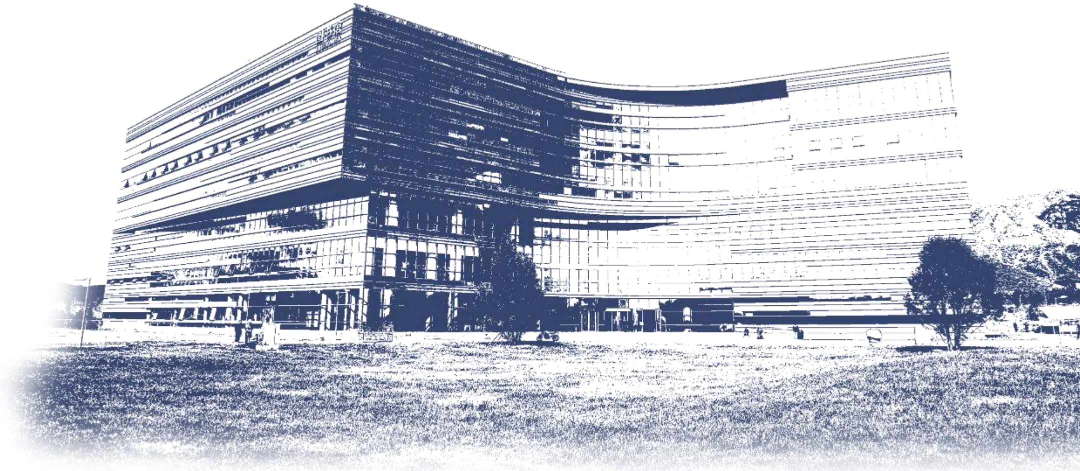
北京化工大学
BEIJING UNIVERSITY OF CHEMICAL TECHNOLOGY

网络理论和时间序列因果算法在数据分析中的应用



2020.07.16

万晓耕

2020年数学建模国赛培训

1. 网络的构建以及相关测度

1.1 Basic concepts of networks

1.2 Adjacency matrix

1.3 Weighted networks

1.4 Directed networks

1.5 Centrality measures

2. 时间序列因果算法与数据分析

2.1 Time series models

2.2 Granger causality(GC)

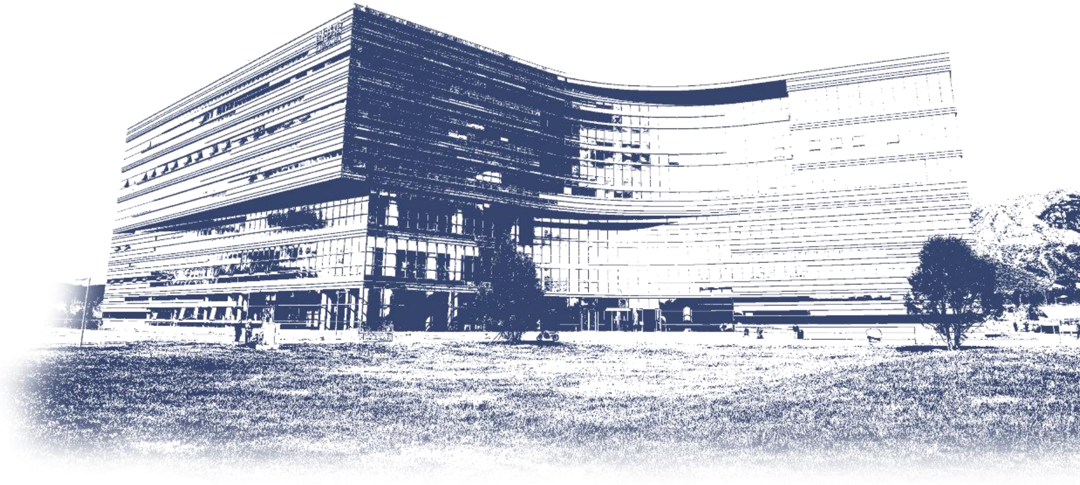
2.3 Partial directed coherence(PDC)

2.4 Mutual information rate (MIR)

2.5 Transfer entropy (TE)

3. 算法与实例

3.1 PDC算法应用实例



1.1 Basic concepts of networks

In this section, we introduce the basic concepts that used to describe and analyze networks, most of which come from graph theory, the branch of mathematics that deals with networks.

A network—also called a *graph* in the mathematical literature—is, as we have said, **a collection of vertices joined by edges**. Vertices and edges are also called ***nodes* and *links*** in computer science, ***sites* and *bonds* in physics**, and *actors* and *ties* in sociology.

Here, **we normally denote the number of vertices in a network by *n* and the number of edges by *m***. Most of the networks have at most a **single edge** between any pair of vertices. In the rare cases where there can be more than one edge between the same pair of vertices we refer to those edges collectively as a ***multiedge***. In most of the networks, there are also **no edges** that connect vertices to themselves, although such edges will occur in a few instances. Such edges are called ***self-edges*** or ***self-loops***. A network that has neither self-edges nor multiedges is called **a *simple network*** or ***simple graph***. A network with multiedges is called a ***multigraph***.

Figure 1.1 shows examples of (a) a simple graph and (b) a non-simple graph having both multiedges and self-edges.

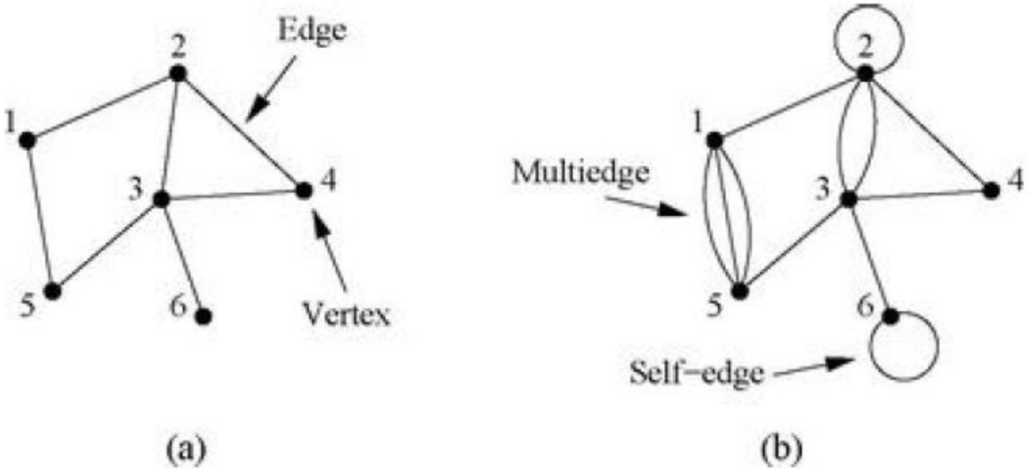


Figure 1: Two small networks. (a) A simple graph, i.e., one having no multiedges or self-edges. (b) A network with both multiedges and self-edges.

1.2 The adjacency matrix

There are a number of different ways to represent a network mathematically. **Consider an undirected network with *n* vertices and let us label the vertices with integer labels 1 . . . *n***, as we have, for instance, for the network in Fig. 1.1(a). It does not matter which vertex gets which label, only that each label is unique, so that we can use the labels to refer to any vertex unambiguously. **If we denote an edge between vertices *i* and *j* by (*i*,*j*) then the complete network can be specified by giving the value of *n* and a list of all the edges.**

For example, the network in Fig. 1.1(a) has *n* = 6 vertices and edges (1,2), (1,5), (2,3), (2,4), (3,4), (3,5), and (3,6). Such a specification is called an ***edge list***. Edge lists are sometimes used to store the structure of networks on computers.

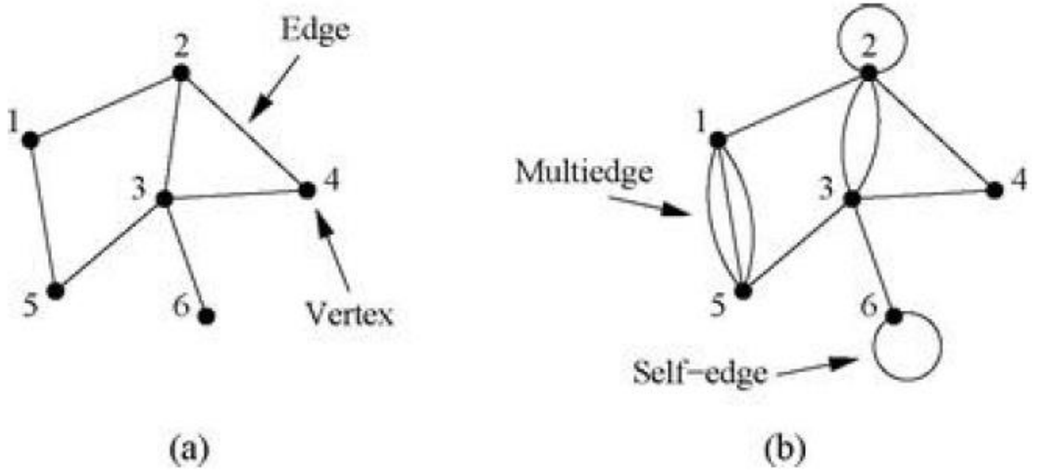


Figure 1.1: Two small networks. (a) A simple graph, i.e., one having no multiedges or self-edges. (b) A network with both multiedges and self-edges.

1 网络的构建以及相关测度

A better representation of a network for present purposes is the ***adjacency matrix***. The adjacency matrix **A** of a simple graph is the matrix with elements $\boldsymbol{A_{ij}}$ such that

$$A_{ij} = \begin{cases} 1 & if\ there\ is\ an\ edge\ between\ vertices\ i\ and\ j, \\ 0 & otherwise. \end{cases} \tag{1.1}$$

For example, the adjacency matrix of the network in Fig. 1(a) is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{1.2}$$

1 网络的构建以及相关测度

Two points to notice about the adjacency matrix are that, **first, for a network with no self-edges the diagonal matrix elements are all zero, and second that it is symmetric, since if there is an edge between *i* and *j* then there is an edge between *j* and *i*.**

It is also possible to represent multiedges and self-edges using an adjacency matrix. A multiedge is represented by setting the corresponding matrix element A_{ij} equal to the multiplicity of the edge. For example, a double edge between vertices *i* and *j* is represented by $A_{ij} = A_{ji} = 2$.

1 网络的构建以及相关测度

A single self-edge from vertex *i* to itself is represented by setting the corresponding diagonal element A_{ii} of the matrix equal to 2. This is because non-self-edges appear twice in the adjacency matrix—an edge from *i* to *j* means that both A_{ij} and A_{ji} are 1. To count edges equally, self-edges should also appear twice, and since there is only one diagonal matrix element A_{ii}, we need to record both appearances. One can also have multiple self-edges (or “multi-self-edges” perhaps). Such edges are represented by setting the corresponding diagonal element of the adjacency matrix equal to twice the multiplicity of the edge.

1 网络的构建以及相关测度

To give an example, the adjacency matrix for the multigraph in Fig. 1.1(b) is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 3 & 0 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix}. \qquad (1.3)$$

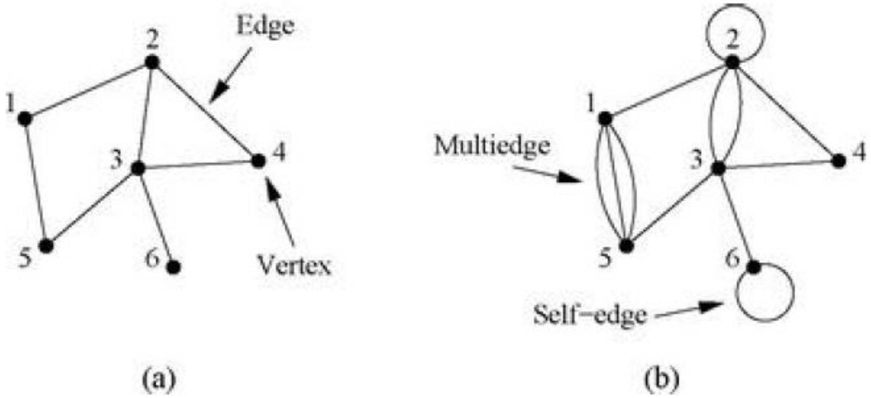


Figure 1.1: Two small networks. (a) A simple graph, i.e., one having no multiedges or self-edges. (b) A network with both multiedges and self-edges.

1.3 WEIGHTED NETWORKS

In some situations, it is useful to represent edges as having a strength, weight, or value to them, usually a real number. Such *weighted* or *valued networks* can be represented by giving the elements of the adjacency matrix values equal to the weights of the corresponding connections.

The adjacency matrix

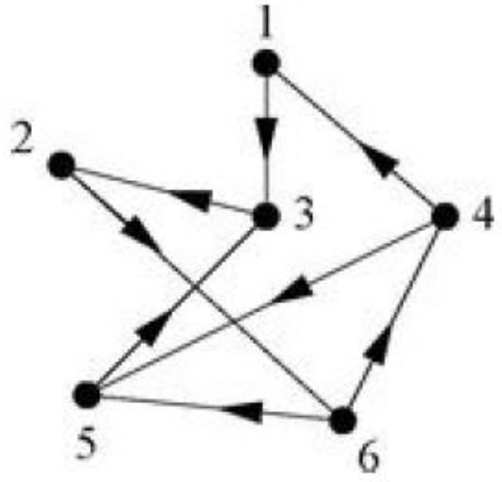
$$A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 0 & 0.5 \\ 1 & 0.5 & 0 \end{pmatrix} \tag{1.4}$$

represents a weighted network in which the connection between vertices 1 and 2 is twice as strong as that between 1 and 3, which in turn is twice as strong as that between 2 and 3.

1.4 DIRECTED NETWORKS

A ***directed network*** or ***directed graph***, also called a ***digraph*** for short, is a network in which each edge has a direction, pointing *from* one vertex *to* another. Such edges are themselves called *directed edges*, and can be represented by lines with arrows on them—see Fig. 1.2.

Figure 1.2: A directed network. A small directed network with arrows indicating the directions of the edges.



1 网络的构建以及相关测度

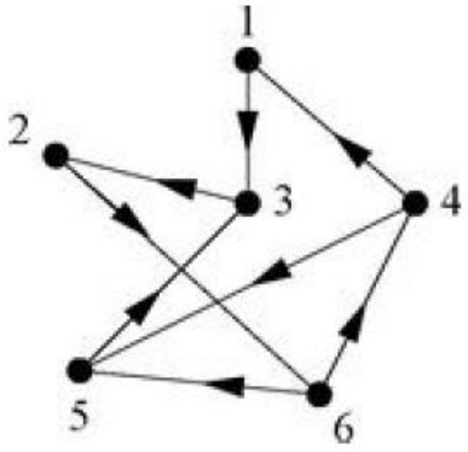


Figure 1.2: A directed network. A small directed network with arrows indicating the directions of the edges.

The adjacency matrix of a directed network has matrix elements

$$A_{ij} = \begin{cases} 1 & \textit{if there is an edge from } j \textit{ to } i, \\ 0 & \textit{otherwise}. \end{cases} \tag{1.5}$$

Notice the direction of the edge here—it runs *from* the second index *to* the first. As an example, the adjacency matrix of the small network in Fig. 1.2 is

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1.6}$$

Note that this matrix is not symmetric. In general the adjacency matrix of a directed network is asymmetric.

1.5 Centrality measures

In this section, we look at some of centrality measures, which can be applied to the analysis of network data from a variety of fields.

1.5.1 DEGREE CENTRALITY

A large volume of research on networks has been devoted to the concept of *centrality*. **The simplest centrality measure in a network is just the degree of a vertex, the number of edges connected to it.** Degree is sometimes called ***degree centrality***, to emphasize its use as a centrality measure. In directed networks, vertices have both an in-degree and an out-degree, and both may be useful as measures of centrality in the appropriate circumstances. Although degree centrality is a simple centrality measure, it can be very illuminating.

1.5.2 EIGENVECTOR CENTRALITY

A natural extension of the simple degree centrality is ***eigenvector centrality***. We can think of degree centrality as awarding one “centrality point” for every network neighbor a vertex has. But not all neighbors are equivalent. In many circumstances a vertex’s importance in a network is increased by having connections to other vertices that are *themselves important*. This is the concept behind eigenvector centrality. Instead of awarding vertices just one point for each neighbor, **eigenvector centrality gives each vertex a score proportional to the sum of the scores of its neighbors**.

The centrality x_i of vertex i is proportional to the sum of the centralities of i's neighbors:

$$x_i = k_1^{-1} \sum_j A_{ij} x_j \,, \tag{1.7}$$

where the k_1 is the largest eigenvalue of **A**. This equation gives the eigenvector centrality the nice property that it can be large either because a vertex has many neighbors or because it has important neighbors (or both).

网络的构建以及相关测度

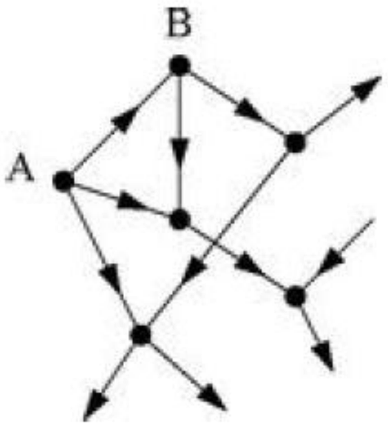


Figure 1.3: A portion of a directed network. Vertex A in this network has only outgoing edges and hence will have eigenvector centrality zero. Vertex B has outgoing edges and one ingoing edge, but the ingoing one originates at A, and hence vertex B will also have centrality zero.

In theory eigenvector centrality can be calculated for either undirected or directed networks. It works best however for the undirected case. In directed networks, the adjacency matrix are asymmetric, there are two sets of eigenvectors, the left eigenvectors and the right eigenvectors, and hence two leading eigenvectors. **In most cases the correct eigenvector centrality uses the right eigenvector.**

The other problem for directed networks is as shown in Fig. 1.3. In Fig 1.3, Vertex A has only outgoing edges and no incoming ones. Such a vertex will always have centrality zero because there are no terms in the sum in Eq. (1.7). This might not seem to be a problem: perhaps a vertex that no one points to *should* have centrality zero. But then consider vertex B, which has one ingoing edge, but that edge originates at vertex A, and hence B also has centrality zero, because the one term in its sum in Eq. (1.7) is zero.

1.5.3 KATZ CENTRALITY

One solution to the problems encountered with ordinary eigenvector centrality in directed networks is to give each vertex a small amount of centrality “for free,” regardless of its position in the network or the centrality of its neighbors. In other words, we define

$$x_i = \alpha \sum_j A_{ij} x_j + \beta \tag{1.8}$$

where α and β are positive constants. The first term is the normal eigenvector centrality term in which the centralities of the vertices linking to i are summed, and the second term is the “free” part, the constant extra term that all vertices receive. By adding this second term, even vertices with zero in-degree still get centrality β, and once they have a non-zero centrality, then the vertices they point to derive some advantage from being pointed to. **This means that any vertex that is pointed to by many others will have a high centrality**.

1 网络的构建以及相关测度

In matrix terms, Eq. (1.8) can be written

$$x = \alpha A x + \beta 1, \tag{1.9}$$

where **1** is the vector (1, 1, 1 ...). Rearranging for $\boldsymbol{x}$, we find that $\boldsymbol{x} = \beta(\boldsymbol{I} - \alpha\boldsymbol{A})^{-1} \cdot \mathbf{1}$. As we have said, **we normally don’t care about the absolute magnitude of the centrality, only about which vertices have high or low centrality values**, so the overall multiplier β is unimportant. For convenience we usually set $\beta = 1$, giving

$$x = (\boldsymbol{I} - \alpha\boldsymbol{A})^{-1} \cdot \mathbf{1}, \tag{1.10}$$

This centrality measure was first proposed by Katz in 1953 and we will refer to it as the ***Katz centrality***.

1.5.4 PAGERANK

The Katz centrality of the previous section has one feature that can be undesirable. If a vertex with high Katz centrality points to many others then those others also get high centrality. A high centrality vertex pointing to one million others gives all one million of them high centrality. This problem can be solved by defining a variation on the Katz centrality in which the centrality is proportional to the Katz centrality *divided by the out-degree of the node*.

In mathematical terms this centrality is defined by

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta \qquad (1.11)$$

This gives problems however if there are vertices in the network with out-degree $k_i^{out} = 0$. If there are any such vertices then the first term in Eq. (1.11) is indeterminate—it is equal to zero divided by zero (because $A_{ij} = 0$ for all i). This problem is easily fixed. In fact, we could set k_i^{out} to any nonzero value and the calculation would give the same answer.

1 网络的构建以及相关测度

In matrix terms, Eq. (1.11), is then

$$x = \alpha AD^{-1}x + \beta 1, \tag{1.12}$$

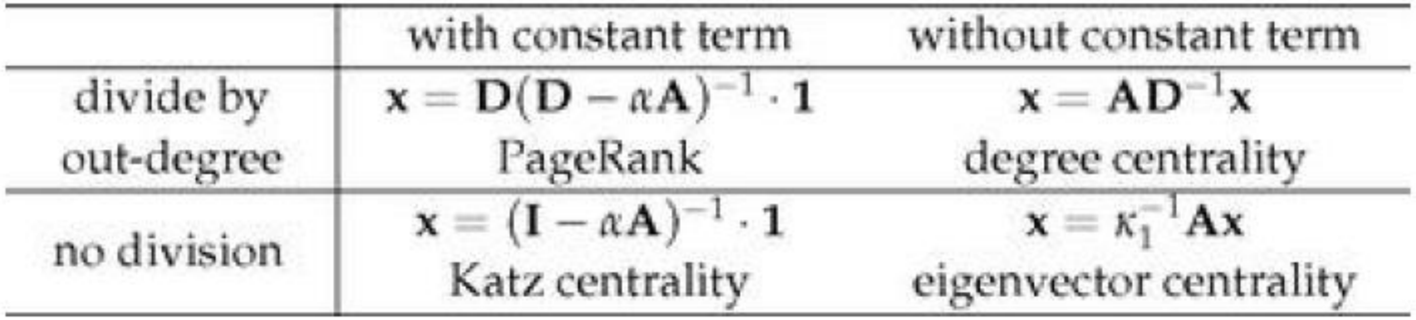
with **1** being again the vector (1, 1, 1, ...) and **D** being the diagonal matrix with elements. Rearranging, we find that $\boldsymbol{x} = \beta(\boldsymbol{I} - \alpha \boldsymbol{AD}^{-\mathbf{1}})^{-1} \cdot 1$, and thus, as before, β plays the role only of an unimportant overall multiplier for the centrality. Conventionally we set $\beta = 1$, giving

$$x = (I - \alpha AD^{-1})^{-1} \cdot 1 = D(\boldsymbol{D} - \alpha \boldsymbol{A})^{-1} \cdot 1 \tag{7.13}$$

This centrality measure is commonly known as ***PageRank***.

1 网络的构建以及相关测度

In Table 1.1 we give a summary of the different matrix centrality measures we have discussed, organized according to their definitions and properties. If you want to use one of these measures in your own, **eigenvector centrality and PageRank are probably the two measures to focus on initially**.



	with constant term	without constant term
divide by out-degree	$\mathbf{x} = \mathbf{D}(\mathbf{D} - \alpha\mathbf{A})^{-1} \cdot \mathbf{1}$ PageRank	$\mathbf{x} = \mathbf{A}\mathbf{D}^{-1}\mathbf{x}$ degree centrality
no division	$\mathbf{x} = (\mathbf{I} - \alpha\mathbf{A})^{-1} \cdot \mathbf{1}$ Katz centrality	$\mathbf{x} = \kappa_1^{-1}\mathbf{A}\mathbf{x}$ eigenvector centrality

Table 1.1: Four centrality measures. Note that the diagonal matrix **D**, which normally has elements $D_{ii} = k_i$, must be defined slightly differently for PageRank, as $D_{ii} = \max(1, k_i)$—see Eq. (1.11) and the following discussion.

1 网络的构建以及相关测度

Reference:

[1] M.E.J. Newman, Networks, An introduction, Oxford University Press, UK, 2010.

2.1 Time series models

In this section, we introduce some classic time series models that play an important role in causality analysis.

White noise process

Denote $\{X_t\}$ as a sequence of **uncorrelated** random variables such that $\forall t$

$$E(X_t) = \mu; Var(X_t) = \sigma^2,$$

whose auto-covariances and autocorrelations are

$$s_\tau = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \text{ and } \rho_\tau = \begin{cases} 1 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

This pure random process is known as the **white noise process**, where $\{X_t\}$ **can be assigned with different distributions e.g. the Gaussian distribution and the exponential distribution to obtain different realizations.**

Moving average process

A q −th order moving average process $\{X_t\}$ can be expressed by

$$X_t = \mu - \theta_{0,q}\epsilon_t - \theta_{1,q}\epsilon_{t-1} - \cdots - \theta_{q,q}\epsilon_{t-q} = \mu - \sum_{j=0}^{q}\theta_{j,q}\epsilon_{t-j}\,, \quad (2.1)$$

Where μ and $\theta_{j,q}$ are constants ($\theta_{0,q} = -1, \theta_{q,q} \neq 0, \theta_{j,q} < \infty, j = 1, \cdots, q$), $\boldsymbol{\{\epsilon_t\}}$ **is a zero-mean white noise process** with variance σ_ϵ^2. Without loss of generality, assume $\boldsymbol{E(X_t) = \mu = 0}$, the auto-covariance for $\tau \geq 0$ is given by

$$s_\tau = Cov(X_t, X_{t+\tau}) = \sum_{j=0}^{q}\sum_{k=0}^{q}\theta_{j,q}\theta_{k,q}E(\epsilon_{t-j}\epsilon_{t+\tau-k})$$

$$= \begin{cases} \sigma_\epsilon^2\sum_{j=0}^{q-\tau}\theta_{j,q}\theta_{j+\tau,q}\;(k = j + \tau) & 0 \leq \tau \leq q \\ 0 & \tau > q \end{cases}$$

where $\boldsymbol{E(\epsilon_t\epsilon_{t+\tau}) = 0, \forall\tau \neq 0}$. Thus $\{X_t\}$ is a stationary process with unconditionally stationary (i.e. independent to t) auto-covariance sequence

$$s_\tau = \begin{cases} \sigma_\epsilon^2\sum_{j=0}^{q-|\tau|}\theta_{j,q}\theta_{j+|\tau|,q}\;(k = j + \tau) & |\tau| \leq q \\ 0 & |\tau| > q \end{cases}$$

Autoregressive process

A p-th order autoregressive process $\{X_t\}$ can be expressed by

$$X_t = \phi_{1,p}X_{t-1} + \phi_{2,p}X_{t-2} + \cdots + \phi_{p,p}X_{t-p} + \epsilon_t, \qquad (2.2)$$

where $\phi_{1,p}, \phi_{2,p}, \cdots, \phi_{p,p}(\phi_{p,p} \neq 0)$ are constants and $\{\epsilon_t\}$ is a zero-mean white noise process with variance σ_ϵ^2. Autoregressive processes are not unconditional stationary, it requires $\{\phi_{k,p}\}$ to satisfy certain conditions to be stationary.

Autoregressive moving average process

Autoregressive moving average (ARMA) process is a combination of the autoregressive and the moving average processes. A (p,q)-th order ARMA process $\{X_t\}$ can be expressed by

$$X_t = \phi_{1,p}X_{t-1} + \phi_{2,p}X_{t-2} + \cdots + \phi_{p,p}X_{t-p} + \epsilon_t - \theta_{0,q}\epsilon_t - \theta_{1,q}\epsilon_{t-1} - \cdots - \theta_{q,q}\epsilon_{t-q}, \text{(2.3)}$$

where $\phi_{j,p}$ *'s* and $\theta_{j,q}$ *'s* are constants coefficients $(\phi_{p,p} \neq 0, \theta_{q,q} \neq 0)$ *and* $\{\epsilon_t\}$ is a zero mean white noise process with variance σ_ϵ^2. ARMA processes need to satisfy certain conditions to be stationary.

Partial spectrum

Recall in time series analysis, the (square integrable) auto-spectrum $S(f)$ of a univariate stationary processes $\{X_t\}$ is the **Fourier transform of the auto-covariance sequence $\{\boldsymbol{s_\tau}\}$**

$$S(\omega) = \sum_{\tau=-\infty}^{\infty} s_\tau e^{-i2\pi f\tau}, \omega \in [-\pi, \pi). \qquad (2.4)$$

Let $\{X_{1,t}, X_{2,t}, \cdots, X_{K,t}\}$ be K zero mean stationary vector processes with

$$S_{X_j}(\omega) = \sum_{\tau=-\infty}^{\infty} s_{X_j,\tau} e^{-i\omega\tau}; |f| \leq \frac{1}{2}, j = 1,2,\cdots,K, \qquad (2.5)$$

Assuming the square summability of the cross-covariance sequence. **The auto-spectrum $\{\boldsymbol{S_{X_j}(\omega)}\}$ (resp. the cross-spectra $\boldsymbol{S_{X_jX_k}(\omega)}$) and the auto-covariances $\{\boldsymbol{s_{X_j,\tau}}\}$ (resp. the corss-covariances $\{\boldsymbol{s_{X_j,X_{k,\tau}}}\}$) form a Fourier transform pair.**

For the K-variate real-valued discrete time stationary process $\{X_{1,t}, X_{2,t}, \cdots, X_{K,t}\}$, the partialized process $\{\eta_{j,n}\}$ associated to X_j is defined as $\eta_{j,n} = X_{j,n} - E\left[X_{j,n} \middle| \{X_{l,m}, l \neq j, m \in \mathbb{Z}\}\right]$, which consists of the residues of the projection of X_j onto the past, the future and the present of the remaining processes.

2 时间序列因果算法与数据分析

For the K-variate real-valued discrete time stationary process $\{X_{1,t}, X_{2,t}, \cdots, X_{K,t}\}$, the partialized process $\{\eta_{j,n}\}$ associated to X_j is defined as

$$\eta_{j,n} = X_{j,n} - E\left[X_{j,n} \middle| \{X_{l,m}, l \neq j, m \in \mathbb{Z}\}\right],$$

which consists of the residues of the projection of X_j onto the past, the future and the present of the remaining processes.

2 时间序列因果算法与数据分析

The auto-spectrum of $\eta_{j,n}$ associated to X_j is defined as the partial spectrum of X_j given X^j:

$$S_{\eta_j\eta_j}(\omega) = S_{X_jX_j}(\omega) - s_{X_jX^j}(\omega) S^{-1}_{X^jX^j}(\omega) s_{X^jX_j}(\omega), \qquad (2.6)$$

where $X^j = [X_{l_1} \cdots X_{l_{K-1}}]^T, \{l_1 \cdots l_{K-1}\} = \{1, \cdots, K\} \backslash \{j\}, s_{X_jX^j}(\omega)$ is the K-1 dimensional vector made up of the cross-spectra between X_j and the remaining K-1 processes and $S_{X^jX^j}(\omega)$ is the spectral density matrix of $\{X_{l_1} \cdots X_{l_{K-1}}\}$.

In this partial spectrum,

$$g_{jj}(\omega) = s_{X_jX^j}(\omega) S^{-1}_{X^jX^j}(\omega), \qquad (2.7)$$

constitutes an optimum Wiener filter whose role in producing $\boldsymbol{\eta_k}$ is to deduct the influence of the other variables from $\boldsymbol{X_j}$ to single out the contribution that is originating from only $\boldsymbol{X_j}$. The partial spectrum plays an important role in developing frequency domain causality measures.

2.2 Granger causality

Granger causality, introduced by Prof. Clive Granger, is an original concept of causality measures. Granger causality is a bivariate measure and it depends on linear autoregressive models. Consider a p-th order bivariate linear autoregressive model of X and Y:

$$X_t = \sum_{j=1}^{p} a_{XX,j} X_{t-j} + \sum_{j=1}^{p} a_{XY,j} Y_{t-j} + E_{X,t}, \qquad (2.8)$$

$$Y_t = \sum_{j=1}^{p} a_{YX,j} X_{t-j} + \sum_{j=1}^{p} a_{YY,j} Y_{t-j} + E_{Y,t},$$

Where $A_j = \begin{bmatrix} a_{XX,j} & a_{XY,j} \\ a_{YX,j} & a_{YY,j} \end{bmatrix}$ denotes the coefficient matrix for the autoregressive model (j=1,2,…,p), E_X and E_Y are residuals (prediction errors) for each time series.

2 时间序列因果算法与数据分析

If the variance of E_X (or E_Y) is reduced by the inclusion of Y (or X) terms in the first (or second) equation, then Y (or X) is said to **Granger causes** X (or Y), i.e. **Y Granger causes X if the coefficients in A_{xy} are jointly significantly from zero**. This can be tested by performing an F-test on the null hypothesis that $a_{xy} = 0$, assuming X and Y are covariance stationary. The magnitudes of Granger causality can be estimated by the logarithm of corresponding F-statistic. The order p of the autoregressive model can be selected by criteria such as the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC).

2.3 Partial directed coherence (PDC)

Partial directed coherence (PDC) is a frequency domain causality measure, which uses the **partial spectrum** to define the causalities in frequency domain. Consider a K-dimensional zero-mean stationary vector process $X_n = [X_{1,n}, \dots, X_{K,n}]^T$ represented by a multivariate autoregressive model

$$X_n = \sum_{l=1}^{\infty} A_l X_{n-l} + \epsilon_n \tag{2.9}$$

$\epsilon_n = [\epsilon_{1,n}, \dots, \epsilon_{K,n}]^T$ is a zero-mean Gaussian stationary innovation vector process with positive definite covariance matrix $\Sigma_\epsilon = E[\epsilon_n \epsilon_n^T]$.

2 时间序列因果算法与数据分析

Assuming the existence of the spectral density functions, a matrix $\bar{A}(\omega)$ was defined in terms of the autoregressive coefficient matrix $A_l = (a_{ij,l})_{K\times K}$

$$\overline{A_{ij}}(\omega) = \begin{cases} 1 - \sum_{l=1}^{\infty} a_{ij,l} e^{-i\omega l}, & if\ i = j \\ -\sum_{l=1}^{\infty} a_{ij,l} e^{-i\omega l}, & otherwise \end{cases} \qquad (2.10)$$

along with a vector $\bar{a}_j(\omega)$ = $[\overline{A_{1j}}(\omega), \ldots, \overline{A_{Kj}}(\omega)]^T$, where $i = \sqrt{-1}$ is the unit for imaginary numbers and $\omega \in [-\pi, \pi)$ is the angular frequency. The information PDC is defined as follows.

Definition 2.1 [Partial Directed Coherence] For the zero mean stationary vector process $X(n)$ defined above, the information PDC from X_j to X_i is defined as

$$\iota\pi_{ij}(\omega) = \frac{\overline{A_{ij}}(\omega)\sigma_{ii}^{-1/2}}{\sqrt{\bar{a}_j^H(\omega)\Sigma_\epsilon^{-1}\bar{a}_j(\omega)}}, \tag{2.11}$$

where $\sigma_{ii} = E\left[\epsilon_{i,n}^2\right](i = 1, \dots, K)$ are the covariances for the innovation processes and the superscript H is for the Hermitian transpose.

The PDC is usually a complex value, one often uses the **square magnitude** of the PDC value as an alternative to judge the directionality of causal inference. **The information PDC is reduced to the generalized PDC (gPDC), if Σ_ϵ is a diagonal matrix with distinct diagonal elements. The generalized PDC is simplified to the original PDC, if Σ_ϵ equals to the identity matrix.** PDC is a linear method depends strictly on linear autoregressive model, and direct due to the multivariate estimation of causalities.

2.3 Partial directed coherence (PDC)

Example 2.1 (Two-dimensional AR(1)) Consider a system of two zero-mean stationary autoregressive processes:

$$\begin{aligned} X_1(n) &= \alpha X_{1,n-1} + \beta X_{2,n-1} + \epsilon_{1,n}, \\ X_2(n) &= \gamma X_{1,n-1} + \delta X_{2,n-1} + \epsilon_{2,n}, \end{aligned} \tag{2.12}$$

Which can be written in matrix form as

$$\begin{bmatrix} X_{1,n} \\ X_{2,n} \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} X_{1,n-1} \\ X_{2,n-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,n} \\ \epsilon_{2,n} \end{bmatrix}, \tag{2.13}$$

where $\epsilon_i, i = 1,2$ are zero-mean Gaussian stationary innovation processes with orthonormal covariances $E[\epsilon_{i,n}\epsilon_{j,m}] = \delta_{nm}\,\delta_{ij}$ (δ_{pq} is the Kronecker delta symbol), $i, j \in \{1,2\}$, and $\mathrm{m}, \mathrm{n} \in \mathbb{Z}$,$\alpha, \beta, \gamma$ and δ are the autoregressive coefficients not all vanish which form the coefficient matrix denoted as A. To verify Theorem 6.3.1, the PDC was calculated from both sides of Identity 6.9.

2 时间序列因果算法与数据分析

Calculation of $\iota\pi_{ij}$ by definition: From the coefficient matrix A, the $\bar{A}$ can be calculated as

$$\bar{A}(\omega) = \begin{bmatrix} \overline{A_{11}}(\omega) & \overline{A_{12}}(\omega) \\ \overline{A_{21}}(\omega) & \overline{A_{22}}(\omega) \end{bmatrix} = \begin{bmatrix} 1-\alpha e^{-i\omega} & -\beta e^{-i\omega} \\ -\gamma e^{-i\omega} & 1-\delta e^{-i\omega} \end{bmatrix}, \qquad (2.14)$$

along with the vectors

$$\overline{a_1}(\omega) = [\overline{A_{11}}(\omega), \overline{A_{21}}(\omega)]^T = [1-\alpha e^{-i\omega}, -\gamma e^{-i\omega}]^T,$$
$$\overline{a_2}(\omega) = [\overline{A_{12}}(\omega), \overline{A_{22}}(\omega)]^T = [-\beta e^{-i\omega}, 1-\delta e^{-i\omega}]^T, \qquad (2.15)$$

and their Hermitian transposes

$$\bar{a}_1^H(\omega) = [1-\alpha e^{i\omega}, -\gamma e^{i\omega}],$$
$$\bar{a}_2^H(\omega) = [-\beta e^{i\omega}, 1-\delta e^{i\omega}], \qquad (2.15)$$

where $i = \sqrt{-1}$ and $\omega \in [-\pi, \pi)$.

2 时间序列因果算法与数据分析

Due to orthonormality of the innovation processes, the covariance matrix Σ_ϵ and its inverse Σ_ϵ^{-1} are identity matrices with $\sigma_{ii} = E\left[\epsilon_{i,n}^2\right] = 1$, for $i \in \{1,2\}$. The PDC values can be calculated by definition as

$$\iota\pi_{11}(\omega) = \frac{\bar{A}_{11}(\omega)\sigma_{11}^{-1/2}}{\sqrt{\bar{a}_1^H(\omega)\Sigma_\omega^{-1}\bar{a}_1(\omega)}} = \frac{1-\alpha e^{-i\omega}}{\sqrt{1+\alpha^2+\gamma^2-\alpha(e^{i\omega}+e^{-i\omega})}},$$

$$\iota\pi_{12}(\omega) = \frac{\bar{A}_{12}(\omega)\sigma_{11}^{-1/2}}{\sqrt{\bar{a}_2^H(\omega)\Sigma_\omega^{-1}\bar{a}_2(\omega)}} = \frac{-\beta e^{-i\omega}}{\sqrt{1+\beta^2+\delta^2-\delta(e^{i\omega}+e^{-i\omega})}},$$

$$\iota\pi_{21}(\omega) = \frac{\bar{A}_{21}(\omega)\sigma_{22}^{-1/2}}{\sqrt{\bar{a}_1^H(\omega)\Sigma_\omega^{-1}\bar{a}_1(\omega)}} = \frac{-\gamma e^{-i\omega}}{\sqrt{1+\alpha^2+\gamma^2-\alpha(e^{i\omega}+e^{-i\omega})}},$$

$$\iota\pi_{22}(\omega) = \frac{\bar{A}_{22}(\omega)\sigma_{22}^{-1/2}}{\sqrt{\bar{a}_2^H(\omega)\Sigma_\omega^{-1}\bar{a}_2(\omega)}} = \frac{1-\delta e^{-i\omega}}{\sqrt{1+\beta^2+\delta^2-\delta(e^{i\omega}+e^{-i\omega})}}.$$

For unidirectionality, let $\alpha = \beta = \delta = 0$ and $\gamma \neq 0$, the model becomes

$$\begin{aligned} X_1(n) &= \epsilon_{1,n}, \\ X_2(n) &= \gamma X_{1,n-1} + \epsilon_{2,n}, \end{aligned} \tag{2.16}$$

And the PDC values become

$$\begin{bmatrix} \iota\pi_{11} & \iota\pi_{12} \\ \iota\pi_{21} & \iota\pi_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{1+\gamma^2}} & 0 \\ \frac{-\gamma e^{-i\omega}}{\sqrt{1+\gamma^2}} & 1 \end{bmatrix}, \tag{2.17}$$

The $\iota\pi_{21} \neq 0$ and $\iota\pi_{12} = 0$ indicate direct causality from $X_1 \to X_2$.

Notation 2.1 The causal direction of PDC is read from the second index to the first index. In matrix form, the causal direction is read from column index to row index. For example, in the matrix of 6.21, $\iota\pi_{21} = \frac{-\gamma e^{-i\omega}}{\sqrt{1+\gamma^2}}$ indicates direct causality from $X_1 \to X_2$, while $\iota\pi_{12} = 0$ means $X_2 \nrightarrow X_1$.

2.4 Mutual information rate (MIR)

The mutual information rates between two time series is given by the formula:

$$\mathrm{I}(X_i; X_j) = \sum_{\alpha,\beta} p(x_i = \alpha, x_j = \beta) \log \frac{p(x_i=\alpha, x_j=\beta)}{p(x_i=\alpha)p(x_j=\beta)}, \quad \mathrm{i} \neq \mathrm{j}, \mathrm{i}, \mathrm{j} = 1,2, \dots, 90. \tag{2.18}$$

when $\mathrm{i} = \mathrm{j}$ the mutual information rate degenerates to the Shannon Entropy:

$$H(X_i) = \sum_{\alpha} p(x_i = \alpha) \log \frac{1}{p(x_i=\alpha)} = -\sum_{\alpha} p(x_i = \alpha) \log p(x_i = \alpha) . \tag{2.19}$$

2.5 Transfer entropy (TE)

Transfer entropy is a fundamental bivariate information transfer measure given by:

$$\mathrm{TE}_{\mathrm{Y}\rightarrow\mathrm{X}} = \sum\nolimits_{\boldsymbol{\alpha},\boldsymbol{\beta},\gamma} \mathrm{p}(X_{\mathrm{n+1}} = \gamma, \mathrm{X}_n^{(k)} = \boldsymbol{\alpha}, \mathrm{Y}_n^{(l)} = \boldsymbol{\beta}) \log \frac{\mathrm{p}(X_{\mathrm{n+1}}=\gamma | \mathrm{X}_n^{(k)}=\boldsymbol{\alpha}, \mathrm{Y}_n^{(l)}=\boldsymbol{\beta})}{\mathrm{p}(X_{\mathrm{n+1}}=\gamma | \mathrm{X}_n^{(k)}=\boldsymbol{\alpha})}, \qquad (2.20)$$

where X and Y are two stationary time series, $X_{\mathrm{n+1}}$ denotes the prediction of X, $\mathrm{X}_n^{(k)} = (X_n, X_{n-1}, \dots, X_{n-k+1})$ and $\mathrm{Y}_n^{(l)} = (Y_n, Y_{n-1}, \dots, Y_{n-l+1})$ are the time lagged variables i.e. the history of X and Y respectively, the l and k are words lengths, where often $l = k$. The summation runs over all possible combinations of the states of $X_{\mathrm{n+1}}, \mathrm{X}_n^{(k)}$, and $\mathrm{Y}_n^{(l)}$.

Time-shifted surrogates

The information measures often use time-shifted surrogates to reducing the bias. Let X and Y be two arbitrary time series, and $\mathrm{TE}_{Y\to X}$ is the TE from Y to X. To obtain a surrogate of X, we shuffle the index of X while keeping Y unchanged. We then apply TE on Y and the surrogate of X, the results are denoted as $\mathrm{TE}_{Y\to X}(q)$, where q is the surrogates' index. The bias-corrected transfer entropy for $Y \to X$ is defined by

$$\mathrm{TE}_{C,Y\to X} = \mathrm{TE}_{Y\to X} - \max_{q}\{\mathrm{TE}_{Y\to X}(q)\} \tag{2.21}$$

We use $q = 10$ in the bias-corrections for all the simulation studies.

2 时间序列因果算法与数据分析

The TE between X and Y are asymmetric, where $\mathrm{TE}_{\mathrm{Y}\to\mathrm{X}}$ evaluates the level of dependence of X on Y. TE a bivariate measure which measures the directed interactions from a time series X to Y no matter the interactions are one-step or via multi-step. In analysis, we often use the nearest neighbor estimator and words length $l = k = 5$ for the computation of TE.

2 时间序列因果算法与数据分析

Reference:

[1]L. A. Baccala, C. S. N. D. Brito, D. Y. Takahashi and K. Sameshima, Unified asymptotic theory for all partial directed coherence forms. Philosophical Transactions of the Royal Society A 371, 1-14, 2012.

[2]L. A. Baccala and K. Sameshima, Comments on ‘Is partial coherence a viable technique for identifying generators of neural oscillations?’, Why the term ‘Gersch Causality’ is inappropriate : Common neural structure inference pitfalls. Biological Cybernetics 95, 135-141, 2006.

[3]L. A. Baccala and K. Sameshima, Partial directed coherence: a new concept in neural structure determination. Biological Cybernetics 84, 463-474, 2001.

[4]T. Schreiber, Measuring information transfer. Physical Review Letters, 85 2, 461-464, 2000

[5]T. Schreiber and A. Schmitz, Surrogate time series. Physica D, 142, 346-382, 2000.

[6]A. Seth, Granger causality. Scholarpedia, 2, 7, 1967.

[7]D. Y. Takahashi, L. A. Baccala and K. Sameshima, Frequency domain connectivity: an information theoretic perspective. The 32nd Annual International Conference of the IEEE EMBS, Buenos Aires, Argentina, 2010.

Reference:

[8]D. Y. Takahashi, L. A. Baccala and K. Sameshima, Information theoretical interpretation of frequency domain connectivity measures. Biological Cybernetics, 103, 463-469, 2010.

[9]D. Y. Takahashia, L. A. Baccala and K. Sameshima, Partial directed coherence asymptotics for VAR processes of infinite order. International Journal of Bioelectromagnetism 10, 1, 31-36, 2008.

3.1 PDC算法应用实例

Example 3.1 (Linear interaction) Consider a 4-dimension linear autoregressive system

$$\begin{aligned}
X_{1,n} &= 0.96\sqrt{2}X_{1,n-1} - 0.9025X_{1,n-2} + \epsilon_{1,n}, \\
X_{1,n} &= 0.5X_{1,n-2} + \epsilon_{2,n}, \\
X_{3,n} &= -0.4X_{1,n-1} + 0.6X_{4,n-2} + \epsilon_{3,n}, \\
X_{4,n} &= -0.5X_{2,n-1} - 0.25\sqrt{2}X_{3,n-1} + \epsilon_{4,n},
\end{aligned} \tag{3.1}$$

Where $\epsilon_i, i = 1,2,3,4$ are are generated standard Gaussian distributed white noise. This model describes the direct causal influence from $X1 \rightarrow X2, X1 \rightarrow X3, X3 \leftrightarrow X4$ and $X2 \rightarrow X4$.

3 算法与实例

We used the recommended parameter choice in this numerical simulation. In consistency to the numerical studies of various causality measures, N = 5000 number of data points were generated per data channel (PDC converges as the time series length increases, but considering the computation time, the recommended time series length N = 5000 is long enough for convergence and suitable for numerical simulations). There are a couple of recommended options of Parameters, e.g. α= 1% orα= 5% (significance level for PDC asymptotic statistics), AIC (or Hannan-Quinn, Schwartz or fixed) order selection criterion, Nutall-Strand (or least square or Vieira Morf) AR model fitting algorithm.

3 算法与实例

We have altered the several recommended options in our analysis. PDC causalities plotted against frequencies were shown for example in Figure 3.1 (α= 1%) and 3.2 (α= 5%). We found that the PDC can identify the correct direct causal directions, but may also introduce false causalities.

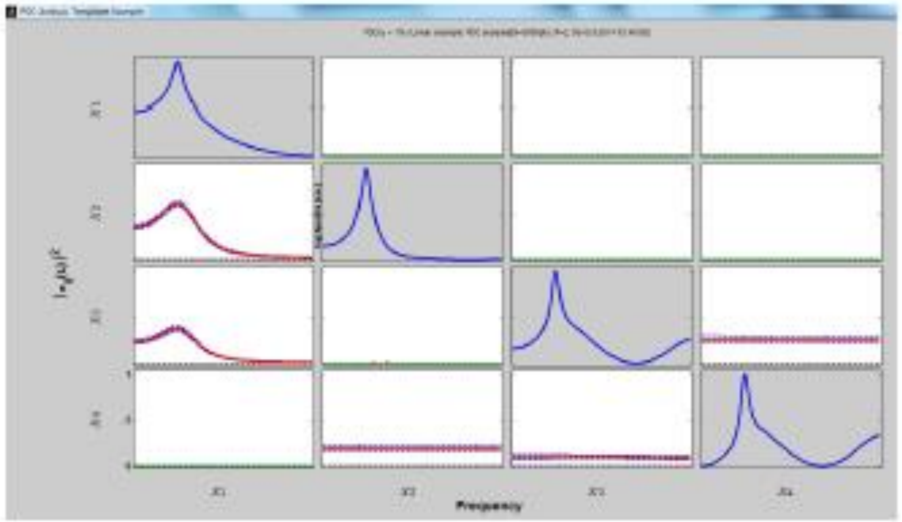


Figure 3.1: Original PDC for the linear example 3.1 (α= 1%). In this figure, the 44 layout plots show the square magnitudes of PDC causalities against frequencies, for the linear example 3.1 with α= 1%. The direction of causalities is read from the column index to the row index of each plot. In these plots, red curves indicate significant PDC causalities at corresponding frequencies, the green curves indicate insignificant PDC causalities, the black curves are the significance thresholds (according to PDC asymptotic statistics), whereas the blue curves on the diagonal plots are the coherences for each data channel. This figure shows that PDC indicates the correct and direct significant causalities from: $X1 \rightarrow X2, X1 \rightarrow X3, X2 \rightarrow X4$ and $X3 \leftrightarrow X4$, along with false causalities from $X2 \rightarrow X3$ at certain frequencies.

3 算法与实例

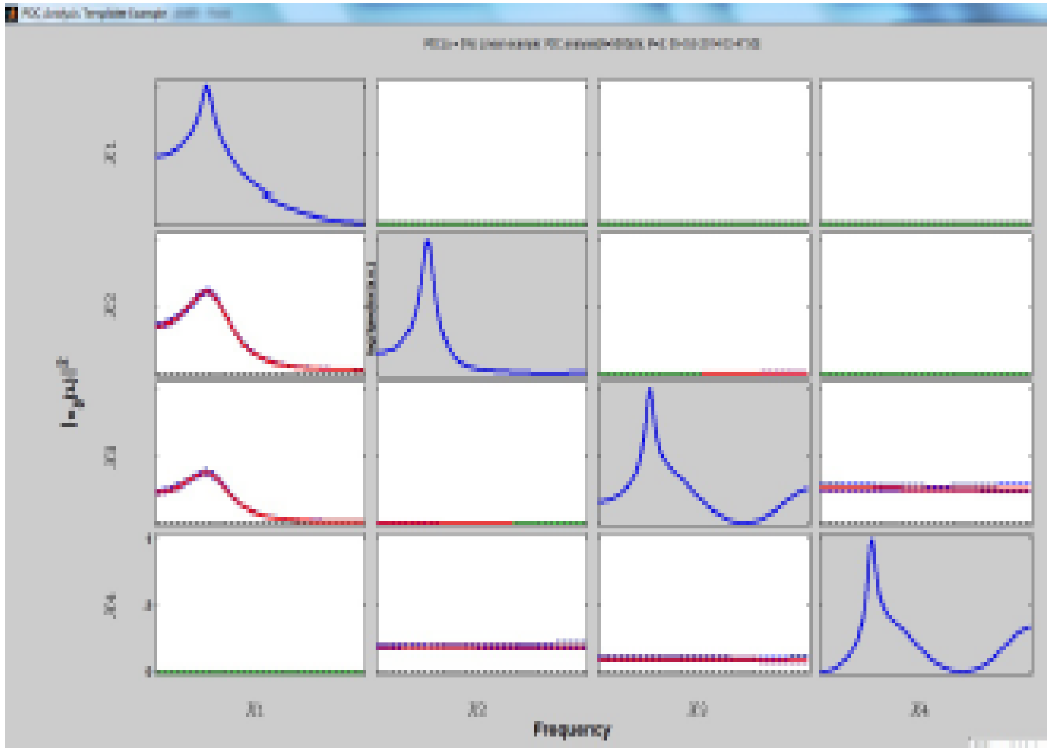


Figure 3.2: Original PDC for the linear example 3.1 (α= 5%). In this figure, the 4×4 layout plots show the square magnitudes of PDC causalities against frequencies, for the linear example 3.1 with α= 5%. The direction of causalities is read from the column index to the row index of each plot. In these plots, red curves indicate significant PDC causalities at corresponding frequencies, the green curves indicate insignificant PDC causalities, the black curves are the significance thresholds (according to PDC asymptotic statistics), whereas the blue curves on the diagonal plots are the coherences for each data channel. This figure shows that PDC indicates the correct and direct significant causalities from: $X_1 \rightarrow X_2, X_1 \rightarrow X_3, X_2 \rightarrow X_4$and $X_3 \leftrightarrow X_4$, along with false causality from $X_2 \rightarrow X_3$ and $X_3 \rightarrow X_2$ at certain frequencies.

Reference:

[1] Koichi Sameshima, Luiz Antonio Baccala. AsympPDC Package 1.0 User Guide. Unitersity of Sao Paulo, Brazil, 2011.