1 Main tools for synchronization analysis

In this chapter, the main tools for the analysis of synchronization between signals will be presented, with a focus on their mathematical definition and their role in the application of *Intebrain synchronization*.
After dealing with the problem of *what* synchronization actually means in this context and which measures *should not* be considered, several valid tools will be proposed: **cross-correlation**, **peak-synchronization**, **angular similarity**, L^2 **error analysis**. Each one of these tools has a specific purpose and meaning, and they will all be considered in the analysis of this work, in order to give different shapes to the concept of synchronization and to try to study it under the largest possible perspective.
Finally, when observing the presence of correlation between two signals, a different (although strongly connected) question may arise: "which is the *relationship* of such connection?", or, in other words, "is one signal determining the behaviour of the other?". To answer such questions, a sophisticated tool will be object of study: the **Granger causality**.

1.1 Understanding synchronization

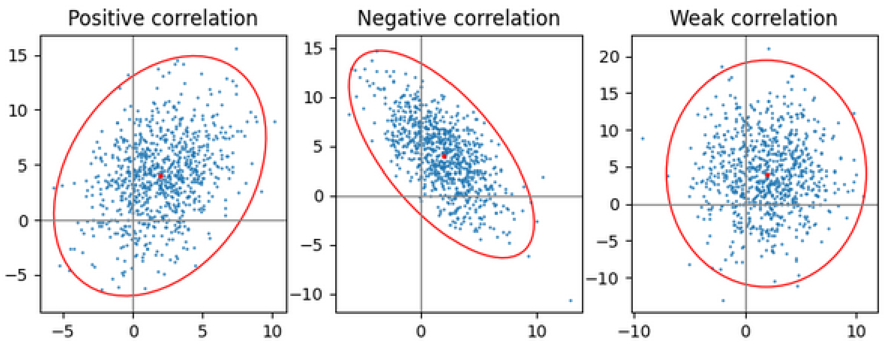

Figure 1: *Geometrical interpretation of the Pearson correlation: such coefficient gives a measure of how well the pairs obtained from the two samples distribute linearly along the diagonal of an ellipse covering the scatterplot.*

One of the main goals of this work will be to perform a data analysis on the Interbrain synchronization of neural signals, recorded in behavioural tasks. In order to do so, first one has to clarify the meaning of the term *synchronization*. In a statistical sense, the synchronization between two time series can be seen as a measure of the **correlation** between them: the more two series are correlated, the more similar and connected they will be. This leads inevitably to the fact that, when talking about synchronization, the definition can't be unique, since the similarity between two series can be assumed in different ways, depending on the particular aspect of interest.
However, a first hint about what to look for or not can be found in the type of signal which is under analysis. In the present case, dealing with single neuron

measurements of the intracellular concentration of $[Ca^{2+}]$ means dealing with *strongly nonlinear and unstable signals*, which exhibit sudden peaks and anre often chaotic and difficult to predict.
For this reason, the most commonly adopted measure of correlation, namely the **Pearson correlation**, is not suited to describe such signals. Given two random samples $\mathbf{x} = \{x_i\}_{i=1}^N$ and $\mathbf{y} = \{y_i\}_{i=1}^N$, the Pearson correlation (PC) between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\rho(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}$$

where:

- $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ is the sample mean
- $Cov(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$ is the sample covariance
- $\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$ is the sample standard deviation

The PC measures the *linear correlation* between two variables. This implies that this type of correlation should be used to inspect a linear relationship between two variables, which have a distribution close to the Gaussian one, and a uniform variance (i.e. reduced presence of outliers) [Applied multivariate statistics book by Johnson]. Unfortunately, with the type of data incoming from the neural recordings, all these hypoteses fail, and it follows that other correlation measures should be inspected instead.

1.2 Cross-correlation

When investigating the similarity between two time-dependent signals, the first tool to be considered is the **cross-correlation**.
Formally, given two functions $f = f(t)$ and $g = g(t)$, we define the cross-correlation between them as

$$[f(t) \star g(t)](\tau) = \int_{-\infty}^{+\infty} f(\tau) g(t + \tau) dt \tag{1}$$

Given that the **convolution** between such functions is defined as

$$[f(t) * g(t)](t) = \int_{-\infty}^{+\infty} f(t) g(t - \tau) d\tau \tag{2}$$

it follows that

$$[f(t) \star g(t)](t) = [f(-t) * g(t)](t) \tag{3}$$

This means that the cross-correlation coincides with a convolution in which one function is considered backward in time. Moreover, in the common form of cross-correlation, as shown in (eq.1), the resulting quantity is not expressed as a function of the time variable t, but as a function of τ, i.e. the *lag* or *delay* between the signals. The interpretation is straightforward: when, for a given value of the delay τ, simultaneous peaks of the two signals are both present, the contribution of their product in the integral will be more relevant for the computed value of cross-correlation correspondent to that specific lag.

As a consequence of this, once computing the cross-correlation between two time-dependent functions, one can identify its maximum value and retrieve the corresponding value of the lag τ, from which it is possible to obtain an estimation of the delay between the two functions. To summarize, a cross-correlation analysis allows to:

1. Compute the cross-correlation between the two signals as a function of the lag value τ
2. Estimate the real delay between such signals, observing when the peak of cross-correlation occurs

When the cross-correlation is computed between the same function, it takes the name of **autocorrelation**. Confronting two identical signals, there will always be a peak of correlation (equal to 1 using *normalized* cross-correlation) corresponding to the lag $\tau = 0$. Moreover, the shape of the cross-correlation will be always symmetrical (see Figure).

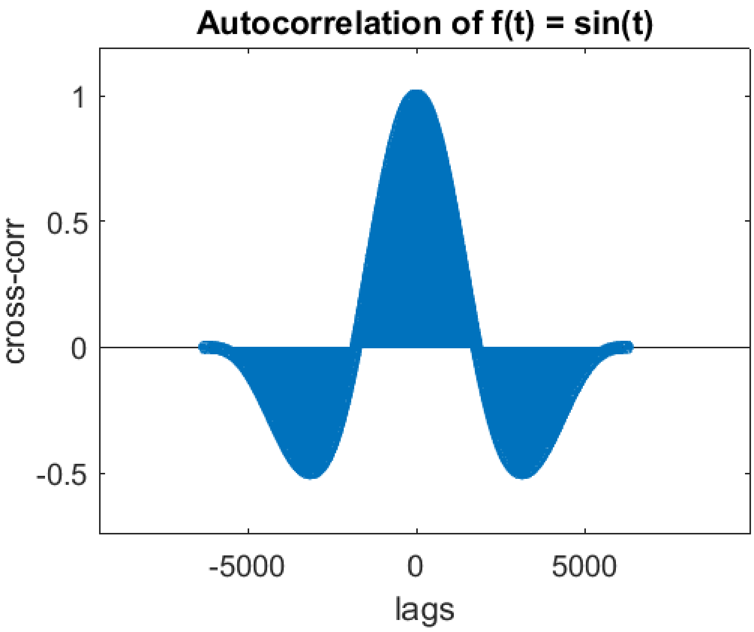

Figure 2: *Aurocorrelation of the function* $f(t) = \sin(t)$

While the autocorrelation represents an ideal case, when dealing with two different functions $f(t)$ and $g(t)$, the significance of their cross-correlation will be given by the amount of shared properties to the case of autocorrelation; however, depending on the current application, it can be reasonable to expect a peak value in correspondence to a nonzero value of the lag, as a representation of an actual physical delay.

Given two time series $\mathbf{x} = \{x_i\}_{i=1}^{N}$ and $\mathbf{y} = \{y_i\}_{i=1}^{N}$, the discrete approximation of (eq.1) reads:

$$[\mathbf{x} \star \mathbf{y}](m) = \sum_{n=1}^{N-1} x_n y_{n+m} \qquad\qquad m_{min} < m < m_{max}$$

where m is the approximate lag value, chosen in an appropriate interval.

1.3 Peak-synchronization

Cross-correlation is a general and widely used tool to quantify the similarity between two signals evolving in time. With the following concept of **peak-synchronization**, the focus of the analysis is restricted to the characteristic type of signal of this work: recording of intracellular calcium activity.
As shown in Chapter 1, a typical recording of the activity of a single neuron is characterized by the presence of rapid and intense *peaks*, which define the neuron as *active*. It follows that a way to intend synchronization between neurons could be related to the presence of *close* or *simultaneous* peaks observed at the same time. In other words, two signals (hence, neurons) are synchronized, under the peak-synchronization point of view, if a *pattern* of simultaneous firing occurs.

In order to quantify the peak-synchronization between two signals, the following steps have to be faced:

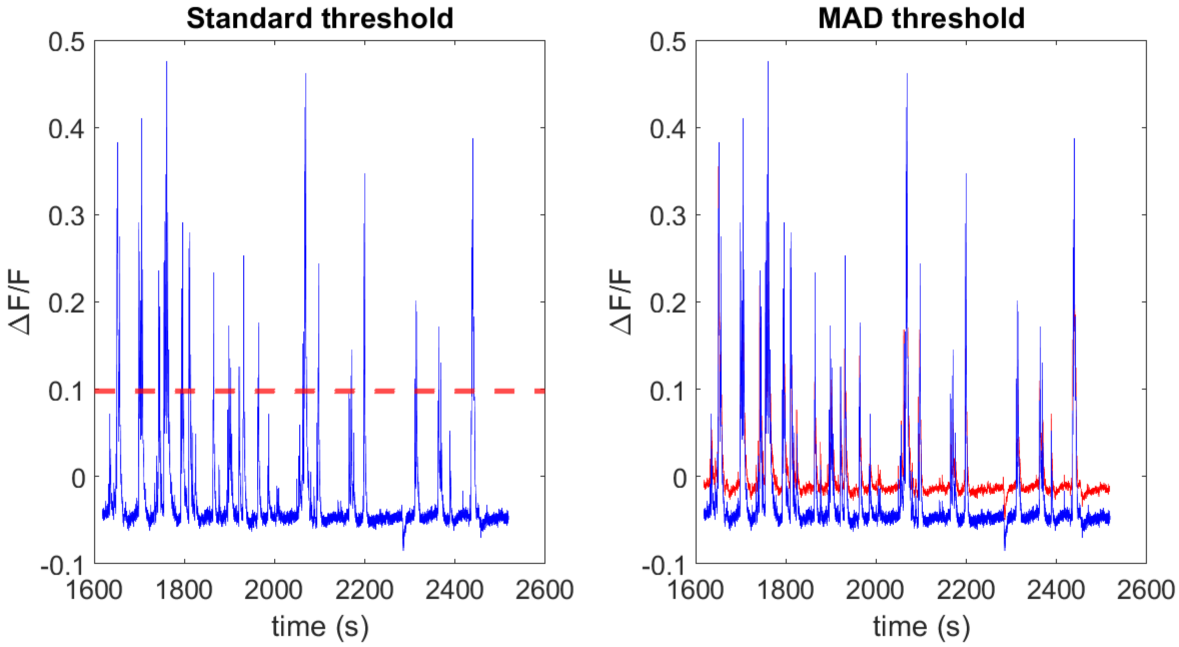

Figure 3: *Example of a calcium signal (blue). The red curves are the thresholds built from the two algorithms: standard threshold (left panel) and MAD threshold (right panel)*

1. Identify when a neuron is *active*, i.e. when we are in presence of a peak

2. Choose an appropriate time interval in which two neurons firing simultaneously can be considered as synchronized

3. Quantify the correlation between the two neurons

The way to solve all of these three steps is not unique, but it is strictly depending on the current biological application under study and by the choice of appropriate algorithms for the activity detection and synchronization quantification. In this work, the following considerations and choices have been made:

1. Given a discrete time series, several algorithms are available or could be designed to detect, when a peak is occurring. A straightforward idea is to establish a threshold for the activity, in such a way that all the points above the treshold are active and all the ones below non active. Therefore, the output of such algorithm is of binary type (1 for activity, 0 for non-activity).

   A naive way to define the threshold could be to consider a horizontal line, for example given by the equation $y = \mu + 2\sigma$, where μ and σ are the mean and the standard deviation of the signal. In this way, all the values higher than y will be considered active and vice versa. This approach is good enough when dealing with "well-behaving" signals, presenting a baseline low activity alternated by huge peaks, however it seems to fail when dealing with more complex and noisy signals.

Algorithm 1 Standard threshold algorithm

```
Consider a horizontal threshold given by $y = \mu + 2\sigma$
Confront every point of the signal with each corresponding point of the threshold
Every point above the threshold is labeled as 1, all the points below as 0
```

   For this reason, a different algorithm can be considered as well [Inscopix manual]: the **MAD threshold** algorithm. The threshold established by this algorithm is not constant, but it "follows" the signal, in order to better capture its dynamics.

Algorithm 2 MAD threshold algorithm

```
Start from a baseline threshold given by $MAD = median(X_i - median(X))$
Identify the points where the slope changes from positive to negative (PN) and from negative to positive (NP)
At every PN point, the threshold value is the MAD value plus the previous NP point's value
The overall threshold is obtained from linear interpolation of the threshold points
Every point above the threshold is labeled as 1, all the points below as 0
```

2. A typical time interval in which neuronal firing occurs usually strongly depends on the specific case. However, in general it is safe to say that the

peak of a neuron usually has a duration of $250 - 750$ ms. Such value will define the reference time window.

3. Finally, once the binary vectors of activations are available, as well as a reference time window, to quantify their synchronization, a common tool adopted is the **Peak-correlation index** [Cutts and Eglen], defined as

$$i_{AB} = \frac{N_{AB}T}{2N_A N_B dT}$$

Here T is the overall signal time window, dT is the synchronization time window, N_A is the number of peaks in signal A, N_B is the number of peaks in signal B and finally $N_{AB} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} I_{[-dT,dT]}(|a_i - b_j|)$ is the sum of simultaneous peaks within each synchronization window.

The peak correlation index gives a representation of how well two signals (neurons) are peak-synchronized. However, it should be noticed that such measure is not normalized, meaning that the value of one index considered by itself has no real meaning, and this measure should be used only as tool to compare the same pair of signals through different phases.

1.4 Angular distance and L^2 distance

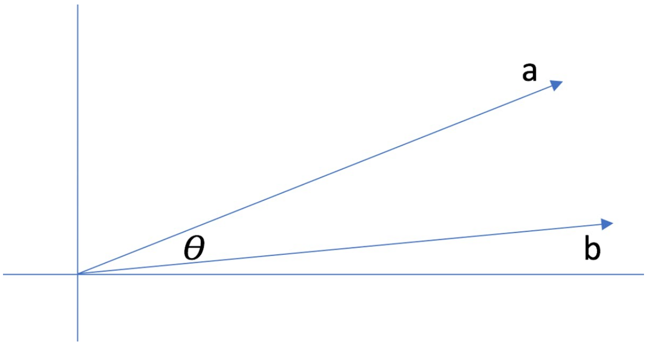

Figure 4: *Angle between two vectors in the euclidean space (2D case)*

Besides the two main tools for the synchronization analysis, cross-correlation and peak-synchronization, other concepts of correlation and similarity can be investigated as well. Here, two supplementary measures are considered: the **angular distance** and the L^2 **distance**.

The purpose of the angular distance analysis is to determine the similarity between two signals in a *geometric sense*. The idea is indeed rather simple: project the signals on a euclidean space and determine the angle between them. The two signals are closer, and thus more similar, if the angle between them is small. A way to determine the angle θ between two vectors is to compute first the **cosine similarity**

$$Sim_C(\mathbf{x}, \mathbf{y}) = \cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||_E ||\mathbf{y}||_E}$$

where $\mathbf{x} \cdot \mathbf{y}$ denotes the scalar product between $\mathbf{x}$ and $\mathbf{y}$ and $|| \cdot ||_E$ denotes the euclidean norm.

Then, the cosine of the angle, the angular distance can be retrieved as

$$d_\theta(\mathbf{x}, \mathbf{y}) = \frac{\arccos(Sim_C(\mathbf{x}, \mathbf{y}))}{\pi}$$

Where the actual angle is divided by a reference π angle.
As for the L^2 error, between two continuous in time signals $f = f(t)$ and $g = g(t)$ defined on an interval $[t_1, t_2]$, the L^2 distance between them is defined as

$$||f - g||_2 = \int_{t_1}^{t_2} |f(t) - g(t)|^2 dt$$

When dealing with time discrete signals $\mathbf{x} = \{x_i\}_{i=1}^N$ and $\mathbf{y} = \{y_i\}_{i=1}^N$, this quantity it actually coincides with the **mean squared error (MSE)**

$$MSE(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|^2$$

The MSE analysis performs a point-by-point comparison of the two signals, penalizing quadratically the differences between the two. It follows that in order to have a realistic estimate of this quantity, it is necessary to confront well-aligned signals. To this purpose, as described in Section 2.2, a cross-correlation analysis can be helpful, since it allows to identify the *delay* between two time series, detected in correspondence of the cross-correlation peak. In conclusion, an alignment based on such value of lag between the two time signals should be performed before computing the L^2 distance.

1.5 Granger causality

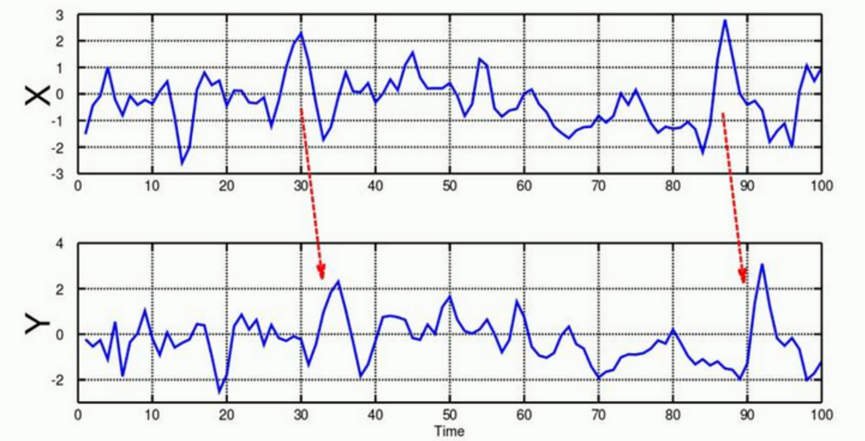

Figure 5: *Example of signal X Granger-predicting signal Y*

As already mentioned in Section 1.5, observing a correlation between two time series is not enough to establish a *relationship* between them. Indeed,a further

indicator of the synchronization between signals is the presence of an underlying *cause-effect* mechanism underlying it all. If such relationship is found, besides observing a synchronization, one can determine also which signal (and, in this case, mouse) is responsible for causing the opponent's one.
In situations like the EEG experiment in [Novembre et al.] discussed in Chapter 1, such relationshipis investigated through a manipulations on the experimental level (via *Multi-brain stimulation*). In the current work, once dealing with data already measured, a different and sophisticated approach is adopted,based on statistical principles: the **Granger causality**.
The Granger causality (**G-causality**) is a method aimed to identify causal relationship between time series data. The word "casuality" is mainly due to historical reasons, since many debates about its correctness are still ongoing, and it would be probably more precise to refer to it as "Granger *prediction*".

Given two time series $\mathbf{x} = \{x_i\}_{i=1}^N$ and $\mathbf{y} = \{y_i\}_{i=1}^N$, the G-causality method is based on the following scheme:

1. Generation of a **vector autoregression model (VAR)** of order p, where $p \geq 1$ is an integer to be determined, for one of the two time series, considering its previous values

$$\hat{x}_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} \tag{4}$$

2. Generation of a second model, in which the values of the second series are added to model (4)

$$\hat{x}_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + b_1 y_{t-1} + \cdots + b_p y_{t-p} \tag{5}$$

3. Comparison between the two models: if model 2 is more significant than model 1 (in a way to be clarified in the following), then signal **y** *Granger-predicts* signal **x**

More formally [Barnett-Seth], given two stochastic processes $\mathbf{X} = X_{i=1}^N$ and $\mathbf{Y} = Y_{i=1}^N$, process Y *does not* G-cause process X if X, conditional its past, is independent by the past of Y. A vector autoregressive model for a process U takes the form

$$\mathbf{U}_t = \sum_{k=1}^{p} A_k \mathbf{U}_{t-k} + \varepsilon_t \tag{6}$$

where p is the **order** of the model, $\{A_k\}k = 1^p$ are the **regression coefficients** and ε_t the **residuals**, assumed normally and independently distributed. The **residual covariance matrix** of the model is defined as $\Sigma = Cov(\varepsilon_t)$ and it is assumed to be stationary. The process U can then be identified both as X and Y. Given a VAR model of the form (6), the **autocovariance sequence** $\{\Gamma_k\}_{k=1}^p$ is defined as $\Gamma_k = Cov(\mathbf{U}_t, \mathbf{U}_{t-k})$, and it is possible to relate this quantity to the autoregression coefficients $\{A_k\}$ thanks to the **Yule-Walker** equations [Anderson,1971]

$$\Gamma_k = \sum_{i=1}^{p} A_i \Gamma_{k-i} + \delta\Sigma \qquad k = 1, \cdots, p \tag{7}$$

Standard VAR theory [Hamilton-Lutkephol] requires the condition $\sum_{k=1}^{N} ||A_k||^2 < \infty$. Moreover, defining the **characteristic polynomial** as

$$\phi_A(z) = \det\left(I - \sum_{k=1}^{p} A_k z^k\right)$$

it must be that the **spectral radius** $\rho(A) := \max_{\phi_A(z)=0} |z|^{-1}$ is strictly less than 1, as a *stability* condition.

Considering now a process in which $\mathbf{U}_t = \begin{bmatrix}\mathbf{X}_t \\ \mathbf{Y}_t\end{bmatrix}$, its VAR formulation reads

$$\mathbf{U}_t = \sum_{k=1}^{p} \begin{bmatrix} A_{xx,k} & A_{xy,k} \\ A_{yx,k} & A_{yy,k} \end{bmatrix} \mathbf{U}_{t-k} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix} \tag{8}$$

and its residual covariance is $\Sigma = Cov\left(\begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$.
This augmented formulation contains both the regression models for process X and Y. For example, its first component reads

$$\mathbf{X}_t = \sum_{k=1}^{p} A_{xx,k}\mathbf{X}_{t-k} + \sum_{k=1}^{p} A_{xy,k}\mathbf{Y}_{t-k} + \varepsilon_{x,t} \tag{9}$$

If the process Y does not G-cause the process X, it follows that the coefficients $\{A_{xy,k}\}_{k=1}^{p}$ are all equal to 0, and the model becomes

$$\mathbf{X}_t = \sum_{k=1}^{p} A'_{xx,k}\mathbf{X}_{t-k} + \varepsilon'_{x,t} \tag{10}$$

Therefore, a statistic test checking the null hypothesis $\{H_0 : Y \text{ does not G-predicts } X\}$ has the form

$$H_0 : A_{xy,1} = A_{xy,2} = \cdots = A_{xy,p} = 0 \tag{11}$$

If $\Sigma'_{xx} = Cov(\varepsilon'_{x,t})$ is the residual covariance matrix of model (10), standard theory [Edwards, 1992] suggests the use of the **(log-)likelihood statistics** to obtain a *maximum-likelihood* estimator of the G-causality between Y and X (here referred as $\mathcal{F}_{Y\to X}$):

$$\mathcal{F}_{Y\to X} = \ln\frac{det(\Sigma'_{xx})}{det(\Sigma_{xx})} \tag{12}$$

Since the determinant of a covariance matrix (i.e. the *generalized variance*) quantifies the *prediction error* of its regression model, the interpretation of (11) is that the G-causality statistics $\mathcal{F}_{Y\to X}$ is a measure of how much the prediction error is reduced when also the process Y is included in the regression model. Clearly, this same procedure applies to the statistics $\mathcal{F}_{X\to Y}$, in which the directionality of the relationship is inverted. It can be proven [Wilks & Wald] that, under the null hypothesis, $(N-p)\mathcal{F}_{Y\to X} \sim \chi^2(d)$, where $d = pN^2$.
To summarize, the typical workflow for G-causality estimation, consists in the following steps:

1. Estimate the model order p via appropriate criterion (such as AIC and BIC)

2. Estimate the autocovariance sequence Γ_k and the VAR coefficients (A_k, Σ) through the Yule-Walker equations (7), both for reduced and augmented models. Verify that $\sum_{k=1}^{N} ||A_k||^2 < \infty$ and $\rho(A) < 1$

3. Computation of the G-statistics $\mathcal{F}_{Y\to X}$ and $\mathcal{F}_{X\to Y}$ through (12)

4. Test the significance of the statistical tests against the null hypotesis, computing the p -values of the two tests