X_k = (x_{1,k}, \ldots, x_{m,k})'$ is an $m \times 1$ observed vector of measurements of brain structure activities at times $t = 1, \ldots, N$; A_i, $i = 1, \ldots, p$, are $m \times m$ coefficient matrices of unknown parameters; and ε_k are i.i.d. normal random variables with mean zeros and $m \times m$ variance-covariance matrix $\hat{\Sigma}_p$. The approximating VAR(p) model, equation 4.1, can be rewritten as

$$\mathbf{U} = \mathbf{A}\mathbf{V} + \varepsilon, \tag{4.2}$$

where

$$\mathbf{U} = \begin{bmatrix} X_{p+1} \\ \vdots \\ X_N \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} X_p & \cdots & X_1 \\ \vdots & \ddots & \vdots \\ X_{N-1} & \cdots & X_{N-p} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} A_1 \\ \vdots \\ A_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_{p+1} \\ \vdots \\ \varepsilon_N \end{bmatrix},$$

and $\mathbf{V}$ is a known $(N-p)m \times p$ design matrix of full column rank. The vector autoregressive modeling equivalent to equation 3.3 is given by

$$\begin{aligned} X_k &= A_1 X_{k-1} + A_2 X_{k-2} + \cdots + A_p X_{k-p} + B_1 Y_{k-1} + \cdots + B_q Y_{k-q} + \varepsilon'_k, \\ k &= \max(p, q) + 1, \ldots, N, \end{aligned} \tag{4.3}$$

where B_i, $i = 1, \ldots, q$, are $m \times m$ coefficient matrices of unknown parameters and ε_k are i.i.d. normal random variables with mean zeros and $m \times m$

variance-covariance matrix $\hat{\Sigma}_{p,q}$. In case $\max(p, q) = p$, the approximating VAR(p,q) model, equation 4.3, can also be rewritten as

$$\mathbf{U} = \mathbf{CW} + \varepsilon', \tag{4.4}$$

where

$$\mathbf{W} = \begin{bmatrix} X_p & \cdots & X_1 & Y_p & \cdots & Y_{p+1-q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{N-1} & \cdots & X_{N-p} & Y_{N-1} & \cdots & Y_{N-q} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} A_1 \\ \vdots \\ A_p \\ B_1 \\ \vdots \\ B_q \end{bmatrix}, \quad \varepsilon' = \begin{bmatrix} \varepsilon'_{p+1} \\ \vdots \\ \varepsilon'_N \end{bmatrix},$$

and $\mathbf{W}$ is a known $(N - p)m \times (p + q)$ design matrix of full column rank.

From equations 4.2 and 4.4, we have

$$\begin{aligned} -2 \ln p(\mathbf{U}|\mathbf{A}) &= (N - p)m \ln(2\pi) + (N - p) \ln(|\hat{\Sigma}_p|) \\ &\quad + \mathrm{tr}\big\{(\mathbf{U} - \mathbf{AV})' \hat{\Sigma}_p^{-1} (\mathbf{U} - \mathbf{AV})\big\} \end{aligned}$$

and

$$\begin{aligned} -2 \ln p(\mathbf{U}|\mathbf{C}) &= (N - p)m \ln(2\pi) + (N - p) \ln(|\hat{\Sigma}_{p,q}|) \\ &\quad + \mathrm{tr}\big\{(\mathbf{U} - \mathbf{CW})' \hat{\Sigma}_{p,q}^{-1} (\mathbf{U} - \mathbf{CW})\big\}, \end{aligned}$$

from which it can be easily established that the directed influence measure, equation 2.2, is given by

$$DI_a(Y \to X) = \frac{N-p}{2} \ln \left(\frac{|\hat{\Sigma}_p|}{|\hat{\Sigma}_{p,q}|} \right). \tag{4.5}$$

This represents the Granger causality (Granger, 1969; Geweke, 1982) up to a multiplication constant in the multivariable case. In this context, if the variability of the error vector ε'_k, $k = p + 1, \ldots, N$ of the VAR(p,q) model, equation 4.3, as measured by the determinant of its variance-covariance matrix $|\hat{\Sigma}_{p,q}|$, is smaller than the variability of the error vector ε_k, of the

VAR(p) model, equation 4.1, as measured by the determinant of its variance-covariance matrix $|\hat{\Sigma}_p|$, then there is an improvement in the prediction of m-dimensional stochastic processes X due to Y. Also if framed in term of discrepancy between conditional probabilities, in the context of vector autoregressive modeling, the proposed directed influence measure, equation 2.2, is equivalent to Granger causality.

Proposition 3. *In asymptotic terms or large sample approximation, note that*

$$E_0\left\{DI_a(Y \to X)\right\} \simeq \frac{m^2 q}{2}, \tag{4.6}$$

where E_0 represents the expectation with respect to the true unknown generating probability density.

The derivation is given in appendix C.

In comparison to equation 3.7, note the presence of m^2 on the right-hand side of equation 4.6. This represents the dimension of the coefficient matrices A_i, $i = 1, \ldots, p$ and B_i, $i = 1, \ldots, q$. In the univariate case, $m = 1$, (22) is equal to equation 3.7. Therefore, $DI_a(Y \to X)$ (and the Granger causality) measures the directed influence of Y on X by asymptotically measuring the complexity by which Y influences X as measured by the number of required new parameters.

The parametric estimation of the proposed measure, equation 2.2, using models 4.1 and 4.3, requires the estimation of the parameters of these models as well as the determination of their orders. The estimation of the parameters can be obtained using the ordinary least squares method or the multivariate Yule-Walker equations (Brockwell & Davis, 1991). The estimation of orders p and q can be obtained using an appropriate model selection criterion (Hurvich & Tsai, 1993; Seghouane, 2006).

5 Simulation Examples

To verify relation 3.7 of proposition 2, we generated 50 data sets of size $N = \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ using two models:

- **Model 1**

$$\begin{cases} x_i = 0.5x_{i-1} + u_i \\ y_i = 0.8x_{i-1} + 0.2y_{i-1} + v_i \end{cases}, \tag{5.1}$$

where v_i and u_i are independent and (i.i.d) $N(0, 1)$

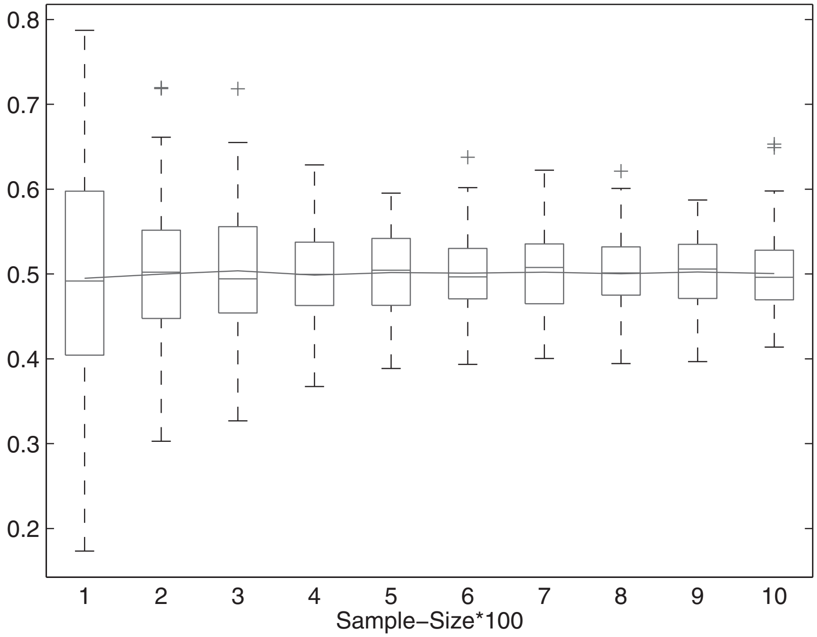

Figure 1: Numerical estimation of $E_0\{DI_a(X \to Y)\}$ for model 1.

- **Model 2**

$$\begin{cases} x_i = 0.3x_{i-1} - 0.1x_{i-1} + u_i \\ y_i = 0.4x_{i-1} - 0.2x_{i-2} + 0.3y_{i-1} - 0.1y_{i-2} + v_i \end{cases}, \tag{5.2}$$

where v_i and u_i are independent and (i.i.d) $N(0, 0.1)$.

For each sample size N, the expectation of relation 3.7 was obtained by averaging relation 3.6 ($DI_a(X \to Y)$) over the 50 realizations. For both models, each N, and each data set, the parameter estimates of the AR and ARX models were obtained by least squares. The results of the numerical estimations closely correspond to the theoretical estimate $\frac{q}{2} = 1/2$ (see Figure 1) for model 1 and $\frac{q}{2} = 1$ (see Figure 2) for model 2.

6 Conclusion

Directed influence measures are finding increasing applications in neuroscience. In this letter, the relations between all current statistics used to make inference about directed influences have been derived. It has been established that these statistics are variants of the same underlying quantity, the optimum statistics under the Neyman-Pearson lemma. In the context

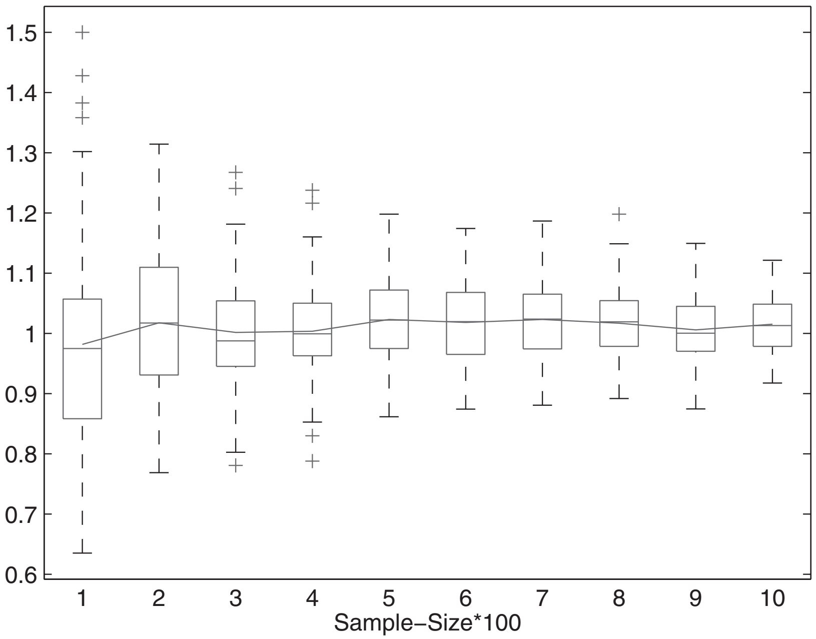

Figure 2: Numerical estimation of $E_0\{DI_a(X \to Y)\}$ for model 2.

of autoregressive modeling, it has been shown that asymptotically, both measures quantify the same information about the directed influence—the complexity of directed influence as measured by the order of the extra-autoregressive part of the model used to characterize the influence of the past of one time series on another. In the multivariate case, this complexity of influence is represented by the order of the extra-autoregressive part of the model used to characterize the influence of the past of one multivariate time series on another, multiplied by the dimension of the coefficient matrices. This corresponds to the number of extra parameters needed to model the influence. There is therefore an attractive complementarity among complexity, information, and prediction when it comes to measuring directed influences from neuroimage time series measurements.

Appendix A: Proof of Proposition 1

From equation 2.5, we have *for* $p = k - 1$,

$$\begin{aligned} I(y_k, x^f | x^p y^p x_k) &= I(y_k, x^f | x_k x_k^{k-1} y_k^{k-1}) \\ &= H(y_k | y_k^{k-1} x_k x_k^{k-1}) - H(y_k | y_k^{k-1} x_k x_k^{k-1} x^f) \end{aligned}$$

$$
\begin{aligned}
&= H\big(y_k | y_k^{k-1} x_k x_k^{k-1}\big) - H\big(y_k | y_k^{k-1} x^N\big) \\
&= H\big(y_k | y_k^{k-1}\big) - H\big(y_k | y_k^{k-1} x^N\big) - I\big(x_k x_k^{k-1}; y_k | x_k^{k-1}\big).
\end{aligned}
\tag{A.1}
$$

Taking the sum over $k = 1$ to N of the last line above gives

$$
\begin{aligned}
DI_c(Y \to X) &= H(y^N) - H(y^N | x^N) - DI_b(X \to Y) \\
&= I(y^N, x^N) - DI_b(X \to Y).
\end{aligned}
\tag{A.2}
$$

Using a similar derivation, we have

$$
DI_c(X \to Y) = I(x^N, y^N) - DI_b(Y \to X).
\tag{A.3}
$$

Subtracting equations A.3 from A.2 generates equation 2.8.

Using the chain rule (Permuter et al., 2011) $p(x^N, y^N) = p(x^N | y^N)$ $p(y^N | x^{N-1})$ and equation A.1 gives

$$
\begin{aligned}
\sum_{k=1}^{N} I\big(y_k, x^f | x_k x_k^{k-1} y_k^{k-1}\big) &= \sum_{k=1}^{N} H\big(y_k | y_k^{k-1} x_k x_k^{k-1}\big) - H\big(y_k | y_k^{k-1} x^N\big) \\
&= H(x^N, y^N) - H(x^N | y^{N-1}) - H(y^N | x^N) \\
&= H(x^N) - H(x^N | y^{N-1}) \\
&= DI_a(Y \to X).
\end{aligned}
\tag{A.4}
$$

Appendix B: Proof of Proposition 2

Equation 3.6 can be rewritten as

$$
\frac{N-p}{2} \ln\left(\frac{\hat{\sigma}_p^2}{\hat{\sigma}_{p,q}^2}\right) = \frac{N-p}{2} \ln\left(\frac{\hat{\sigma}_p^2}{\sigma_0^2}\right) + \frac{N-p}{2} \ln\left(\frac{\sigma_0^2}{\hat{\sigma}_{p,q}^2}\right),
\tag{B.1}
$$

where σ_0^2 is the variance of the unknown true noise.

From Koltz and Johnson (1982), we have

$$
E_0\left\{\ln(N-p)\frac{\hat{\sigma}_{p,q}^2}{\sigma_0^2}\right\} = \psi\left(\frac{(N-p)-(p+q)}{2}\right) + \ln 2,
\tag{B.2}
$$

where ψ is the Euler's psi function, and from Seghouane and Bekara (2004), we have

$$\psi\left(\frac{(N-p)-(p+q)}{2}\right) \simeq \ln\left(\frac{N-p}{2}\right) - \frac{(p+q)}{N-p} - \frac{1}{(N-p)-(p+q)} + o\left(\frac{1}{((N-p)-(p+q))^2}\right) + o\left(\left(\frac{(p+q)}{(N-p)}\right)^2\right). \tag{B.3}$$

Therefore,

$$\begin{aligned}(N-p)E_0\left\{\ln\left(\frac{\hat{\sigma}^2_{p,q}}{\sigma^2_0}\right)\right\} &= (N-p)E_0\left\{\ln\left((N-p)\frac{\hat{\sigma}^2_{p,q}}{\sigma^2_0}\right)\right\} \\ &\quad -(N-p)\ln(N-p) \\ &= (N-p)\psi\left(\frac{(N-p)-(p+q)}{2}\right) \\ &\quad -(N-p)\ln\left(\frac{(N-p)}{2}\right) \\ &\simeq -(p+q) - \frac{(N-p)}{(N-p)-(p+q)}. \end{aligned} \tag{B.4}$$

Similarly we have

$$(N-p)E_0\left\{\ln\left(\frac{\hat{\sigma}^2_{p}}{\sigma^2_0}\right)\right\} \simeq -p - \frac{(N-p)}{(N-p)-p}. \tag{B.5}$$

From equations B.3 and B.4, we have

$$E_0\left\{\frac{N-p}{2}\ln\left(\frac{\hat{\sigma}^2_{p}}{\hat{\sigma}^2_{p,q}}\right)\right\} = \frac{q}{2} + \frac{(N-p)}{2((N-p)-p-q)} - \frac{(N-p)}{2((N-p)-p)}, \tag{B.6}$$

from which for large sample approximation ($N \to \infty$), we obtain equation 3.7.

Appendix C: Proof of Proposition 3

Equation 4.5 can be rewritten as

$$\frac{N-p}{2}\ln\left(\frac{|\hat{\Sigma}_p|}{|\hat{\Sigma}_{p,q}|}\right)=\frac{N-p}{2}\ln\left(\frac{|\hat{\Sigma}_p|}{|\Sigma_0|}\right)+\frac{N-p}{2}\ln\left(\frac{|\Sigma_0|}{|\hat{\Sigma}_{p,q}|}\right),$$

where $|\Sigma_0|$ is the variance of the unknown true noise.

Since $(N-p)\hat{\Sigma}_{p,q}$ is asymptotically distributed as Wishart $W_m(\Sigma_0, (N-p)-(p+q)m)$, the distribution of the determinant $|(N-p)\hat{\Sigma}_{p,q}|$ is the distribution of a product of χ^2 random variables (Muirhead, 1982),

$$|(N-p)\hat{\Sigma}_{p,q}| \sim |\Sigma_0|\prod_{i=1}^{m}\chi^2_{(N-p)-m(p+q+1)+i},$$

and it follows that

$$\ln|(N-p)\hat{\Sigma}_{p,q}| \sim \ln|\Sigma_0|+\sum_{i=1}^{m}\ln\chi^2_{(N-p)-m(p+q+1)+i}.$$

From Koltz and Johnson (1982), we have

$$E\{\ln\left|(N-p)\hat{\Sigma}_{p,q}\right|\}$$
$$=\ln|\Sigma_0|+m\ln(2)+\sum_{i=1}^{m}\psi\left(\frac{(N-p)-m(p+q+1)+i}{2}\right). \tag{C.1}$$

Therefore,

$$\begin{aligned}(N-p)E_0\left\{\ln\left(\frac{|\hat{\Sigma}_{p,q}|}{|\Sigma_0|}\right)\right\} &= (N-p)E_0\left\{\ln\left(\frac{|(N-p)\hat{\Sigma}_{p,q}|}{|\Sigma_0|}\right)\right\}\\ &\quad -(N-p)m\ln(N-p)\\ &=\sum_{i=1}^{m}(N-p)\psi\left(\frac{(N-p)-m(p+q+1)+i}{2}\right)\\ &\quad -(N-p)m\ln\left(\frac{(N-p)}{2}\right),\end{aligned} \tag{C.2}$$

and with Seghouane (2006),

$$\begin{aligned}
&\sum_{i=1}^{m}(N-p)\psi\left(\frac{(N-p)-m(p+q+1)+i}{2}\right)\\
&\quad = (N-p)m\ln\left(\frac{(N-p)}{2}\right) - \frac{2m^2(p+q)+m^2-m}{2}\\
&\quad = -\frac{(N-p)m}{(N-p)-m(p+q)-\frac{(m-1)}{2}},
\end{aligned}$$

the relation B.5 becomes

$$\begin{aligned}
(N-p)E_0\left\{\ln\left(\frac{|\hat{\Sigma}_{p,q}|}{|\Sigma_0|}\right)\right\} &= -\frac{2m^2(p+q)+m^2-m}{2}\\
&\quad -\frac{(N-p)m}{(N-p)-m(p+q)-\frac{(m-1)}{2}}. \qquad \text{(C.3)}
\end{aligned}$$

Using the same lines as above, we have

$$(N-p)E_0\left\{\ln\left(\frac{|\hat{\Sigma}_{p}|}{|\Sigma_0|}\right)\right\} = -\frac{2m^2p+m^2-m}{2} - \frac{(N-p)m}{(N-p)-mp-\frac{(m-1)}{2}}. \qquad \text{(C.4)}$$

From equations C.3 and C.4, we have

$$\begin{aligned}
E_0\left\{\frac{N-p}{2}\ln\left(\frac{|\hat{\Sigma}_{p}|}{|\hat{\Sigma}_{p,q}|}\right)\right\} &= \frac{m^2q}{2} + \frac{(N-p)m}{2\left((N-p)-m(p+q)-\frac{(m-1)}{2}\right)}\\
&\quad -\frac{(N-p)m}{2\left((N-p)-mp-\frac{(m-1)}{2}\right)}.
\end{aligned}$$

from which for large sample approximation ($N \to \infty$), we obtain equation 3.7.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions, which helped to improve the quality and presentation of this letter. NICTA is funded by the Australian government as represented by the Department of Broadband, Communications and the Digital Economy and the

Australian Research Council through the ICT Centre of Excellence program. This work was partly supported by the Human Frontier Science Program (grant RGY 80/2008) and the Japanese Society for Promoting Science Fellowship (number S-10738).

References

Al-Khassaweneh, M., & Aviyente, S. (2008). The relationship between two directed information measures. *IEEE Signal Processing Letters*, *15*, 801–804.

Amblard, P. O., & Michel, O. J. J. (2011). On directed information theory and Granger causality graphs. *Journal of Computational Neuroscience*, *30*, 7–16.

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

Brockwell, P., & Davis, R. (1991). *Time series: Theory and methods*. New York: Springer-Verlag.

Friston, K. J. (1998). The disconnection hypothesis. *Schizophrenia Research*, *30*, 115–125.

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

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral models. *Econometrica*, *37*, 424–438.

Granger, C. W. J., & Newbold, P. (1986). *Forecasting economic time series*. Orlando, FL: Academic Press.

Hesse, W., Moller, E., Arnold, M., & Schack, B. (2003). The use of time-varying EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, *124*, 27–44.

Hinrichs, H., Heinze, H. J., & Schoenfeld, M. A. (2006). Causal visual interactions as revealed by an information theoretic measure and fMRI. *NeuroImage*, *31*, 1051–1060.

Hurvich, C. M., & Tsai, L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, *14*, 271–279.

Kaminski, M., Ding, M. A., Truccolo, W. A., & Bressler, D. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernectics*, *85*, 145–157.

Kamitake, T., Harashima, H., & Miyakawa, H. (1984). A time-series analysis method based on directed transinformation. *Electron. Commun. Japan*, *67*, 1–9.

Kim, S., Putrino, D., Ghosh, S., & Brown, E. N. (2011). A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS Computational Biology*, *7*, e1001110.

Koltz, S., & Johnson, N. L. (1982). *Encyclopedia of statistical sciences*. New York: Wiley.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematics Statistics*, *22*, 76–86.

Massey, J. L. (1990). Causality, feedback and directed information. In *IEEE International Symposium on Information Theory and Its Applications* (pp. 303–305). Piscataway, NJ: IEEE.

Massey, J. L., & Massey, P. C. (2005). Conservation of mutual and directed information. In *IEEE International Symposium on Information Theory* (pp. 157–158). Piscataway, NJ: IEEE.

Muirhead, R. J. (1982). *Aspect of multivariate statistical theory*. New York: Wiley.

Permuter, H. H., Kim, Y.-H., & Weissman, T. (2011). Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Transactions on Information Theory*, *57*, 3248–3259.

Quinn, C., Coleman, T. P., & Kiyavash, N. (2011). A generalized prediction framework for Granger causality. In *IEEE International Workshop on Network Science for Communication Networks (NetSciCom)* (pp. 923–928). Piscataway, NJ: IEEE.

Quinn, C., Coleman, T. P., Kiyavash, N., & Hatsopoulos, N. G. (2011). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of Computational Neuroscience*, *30*, 17–44.

Roebroeck, A., Formissamo, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*, *25*, 230–242.

Schreider, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*, 461–464.

Seghouane, A. K. (2006). Vector autoregressive model-order selection from finite samples using Kullback's symmetric divergence. *IEEE Transactions on Circuits and Systems–I: Regular Papers*, *53*, 2327–2335.

Seghouane, A. K. (2010). Asymptotic bootstrap corrections of AIC for linear regression models. *Signal Processing*, *90*, 217–224.

Seghouane, A. K. (2011). Quantifying information flow in fMRI using the Kullback-Leibler divergence. In *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (pp. 1569–1572). Piscataway, NJ: IEEE.

Seghouane, A. K., & Amari, S. I. (2007). The AIC criterion and symmetrizing the Kullback-Leibler divergence. *IEEE Transactions on Neural Networks*, *18*, 97–106.

Seghouane, A. K., & Bekara, M. (2004). A small sample model selection criterion based on Kullback's symmetric divergence. *IEEE Transactions on Signal Processing*, *52*, 3314–3323.

Zhou, Z., Chen, Y., Ding, P., Lu, W. Z., & Liu, Y. (2009). Analyzing brain networks with PCA and conditional Granger causality. *Human Brain Mapping*, *30*, 2197–2206.

Received March 1, 2011; accepted December 6, 2011.

This article has been cited by: