# 3. Multivariate Autoregressive Modeling

Assume we have an *M*-dimensional time-series of length *T* (e.g., *M* channels of EEG data, with *T* time points per channel): $X := x_1 \ldots x_T$ where $x_t = [x_{1t} \ldots x_{Mt}]'$. We can represent the multivariate process at time *t* as a stationary, stable vector autoregressive (VAR, MVAR, MAR) process of order *p* (Henceforth we will denote this as a VAR[*p*] process):

$$x_t = v + \sum_{k=1}^{p} A_k x_{t-k} + u_t \qquad \text{(Eq 3.1)}$$

Here $v = [v_1 \ldots v_M]'$ is an (*M* x 1) vector of intercept terms (the mean of *X*), $A_i$ are (*M* x *M*) model coefficient matrices and $u_t$ is a zero-mean white noise process with nonsingular covariance matrix $\Sigma$.

## 3.1. Stationarity and Stability

We assume two basic conditions regarding the data *X* and its associated VAR[*p*] model: *stationarity* and *stability*. A stochastic process *X* is *weakly stationary* (or *wide-sense stationary* (WSS)) if its first and second moments (mean and covariance) do not change with time. In other words $E(x_t) = \mu$ for all *t* and $E[(x_t - \mu)(x_{t-h} - \mu)'] = \Gamma(h) = \Gamma(-h)'$ for all *t* and *h*=0,1,2, ... where *E* denotes expected value. A VAR[*p*] process is considered *stable* if its reverse characteristic polynomial has no roots in or on the complex unit circle. Formally, $x_t$ is stable if

$$\det(I_{Mp} - Az) \neq 0 \text{ for } |z| \leq 1 \text{ where}$$

$$\mathbf{A} := \underset{(Mp \times Mp)}{\begin{bmatrix} A_1 & A_2 & \cdots & A_p \\ I_M & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & I_M & 0 \end{bmatrix}}$$

Equivalently, $x_t$ is stable if all eigenvalues of **A** have modulus less than 1 (Lütkepohl, 2006). A stable process is one that will not diverge to infinity ("blow up"). An important fact is that stability implies stationarity – thus it is sufficient to test for stability to ensure that a VAR[*p*] process is both stable and stationary. SIFT performs a stability test by analyzing the eigenvalues of **A**.

## 3.2. The Multivariate Least-Squares Estimator

A parametric VAR model can be fit using a number of approaches including multivariate least-squares approaches (e.g., MLS, ARFIT), lattice algorithms (e.g., Vieira-Morf), or state-

space models (e.g., Kalman filtering). Here we will briefly outline the multivariate least-squares algorithm (multichannel Yule-Walker) and encourage the interested reader to consult (Schlögl, 2000; Lütkepohl, 2006; Schlögl, 2006) for more details on this and other algorithms (several of which are implemented in SIFT).

To derive the multivariate least-squares estimator, let us begin with some definitions:

$$X := (x_1, \ldots, x_T) \qquad (M \times T),$$
$$B := (v, A_1, \ldots, A_p) \quad (M \times (Mp+1)),$$
$$Z_t := \begin{bmatrix} 1 \\ x_t \\ \vdots \\ x_{t-p+1} \end{bmatrix} \qquad ((Mp+1) \times 1),$$
$$Z := (Z_0, \ldots, Z_{T-1}) \qquad ((Mp+1) \times T),$$
$$U := (u_1, \ldots, u_T) \qquad (M \times T)$$

Our VAR[$p$] model (Eq 3.1) can now be written in compact form:

$$X = BZ + U \qquad \text{(Eq 3.2)}$$

Here $B$ and $U$ are unknown. The multivariate (generalized) least-squares (LS, GLS) estimator of $B$ is the estimator $\hat{B}$ that minimizes the variance of the innovation process (residuals) $U$. Namely,

$$\hat{B} = \arg\min_B S(B)$$

where $S(B) = \mathrm{tr}[(X - BZ)'\Sigma^{-1}(X - BZ)]$.

It can be shown (Lütkepohl, 2006) that the LS estimator can be obtained by

$$\hat{B} = XZ'(ZZ')^{-1} \qquad \text{(Eq 3.3)}$$

This result can be derived in several ways, however a simple approach follows from post-multiplying

$$x_t = BZ_{t-1} + u_t$$

by $Z'_{t-1}$ and taking expectations:

$$E(x_t Z'_{t-1}) = BE(Z_{t-1} Z'_{t-1}) \qquad \text{(Eq 3.4)}$$

Estimating $E(x_t Z'_{t-1})$ by

$$\frac{1}{T}\sum_{t=1}^{T} x_t Z'_{t-1} = \frac{1}{T} XZ'$$

we obtain the normal equations

$$\frac{1}{T} XZ' = \hat{B} \frac{1}{T} ZZ'$$

and thus, $\hat{B} = XZ'(ZZ')^{-1}$.

The reader may note that $\hat{B}$ is simply the product of *X* and the Moore-Penrose pseudoinverse of *Z*: $\hat{B} = XZ^{\dagger}$ where $Z^{\dagger} = \mathrm{pinv}(Z)$. The reader familiar with univariate autoregressive model fitting might also note that (Eq 3.4) is very similar to the well-known system of Yule-Walker equations. Hence, this can be considered an extension to the multivariate case of the Yule-Walker algorithm for univariate AR model fitting.

Although asymptotically optimal, the LS algorithm often suffers from sub-optimal performance when even moderate sample sizes are available, as compared to more robust modified LS algorithms (e.g., the stepwise least-squares ARFIT algorithm) or non-LS algorithms (e.g., the Vieira-Morf lattice algorithm). A detailed empirical performance comparison of these and other algorithms can be found in (Schlögl, 2006). For this reason, SIFT abandons the LS algorithm in favor of these more robust algorithms. The SIFT functions **`pop_est_fitMVAR()`** and **`est_fitMVARKalman()`** provide access to various model-fitting approaches.

## 3.3. Frequency-Domain Representation

Electrophysiological processes generally exhibit oscillatory structure, making them well suited for frequency-domain analysis (Buzsaki, 2006). A suitably fit autoregressive model provides an idealized model for the analysis of oscillatory structure in stochastic time series (Burg, 1967; Zetterberg, 1969; Burg, 1975; Neumaier and Schneider, 2001). From the AR coefficients, we can obtain a number of useful quantities including the *spectral density matrix* and the *transfer function* of the process. From these and related quantities we can obtain power spectra, coherence and partial coherence, Granger-Geweke causality, directed transfer function, partial directed coherence, phase-locking value, and a number of other quantities increasingly being used by the neuroscience community to study synchronization and information flow in the brain (Pereda et al., 2005; Schelter et al., 2006).

To obtain our frequency-domain representation of the model, we begin with our VAR[*p*] model from (Eq 3.1). For simplicity, we will assume the process mean is zero:

$$x_t = \sum_{k=1}^{p} A_k x_{t-k} + u_t$$

Rearranging terms we get

$$u_t = \sum_{k=0}^{p} \hat{A}_k x_{t-k} \text{ where } \hat{A}_k = -A_k \text{ and } \hat{A}_0 = -I$$

Z-transforming both sides yields:

$U(f) = A(f)X(f)$ where

$$A(f) = \sum_{k=0}^{p} \hat{A}_k e^{-i2\pi fk}$$

Premultiplying by $A(f)^{-1}$ and rearranging terms we obtain:

$$X(f) = A(f)^{-1}U(f) = H(f)U(f)$$

Here *X(f)* is the (*M* x *M*) spectral matrix of the multivariate process, *U(f)* is a matrix of random sinusoidal shocks and $A(f)^{-1} = H(f)$ is the *transfer matrix* of the system. Note that *H(f)* transforms the noise input (*U*) into the structured spectral matrix. This should give us a hint that analysis of *H(f)* (and *A(f)*) will help us in identifying the structure of the modeled system (including information flow dynamics). The spectral density matrix of the process (which contains the auto-spectrum of each variable (at frequency *f*) on the diagonals and the cross-spectrum on the off-diagonals) is given by:

$$S(f) = X(f)X(f)^* = H(f)\Sigma H(f)^*$$

As we shall see in Section 4.3. , from $S(f), A(f), H(f)$ and $\Sigma$, we can derive a number of frequency-domain quantities relevant to the study of oscillations, information flow, and coupling in neural systems.

## 3.4. Modeling non-stationary data using adaptive VAR models

In section 3. we stated that data stationarity is a necessary precondition for accurate VAR estimation. However, it is well-known that neural data, including EEG and Local Field Potentials (LFPs), can be highly non-stationary, exhibiting large fluctuations in both the mean and variance over time. For instance, a record of EEG data containing evoked potentials (EPs) is a classic example of a non-stationary time series (both the mean and variance of the series changes dramatically and transiently during the evoked response). Another example would be EEG data collected during slow-wave sleep, which exhibits slow fluctuations in the mean EEG voltage over time. A number of algorithms have been proposed for fitting VAR models to non-stationary series. In the neuroscience community the most popular approaches include segmentation (overlapping sliding-window) approaches (Jansen et al., 1981; Florian and Pfurtscheller, 1995; Ding et al., 2000), state-space (Kalman filtering) approaches (Schlögl, 2000; Sommerlade et al., 2009), and non-parametric methods based on minimum-phase spectral matrix factorization (Dhamala et al., 2008). All of these approaches are currently – or soon to be made – accessible in SIFT. Here we will briefly outline the concepts behind each modeling approach.

### 3.4.1. Segmentation-based Adaptive VAR (AMVAR) models

A segmentation-based AMVAR adopts an approach rather similar to the concept behind short-time fourier transforms or other windowing techniques. Namely, we extract a sliding window of length $W$ from the multivariate dataset, and fit our VAR[$p$] model to this data. We then increment the window by a (small) quantity $Q$ and repeat the procedure until the start of the window is greater than $T$-$W$. This produces *floor((T-W)/Q+1)* VAR coefficient matrices which describe the evolution of the VAR[$p$] across time. The concept here is that by using a sufficiently small window, the data will be *locally stationary* within the window and suitable for VAR modeling. By using highly overlapping windows (small $Q$) we can obtain coefficients that change relatively smoothly with time. Figure 1 shows a schematic of the sliding-window AMVAR approach.

One concern here is whether sufficient data points are available to accurately fit the model. In the general case, we have $M^2p$ coefficients (free parameters) to estimate, which requires a minimum of $M^2p$ data samples. However, in practice, we would like to have at least 10 times as many data points as free parameters (Schlögl and Supp, 2006; Korzeniewska et al., 2008). When multiple realizations (e.g., experimental trials) are available, we can assume that each trial is a random sample from the same stochastic process and average covariance matrices across trials to reduce the bias of our model coefficient estimator (Ding et al., 2000). For the LS algorithm, explained in section 3. , this yields the modified estimator:

$$\hat{B} = E(X^{(i)}Z'^{(i)})E(Z^{(i)}Z'^{(i)})^{-1} \qquad \text{(Eq 3.5)}$$

Where $X^{(i)}$ and $Z^{(i)}$ denote matrices $X$ and $Z$ for the $i^{\text{th}}$ single-trial and the expected value is taken across all trials. This approach effectively increases the number of samples available for a sliding window of length $W$ from $W$ to $WN$, where $N$ is the number of trials/realizations. This allows us to potentially use very small windows (containing as few as $p+1$ sample points) while still obtaining a good model fit.

When using short windows with multi-trial data, an important preprocessing step is to pointwise subtract the ensemble mean and divide by the ensemble standard deviation (ensemble normalization). This ensures that the ensemble mean is zero and the variance is one, at every time point. This can dramatically improve the local stationarity of the data (Ding et al., 2000). An important result of this is that we are essentially modeling dependencies in the residual time-series after removing the event-related potential (ERP) from the data. The fact that this preprocessing step has become common practice in published applications of AMVAR analysis to neural data suggests that there is, in fact, rich task-relevant information present in the so-called "residual noise" component of the EEG which cannot be inferred from the ERP itself (Ding et al., 2000; Bressler and Seth, 2010). This fits under the model that mean-field electrophysiological measures such as LFPs and EEG measure a sum of (potentially oscillatory) ongoing activity and evoked responses where the amplitude and phase of the evoked response depends largely on the phase of the ongoing oscillations (Kenet et al., 2005; Wang et al., 2008). Analyzing the phase structure of the stationary ongoing oscillations may provide a deeper insight into the state of the underlying neural system than the analysis of the evoked responses themselves.

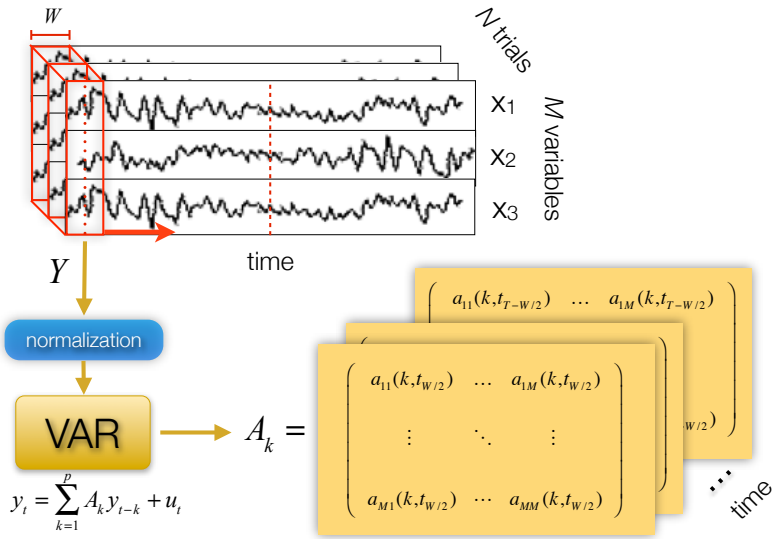


**Figure 1. Schematic of sliding-window AMVAR modeling. W is the window length, T is the length of each trial in samples, N is the number of trials.**

## 3.5. Model order selection

Parametric VAR model fitting really involves only one parameter: the model order. The most common approach for model order selection involves selecting a model order that minimizes one or more *information criteria* evaluated over a range of model orders. Commonly used information criteria include, Akaike Information Criterion (AIC), Schwarz-Bayes Criterion (SBC) – also known as the Bayesian Information Criterion (BIC) – Akaike's Final Prediction Error Criterion (FPE), and Hannan-Quinn Criterion (HQ). A detailed comparison of these criteria can be found in Chapter 4.3 of (Lütkepohl, 2006). In brief, each criterion is a sum of two terms, one that characterizes the entropy rate or prediction error of the model, and a second term that characterizes the number of freely estimated parameters in the model (which increases with increasing model order). By minimizing both terms, we seek to identify a model that is both parsimonious (does not overfit the data with too many parameters) while also accurately modeling the data. The criteria implemented in SIFT are defined in Table 2.

**Table 2. Information criteria for model order selection implemented in SIFT. Here $\hat{T} = TN$ is the total number of samples (data points) used to fit the model**

| Estimator | Formula |
|---|---|
| **Schwarz-Bayes Criterion (Bayesian Information Criterion)** | $SBC(p) = \ln\left\|\tilde{\Sigma}(p)\right\| + \frac{\ln(\hat{T})}{\hat{T}} pM^2$ |
| **Akaike Information Criterion** | $AIC(p) = \ln\left\|\tilde{\Sigma}(p)\right\| + \frac{2}{\hat{T}} pM^2$ |

| | |
|---|---|
| **Akaike's Final Prediction Error** | $$FPE(p) = \left\|\tilde{\Sigma}(p)\right\| + \left(\frac{\hat{T} + Mp + 1}{\hat{T} - Mp - 1}\right)^{M}$$ and its logarithm (used in SIFT) $$\ln(FPE(p)) = \ln\left\|\tilde{\Sigma}(p)\right\| + M \ln\left(\frac{\hat{T} + Mp + 1}{\hat{T} - Mp - 1}\right)$$ |
| **Hannan-Quinn Criterion** | $$HQ(p) = \ln\left\|\tilde{\Sigma}(p)\right\| + \frac{2\ln(\ln(\hat{T}))}{\hat{T}} pM^2$$ |

For a given information criterion, *IC*, we select the model order that minimizes *IC*:

$$p_{sel} = \arg\min_{p} IC(p)$$

Here, the first term, $\ln\left|\tilde{\Sigma}(p)\right|$ is the logarithm of the determinant of the estimated noise covariance matrix (prediction error) for a VAR model of order $p$ fit to the $M$-channel data, where $TN$ is the total number of datapoints used to fit the model ($T$ samples per trial x $N$ trials). The key difference between the criteria is how severely each penalizes increases in model order (the second term). AIC and SBC are the most widely used criteria, but SBC more heavily penalizes larger model orders. For moderate and large $TN$, FPE and AIC are essentially equivalent (see Lutkepohl (2006) p. 148 for a proof); however, FPE may outperform AIC for very small sample sizes. HQ penalizes high model orders more heavily than AIC but less than SBC. Both SBC and HQ are *consistent* estimators, which means that $\lim_{N\to\infty} \Pr\{p_{sel} = p_{true}\} = 1$. This cannot be said of AIC and FPE. However, under small sample conditions (small $N$), AIC/FPE may outperform SBC and/or HQ in selecting the true model order (Lütkepohl, 2006). When modeling EEG data, it is common for AIC and FPE to show no clear minimum over a reasonable range of model orders. In this case, there may be a clear "elbow" in the criterion plotted as a function of increasing model order, which may suggest a suitable model order.

When selecting a model order for neural connectivity analysis, it is important to consider the dynamics of the underlying physiological system. In particular, one should consider the maximum expected time lag between any two variables included the model. If we have reason to expect a time lag of $\tau$ seconds between any two brain processes, we should make sure to select a model order of $p \geq \tau F_s$ where $F_s$ is the process sampling rate in Hz. Additionally, we should consider that the multivariate spectrum of a $M$-dimensional VAR[$p$] model has $Mp/2$ frequency components (peaks) distributed amongst the $M$ variables (there are $Mp$ complex-conjugate roots of the characteristic equation of the model). This means that we can observe $p/2$ frequency peaks between each pair of variables (Florian and Pfurtscheller, 1995; Schlögl and Supp, 2006). Thus a reasonable lower bound on the model order might be twice the number of expected frequencies plus one (for the zero-Hz peak). Tests performed by Jansen (1981) and Florian and Pfurtscheller (1995) demonstrated that a potentially optimal model order for modeling EEG spectra was p=10, although little

spectral differences were identified for model orders between 9 and 13. A key point, however, is that this was identified for a sampling rate of 128 Hz and it is known that the optimal model order depends significantly on the sampling rate of the process (Zetterberg, 1969).

The principle motivation behind heavy penalization of high model orders in an information criterion is to improve forecasting performance by reducing over-fitting. However, forecasting is not necessarily the ultimate goal of our neural modeling approach. Furthermore, selecting a too-small model order can severely impair our frequency resolution (merging peaks together) as well as our ability to detect coupling over long time lags. Where there is a question as to a suitable model order, it is often better to err on the side of selecting a larger model order. As such, a criterion such as HQ, which often shows a clear minimum but affords intermediate penalization between AIC and SBC may represent an optimal choice for neural data.

In general, it is good practice to select a model order by examining multiple information criteria and combining this information with additional expectations and knowledge specific to the physiological properties of the neural system being analyzed. When possible spectra and coherence obtained from fitted VAR models should be compared with those obtained from non-parametric methods (such as wavelets) to validate the model. Model order selection is often an iterative process wherein, through model validation, we determine the quality of our model fit, and, if necessary, revise our model specification until the data is adequately modeled.

Model order selection is implemented in SIFT using **`pop_est_selModelOrder()`**.

## 3.6. Model Validation

There a number of criteria which we can use to determine whether we have appropriately fit our VAR model. SIFT implements three commonly used categories of tests: (1) checking the residuals of the model for serial and cross-correlation (whiteness tests), (2) testing the consistency of the model, and (3) check the stability/stationarity of the model. These can be accessed through the SIFT GUI using **`pop_est_validateMVAR()`**

### 3.6.1. Checking the whiteness of the residuals

Recall the compact model definition from (Eq 3.2): $X = BZ + U$. Here we can regard the VAR[*p*] model coefficients *B* as a filter which transforms innovations (random white noise), *U*, into observed, structured data *X*. Consequently, for coefficient estimates $\hat{B}$, we can obtain the residuals $\hat{U} = X - \hat{B}Z$. If we have adequately modeled the data, the residuals should be small and uncorrelated (white). Correlation structure in the residuals means there is still some correlation structure in the data that has not been described by our model. Checking the whiteness of residuals typically involves testing whether the residual autocorrelation coefficients up to some desired lag *h* are sufficiently small to ensure that we cannot reject the null hypothesis of white residuals at some desired significance level.

#### *3.6.1.1. Autocorrelation Function (ACF) Test*

The ($M$ x $M$) lag $l$ autocovariance matrix of the residuals is given by $C_l = E[\hat{u}_t \hat{u}'_{t-l}]$. We denote the autocovariances up to lag $l$ as $\mathbf{C}_h = (C_1, \ldots, C_h)$. The lag $l$ autocorrelation matrix is given by $R_l = D^{-1} C_l D^{-1}$ where D is a ($M$ x $M$) diagonal matrix, the diagonal elements being the square root of the diagonal elements of $C_0$. We are generally interested in testing the (white noise) null hypothesis $H_0 : \mathbf{R}_h = (R_1, \ldots, R_h) = 0$ against the alternative $H_1 : \mathbf{R}_h \neq 0$. A simple test, based on asymptotic properties of univariate white noise processes, involves rejecting the hypothesis that $\hat{U}$ is white noise at the 5% level if $|R_l| > \pm 2/\sqrt{\hat{T}}$ for any lag $l$ (excluding the diagonal elements of $R_0$ which are always 1). $\hat{T} = TN$ is the total number of samples used in estimating the covariance. However, since this is a pointwise significance test at the 5% level, in practice we expect one in twenty coefficients to exceed $2/\sqrt{\hat{T}}$ in absolute value even if $\hat{U}$ is white. A reasonable corrected statistic is thus the probability of a coefficient exceeding the 5% significance bounds:

$$\rho = \frac{\text{count}\left(|\mathbf{R}_h| > \pm 2/\sqrt{\hat{T}}\right)}{\text{count}\left(\mathbf{R}_h\right)} = \frac{\text{count}\left(|\mathbf{R}_h| > \pm 2/\sqrt{\hat{T}}\right)}{M^2(h+1) - M}$$

If $\rho < 0.05$, or equivalently $1\text{-}\rho > 0.95$, then we cannot reject the null hypothesis at the 5% level and we accept that the residuals are white.

Due to its simplicity, this sort of test enjoys much popularity. However, it is important to bear in mind that the 5% confidence intervals apply to individual coefficients (i.e., for univariate models) and although the $R_i$ and $R_j$ are asymptotically uncorrelated for $i \neq j$ this is not necessarily true for the elements of $R_i$. As such, this test may be misleading when considering the coefficients of a multivariate model as a group. Additionally, in small sample conditions (small $\hat{T}$), this test may be overly conservative such that the null hypothesis is rejected (residuals indicated as non-white) less often than indicated by the chosen significance level (Lutkepohl, 2006).

#### *3.6.1.2. Portmanteau Tests*

In the previous section, we noted that the simple asymptotic ACF test may yield misleading results when the coefficients are considered independently rather than as a group, derived from a multivariate process. In contrast, *portmanteau* tests are a powerful class of test statistics explicitly derived to test $H_0$ up to some lag $h$. SIFT implements three portmanteau test statistics: *Box-Pierce (BPP)*, *Ljung-Box (LBP),* and *Li-McLeod (LMP)*. Under the null hypothesis, for large sample size and $h$, each of these test statistics approximately follow a $\chi^2$-distribution with $M^2(h-p)$ degrees of freedom. A $\rho$-value can thus be obtained by comparing the test statistic with the c.d.f. of this distribution. If $1\text{-}\rho$ is greater than some value $\alpha$ (e.g., 0.05 for a 5% significance level), we cannot reject the null hypothesis and we accept that the residuals are white. Table 3 lists the three tests implemented in SIFT along with their test statistics and practical notes.

**Table 3. Popular portmanteau tests for whiteness of residuals, implemented in SIFT. Here $\hat{T} = TN$ is the total number of samples used to estimate the covariance**

| Portmanteau Test | Formula (Test Statistic) | Notes |
|---|---|---|
| Box-Pierce (BPP) | $Q_h := \hat{T}\sum_{l=1}^{h} \mathrm{tr}\left(C_l' C_0^{-1} C_l C_0^{-1}\right)$ | The original portmanteau test. Potentially overly-conservative. Poor small-sample properties. |
| Ljung-Box (LBP) | $Q_h := \hat{T}(\hat{T}+2)\sum_{l=1}^{h} (\hat{T}-l)^{-1} \mathrm{tr}\left(C_l' C_0^{-1} C_l C_0^{-1}\right)$ | Modification of BPP to improve small-sample properties. Potentially inflates the variance of the test statistic. Slightly less conservative than LMP with slightly higher (but nearly identical) statistical power. |
| Li-McLeod (LMP) | $Q_h := \hat{T}\sum_{l=1}^{h} \mathrm{tr}\left(C_l' C_0^{-1} C_l C_0^{-1}\right) + \frac{M^2 h(h+1)}{2\hat{T}}$ | Further modification of BPP to improve small-sample properties without variance inflation. Slightly more conservative than LBP. Probably the best choice in most conditions. |

BPP is the classical portmanteau test statistic. It can be shown that in small sample conditions (small $\hat{T}$) its distribution under the null hypothesis diverges from the asymptotic $\chi^2$ distribution. This can render it overly-conservative leading us to reject the null hypothesis of white residuals even when the model was appropriately fit.

The LBP statistic attempts to improve the small-sample properties of the test statistic. By adjusting each covariance coefficient by its asymptotic variance, it can be shown that under the null hypothesis, the LBP statistic has a small-sample distribution much closer to the asymptotic distribution than the BPP statistic. However, it can also be shown that the variance of the LBP statistic can be inflated to substantially larger than its asymptotic distribution.

Like LBP, the LMP statistic has better small-sample properties than BPP. However, unlike LBP, it does so without inflating its variance. Although less popular than LBP, it has been demonstrated that the variance of LMP is closer to its asymptotic variance whereas LBP is more sensitive with significance levels somewhat larger than expected when $\hat{T}$ is large. LMP is slightly conservative but the statistical power for LMP and LBP are nearly identical. Since, in practice, it is preferable to select the more conservative test among tests with comparable power, LMP may represent an ideal choice of test statistic for most applications.

The interested reader should consult (Lutkepohl, 2006) and (Arranz, n.d.)for additional details and references concerning checking the whiteness of residuals. The whiteness of residuals can be tested in SIFT using **`est_checkMVARWhiteness()`**

### 3.6.2. Checking the consistency of the model

To address the question of what fraction of the correlation structure of the original data is captured by our model, we can calculate the *percent consistency* (Ding et al., 2000). We generate an ensemble, of equal dimensions as the original data, using simulated data from the VAR model. For both the real and simulated datasets, we then calculate all auto- and cross-correlations between all variables, up to some predetermined lag. Letting $\mathbf{R_r}$ and $\mathbf{R_s}$ denote the vectorized correlation matrices of the real and simulated data, respectively, the percent consistency index is given by

$$\mathrm{PC} = \left(1 - \frac{\|\mathbf{R}_s - \mathbf{R}_r\|}{\|\mathbf{R}_r\|}\right) \times 100 \quad \text{where } \|\cdot\| \text{ denotes the Euclidean } (L_2) \text{ norm.}$$

A PC value near 100% would indicate that the model is able to generate data that has a nearly identical correlation structure as the original data. A PC value near 0% indicates a complete failure to model the data. While determining precisely what constitutes a sufficiently large PC value is an area for future research, a rule of thumb is that is a value of PC > 85% suggests the model is adequately capturing the correlation structure of the original data. The percent consistency can be calculated in SIFT using **`est_checkMVARConsistency()`.**

### 3.6.3. Checking the stability and stationarity of the model

In section 3.1. we provided a condition for the stability of a VAR[$p$] process. Namely, an $M$-dimensional VAR[$p$] process is stable if all the eigenvalues of the ($Mp$ x $Mp$) augmented coefficient matrix **A** have modulus less than 1. Thus, a useful stability index is the log of the largest eigenvalue $\lambda_{max}$ of **A**:

$$\mathrm{SI} = \ln|\lambda_{\max}|$$

A VAR[$p$] process is stable if and only if SI < 0. The magnitude of the SI can be loosely interpreted as an estimate of the degree to which the process is stable. As mentioned in section 3.1. , a stable process is a stationary process. Thus it is sufficient to test for stability of the model to guarantee that the model is also stationary. If the model is not stable, additional tests such as the Augmented Dickey-Fuller test may be used to separately evaluate the stationarity of the data. However, since we are generally interested modeling *stable* processes, these additional stationarity tests are not implemented in SIFT. The stability index of a fitted model can be calculated in SIFT using **`est_checkMVARStability()`.**

### 3.6.4. Comparing parametric and nonparametric spectra and coherence

Another approach sometimes used to validate a fitted VAR model is to compare the spectra and/or pairwise coherence estimated from the parametric models with those derived from a robust nonparametric approach such as multitapers or wavelets. Using an equation similar to percent consistency, we can estimate the fraction of the nonparametric spectrum or coherence that is captured by our VAR model. Of course, here we assuming the

nonparametric spectra are optimal estimates of the true spectra ("ground truth"), which may not be the case (interestingly, Burg (1967; 1975) demonstrated that, if the data is generated by an AR process and the true model order is known, AR spectral estimation is a maximum-entropy method which means that it represents an optimal spectral estimator). Nevertheless, if the nonparametric quantities are carefully computed, this can be a useful validation procedure. An upcoming release of SIFT will include routines for computing this spectral consistency index.

# 4. Granger Causality and Extensions

Granger causality (GC) is a method for inferring certain types of causal dependency between stochastic variables based on reduction of prediction error of a putative effect when past observations of a putative cause are used to predict the effect, in addition to past observations of the putative effect. The concept was first introduced by Norbert Wiener in 1956 and later reformulated and formalized by C.W. Granger in the context of bivariate linear stochastic autoregressive models (Weiner, 1956; Granger, 1969). The concept relies on two assumptions:

**Granger Causality Axioms**

1. **Causes must precede their effects in time**
2. **Information in a cause's past must improve the prediction of the effect above and beyond information contained in the collective past of all other measured variables (including the effect).**

Assumption (1) is intuitive from basic thermodynamical principles: the arrow of causation points in the same direction as the arrow of time – the past influences the future, but not the reverse. Assumption (2) is also intuitive: for a putative cause to truly be causal, removal of the cause should result in some change in the future of the putative effect – there should be some shared information between the past of the cause and the future of the effect which cannot be accounted for by knowledge of the past of the effect.

The theory and application of GC (and its extensions) to neural system identification has been elaborated in a number of other articles and texts (Kaminski, 1997; Eichler, 2006; Blinowska and Kaminski, 2006; Ding et al., 2006; Schlögl and Supp, 2006; Bressler and Seth, 2010). As such, here we will only briefly introduce the theory and focus primarily on multivariate extensions of the granger-causal concept, including the partial directed coherence (PDC) and direct directed transfer function (dDTF).

## 4.1. Time-Domain GC

Suppose we wish to test whether a measured EEG variable *j* Granger-causes another variable *i* conditioned on all other variables in the measured set. Let *V* represent the set of

all measured variables (e.g., all available EEG sources/channels): $V = \{1, 2, \ldots, M\}$. Our complete (zero-mean) VAR[$p$] model is specified as:

$$x_t^{(V)} = \sum_{k=1}^{p} A_k x_{t-k}^{(V)} + u_t$$

We fit the full model and obtain the mean-square prediction error when $x^{(i)}$ is predicted from past values of $x^{(V)}$ up to the specified model order:

$\mathrm{var}(x_t^{(i)} \mid x_{(\bullet)}^{(V)}) = \mathrm{var}(u_t^{(i)}) = \Sigma_{ii}$ where $x_{(\bullet)}^{(V)} = \{x_{t-k}^{(V)}, k \in \{1,\ldots,p\}\}$ denotes the past of $x^{(V)}$

Now, suppose we exclude $j$ from the set of variables (denoted $V \backslash j$) and re-fit the model

$$x_t^{(V\backslash j)} = \sum_{k=1}^{p} \bar{A}_k x_{t-k}^{(V\backslash j)} + \bar{u}_t$$

and again obtain the mean-square prediction error for $x^{(i)}$.

$$\mathrm{var}(x_t^{(i)} \mid x_{(\bullet)}^{(V\backslash j)}) = \mathrm{var}(\bar{u}_t^{(i)}) = \bar{\Sigma}_{ii}$$

In general, $\Sigma_{ii} \geq \bar{\Sigma}_{ii}$ and $\Sigma_{ii} = \bar{\Sigma}_{ii}$ if and only if the best linear predictor of $x^{(i)}$ based on the full past $x^{(V)}$ does not depend on $x^{(j)}$. This leads us to the following definition for multivariate GC (Eichler, 2006):

**DEFINITION 1**

**Let *I* and *J* be two disjoint subsets of *V*. Then $x^{(J)}$ Granger-causes $x^{(I)}$ conditioned on $x^{(V)}$ if and only if the following two equivalent conditions hold:**

1. $\Sigma_{ii} \gg \bar{\Sigma}_{ii}$
2. $A_{k,ij} \gg 0$ **for some** $k \in \{1,\ldots,p\}$

**Here $\gg$ means "significantly greater than." In other words, inferring conditional GC relationships in the time domain amounts to identifying non-zero elements of a VAR[*p*] coefficient matrix fit to all available variables.**

Granger (1969) quantified DEFINITION 1 for strictly bivariate processes in the form of an F-ratio:

$$F_{ij} = \ln\left(\frac{\bar{\Sigma}_{ii}}{\Sigma_{ii}}\right) = \ln\left(\frac{\mathrm{var}(x_t^{(i)} \mid x_{(\bullet)}^{(i)})}{\mathrm{var}(x_t^{(i)} \mid x_{(\bullet)}^{(i)}, x_{(\bullet)}^{(j)})}\right) \qquad \text{(Eq 4.1)}$$

Here, $F_{ij}$ denotes the GC from process *j* to process *i*. This quantity is always non-negative and increases away from zero proportionate to the degree to which the past of process *j* conditionally explains ("granger-causes") the future of process *i*.

## 4.2. Frequency-Domain GC

In the frequency domain a very similar definition holds for GC as in the time domain. If we obtain the Fourier-transform of our VAR[*p*] coefficient matrices *A(f)* as in section 3.3. , based on the time-domain definition of GC we can derive the following definition for GC in the frequency-domain (Eichler, 2006):

**DEFINITION 2**

**Let *I* and *J* be two disjoint subsets of *V*. Then $x^{(I)}$ Granger-causes $x^{(J)}$ conditioned on $x^{(V)}$ if and only if the following condition holds:**

$A_{ij}(f) \gg 0$ **for some frequency *f***

DEFINITION 2 suggests a simple method for testing multivariate (conditional) GC at a given frequency *f*: we simply test for non-zero coefficients of $|A(f)|$. This approach yields a class of GC estimators known as Partial Directed Coherence (PDC) measures (Baccalá and Sameshima, 2001).

A slightly different approach, due to Granger (1969) and later refined by Geweke (1982), provides an elegant interpretation of frequency-domain GC as a decomposition of the total spectral interdependence between two series (based on the bivariate spectral density matrix, and directly related to the coherence) into a sum of "instantaneous", "feedforward" and "feedback" causality terms. However, this interpretation was originally derived only for bivariate processes and, while this has been recently been extended to trivariate (and block-trivariate) processes (Chen et al., 2006; Wang et al., 2007), it has not yet been extended to the true multivariate case. An implementation of the Granger-Geweke formulation for bivariate processes is provided in SIFT as the "GGC" connectivity estimator. The interested reader should consult (Ding et al., 2006) for an excellent tutorial on the Granger-Geweke approach.

There is a direct relationship between bivariate time-domain and frequency-domain GC. If $F_{ij}$ is the time-domain GC estimator ((Eq 4.1) and $W(f)_{ij}$ is the frequency-domain Granger-Geweke estimator, then the following equivalency holds:

$$F_{ij} = \int_0^{F_s/2} W(f)_{ij}\, df$$

It is unknown whether a similar equivalency exists for other multivariate GC estimators, such as the PDC and dDTF. However, in practice, integrating these estimators over a range of frequencies provides a simple way to obtain a general time-domain representation of the estimator.

## 4.3. A partial list of VAR-based spectral, coherence and GC estimators

Table 4 contains a list of the major spectral, coherence, and GC/information flow estimators currently implemented in SIFT. Each estimator can be derived from the quantities $S(f), A(f), H(f),$ and $\Sigma$ obtained in section 3.3. , with the exception of the renormalized PDC (rPDC). The rPDC requires estimating the [$(Mp)^2$ x $(Mp)^2$] inverse cross-covariance matrix of the VAR[$p$] process. SIFT achieves this using an efficient iterative algorithm proposed in (Barone, 1987) and based on the doubling algorithm of (Anderson and Moore, 1979). These estimators and more can be computing using the SIFT's functions **`pop_est_mvarConnectivity()`** or the low-level function **`est_mvtransfer()`**.

**Table 4. A partial list of VAR-based spectral, coherence, and information flow / GC estimators implemented in SIFT.**

| | Estimator | Formula | Primary Reference and Notes |
|---|---|---|---|
| Spectral M. | Spectral Density Matrix | $S(f) = X(f)X(f)^* = H(f)\Sigma H(f)^*$ | (Brillinger, 2001)<br><br>$S_{ii}(f)$ is the spectrum for variable *i*. $S_{ij}(f) = S_{ji}(f)^*$ is the cross-spectrum between variables *i* and *j*. |
| Coherence Measures | Coherency | $C_{ij}(f) = \frac{S_{ij}(f)}{\sqrt{S_{ii}(f)S_{jj}(f)}}$<br>$0 \le \lvert C_{ij}(f)\rvert^2 \le 1$ | (Brillinger, 2001)<br><br>Complex quantity. Frequency-domain analog of the cross-correlation. The magnitude-squared coherency is the *coherence* $Coh_{ij}(f) = \lvert C_{ij}(f)\rvert^2$. The phase of the coherency can be used to infer lag-lead relationships, but, as with cross-correlation, this should be treated with caution if the coherence is low, or if the system under observation may be open-loop. |
| | Imaginary Coherence (iCoh) | $iCoh_{ij}(f) = \mathrm{Im}(C_{ij}(f))$ | (Nolte et al., 2004)<br><br>The imaginary part of the coherency. This was proposed |

| | | | |
|---|---|---|---|
| | | | as a coupling measure invariant to linear instantaneous volume-conduction. $iCoh_{ij}(f) > 0$ only if the phase lag between *i* and *j* is non-zero, or equivalently, $0 < \text{angle}(C_{ij}(f)) < 2\pi$ |
| | Partial Coherence (pCoh) | $P_{ij}(f) = \frac{\hat{S}_{ij}(f)}{\sqrt{\hat{S}_{ii}(f)\hat{S}_{jj}(f)}}$<br>$\hat{S}(f) = S(f)^{-1}$<br>$0 \le \lvert P_{ij}(f)\rvert^2 \le 1$ | (Brillinger, 2001)<br>The partial coherence between *i* and *j* is the remaining coherence which cannot explained by a linear combination of coherence between *i* and *j* and other measured variables. Thus, $P_{ij}(f)$ can regarded as the *conditional* coherence between *i* and *j* with respect to all other measured variables. |
| | Multiple Coherence (mCoh) | $G_i(f) = \sqrt{1 - \frac{\det(S(f))}{S_{ii}(f)\mathbf{M}_{ii}(f)}}$<br>$\mathbf{M}_{ii}(f)$ is the **minor** of $S(f)$ obtained by removing the i[th] row and column of $S(f)$ and returning the determinant. | (Brillinger, 2001)<br>Univariate quantity which measures the total coherence of variable *i* with all other measured variables. |
| **Partial Directed Coherence Measures** | Normalized Partial Directed Coherence (PDC) | $\pi_{ij}(f) = \frac{A_{ij}(f)}{\sqrt{\sum_{k=1}^{M} \lvert A_{kj}(f)\rvert^2}}$<br>$0 \le \lvert \pi_{ij}(f)\rvert^2 \le 1$<br>$\sum_{j=1}^{M} \lvert \pi_{ij}(f)\rvert^2 = 1$ | (Baccalá and Sameshima, 2001)<br>Complex measure which can be interpreted as the conditional granger causality from *j* to *i* normalized by the total amount of causal outflow from *j*. Generally, the magnitude-squared PDC $\lvert \pi_{ij}(f)\rvert^2$ is used. |
| | Generalized PDC (GPDC) | $\bar{\pi}_{ij}(f) = \frac{\frac{1}{\Sigma_{ii}} A_{ij}(f)}{\sqrt{\sum_{k=1}^{M} \frac{1}{\Sigma_{ii}^2} \lvert A_{kj}(f)\rvert^2}}$<br>$0 \le \lvert \bar{\pi}_{ij}(f)\rvert^2 \le 1$<br>$\sum_{j=1}^{M} \lvert \bar{\pi}_{ij}(f)\rvert^2 = 1$ | (Baccalá and Sameshima, 2007)<br>Modification of the PDC to account for severe imbalances in the variance of the innovations. Theoretically provides more robust small-sample estimates. As with PDC, the squared-magnitude $\lvert \bar{\pi}_{ij}(f)\rvert^2$ is typically used |

| | | | |
|---|---|---|---|
| | | | |
| | Renormalized PDC (rPDC) | $\lambda_{ij}(f) = Q_{ij}(f) * V_{ij}(f)^{-1} Q_{ij}(f)$<br>where<br>$Q_{ij}(f) = \begin{pmatrix} \mathrm{Re}[A_{ij}(f)] \\ \mathrm{Im}[A_{ij}(f)] \end{pmatrix}$ and<br>$V_{ij}(f) = \sum_{k,l=1}^{p} R_{jj}^{-1}(k,l)\Sigma_{ii} Z(2\pi f,k,l)$<br>$Z(\omega,k,l) = \begin{pmatrix} \cos(\omega k)\cos(\omega l) & \cos(\omega k)\sin(\omega l) \\ \sin(\omega k)\cos(\omega l) & \sin(\omega k)\sin(\omega l) \end{pmatrix}$<br>$\boldsymbol{R}$ is the $[(Mp)^2 \times (Mp)^2]$ covariance matrix of the VAR[$p$] process (Lütkepohl, 2006) | (Schelter et al., 2009)<br>Modification of the PDC. Non-normalized PDC is renormalized by the inverse covariance matrix of the process to render a scale-free estimator (does not depend on the unit of measurement) and eliminate normalization by outflows and dependence of statistical significance on frequency. To our knowledge SIFT is the first publically available toolbox to implement this estimator. |
| **Directed Transfer Function Measures** | Normalized Directed Transfer Function (DTF) | $\gamma_{ij}(f) = \frac{H_{ij}(f)}{\sqrt{\sum_{k=1}^{M} \lvert H_{ik}(f)\rvert^2}}$<br>$0 \le \lvert\gamma_{ij}(f)\rvert^2 \le 1$<br>$\sum_{j=1}^{M} \lvert\gamma_{ij}(f)\rvert^2 = 1$ | (Kaminski and Blinowska, 1991; Kaminski et al., 2001)<br>Complex measure which can be interpreted as the total information flow from $j$ to $i$ normalized by the total amount of information inflow to $i$. Generally, the magnitude-squared DTF $\lvert\gamma_{ij}(f)\rvert^2$ is used and, in time-varying applications the DTF should not be normalized. |
| | Full-Frequency DTF (ffDTF) | $\eta_{ij}^2(f) = \frac{\lvert H_{ij}(f)\rvert^2}{\sum_f \sum_{k=1}^{M} \lvert H_{ik}(f)\rvert^2}$ | (Korzeniewska, 2003)<br>A different normalization of the DTF which eliminates the dependence of the denominator on frequency allowing more interpretable comparison of information flow at different frequencies. |
| | Direct DTF (dDTF) | $\delta_{ij}^2(f) = \eta_{ij}^2(f) P_{ij}^2(f)$ | (Korzeniewska, 2003)<br>The dDTF is the product of the ffDTF and the pCoh. Like the PDC, it can be interpreted as frequency-domain conditional GC. |

| | | | |
|---|---|---|---|
| Granger-Geweke | Granger-Geweke Causality (GGC) | $F_{ij}(f) = \frac{\left(\Sigma_{jj} - (\Sigma_{ij}^2 / \Sigma_{ii})\right)\left\|H_{ij}(f)\right\|^2}{S_{ii}(f)}$ | (Geweke, 1982; Bressler et al., 2007)<br><br>For bivariate models ($M$ = 2), this is identical to Geweke's 1982 formulation. However, it is not yet clear how this extends to multivariate models ($M$ > 2). |

## 4.4. Time-Frequency GC

In section 3.4. we discussed using adaptive VAR models to model nonstationary time series. These methods allow us to obtain a sequence of time-varying VAR coefficient matrices. A time-frequency representation of the spectrum, coherence or information-flow/GC can thus easily be obtained by computing one or more of the estimators in Table 4 for each coefficient matrix. Figure 2 shows an example of a time-frequency image of dDTF information flow between two neural processes. Each column of the image corresponds to the dDTF "spectrum" at a given point in time.

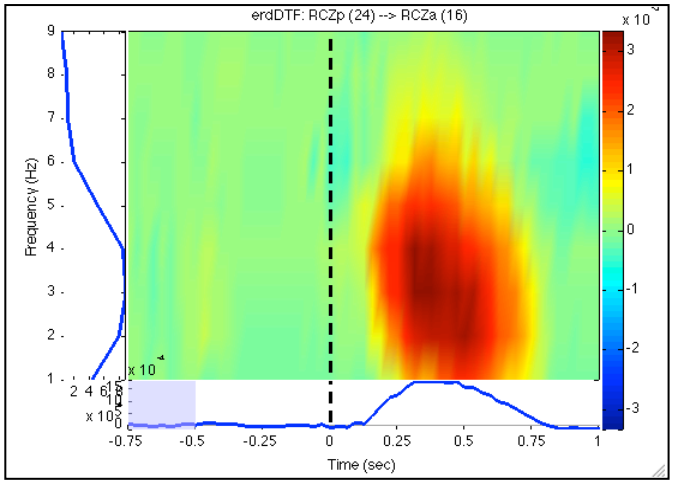


**Figure 2. A time-frequency image showing the dDTF between two processes for a selected range of frequencies and times. Frequency is on the y-axis and Time on the x-axis. Red (blue) indicates more (less) information flow, relative to a baseline period (purple shaded region).**

## 4.5. (Cross-) correlation does not imply (Granger-) causation

An important result of the definition of granger causality is that it provides a much more stringent criterion for causation (or information flow) than simply observing high correlation with some lag-lead relationship. A common approach for inferring information flow is to compute the cross-correlation (or cross-partial-correlation) between two variables for a range of time lags and determine whether there exists a peak in the correlation at some non-zero lag. From this we might infer that the leading variable "causes"

– or transmits information to – the lagged variable. However, using such an approach to infer causation, or even a direction of information flow, can be quite misleading for several reasons.

Firstly, the cross-correlation is a symmetric measure and is therefore unsuitable for identifying lag-lead relationships in systems with feedback (closed-loop systems) (Chatfield, 1989). It is currently understood that many neural systems exhibit feedback, albeit potentially on a large enough time scale that they system may appear locally open-loop.

Secondly, even if the system under observation is open-loop, a clear peak in the cross-correlation at some non-zero lag would satisfy Assumption 1 of GC (causes must precede effects in time) but not Assumption 2 (the past of a cause must share information with the future of the effect that cannot be explained by the past of all other measured variables, including the effect). In this regard it is fundamentally different than GC. As it turns out, the ability for GC to test Assumption 2 is what makes it such a powerful tool for causal inference, in contrast to simple correlative measures.

To illustrate: suppose we are observing two ants independently following a pheromone trail towards some tasty morsel. Ant 1 started the journey two minutes before Ant 2 and so he appears to be "leading" Ant 2. If we compute the cross-correlation between the two ants' trajectories for a range of time lags we would find a high correlation between their trajectories and, furthermore, we would find the correlation was peaked at a non-zero lag with Ant 1 leading Ant 2 by a lag of two minutes. But it would be foolish to say that Ant 1 was "causing" the behavior of Ant 2. In fact, not only is there no causal relationship whatsoever between the two, but there is not even any information being transmitted between the two ants. They are conditionally independent of each other, given their own past history and given the fact that each is independently following the pheromone trail (this is the "common (exogenous) cause" that synchronizes their behavior). If we were to intervene and remove Ant 2 (Ant 1), Ant 1 (Ant 2) would continue on his way, oblivious to the fact that his comrade is no longer in lock-step with him. Consequently, if we calculate the Granger-causality between the two trajectories we will find that the GC is zero in both directions: there is no information in the history of either ant that can help predict the future of the other ant above and beyond the information already contained in each ant's respective past.

Because the spectral coherence is simply the Fourier transform of the cross-correlation (and therefore the frequency-domain representation of the cross-correlation), the same limitations hold for coherence as for cross-correlation regarding inference of directionality of information flow or causation. Namely, using the phase of coherence to infer directionality of information flow in some frequency (as is often done in the neuroscience community) may be highly misleading if there is even moderate feedback in the system (or if the coherence is low). Coherence is not necessarily a measure of information flow, but rather correlation between two processes at a particular frequency (a useful analogy here, similar to the ants, is to consider two pendulums on opposite sides of the globe swinging in synchrony at the same frequency, with one pendulum started ¼ cycle before the other – their behavior is coherent, but is there information flow between them?). In contrast, frequency-domain extensions of Granger-causality condition on the past history of the processes and, assuming all relevant variables have been included in the model, can correctly distinguish between such spurious forms of information flow or causation and "true" information flow.