Granger Causality

Moo K. Chung
University of Wisconsin-Madison
mkchung@wisc.edu

1 Introduction

Causality is a fundamental concept in science, aiming to determine whether one event influences another. In time series analysis, **Granger causality**, introduced by Clive Granger in 1969, provides a statistical framework for assessing whether past values of one time series help predict another. Unlike traditional notions of causality in physics or philosophy, Granger causality is based on predictability if including past values of one time series improves the forecast of another. Granger causality has wide applications in economics, neuroscience, climate science, and engineering. In a *multivariate time series*, where multiple variables evolve over time, Granger causality extends naturally to examine interactions among multiple time-dependent signals. This is most often used causal model in brain network analysis. Granger causality is usually built on top of Vector Autoregressive (VAR) models.

2 Vector Autoregressive (VAR) Model

The **Vector Autoregressive (VAR)** model is a fundamental framework for modeling multivariate time series, where multiple variables evolve over time and potentially influence each other. Unlike univariate autoregressive models that describe the evolution of a single time series, the VAR model extends this concept to a system of interdependent time series. This makes it particularly useful for capturing dynamic relationships in fields such as economics, neuroscience, and climate science.

Consider a system of N time series denoted as

$$X(t) = [X_1(t), X_2(t), \ldots, X_N(t)]^\top,$$

where $X_i(t)$ represents the time series of the i-th region at time t. The VAR model of order P, denoted as VAR(P), expresses each time series as a linear combination of its own past values and the past values of all other variables in the system. $X(t) \in \mathbb{R}^N$ is a time-varying vector. This model is formulated as

$$X(t) = \sum_{p=1}^{P} A_p X(t-p) + \varepsilon(t), \tag{1}$$

where A_p are $N \times N$ coefficient matrices that capture the influence of past values at lag p, and $\varepsilon(t)$ is an N-dimensional vector of error terms, assumed to be normally distributed with zero mean and a covariance matrix Σ, i.e.,

$$\varepsilon(t) \sim \mathcal{N}(0, \Sigma).$$

The number of lags P is a crucial hyperparameter that determines how far back in time the model considers past information. Expanding each component, we can write

$$X_i(t) = \sum_{p=1}^{P} \sum_{j=1}^{N} A_{ij}^{(p)} X_j(t-p) + \varepsilon_i(t), \tag{2}$$

where $A_{ij}^{(p)}$ represents the effect of the j-th variable at lag p on the i-th variable. Note $A_p = (A_{ij}^{(p)})$. If $A_{ij}^{(p)}$ is significantly different from zero, this indicates a directional relationship where past values of $X_j(t)$ contribute to predicting $X_i(t)$, which is the foundation for testing Granger causality.

The estimation of VAR model parameters is typically performed using the least squares estimation to each equation in the system. Given a dataset of T time points, the model parameters are estimated by minimizing the sum of squared residuals

$$\widehat{A} = \arg\min_{A} \sum_{t=P+1}^{T} \|X(t) - \sum_{p=1}^{P} A_p X(t-p)\|^2. \tag{3}$$

The Vector Autoregressive (VAR) model requires past values of the time series $X(1), \ldots, X(P)$ to make predictions. The selection of the optimal lag order P is an essential step in model fitting. Choosing too small a value may result in missing important dependencies, whereas an excessively large P may introduce overfitting. The lag order is typically determined using information criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which balance model complexity with goodness of fit.

Despite its flexibility and interpretability, the VAR model assumes stationarity and its statistical properties of the time series remain constant over time. Additionally, VAR models assume linear relationships, which may not fully capture nonlinear dependencies present in brain network data.

The VAR framework provides a powerful foundation for analyzing multivariate time series and forms the basis for testing Granger causality. By capturing interactions among multiple variables, it serves as an essential tool for investigating directional dependencies and understanding the underlying structure of complex dynamic systems.