Topological Properties in Identification and Modeling Techniques

G. W. Innocenti [*,**] **D. W. Materassi** [*,***]

* Dipartimento di Sistemi e Informatica
Centro per lo Studio di Dinamiche Complesse - CSDC
Universitá di Firenze,
via di S. Marta 3, 50139 Firenze, Italy
tel: +39 055 4796 360.
** E-mail: giacomo.innocenti@gmail.com
*** E-mail: materassi@dsi.unifi.it

Abstract: the paper deals with the problem of finding models and making predictions within a large set of time series or random processes. Nothing is assumed about their mutual influence and dependence. The problem can not be tackled efficiently, starting from a classical system identification approach. Indeed, the general optimal solution would provide a large number of models, since it would consider every possible interdependence. Then a suboptimal approach will be developed. The proposed technique will present interesting modeling properties which can interpreted in terms of graph theory. The application of this procedure will also be exploited as a tool to provide a clusterization of time series. Finally, we will show that it turns out to be a dynamical generalization of other techniques described in literature.

1. INTRODUCTION

Deriving information from data is a crucial problem in science. During the years a variety of contributions has been developed with different objectives in many fields, such as engineering, physics, biology and economy (see e.g. Eisen et al. [1998] and Tumminello et al. [2007]). The modeling approach can be seen as the attempt to derive a mathematical relation describing the dynamics of measured signals. Many techniques have been developed in scientific literature, especially in the econometrics and engineering fields. The Auto-Regressive Moving-Average (ARMAX) and Box-Jenkins models are among the most common approaches in the linear framework (see Kailath et al. [2000] and Alexander [2001]). The problem of the clusterization, instead, deals with the search of similarities and relations inside the original values, trying to catch the internal connections which divide the data into homogeneus subsets. The artificial neural network approach is based on the definition of a learning algorithm capable of detecting analogies and differences in the processed data set (see Rojas [1996] for a general overview). Data mining relies on methodologies to sort large amounts of data where the information is assumed implicit and previously unknown. Though these tools were developed in different fields and with different objectives, it is remarkable how "close" to each other they often result. In this paper we propose a modeling technique based on frequency analysis, which also provides a clusterization of the data set. Introducing the hypothesis of linear dynamical connections among the original values, a simplified model is first derived. Then, the result is interpreted in terms of graph theory and some clusterization properties are presented. The proposed approach is compared to some results in literature. It is remarkable that it turns out to be the dynamical extension of a well-known technique succesfully employed in the economic field.

Notation

- $\mathcal{Z}(\cdot)$ denotes the Zeta-transform of a signal
- $E[\cdot]$ denotes the mean operator
- $R_{XY}(\tau) := E[X(t)Y(t+\tau)]$ is the cross-correlation function which does not depend upon t if the processes X and Y are stationary.
- $R_X(\tau) := R_{XX}(\tau)$ is the autocorrelation
- $\Phi_{XY}(z) := \mathcal{Z}(R_{XY}(\tau))$ is the cross-power spectral density
- $\Phi_X(z) := \Phi_{XX}(z)$ is the power spectral density
- With abuse of notation $\Phi_X(\omega) = \Phi_X(e^{i\omega})$
- Causal truncation $\{\cdot\}_C$
- $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the ceiling and the floor function respectively

2. PROBLEM FORMULATION

Let us consider a set of N scalar time series $\{S_i\}_{i=1,...,N}$. Assume that it is possible to remove from them any deterministic component in order to obtain N stochastic processes $\{X_i\}_{i=1,...,N}$ which are wide sense stationary and with zero mean (see Shiryaev [1995]). We intend to derive a mathematical model describing, in a quantitative way, the possible connections and the mutual influences among the time series. We decide to depict each stochastic process X_i as the superposition of linear dynamical transformations of the other processes' outputs:

$$X_j(t) = e_j(t) + \sum_{j=1, j \neq i}^{N} W_{ji}(z)X_i(t) \qquad (1)$$

where $W_{ji}(z)$ is a suitable transfer function and e_j is the model error. In this framework, it can be considered

interesting to find the set of $\{W_{ji}(z)\}_{i,j=1...N,i\neq j}$ which allows us to best describe the time series according to the least squares criterion

$$\min \sum_j E\left[(Q_j(z)e_j)^2\right] \qquad (2)$$

where $Q_j(z)$ are dynamical weight functions. The description provided by (1) according to (2) could be exploited in order to make predictions or detect dynamical relations between random processes. The problem with such an approach is due to the complexity of the final model since it may be given by an apriori large number of transfer functions, namely $N(N-1)$. Hence, it is quite natural to develop a suboptimal strategy to reduce its complexity. A very intuitive approach is to consider only those transfer functions providing a most significative reduction of the cost. In this paper, for the sake of simplicity, we limit ourselves to the case of just a transfer function per process. Even though this is the simplest case, it still may provide useful insights about the connection topology and interesting theoretical properties.

3. NON CAUSAL SCENARIO

Given two stochastic processes X_i, X_j and a transfer function $W(z)$, let us consider the quadratic cost

$$E\left[(\varepsilon_Q)^2\right] \qquad (3)$$

where

$$\varepsilon_Q := Q(z)(X_j - W(z)X_i) \qquad (4)$$

being $Q(z)$ an arbitrary stable and causally invertible function weighting the error $X_j - W(z)X_i$. Then, the problem of evaluating the transfer function $\hat{W}(z)$ such that the quadratic cost (3) is minimized is well-known in scientific literature and its solution is referred to as the Wiener filter (see Kailath et al. [2000]).

Proposition 1. (Wiener filter). The Wiener filter modeling X_j by X_i is the linear stable filter $\hat{W}ji$ minimizing the filtered quantity (3). Its expression is given by

$$\hat{W}_{ji}(z) = \frac{\Phi_{X_iX_j}(z)}{\Phi_{X_i}(z)} \qquad (5)$$

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

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

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 need a causal filter in order to make predictions.

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

$$\Phi_{X_j}(z) = F_j^{-1}(z)(F_j^{-1}(z))^* \qquad (6)$$

with $F_j(z)$ stable and causally invertible (see Sayed and Kailath [2001]). In such a case the minimum cost assumes the value

$$minE[\varepsilon_{F_j}^2] = \int_{-\pi}^{\pi}\left(1 - \frac{|\Phi_{X_jX_i}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)}\right)d\omega. \qquad (7)$$

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

$$C_{X_iX_j}(\omega) := \frac{|\Phi_{X_jX_i}(\omega)|^2}{\Phi_{X_i}(\omega)\Phi_{X_j}(\omega)} \qquad (8)$$

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 an adimensional cost function not depending on the power of the signals as in (7).

Let us consider, now, a set of discrete zero mean and wide sense stationary stochastic processes Θ. The cost obtained by the minimization of the error ε_{F_j} using the Wiener filter as before allows us to define on Θ the binary function

$$d(X_i, X_j) := \left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(1 - C_{X_iX_j}(\omega)\right)d\omega\right]^{1/2} \qquad (9)$$
$$\forall X_i, X_j \in \Theta$$

Proposition 2. The function $d(\cdot,\cdot)$, as defined in (9) is a metric on Θ, that is

- $d(X_1, X_2) \geq 0$
- $d(X_1, X_2) = 0 \Leftrightarrow X_1 = X_2$
- $d(X_1, X_2) = d(X_2, X_1)$
- $d(X_1, X_3) \leq d(X_1, X_2) + d(X_2, X_3)$

for all $X_1, X_2, X_3 \in \Theta$

Proof. The only non trivial property to prove is the triangle inequality (or sublinear additivity). Let $\hat{W}_{ji}(z)$ be the Wiener filter between $X_i, X_j \in \Theta$ computed according to (5) 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

$$\Phi_{e_{31}}(\omega) \leq \Phi_{e_{32}}(\omega) + |\hat{W}_{32}(\omega)|^2\Phi_{e_{21}}(\omega) +$$
$$+ \Phi_{e_{32}e_{21}}(\omega)\hat{W}_{32}^*(\omega) + \hat{W}_{32}(\omega)\Phi_{e_{21}e_{32}}(\omega) =$$
$$\leq \Phi_{e_{32}}(\omega) + |\hat{W}_{32}(\omega)|^2\Phi_{e_{21}}(\omega) +$$
$$+ 2|\Phi_{e_{32}e_{21}}(\omega)||\hat{W}_{32}(\omega)| \leq$$
$$\leq (\sqrt{\Phi_{e_{32}}(\omega)} + |\hat{W}_{32}(\omega)|\sqrt{\Phi_{e_{21}}(\omega)})^2 \quad \forall\,\omega \in \mathbb{R}$$

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

$$\frac{\Phi_{e_{31}}}{\Phi_{X_3}} \leq \frac{1}{\Phi_{X_3}}(\sqrt{\Phi_{e_{32}}} + |\hat{W}_{32}|\sqrt{\Phi_{e_{21}}})^2$$
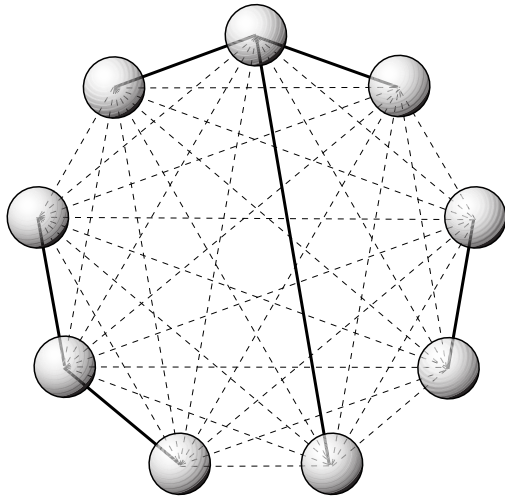
and considering the 2-norm properties

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

$$\left(\int_{-\pi}^{\pi} \frac{\Phi_{e_{31}}}{\Phi_{X_3}} d\omega\right)^{\frac{1}{2}} \leq$$

$$\leq \left(\int_{-\pi}^{\pi} \frac{\Phi_{e_{32}}}{\Phi_{X_3}} d\omega\right)^{\frac{1}{2}} + \left(\int_{-\pi}^{\pi} |\hat{W}_{32}|^2 \frac{\Phi_{e_{21}}}{\Phi_{X_3}} d\omega\right)^{\frac{1}{2}} =$$

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

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

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

we find

$$d(X_1, X_3) \leq d(X_1, X_2) + d(X_2, X_3).$$

□

According to what we have described in Section 2, we would like to model every single process X_j as the output of a SISO linear system whose input is the process X_i providing the largest reduction on the cost function (2). Considering the previous results, the solution to this problem can be provided by the following algorithm
Algorithm A:

1. initialize the set $A = \emptyset$
2. for every process X_j $(j = 1, ..., N)$
   2a. for every $i = 1, ..., N, i \neq j$
       compute the distance $d_{ij} := d(X_i, X_j)$;
   2b. define the set $M(j) := \{k | d_{kj} = \min_i d_{ij}\}$
   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.

It is possible to depict the previous procedure in terms of a graph theory intepretation (see Diestel [2006]) as illustrated in Figure 1. Represent every process X_j as a node and for each couple $(j, m(j))$ connect the j-th and $m(j)$-th nodes with an arc whose weight is $d_{m(j)j}$.

Proposition 3. The graph resulting from the algorithm A 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. The proof of the first property is straightforward because for every node the algorithm considers an incident arc. Let us suppose there is a cycle and be k the number of nodes $n_1, ..., n_k$ and arcs $a_1, ..., a_k$ of such a cycle. Every arc $a_1, ..., a_k$ has been chosen at the step 2e when the algorithm was taking into account one of the nodes $n_1, ..., n_k$. Conversely, every node $n_1, ..., n_k$ is also responsible for one of the arcs $a_1, ..., a_k$. Indeed, if a node n_i causes the selection of an arc $\hat{a} \notin \{a_1, ..., a_k\}$, then we are left with the 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 and linking it to the node n_2. According to the previous results, n_2 can not 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 $d_2 \leq d_1$. 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 $d_i \leq d_{i-1}$, for $i = 2, ..., k-1$. Finally consider n_k. It must be responsible for a_k which has to connect it to n_1 with cost $d_k \leq d_{k-1}$. Since d_k is incident to n_1 it holds that $d_1 \leq d_k$ Therefore $d_1 \leq d_k \leq d_{k-1} ... \leq d_2 \leq d_1$ 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 forces 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 □

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

Remark 4. 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 (see Shiryaev [1995]). Consequently, in such a case the expected topology of the graph is a forest (a graph with no cycles).

Remark 5. In general, nothing can be said about the connectivity. Therefore, the modeling procedure depicted by the algorithm A provides a clusterization of the original processes X_i which, for every node, minimizes the cost (3) according to the criterion of linear dynamic dependency. 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 can be easily replaced by a minimum spanning tree search.

Remark 6. The modelization we have derived makes use of non causal Wiener filters, thus it can be useful to detect

linear dependencies of any sort between the elements of the set Θ.

Unfortunately, the adoption of non causal filters can not be employed to make predictions. To this aim, the causal approach is faced in the next section.

4. CAUSAL FORMULATION

Given two stochastic processes X_i, X_j and a transfer function $W(z)$, let us consider again the quadratic cost (3). Analogously to the non causal case it is possible to derive a causal linear filter minimizing (3). The causal filter providing such a minimization is referred to as causal Wiener filter (see Kailath et al. [2000]).

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

$$\hat{W}_{ij}^C(z) = F_j^{-1}(z) \left\{ F_j(z) \frac{\Phi_{X_i X_j}(z)}{\Phi_{X_i}(z)} \right\}_C \qquad (10)$$

and it does not depend upon $Q(z)$.

Since the weighting function $Q(z)$ does not affect the Wiener filter, but only the energy of the filtered error, we can choose again $Q(z)$ equal to $F_j(z)$, the inverse of the spectral factor of $\Phi_{X_j}(z)$. This choice simply operates a normalization of the error spectrum at every frequency. Thus, similarly to what we have done in the non causal framework, we can define the function

$$d_C(X_i, X_j) := \left\{ E\left[(F(z)(X_j - \hat{W}_{ij}^C(z)X_i))^2 \right] \right\}^{\frac{1}{2}} \qquad (11)$$

to represent the modeling error on the process X_i due to the application of the causal Wiener filter to X_j. To highlight the properties of the present approach, let us derive first the main defference between d_C and the function d defined by (9). It is straightforward to observe that d_C is not symmetric. Therefore, d_C can not be a metric and that provides us with a less general tool to handle the connections among the processes. However, (11) still defines a quantitative description of the modeling errors in terms of their minimum powers in the causal framework. Hence, the original problem can be solved by minimizing the cost (2) applying the following algorithm
Algorithm B:

1. define the set $A := \emptyset$
2. for every process X_j $(j = 1, ..., N)$
   2a. for every $i = 1, ..., N, i \neq j$
        compute $d_{ij} := d(X_i, X_j)$;
   2b. define the set $M(j) := \{k | d_{kj} = \min_i d_{ij}\}$
   2c. choose any $m(j) \in M(j)$
   2d. choose the model
        $X_j = \hat{W}_{jm(j)}(z)X_{m(j)} + e_{jm(j)}$
   2e. add the ordered couple $(j, m(j))$ to A

As in the non causal case, this algorithm admits a usefull and more evocative graphical interpretation as depicted in Figure 2. For every process X_i define a node n_i. Then, for each couple $(j, m(j)) \in A$ take the directed arc linking the node $n_{m(j)}$ to n_j. As expected form the loss of symmetry, the causal model is depicted by a digraph. Therefore, the causality constraint, introduced to provide a prediction
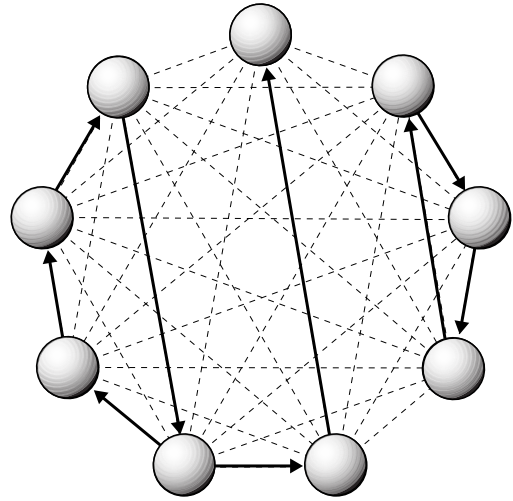

Fig. 2. The figures illustrates in a nine-node network all the possible connections between two nodes (dashed lines). The solid lines depict the clusters as they were the result after the application of the algoritm B. Each cluster has exactly one cycle with possible detours.

model, led us to represent the processes through a more complex network. However, the digraph obtained by the causal approach has still some important properties that we want to highlight.

Proposition 8. The graph resulting from the algorithm B has the following properties:

- on every node there is exactly an entering arc
- the graph has always at least a cycle
- every cycle is a directed cycle
- every connected component of the graph has exactly one cycle

Proof. The first property is a direct consequence of the setting up procedure which takes exactly an entering arc for each node.
For the second, observe that according the rule 2c it is impossible to choose the same directed arc evaluating different nodes. Thus, every node is responsible for the choice of a different arc, which implies that there is at least a cycle, because the graph has always exactly N nodes and N arcs.
The third property can be proved by contradiction. Consider a cycle and suppose that it is not a directed loop. Then, there must be at least one node with two arcs directed to it, but according to the rule 2d each node has exactly one arc directed to it, which is a contraddiction.
Finally, we will prove the forth property in two steps. First, consider a connected component of the graph and be k the number of its nodes. Because every node is responsible for the choice of just one different directed arc, there are also k arcs. This implies the presence of a cycle, that turns out to be a directed loop according to the third property. Now, let us proceed by contradiction. Suppose that there exist at least two different cycles $\mathcal{C}_1$, $\mathcal{C}_2$ in the same connected component. Because of the connectivity property, the two loops must share at least a common node. Indeed, if they do not, there must be a path linking them. Then, consider the two terminal points of this path. Clearly, each of them belongs to a cycle, so there is an arc of the cycle entering it.

Thus, their connections to the path must start from them, which implies that there must be a node in the linking path with two entering arcs. That can not be possible, then the two loops must have a common node. Therefore, consider $\mathcal{C}_1$ and find its node n_i which does not belong to $\mathcal{C}_2$, but that is connected with a directed arc strarting from it to a node n_j, shared by the two cycles. Because n_j belongs also to $\mathcal{C}_2$, there are two arcs of this cycle which are incident to n_j. According to the third property, one of them enters the node, but n_j has already an arc directed to it and starting from n_i and so it turns out to be a contraddiction. Hence, every connected component of the graph has exactly one cycle. $\qquad\square$

After the application of the algorithm, we obtain N causal models which can be exploited on order to make predictions. Unfortunately, the fact that we have to deal with a digraph prevents us to use some common tools to analyze or decompose the graph extracting topological information. However, we must consider that the algorithm itself ends providing a clusterization of the set Θ. In fact, the nodes are grouped together according to the prediction performances they can reciprocally provide.

5. RESTRICTION TO GAIN MODELS

In the previous two sections we have considered modeling and prediction problems for a set of random processes Θ. The best models and predictors (in the mean square sense) have been found in the space of linear transfer functions. We have observed that, in the non causal case, an apt choice of the error function allows one to define a metric on Θ. Conversely, in order to have useful predictors we need to limit ourselves to the space of causal systems. In this way, though, the cost function loses some interesting properties which would provide an easier topological analysis. In particular, we have seen that in the causal case the graph is more difficult to handle since it is oriented.
In Mantegna and Stanley [1995] and Mantegna and Stanley [1996] a technique to derive topological information from a set of random processes is described. 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

$$d_{ij} := \sqrt{2(1 - \rho_{ij})} \ . \tag{12}$$

Then, a graph is defined where every node represents a random process and the arc linking two nodes is weigthed according to (12). Eventually, the minimum spanning tree is extracted by the graph. This procedure has been successfully exploited to provide a quantitative and topological analysis of time series, expecially in the economic field (see Mantegna [1999], Tumminello et al. [2007] and Naylora et al. [2007]). It is worth considering that such a technique can be seen as a special case of the algorithms described in this paper. Indeed, it can be obtained by considering the set of all static gains as the model space along with the connectivity constraint. Consider the problem of modeling a process X_j by scaling another process X_i with a suitable constant α_{ji}. Choosing

$$\alpha_{ji} = \sqrt{\frac{E[X_j^2]}{E[X_i^2]}} \ , \tag{13}$$

we find that

$$E\left[(X_j - \alpha_{ji}X_i)^2\right] =$$
$$= E[X_j^2] \cdot E\left[\left(\frac{X_j}{\sqrt{E[X_j^2]}} - \frac{X_i}{\sqrt{E[X_i^2]}}\right)^2\right]$$
$$= 2 \ E[X_j^2](1 - \rho_{ij}) \ .$$

Hence, the distance (12) can be interpreted as the modeling error, properly normalized, when the simple gain (13) is used. It is important to remark that the choice of (13) is not optimal even in the space of models given by constant gains. Indeed, it is immediate to prove that the best choice of α_{ji}, in the sense of the least square error, is given by

$$\hat{\alpha}_{ji} = \frac{R_{X_j X_i}}{R_{X_i}} \tag{14}$$

and the relative quadratic error amounts to

$$E[e_{ji}^2] = R_{X_j} - \frac{R_{X_j X_i}^2}{R_{X_i}} \tag{15}$$

(see Kailath et al. [2000]). The choice of (14) is equivalent to the approch we followed in previous sections, but just narrowing the model space to the set of constant gains only. Indeed, let us proceed as in Section 3 in order to define an adimensional cost function not depending on the signal powers. To this aim, we consider again the square root of the normalized mean square error, obtaining

$$d(X_i, X_j) := \sqrt{\frac{E[e_{ji}^2]}{R_{X_j}}} =$$
$$= \sqrt{1 - \frac{R_{X_i X_j}^2}{R_{X_i} R_{X_j}}} = \sqrt{1 - \rho_{X_i X_j}^2} \ . \tag{16}$$

It is worth observing that (16) satisfies the same properties of (9), so it is a distance exactly as (12).

Proposition 9. The function (12) represents a distance on Θ.

Proof. The only non trivial property to show is the triangle inequality. Consider the following relations involving the optimal gains $\hat{\alpha}_{31}$, $\hat{\alpha}_{32}$, $\hat{\alpha}_{21}$

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

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

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

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

$$\sqrt{1 - \rho_{X_1 X_3}^2} \leq \sqrt{\frac{1}{R_{X_3}} \left(\sqrt{E[e_{32}^2]} + |\hat{\alpha}_{32}| \sqrt{E[e_{21}^2]} \right)^2} \leq$$

$$\leq \sqrt{\frac{E[e_{32}^2]}{R_{X_3}}} + |\hat{\alpha}_{32}| \sqrt{\frac{E[e_{21}^2]}{R_{X_3}}} =$$

$$= \sqrt{\frac{E[e_{32}^2]}{R_{X_3}}} + \rho_{X_2 X_3} \sqrt{\frac{E[e_{21}^2]}{R_{X_2}}}$$

Since $|\rho_{X_2 X_3}| \leq 1$, we have the assertion. $\qquad \square$

In Mantegna and Stanley [1995], the minimum spanning tree is extracted from the graph, according to the weights (12). This is equivalent to define a hierarchical structure of the time series relying on the adoption of linear gain models (13) between the processes. We stress that (13) represents a normalization of the random process energies, before applying the clusterization algorithm. On the other hand the choice of (14) allows one to set up the best gain model in the sense of least squares. 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. The use of Wiener filters we propose allows one to take into account more complex behaviors such as the presence of delays or even auto-regressive moving average dynamics. Hence, the more general procedure developed in this paper is expected to capture a larger amount of information. In particular, due to the filtering action, it is likely that the high frequency behaviour could be more accurately described (for an alternative approach see Bonanno et al. [2001]).

6. CONCLUSIONS

In this paper, we started from the problem of describing dynamical relations or making predictions within a large set of time series or random processes. No assumptions has been made about the relative influences and dependencies. The general optimal solution would provide a too large number of models since it would consider every possible dependence. Then a suboptimal approach has been taken into account. Every process has been described as the output of a SISO linear system driven by the process providing the "best" model. The procedure stems from the classic modeling/identification point of view, but an interpretation in terms of graph theory has been provided. Such an interpretation has allowed one to compare the technique with other ones in literature. It particular it has been shown that it generalizes some analysis techniques which have been successfully employed especially in the economic and financial fields. The main novel contribution of paper is in the fact that the proposed technique attempts to capture a topological structure describing not only static, but also dynamic relations between the time series. Only a theoretical framework has been taken into account leaving the application to real data as an object of future research.

ACKNOWLEDGEMENTS

REFERENCES

C. Alexander. *Market Models: A Guide to Financial Data Analysis*. Wiley, 2001.

G. Bonanno, F. Lillo, and R. N. Mantegna. High frequency cross-correlation in a set of stocks. *Quant. Fin.*, 1:96–104, 2001.

R. Diestel. *Graph Theory*. Springer-Verlag, Berlin, Germany, 2006.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

R. N. Mantegna. Hierarchical structure in financial markets. *Eur. Phys. J. B*, 11:193–197, 1999.

R. N. Mantegna and H. E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376:46–49, 1995.

R. N. Mantegna and H. E. Stanley. Turbulence and financial markets. *Nature (Scientific Correspondence)*, 383:587–588, 1996.

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

Raúl Rojas. *Neural networks: a systematic introduction*. Springer-Verlag New York, Inc., New York, NY, USA, 1996. ISBN 3-540-60505-3.

A. H. Sayed and T. Kailath. A survey of spectral factorization methods. *Numerical Linear Algebra with Applications*, 8:467–469, 2001.

A. N. Shiryaev. *Probability*. Springer-Verlag, New York, 1995.

M. Tumminello, C. Coronnello, F. Lillo, S. Micciche', and R. N. Mantegna. Spanning trees and bootstrap reliability estimation in correlation-based networks. *Int. J. of Bifurcations & Chaos*, 17:2319–2329, 2007.