IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

A modeling approach to multivariate analysis and clusterization theory

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2008 J. Phys. A: Math. Theor. 41 205101

(http://iopscience.iop.org/1751-8121/41/20/205101)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.148
The article was downloaded on 03/06/2010 at 06:49

Please note that terms and conditions apply.

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **41** (2008) 205101 (13pp) doi:10.1088/1751-8113/41/20/205101

A modeling approach to multivariate analysis and clusterization theory

G Innocenti and D Materassi

Dipartimento di Sistemi e Informatica, Centro per lo Studio di Dinamiche Complesse—CSDC, Università di Firenze, via di S. Marta 3, 50139 Firenze, Italy

Received 20 January 2008, in final form 24 March 2008
Published 23 April 2008
Online at stacks.iop.org/JPhysA/41/205101

Abstract

The paper deals with the problem of identifying the internal dependences and similarities among a large number of random processes. Linear models are considered to describe the relations among the time series, and the energy associated with the corresponding modeling error is the criterion adopted to quantify their similarities. Such an approach is interpreted in terms of graph theory suggesting a natural way to group processes together when one provides the best model to explain the others. Moreover, the clustering technique introduced in this paper will turn out to be the dynamical generalization of other multivariate procedures described in the literature.

PACS numbers: 02.50.Sk, 02.10.Ox

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Deriving information from data is a crucial problem in science, and it has been widely investigated in the literature. A large variety of contributions has been developed in many fields, such as engineering, physics, biology and economy, providing several methods and procedures which accomplish different objectives. In particular, in the study of complex systems, the comprehension of the internal connections, which define the hierarchical structure of the process, turns out to play a key role in fully understanding its dynamics [1]. Just to cite possible and very different applications, in [2] the modular organization of metabolic processes of cells has been detected and studied, in [3] models for internal interactions in glassy materials have been suggested, and in [4] the identification of a hierarchical structure of stocks in the financial market is proposed in order to check how diversified a portfolio is. Techniques to distinguish topological interconnections in complex systems are especially useful in the presence of a multivariate data set, because this kind of sample is usually the result of a process intrinsically organized into modular subsystems [5]. Therefore, the recognition of the system structure is a critical step for the definition of a suitable model. In particular, a clusterization problem can be solved to divide the source data set into interconnected homogeneous groups

 © 2008 IOP Publishing Ltd Printed in the UK

describing different subsystems [6]. This approach deals with the search for similarities and relations inside the original samples, trying to catch their internal connections and providing a schematic representation of hierarchies. Recently, new clustering techniques based on a correlation matrix have been proposed for the analysis of data sets made up by a large variety of time series [7, 8]. However, these procedures are able to detect only the 'static' relations among the samples, since they capture the similarities just at the current time [9–12].

In this paper, we propose a clustering technique based on a modeling approach. Indeed, since the original time series are dynamically interconnected, we intend to derive their hierarchy in terms of mathematical laws, which provide a structured description of the internal mechanism. To this end, we settle the clustering problem into the framework of the system identification theory [13, 14]. Hence, exploiting the modeling errors to quantify the similarities among the original signals, we realize a clustering technique, defined as the solution of a minimization problem. Therefore, a modeling interpretation of the procedures based on the correlation matrix is first introduced. In particular, they turn out to be a non-optimal choice with respect to the modeling error. Then, the approach is developed taking into account dynamic dependences among the time series. In this respect, the identification step is realized introducing the hypothesis of linear dynamic connections, represented by single input–single output (SISO) local models. Moreover, since the clusters are internally organized by means of transfer functions, the final model can be interpreted as a dynamical network of interconnected systems and its structure as the related topology.

Notation.

The symbol $\doteq$ *denotes a definition*

$E[\cdot]$: *mean operator*;

$R_{XY}(\tau) \doteq E[X(t)Y(t+\tau)]$: *cross-covariance function of stationary processes*;

for the sake of simplicity, $R_{XY} \doteq R_{XY}(0)$;

$R_X(\tau) \doteq R_{XX}(\tau)$: *autocovariance*;

$\rho_{XY} \doteq \dfrac{R_{XY}}{\sqrt{R_X R_Y}}$: *correlation index*;

$\mathcal{Z}(\cdot)$: *zeta-transform of a signal*;

$\Phi_{XY}(z) \doteq \mathcal{Z}(R_{XY}(\tau))$: *cross-power spectral density*;

$\Phi_X(z) \doteq \Phi_{XX}(z)$: *power spectral density*;

with abuse of notation, $\Phi_X(\omega) = \Phi_X(e^{i\omega})$;

$\lceil\cdot\rceil$ *and* $\lfloor\cdot\rfloor$: *ceiling and floor function respectively*;

$(\cdot)^*$: *complex conjugate*.

2. A modeling perspective

In [9], a procedure to obtain a hierarchical structure of a set of time series is proposed. N realizations of N random processes X_i are considered. First, an estimation of the correlation index ρ_{ij} related to every couple (X_i, X_j) is computed, along with the associated distances [4]

$$d_{ij} \doteq \sqrt{2(1-\rho_{ij})}. \tag{1}$$

Then, a graph is defined where every node represents a random process and the arc linking two nodes is weighted according to (1). Eventually, the minimum spanning tree (MST) is extracted by the graph. This procedure has been successfully exploited to provide a quantitative and topological analysis of time series, especially in the economic field (see [4, 8, 11]). It is

worth considering that such a technique can be interpreted in terms of a modeling procedure. Consider the problem of describing a process X_j by scaling another process X_i with a suitable real constant α_{ji}. Choosing

$$\alpha_{ji} = \sqrt{\frac{E\left[X_j^2\right]}{E\left[X_i^2\right]}} = \sqrt{\frac{R_{X_j}}{R_{X_i}}}, \tag{2}$$

we find that

$$E[(X_j - \alpha_{ji} X_i)^2] = E\left[X_j^2\right] d_{ij}^2.$$

Hence, the distance (1) can be interpreted as the root of the mean square error, properly normalized by the variance of X_j, when the simple gain (2) is used. Such a normalization is necessary since we are interested into capturing similar trends between the processes regardless of their amplitudes. However, we remark that the choice of (2) can be considered arbitrary. Conversely, we would like to evaluate the closeness of two processes according to the information which can be inferred about one of them assuming to know the other [15]. From this point of view, (2) does not satisfy any optimality criterion. Indeed, considering two anticorrelated time series ($\rho_{ij} = -1$) it is possible to perfectly reconstruct one from the other. Thus the information in the two signals is the same, while their distance (1) makes them the farthest. Let us define

$$e_{ji} = X_j - \alpha_{ji} X_i; \tag{3}$$

then, it is possible to adopt the least squares criterion in order to evaluate the 'best' constant α_{ji}. In this case, it is immediate to prove that the optimal choice is given by

$$\hat{\alpha}_{ji} = \frac{R_{X_j X_i}}{R_{X_i}} \tag{4}$$

and the relative quadratic error amounts to

$$E\left[e_{ji}^2\right] = R_{X_j} - \frac{R_{X_j X_i}^2}{R_{X_i}} \tag{5}$$

[14]. In order to obtain a dimensionless quantity, we can normalize (5) with respect to the power of X_j and define the binary function

$$\mathrm{d}(X_i, X_j) \doteq \sqrt{\frac{E\left[e_{ji}^2\right]}{R_{X_j}}} = \sqrt{1 - \rho_{X_i X_j}^2}. \tag{6}$$

It is worth observing that (6) is a distance exactly as (1).

Proposition 1. *The function $d(\cdot, \cdot)$, as defined in (6), is a metric.*

Proof. See the appendix. □

In [9], the MST is extracted from the graph, according to the weights (1). This is equivalent to define a hierarchical structure of the time series relying on the adoption of linear gain models (2) between the processes and considering the relative modeling error as a distance function.

The choice of a MST can be justified from a modeling point of view as an attempt to define a connected network which minimizes the modeling error on every node.

Substituting (1) with (6), we apply the same topological strategy, but we structure the data according to the best gain model in the sense of the least squares.

Remark 2. From a system theory point of view, it can be said that both the approaches are 'static'. Indeed, the models do not have a state; thus they do not have any dynamics. They simply capture a direct relation between two process samples at the same time instant. However, the optimal approach we have followed can be extended to a more general case.

3. Dynamic modeling using Wiener filters

We consider a model to be 'static' (or 'memoryless') when, at every time instant t, its output is a function of its input at the very same time instant. Conversely, the output of a 'dynamic' model also depends on the input values it receives at instants different from t. In this general sense, we say that it has a 'memory' (or, equivalently, a 'state'). Constant gains as (2) or (4) are linear static models offering an extremely simple proportional relation between two processes. We propose a dynamic extension of the linear approach just described in the previous section based on the well-known Wiener filter.

Given two stochastic processes X_i, X_j and a time discrete transfer function $W_{ji}(z)$ (that is, the zeta-transform of its impulse response), let us consider the quadratic cost

$$E[(\varepsilon_Q)^2], \tag{7}$$

where

$$\varepsilon_Q \doteq Q(z)(X_j - W_{ji}(z)X_i), \tag{8}$$

$Q(z)$ being an arbitrary stable and causally invertible time-discrete transfer function weighting the error

$$e_{ji} = X_j - W_{ji}(z)X_i. \tag{9}$$

Then, the problem of evaluating the transfer function $\hat{W}(z)$ such that the quadratic cost (7) is minimized is well known in the scientific literature and its solution is referred to as the Wiener filter [14].

Proposition 3 (Wiener filter)**.** *The Wiener filter modeling X_j by X_i is the linear stable filter $\hat{W}_{ji}$ minimizing the filtered quantity (7). Its expression is given by*

$$\hat{W}_{ji}(z) = \frac{\Phi_{X_iX_j}(z)}{\Phi_{X_i}(z)}, \tag{10}$$

and it does not depend upon $Q(z)$. Moreover, the minimized cost is equal to

$$\min E[(Q(z)\varepsilon)^2] = \frac{1}{2\pi}\int_{-\pi}^{\pi} |Q(\omega)|^2\big(\Phi_{X_j}(\omega) - |\Phi_{X_jX_i}(\omega)|^2\Phi_{X_i}^{-1}(\omega)\big)\,\mathrm{d}\omega.$$

Proof. See, for example, [14] □

Observe that the stable implementation of the Wiener filter $\hat{W}_{ji}(z)$ is non-causal, in general. That is, its output $\hat{W}_{ji}(z)X_i$ depends on both past and future values of the input process X_i. The Wiener filter, in this formulation, is interesting from an information and modeling point of view, but, of course, we would rather need a causal filter, if we were to make predictions (aim which is beyond the scope of this paper).

Since the weighting function $Q(z)$ does not affect the Wiener filter, but only the energy of the filtered error, we choose $Q(z)$ equal to $F_j(z)$, the inverse of the spectral factor of $\Phi_{X_j}(z)$, that is

$$\Phi_{X_j}(z) = F_j^{-1}(z)\big(F_j^{-1}(z)\big)^*, \tag{11}$$

with $F_j(z)$ being stable and causally invertible [16]. In such a case, the minimum cost assumes the value

$$\min E\big[\varepsilon_{F_j}^2\big] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(1 - \frac{|\Phi_{X_jX_i}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)}\right)\mathrm{d}\omega. \tag{12}$$

This specific choice of $Q(z)$ makes the cost depend explicitly on the coherence function of the two processes

$$C_{X_iX_j}(\omega) \doteq \frac{|\Phi_{X_jX_i}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)}, \tag{13}$$

which turns to be non-negative and symmetric with respect to ω. It is also well known that the cross-spectral density satisfies the Schwartz inequality. Hence, the coherence function is limited between 0 and 1. The choice $Q(z) = F_j(z)$ can be now understood as motivated by the necessity to achieve a dimensionless cost function not depending on the power of the signals as in (12).

The cost obtained by the minimization of the error ε_{F_j} using the Wiener filter as before allows us to define the binary function

$$\mathrm{d}(X_i, X_j) \doteq \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(1 - C_{X_iX_j}(\omega)\right)\mathrm{d}\omega\right]^{1/2}. \tag{14}$$

Proposition 4. *The function $d(\cdot,\cdot)$, as defined in (14), is a metric.*

Proof. See the appendix. □

The metric (14) can now be used to derive a MST and obtain a hierarchical structure of the processes X_i. Such an approach generalizes the results in [9] to the linear dynamic case. We remark that the choice of a tree to describe the topology of the data is a very reasonable but arbitrary solution. In order to capture influences and similarities among the processes X_j, we intend to propose a more flexible modeling technique to extract topological information from the data. Every X_j can be described as the output of a linear SISO dynamical system, whose input is one of the other $N-1$ processes. Thus, for every time series X_j it is natural to choose the model $\hat{W}_{jm(j)}(z)$ with input $X_{m(j)}$, such that it provides the best description according to (12), dropping the others. The application of this procedure results in a set of N interconnected systems, each of them minimizing $\min_i E[(Q_j e_{ji})^2]$. Since the choice of every model $\hat{W}_{jm(j)}(z)$ does not affect the selection of the others, the overall cost function

$$\min_{m(\cdot)} \sum_j E[(Q_j e_{jm(j)})^2] \tag{15}$$

turns out to be minimized, as well. The following algorithm performs such a task.

Algorithm (Clusterization Algorithm).

1. initialize the set $A = \emptyset$
2. for every process $X_j (j = 1, \ldots, N)$
   - 2a. for every $i = 1, \ldots, N, i \neq j$
     compute the distance $d_{ij} \doteq d(X_i, X_j)$;
   - 2b. define the set $M(j) \doteq \{k | d_{kj} = \min_i d_{ij}\}$ with $i \neq j$
   - 2c. choose, if possible, $m(j) \in M(j)$ such that $(m(j), j) \notin A$
   - 2d. choose the model
     $X_j = \hat{W}_{jm(j)}(z)X_{m(j)} + e_{jm(j)}$
   - 2e. add the couple $(j, m(j))$ to A.

The resulting network of processes has an appealing graphical interpretation. Indeed, its topological structure can be seen as a weighted graph where every process X_j is a node, the arc linking X_i to X_j represents the Wiener filter describing the 'output' X_j in terms of the 'input' X_i, and the weights on the arcs are given by (14). Because of the symmetry property

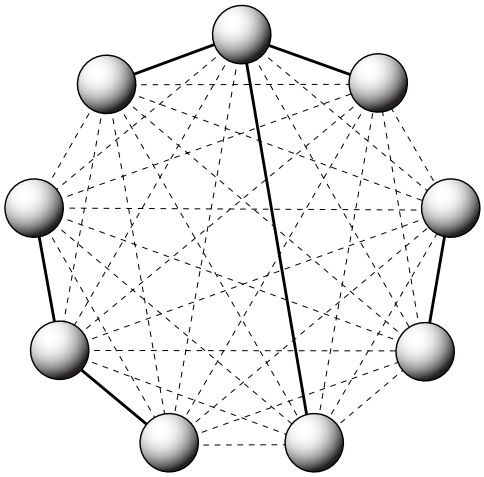

Figure 1. The figures illustrates all the possible connections between two nodes (dashed lines) in a nine-node network. The solid lines depict a forest as it were the result after the application of the algorithm.

of (14), there is no actual need to consider an oriented graph. Hence, the presence of both the arcs (i, j) and (j, i) boils down into just a single link. Following this interpretation, the algorithm determines a graph designed to keep, for every node, the incident arc with the least cost (see figure 1).

Proposition 5. *The graph resulting from the proposed algorithm has the following properties:*

- *on every node, there is at least an incident arc;*
- *if there is a cycle, then all the arcs of the cycle have the same weight;*
- *there are at least* $\lceil N/2 \rceil$ *and at most N arcs.*

Proof. See the appendix. □

The presence of cycles in the resulting graph is a pathological situation as stressed in the following remark.

Remark 6. A necessary condition of existence for a cycle is the presence of more than two nodes with common multiple minimum cost arcs. Therefore, a mild sufficient condition in order to avoid cycles in the graph is to assume that every node has a unique minimum cost arc. If the costs of the arcs are obtained by estimation from real data, the probability to obtain a cycle is zero almost everywhere [17]. Consequently, in such a case the expected topology of the graph is a forest (a graph with no cycles).

Remark 7. If there are no cycles, the graph resulting from the algorithm is a subgraph of the MST.

Remark 8. In general, nothing can be said about the connectivity. Therefore, the modeling procedure depicted by the algorithm provides a clusterization of the original processes X_i which, for every node, minimize the cost (14) according to the criterion of linear dynamic dependence. It is possible to modify the procedure in order to suitably satisfy other constraints about the graph topology. For instance, if we deal with a connectivity condition the algorithm

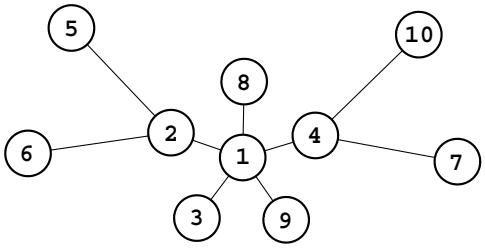

Figure 2. The figure illustrates the topology of the ten-node network analyzed in section 4. Each node represents a process X_j, while the arcs describe the connections among them, according to the linear SISO model (16). For the data generation, we have considered only transfer functions of at most the second order. The noises N_j have been assumed to provide half the power of the affected processes. The samples have been collected over 1000 time steps.

can be easily replaced by a MST search. Therefore, the approach followed in [9] to obtain topological information from the time series results in a constrained optimization of (15).

Remark 9. The modelization we have derived makes use of non-causal Wiener filters; thus it can be useful to detect linear dependences of any sort between the processes X_i.

Unfortunately, the adoption of non-causal filters cannot be employed to make predictions.

4. Numerical example

It is intended to show, by means of numerical examples, the main advantages of the technique described in the previous sections. In particular, we want to evaluate the performance of our procedure when identifying an unknown topology. First, we realized several simulations of ten randomly generated processes X_j, designed as follows. They have been hierarchically structured in a tree topology, where the interconnections were linear, randomly generated, at most second-order transfer functions W_{ji} with external noises N_j:

$$X_j = W_{ji} X_i + N_j. \tag{16}$$

Since all simulations present strong analogies, we are showing just one of them, whose topology is depicted in figure 2. Note that the simulated network involves linear dependences only, so it satisfies the theoretical conditions of the approach based on Wiener filters introduced in the previous sections. On every node X_j (but the root), the deterministic component $W_{ji}(z)X_i$ and the stochastic disturbance are equal in power. A simulation horizon of 1000 steps has been taken into account where the noise components have been generated by pseudorandom number algorithms. The hypothesis of uncorrelated disturbances has been numerically checked, providing a marginally satisfactory result. In order to make a meaningful comparison, we have also considered the technique described in [4], which has given useful insights into the analysis of a multivariate data set in econometrics and in biology. Applying it, we found the distance matrix reported in table 1 and the corresponding MST depicted in figure 3, where every node occupies the same position. We note that the topology is not correctly identified by such a procedure, even though similarities can be identified. On the other hand, the application of the clusterization algorithm introduced by us provides the distances of table 2 with the graph of figure 4. We stress that the choice of an original tree topology is due to the comparison with the widely used technique introduced in [4].

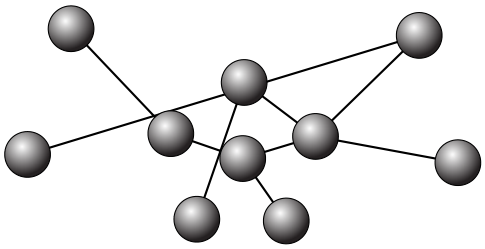

Figure 3. The MST obtained using the correlation-based distance of table 1. Every node is placed in the same position of figure 2 for a direct visual comparison. The actual topology has not been correctly identified, though some analogies with the right structure can be observed. The procedure described in [4] reveals strong limitation in capturing the nature of the network even when the actual topology is exactly a tree.

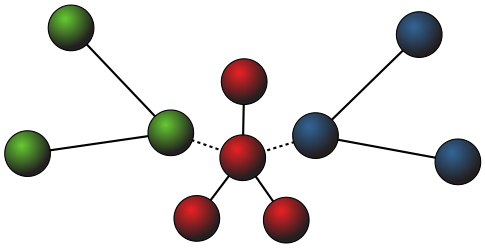

Figure 4. The figure illustrates the MST obtained by using the coherence-based distance (solid+dashed lines). The nodes are left in the same positions as in the previous figures. Notably, it is same as actual topology. The application of the proposed clustering algorithm provides a forest (solid lines): each cluster is virtually connected to the others by the arcs of the MST, which have not been chosen by the algorithm (dashed lines). The use of different colors highlights the modular structure resulting from the clusterization. It is worth noting that the actually topology is exactly identified under the connectivity constraint.

Table 1. Correlation-based distance matrix.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	0	0.9946	1.3763	1.0624	1.1027	1.2393	1.2719	1.3747	0.7306	1.4589
X_2	0.9946	3.3e-8	1.1130	1.0674	0.7723	1.0082	1.2004	1.1269	1.1132	1.4575
X_3	1.3763	1.1130	0	1.1487	1.2217	1.2877	1.1645	0.9 965	1.3507	1.4124
X_4	1.0624	1.0674	1.1487	4.2e-8	1.1727	1.1805	0.9296	1.1455	1.1491	1.3433
X_5	1.1027	0.7723	1.2217	1.1727	3.9e-8	1.1491	1.2418	1.2353	1.1898	1.4587
X_6	1.2393	1.0082	1.2877	1.1805	1.1491	4.9e-8	1.2123	1.2984	1.2858	1.3227
X_7	1.2719	1.2004	1.1645	0.9296	1.2418	1.2123	0	1.1815	1.3003	1.3334
X_8	1.3747	1.1269	0.9965	1.1455	1.2353	1.2984	1.1815	0	1.3542	1.4389
X_9	0.7306	1.1132	1.3507	1.1491	1.1898	1.2858	1.3003	1.3542	7.3e-8	1.4450
X_{10}	1.4589	1.4575	1.4124	1.3433	1.4587	1.3227	1.3334	1.4389	1.4450	0

Moreover, it is worth noting that the link structure is perfectly reconstructed by our procedure, if the connectivity constraint is imposed, and that only real connections are chosen by the proposed clustering algorithm.

Further, we repeated the same procedure with a larger number of processes ($N = 50$). Again, the results showed many similarities; so we are presenting just one case with the

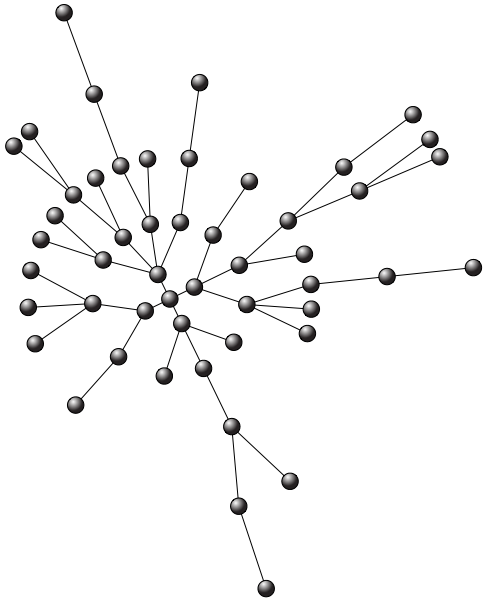

Figure 5. The 50-node network of section 4. The figure provides the actual topology. The example has been designed according to the same assumptions of the network of figure 2.

Table 2. Coherence-based distance matrix.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	0	0.7299	0.6675	0.7351	0.8316	0.8542	0.8297	0.7055	0.6549	0.8298
X_2	0.7299	0	0.8065	0.8353	0.6934	0.7358	0.8786	0.8483	0.8299	0.8717
X_3	0.6675	0.8065	0	0.8216	0.8744	0.8807	0.8750	0.8262	0.7841	0.8821
X_4	0.7351	0.8353	0.8216	0	0.8662	0.8722	0.7404	0.8502	0.8198	0.7039
X_5	0.8316	0.6934	0.8744	0.8662	0	0.8540	0.8919	0.8995	0.8730	0.8846
X_6	0.8542	0.7358	0.8807	0.8722	0.8540	0	0.8934	0.8984	0.8796	0.8944
X_7	0.8297	0.8786	0.8750	0.7404	0.8919	0.8934	0	0.8838	0.8694	0.8346
X_8	0.7055	0.8483	0.8262	0.8502	0.8995	0.8984	0.8838	0	0.8167	0.8908
X_9	0.6549	0.8299	0.7841	0.8198	0.8730	0.8796	0.8694	0.8167	0	0.8715
X_{10}	0.8298	0.8717	0.8821	0.7039	0.8846	0.8944	0.8346	0.8908	0.8715	0

topology depicted in figure 5. Analogously, the correlation-based approach of [4] provides the MST of figure 6 while our coherence-based algorithm identifies the graph of figure 7. However, it is worth noting that our technique detects links actually present in the topology with no mistakes and that the original topology is correctly reconstructed under the connectivity constraint.

These simple examples highlight a better capability of our technique into capturing relationships and dependences among time series. In particular, remarkable improvements should be expected in the presence of strong dynamical interconnections and significant delays in the actual network. Indeed, the correlation approach is not able to detect similarities among time series when time shift delays are present. Conversely, the coherence distance may capture them since it relies on a dynamical modeling of the processes.

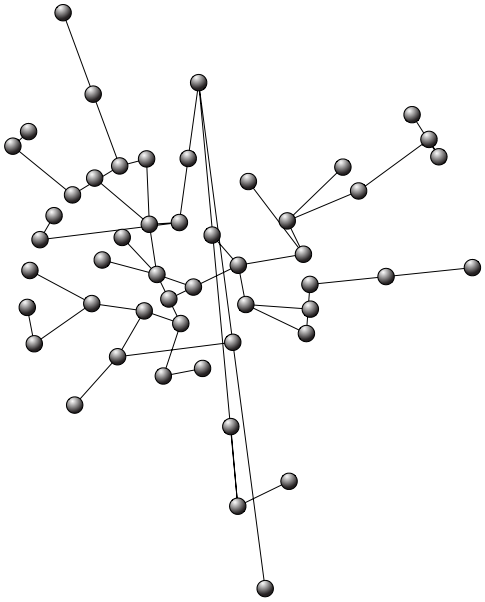

Figure 6. The figure illustrates the MST obtained by means of the correlation-based distance. The nodes are settled in the same positions of the original tree in figure 5 for a direct visual comparison. Though the original topology is a tree, a quite significant number of connections have not been correctly reconstructed. A limited number of similarities with the actual network can be observed.

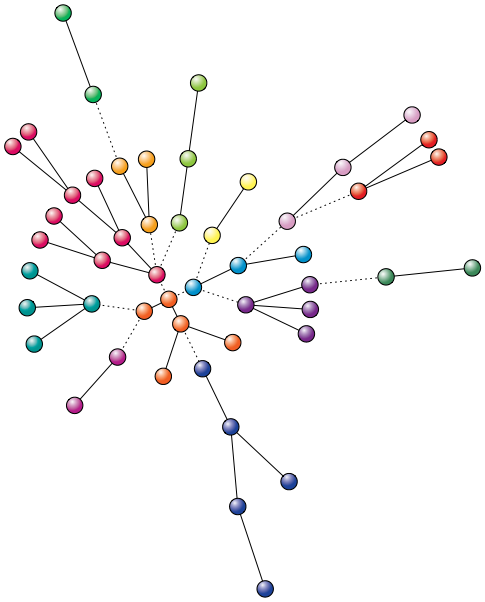

Figure 7. The figure shows the MST (solid+dashed lines) obtained by applying the coherence-based distance (16) to the processes produced by the network depicted in figure 5, assuming the node positions unchanged. Notably, the original structure has been correctly reconstructed. The forest resulting from the application of the clusterization algorithm is also reported (solid lines). The clusters are connected by the remaining arcs of the MST (dashed lines).

5. Conclusions

In this paper we have introduced a novel approach to the clusterization problem. In particular, the similarities among the time series of a multivariate data set have been analyzed from the modeling point of view and their interconnections have been interpreted as functional dependences. Hence, linear SISO transfer functions have been proposed to describe the relations among the processes and the associated modeling errors have been exploited to quantify their similarities. In turn, such a distance introduces a natural way of grouping the time series, since it is very reasonable to place two processes in the same subset, when one provides the best model to explain the other.

Notably, the proposed distance can be directly computed exploiting the coherence function, requiring no identification step. In this respect, for example, in [18] the original time series are reconstructed by means of a fixed nonlinear differential equation system, obtaining the parametric sets which best fit them. The corresponding points in the parameter space are then used for the clusterization. It is worth noting that such a procedure is not suitable to group dynamically linked time series, which instead is an advantage of our technique.

Further, our novel approach has been compared to the clustering technique proposed in [4] and formulated as an extension of the multivariate analysis of [8]. In particular, our coherence-based distance turns out to be the dynamical generalization of the correlation-based metric in [9]. Therefore, it provides an improved capability in capturing the internal topology among the processes, especially when their functional dependences turn out to be dynamical laws. Some numerical examples have finally been presented to illustrate the expected improvements, due to our distance, and to provide a validation for our clustering algorithm.

Appendix

Proof of proposition 1. Note that we consider two processes to be equivalent also when they are anticorrelated since they are identical from an information point of view. Thus, the only non-trivial property to show is the triangle inequality. Consider the following relations involving the optimal gains $\hat{\alpha}_{31}, \hat{\alpha}_{32}, \hat{\alpha}_{21}$, defined as in (4):

$$\begin{aligned} X_3 &= \hat{\alpha}_{31} X_1 + e_{31} \\ X_3 &= \hat{\alpha}_{32} X_2 + e_{32} \\ X_2 &= \hat{\alpha}_{21} X_1 + e_{21}. \end{aligned}$$

Since $\hat{\alpha}_{31}$ is the best constant model, we have that it must perform better than any other constant model (in particular $\hat{\alpha}_{32}\hat{\alpha}_{21}$):

$$R_{X_3} - \frac{R^2_{X_3X_1}}{R_{X_1}} \leqslant E[(e_{32} + \hat{\alpha}_{32} e_{21})^2] \leqslant \left(\sqrt{E[e_{32}^2]} + |\hat{\alpha}_{32}|\sqrt{E[e_{21})^2]}\right)^2.$$

Normalize with respect to R_{X_3} and consider the square root

$$\begin{aligned} \sqrt{1 - \rho^2_{X_1X_3}} &\leqslant \sqrt{\frac{1}{R_{X_3}}\left(\sqrt{E\left[e_{32}^2\right]} + |\hat{\alpha}_{32}|\sqrt{E\left[e_{21}^2\right]}\right)^2} \\ &\leqslant \sqrt{\frac{E\left[e_{32}^2\right]}{R_{X_3}}} + |\rho_{X_2X_3}|\sqrt{\frac{E\left[e_{21}^2\right]}{R_{X_2}}}. \end{aligned}$$

Since $|\rho_{X_2X_3}| \leqslant 1$, we have the assertion. □

Proof of proposition 4. The only non-trivial property to prove is the triangle inequality. Let $\hat{W}_{ji}(z)$ be the Wiener filter between X_i and X_j computed according to (10) and e_{ji} the relative error. The following relations hold: □

$$X_3 = \hat{W}_{31}(z)X_1 + e_{31}$$
$$X_3 = \hat{W}_{32}(z)X_2 + e_{32}$$
$$X_2 = \hat{W}_{21}(z)X_1 + e_{21}.$$

Since $\hat{W}_{31}(z)$ is the Wiener filter between the two processes X_1 and X_3, it performs better at any frequency than any other linear filter, such as $\hat{W}_{32}(z)\hat{W}_{21}(z)$. So we have

$$\begin{aligned}\Phi_{e_{31}}(\omega) &\leqslant \Phi_{e_{32}}(\omega) + |\hat{W}_{32}(\omega)|^2\Phi_{e_{21}}(\omega)\\ &\quad + \Phi_{e_{32}e_{21}}(\omega)\hat{W}_{32}^*(\omega) + \hat{W}_{32}(\omega)\Phi_{e_{21}e_{32}}(\omega)\\ &\leqslant (\sqrt{\Phi_{e_{32}}(\omega)} + |\hat{W}_{32}(\omega)|\sqrt{\Phi_{e_{21}}(\omega)})^2 \qquad \forall\omega\in\mathbb{R}.\end{aligned}$$

For the sake of simplicity, we neglect to explicitly write the argument ω in the following passages. Normalizing with respect to Φ_{X_3}, we find

$$\frac{\Phi_{e_{31}}}{\Phi_{X_3}} \leqslant \frac{1}{\Phi_{X_3}}\left(\sqrt{\Phi_{e_{32}}} + |\hat{W}_{32}|\sqrt{\Phi_{e_{21}}}\right)^2$$

and considering the 2-norm properties

$$\left(\int_{-\pi}^{\pi}\frac{\Phi_{e_{31}}}{\Phi_{X_3}}\,\mathrm{d}\omega\right)^{\frac{1}{2}} \leqslant \left(\int_{-\pi}^{\pi}\frac{\Phi_{e_{32}}}{\Phi_{X_3}}\,\mathrm{d}\omega\right)^{\frac{1}{2}} + \left(\int_{-\pi}^{\pi}\frac{|\Phi_{X_3X_2}|^2}{\Phi_{X_3}\Phi_{X_2}}\frac{\Phi_{e_{21}}}{\Phi_{X_2}}\,\mathrm{d}\omega\right)^{\frac{1}{2}},$$

where we have substituted the expression of $\hat{W}_{32}$. Finally, considering that

$$0 \leqslant \frac{|\Phi_{X_3X_2}|^2}{\Phi_{X_3}\Phi_{X_2}} \leqslant 1,$$

we find

$$\mathrm{d}(X_1, X_3) \leqslant \mathrm{d}(X_1, X_2) + \mathrm{d}(X_2, X_3).$$

Proof of proposition 5. The proof of the first property is straightforward because for every node, the algorithm considers an incident arc. Let us suppose that there is a cycle and k be the number of nodes $n_1, \ldots, n_k$ and arcs $a_1, \ldots, a_k$ of such a cycle. Every arc $a_1, \ldots, a_k$ has been chosen at step `2e` when the algorithm was taking into account one of the nodes $n_1, \ldots, n_k$. Conversely, every node $n_1, \ldots, n_k$ is also responsible for one of the arcs $a_1, \ldots, a_k$. Indeed, if a node n_i causes the selection of an arc $\hat{a} \notin \{a_1, \ldots, a_k\}$, then we are left with k arcs which cannot all be chosen by $k-1$ nodes.

Let us consider the node n_1. Without loss of generality, assume that it is responsible for the selection of the arc a_1 with weight d_1 linking it to the node n_2. According to the previous results, n_2 cannot be responsible for the choice of a_1. Let a_2 be the arc selected because of n_2 with weight d_2 and connecting it to n_3. Observe that necessarily $d_2 \leqslant d_1$. We may repeat this process till the node n_{k-1}. Hence, we obtain that every node n_i is connected to n_{i+1} by the arc a_i whose cost is $d_i \leqslant d_{i-1}$, for $i = 2, \ldots, k-1$. Finally consider n_k. It must be responsible for a_k which has to connect it to n_1 with cost $d_k \leqslant d_{k-1}$. Since d_k is incident to n_1, it holds that $d_1 \leqslant d_k$ Therefore, $d_1 \leqslant d_k \leqslant d_{k-1} \cdots \leqslant d_2 \leqslant d_1$ and we have the assertion of the second property.

About the third property, the upper bound N follows from the consideration that every node causes the choice of at most a new arc. In step `2c` of the algorithm, it may happen at most $\lfloor N/2 \rfloor$ times that we are forced to pick up an arc which is already in A. So we have at least $N - \lceil N/2 \rceil = \lfloor N/2 \rfloor$ arcs. □

References

[1] Blatt M, Wiseman S and Domany E 1996 Super-paramagnetic clustering of data *Phys. Rev. Lett.* **76** 3251

[2] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabási A L 2002 Hierarchical organization of modularity in metabolic networks *Science* **297** 1551

[3] Palmer R G, Stein D L, Abrahams E and Anderson P W 1984 Models of hierarchically constrained dynamics for glassy relaxation *Phys. Rev. Lett.* **53** 958–61

[4] Mantegna R N 1999 Hierarchical structure in financial markets *Eur. Phys. J.* B **11** 193–7

[5] Mardia K V, Kent J T and Bibby J 1979 *Multivariate Analysis* (London: Academic)

[6] Anderberg M 1973 *Cluster Analysis for Applications* (New York: Academic)

[7] Eisen M B, Spellman P T, Brown P O and Botstein D 1998 Cluster analysis and display of genome-wide expression patterns *Proc. Natl Acad. Sci.* **95** 14863–8

[8] Tumminello M, Coronnello C, Lillo F, Micciché S and Mantegna R N 2007 Spanning trees and bootstrap reliability estimation in correlation-based networks *Int. J. Bifurcations Chaos* **17** 2319–29

[9] Mantegna R N and Stanley H E 1995 Scaling behaviour in the dynamics of an economic index *Nature* **376** 46–9

[10] Gopikrishnan P, Plerou V, Amaral L A N, Meyer M and Stanley H E 1999 Scaling of the distributions of fluctuations of financial market indices *Phys. Rev.* E **60** 5305–16

[11] Naylora M J, Roseb L C and Moyle B J 2007 Topology of foreign exchange markets using hierarchical structure methods *Physica* A **382** 199–208

[12] Tumminello M, Lillo F and Mantegna R N 2007 Hierarchically nested factor model from multivariate data *Europhys. Lett.* **78** 30006

[13] Ljung L 1999 *System Identification: Theory for the User* 2nd edn (Upper Saddle River, NJ: Prentice-Hall)

[14] Kailath T, Sayed A H and Hassibi B 2000 *Linear Estimation* (Upper Saddle River, NJ: Prentice-Hall)

[15] Granger Clive W J 1969 Investigating causal relations by econometric models and cross-spectral methods *Econometrica* **37** 424–38

[16] Sayed A H and Kailath T 2001 A survey of spectral factorization methods *Numer. Linear Algebra Appl.* **8** 467–9

[17] Shiryaev A N 1995 *Probability* (New York: Springer)

[18] Friedrich R and Uhl C 1996 Spatio-temporal analysis of human electroencephalograms: petit-mal epilepsy *Physica* D **98** 171–82