Learning Causal Dependencies to Detect and Diagnose Faults in Sensor Networks

Cesare Alippi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Milano, Italy
cesare.alippi@polimi.it

Manuel Roveri
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Milano, Italy
manuel.roveri@polimi.it

Francesco Trovò
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Milano, Italy
francescol.trovo@polimi.it

Abstract—Exploiting spatial and temporal relationships in acquired datastreams is a primary ability of Cognitive Fault Detection and Diagnosis Systems (FDDSs) for sensor networks. In fact, this novel generation of FDDSs relies on the ability to correctly characterize the existing relationships among acquired datastreams to provide prompt detections of faults (while reducing false positives) and guarantee an effective isolation/identification of the sensor affected by the fault (once discriminated from a change in the environment or a model bias). The paper suggests a novel framework to automatically learn temporal and spatial relationships existing among streams of data to detect and diagnose faults. The suggested learning framework is based on a theoretically grounded hypothesis test, able to capture the Granger causal dependency existing among datastreams. Experimental results on both synthetic and real data demonstrate the effectiveness of the proposed solution for fault detection.

I. INTRODUCTION

Detecting faults in sensor networks is a challenging and relevant activity, since harsh environmental conditions might induce faults and ageing effects in the embedded electronics or the sensors. In such scenarios the prompt detection of faults is critical to avoid their propagation to the rest of the system (cascade effects) and prevent that any subsequent application processes faulty or incorrect data.

Fault Detection and Diagnosis Systems (FDDSs) are systems able to detect, isolate, identify and possibly accommodate faults: they have been widely studied in the literature (e.g., see [1] and [2] for comprehensive reviews). To guarantee promptness in detection and effectiveness in the isolation/identification phases, these systems generally require (at least partial) information about the data-generating process in faulty-free conditions as well as information about the faults that might occur (e.g., the fault signature or magnitude).

Recently, a new generation of FDDSs has been proposed to overcome the limits of traditional FDDSs. These systems, which are generally referred to as Cognitive FDDSs [3], [4], [5], [6], do not require any a priori information about the data-generating process or the nature of occurring faults and are characterized by the ability to exploit the temporal and spatial dependencies that are present in the datastreams acquired by the sensor network. For instance, [4] presents a FDDS for sensor networks exploiting Hidden Markov Models (HMM) working in the parameter space of linear models approximating sensor-to-sensor relationships. [5] and [6] propose fault diagnosis systems able to capture spatial and temporal dependencies between couples of sensors. These systems work in the parameter and functional spaces of estimated predictive models, respectively, to create the fault dictionary and distinguish among different fault signatures.

Since a fundamental aspect of Cognitive FDDSs is the ability to exploit temporal and spatial dependencies existing among the acquired datastreams, it is of central importance to consider only those functional relationships that provide useful information for detection and diagnosis purposes. Thus, a reliable method to infer the dependency graph, defined as the directed graph modeling functional dependencies existing among sensors in the network, is necessary. In such a graph, nodes represent the network sensors, while edges model the relationships existing among a set of datastreams. To address this problem, [4] suggested a dependency graph based on the crosscorrelation analysis: an edge between two sensors exists when the maximum value of the crosscorrelation between the two datastreams is larger than a user-defined threshold. The main drawbacks of the aforementioned approach are the lack of a clear characterization of the user-defined threshold (it is difficult to set it at design time), the capability to assess relationships only between couples of datastreams (a relationship involving more than two datastreams cannot be evaluated through cross-correlation) and the fact that results are heavily dependent on noise (as commented in the experimental section). A different approach for modeling relationships in sensor networks encompasses Bayesian networks [7]. The main drawback of such an approach is the fact that information-theoretic mechanisms usually require information or assumptions about the data generating process [8].

In this paper, we extend the dependency graph presented in [4], by introducing a statistical framework able to learn temporal and spatial dependencies in sensor networks. The framework is based on the Granger causal dependency test [9], which, given a confidence level and a dataset in faulty-free conditions, is able to evaluate whether data coming from one stream can be used to improve the prediction of another one. In particular, we rely on the concept of multivariate conditioned Granger causality [10], which operates on multivariate datastreams to assess the causality between two datastreams (conditioned to the information provided by the other sensors in the network). Interestingly, the Granger causality has been successfully applied to different fields like economic [11] and medical applications (e.g., fMRI data [12] and EEG analyses.
[13], [14]). In this paper, for the first time in the literature, the Granger causality is considered for fault detection/diagnosis. The proposed novel Granger-based learning framework for the creation of dependency graphs in sensor networks is characterized by the following advantages:

- a statistically sound hypothesis test to assess the temporal and spatial dependency in datastreams;
- the ability to take into account multiple input relationships;
- the performance of the proposed technique is easily tunable by the confidence level of the Granger causal dependency test.

The rest of the paper is structured as follows: a brief introduction about Cognitive FDDSs is provided in Section II, while Section III presents an overview of the Granger causality and its application to fault detection/diagnosis. Section IV shows the use of the proposed Granger-based statistical framework in a real Cognitive FDDS, while in Section V experiments on both synthetic and real data are provided. Finally, conclusions are drawn in Section VI.

II. COGNITIVE FAULT DETECTION/DIAGNOSIS SYSTEMS FOR SENSOR NETWORKS

Let us consider a sensor network composed by \( n \in \mathbb{N} \) sensors \( s_1, \ldots, s_n \), monitoring a process \( \mathcal{X} \), whose model is unknown. At each time instant \( t \in \mathbb{N} \) the sensors provide the measurements column vector:

\[
X(t) = [x_1(t), \ldots, x_n(t)]^T \in \mathbb{R}^n,
\]

where \( x_i(t) \in \mathbb{R} \) is the scalar measurement acquired at time \( t \) by sensor \( s_i \). Define \( x_i = \{x_i(t), t \in \mathbb{N}\} \) to be the datastream acquired by sensor \( s_i \). Cognitive FDDSs rely on the ability to model functional relationships existing among the data acquired by sensors by analyzing the datastream vector \( \mathcal{X} = \{x_1, \ldots, x_n\} \). Each binary relationship \( f_{ij} \) introduces a constraint between the data acquired by sensor \( s_i \) and sensor \( s_j \). A dependency graph \( \mathcal{G} \) modeling the relationships existing among datastreams \( x_i, x_j \), is defined as:

\[
\mathcal{G} = (V, E),
\]

where \( V = \{s_1, \ldots, s_n\} \) is the set of network sensors and \( E \) is a set of directed edges connecting sensors. More specifically, the directed edge \( e_{ij} = (s_j, s_i) \) is included in \( E \) only when the relationship \( f_{ij} \) between datastreams \( x_i \) and \( x_j \) exists. An example of dependency graph for a sensor network is shown in Figure 1.

The general architecture of Cognitive FDDSs is depicted in Figure 2. These systems initially learn the dependency graph \( \mathcal{G} \) and the functional relationships corresponding to edges in \( E \) by relying on a nominal (faulty-free) training sequence. Both the functional relationships and the dependency graph represent the nominal concept of the sensor network. Deviations from this nominal concept represent anomalous situations that must be detected and inspected by the Cognitive FDDS. In fact, during the operational life, each functional relationship is inspected to detect variations with respect to the nominal concept. When a change in \( f_{ij} \) is detected, this might be induced by a fault (either in sensor \( s_i \) or \( s_j \)), by a change in the environment in which the sensor network operates or by a false detection, e.g., induced by a model bias in the estimation of relationship \( f_{ij} \). To discriminate among these three cases Cognitive FDDSs exploit the dependency graph \( \mathcal{G} \) as follows [4], [3]: if a fault affects sensor \( s_i \) or \( s_j \), the relationships that are connected to that sensor will perceive a change; if there is change in the environment, all the relationships in \( E \) would be affected; if there is model bias in \( f_{ij} \), none of the other relationships in \( E \) would perceive a change.

Since the dependency graph is at the basis of the fault detection, isolation and identification activities of Cognitive FDDSs, the ability to learn it correctly is fundamental to guarantee the fault detection/diagnosis performance. In particular, the ability to correctly learn the dependency graph is crucial, since it allows the Cognitive FDDS to rely only on the relevant relationships (i.e., those relationships that really exist among streams of data). This allows the Cognitive FDDS to increase the fault detection/isolation performance for two main reasons. At first, if we keep not relevant relationships, false positives...
in detection will increase, since all relationships are inspected for changes in parallel. Second, if a relationship providing information is not included in $E$, we can not use it in the isolation phase to distinguish between the occurrence of a fault, a change in the environment or a model bias.

III. LEARNING CAUSAL DEPENDENCIES WITH GRANGER CAUSALITY

In this section, we describe the proposed statistical framework based on Granger causality to learn the dependency graph. The aim of this framework is to automatically select those relationships presenting causal dependency, given a faulty-free set of measurements coming from the sensor network and a predefined level of confidence regulating the probability of including not relevant relationships in $E$. In particular, in Section III-A we describe the technique for assessing the causal dependency of a single relationship, while in Section III-B we extend this learning mechanism to the entire dependency graph.

A. Modeling the relationship between a couple of datastreams

To model the dependency between two datastreams we rely on causal dependency as defined by Granger (1969) in [15]:

**Definition 1 (Granger Causality).** Consider a bivariate discrete-time stochastic vector $(x_i, x_j)$. We say that $x_j$ is Granger causing $x_i$ if the presence of $x_j$ allows for better prediction performance of $x_i$ at time $t$.

The above definition is quite general and relies on the comparison between two different reconstruction abilities: with and without $x_j$ as additional source of information. In sensor networks scenario, a more specific definition of Granger causality can be considered [10]:

**Definition 2 (Multivariate Conditioned Granger Causality).** Consider a multivariate discrete-time stochastic vector $X$, we say that $x_j \in X$ is Granger causing $x_i \in X, i \neq j$, conditioned to $X \setminus \{x_i, x_j\}$, if we are better able to predict $x_i$ by using $X$ instead of using $X \setminus \{x_j\}$.

Starting from the last definition, it is possible to derive a statistical test able to assess the multivariate conditioned Granger causality between two datastreams. To assess if $x_j$ Granger causes $x_i$, we need to model the relationship existing among all datastreams provided by the sensor network as a linear Vector AutoRegressive (VAR) model [16], [8]. More specifically, following Definition 2, we want to assess the influence of $x_j$ in the prediction of $x_i$. To achieve this goal we consider two different versions of the VAR model: $\mathcal{M}_f^{(i)}$ (full model), i.e., the one that considers all the available datastreams $X$ to predict $x_i$, and $\mathcal{M}_r^{(i)}$ (reduced model), i.e., the one considering as predictors only $X \setminus \{x_j\}$. More formally, the two predictive models assume forms:

$$\mathcal{M}_f^{(i)}: x_i(t) = \sum_{k=1}^{\tau} \sum_{h=1}^{n} a_{ihk} x_h(t-k),$$

$$\mathcal{M}_r^{(i)}: x_i(t) = \sum_{k=1}^{\tau} \sum_{h=1, h \neq j}^{n} a'_{ihk} x_h(t-k),$$

where $a_{ihk} \in \mathbb{R}$ and $a'_{ihk} \in \mathbb{R}$ are the regression coefficients modeling the linear relationship between measurements coming from sensors $s_i$ and $s_h$ at the $k$-th time lag for model $\mathcal{M}_f^{(i)}$ and $\mathcal{M}_r^{(i)}$, respectively, and $\tau \in \mathbb{N}$ is the order of the model.

It is possible to show that the problem of assessing the multivariate conditional Granger causality can be formulated as a loglikelihood ratio test between the full model $\mathcal{M}_f^{(i)}$ and the reduced model $\mathcal{M}_r^{(i)}$ [18], where the null hypothesis is that $x_j$ is not Granger causing $x_i$ conditioned to $X \setminus \{x_i, x_j\}$. More specifically, given a training sequence

$$Z_N = \{X(t)\}_{t=1}^{N} \quad N \in \mathbb{N},$$

where $X(t)$ is a realization of the stochastic process $X$ at time $t$, we estimate parameters $\hat{a}_{ihk}$ and $\hat{a}'_{ihk}$ by means of a Least Square (LS) procedure applied to $Z_N$. The loglikelihood ratio test is based on the computation of the test statistic:

$$F_{ij} = \frac{p_f - p_r}{N - p_f - 1} \frac{SSR_{r}^{(i)}(\hat{\alpha}) - SSR_{f}^{(i)}(\hat{\alpha})}{SSR_f^{(i)}} \sim F(p_f - p_r, N - p_f - 1),$$

(3)

where $p_f = n\tau$ is the number of parameters of the full model $\mathcal{M}_f^{(i)}$, $p_r = (n-1)\tau$ is the number of parameters of the reduced one $\mathcal{M}_r^{(i)}$, $F(p_f - p_r, N - p_f - 1)$ is the Fisher distribution with $p_f - p_r$ and $N - p_f - 1$ degrees of freedom, $SSR_f^{(i)}$ is the sum of squared residual of the model $\mathcal{M}_f^{(i)}$ on $Z_N$ and $SSR_r^{(i)}$ is the sum of squared residual of the reduced model $\mathcal{M}_r^{(i)}$ on the same training set. In the test, the critical region of level $\alpha$ for the statistic is:

$$R_{\alpha}^{ij} = \{ F \in \mathbb{R}^+ \mid F \geq F_{\alpha}(p_f - p_r, N - p_f - 1) \},$$

(4)

where $F_{\alpha}(p_f - p_r, N - p_f - 1)$ is the quantile of order $1 - \alpha$ of the Fisher’s distribution with $p_f - p_r$ and $N - p_f - 1$ degrees of freedom. When there is statistical evidence for rejecting the null hypothesis with confidence $\alpha$ (i.e., $F_{ij} \in R_{\alpha}^{ij}$), we say $F_{ij} \rightarrow_{a.s.} x_i \rightarrow z$, i.e., the datastream $x_j$ Granger causes $x_i$ conditioned to $z = X \setminus \{x_i, x_j\}$, since the datastream $x_j$ improves the prediction of $x_i$.

It is worth noting that, in principle, to evaluate the contribution in prediction abilities of datastreams $x_j$ to stream $x_i$, one could inspect the exogenous coefficients corresponding to $x_j$ in the full model $\mathcal{M}_f^{(i)}$, i.e., $a_{ijk} \forall k \in \{1, \ldots, \tau\}$. When these coefficients are statistically null (meaning that stream $x_j$ does not provide improvement in the prediction of stream $x_i$), there is no causal dependency of datastreams $x_i$ from $x_j$. On the contrary, when at least one coefficient differs from zero, a causal dependency relation between the two datastreams exists. Interestingly, it is shown in [18] that the problem of identifying a causal dependency between datastreams $x_i$ and $x_j$, formulated in Equation (4), is equivalent to a hypothesis test which assesses if at least one of the coefficients $a_{ijk}, k \in \{1, \ldots, \tau\}$ is statistically different from zero, i.e.:

$$H_0 : a_{ijk} = 0 \ \forall k \quad \text{vs.} \quad H_1 : \exists k \mid a_{ijk} \neq 0.$$

1Albeit VAR models are linear models, it is possible to show that, under mild assumptions about the data-generating process (e.g., stationarity of the covariance), they are general enough to model several nonlinear multivariate time series [17].
B. Creating the Granger-based dependency graph

Following the definition of Granger causality and the statistical test described above, it is possible to derive a statistical framework to learn the dependency graph for the whole sensor network. A detailed description of the proposed statistical framework is provided in Algorithm 1.

Algorithm 1 Granger-based dependency graph learning algorithm

1: Input: faulty-free dataset \( Z_N = \{X(t)\}_{t=1}^N \), confidence level \( \alpha \)
2: Output: dependency graph \( G \)
3: Set an empty dependency graph \( G \), (Line 1) from data \( Z_N \);
4: for \( i \in \{1, \ldots, n\} \) do
5: Estimate coefficients \( \hat{a}_{ihk} \) of VAR model \( M_{(i)} \) in Equation (1) from data \( Z_N \);
6: for \( j \in \{1, \ldots, n\} \) do
7: if \( j \neq i \) then
8: Compute coefficients \( \hat{a}_{ijh} \) of VAR model \( M_{(ij)} \) in Equation (2) from data \( Z_N \);
9: Compute \( F_{ij} \) as in Equation (3);
10: if \( F_{ij} \geq F_{\alpha}(p_f - p_r, N - p_f - 1) \) then
11: \( E \leftarrow E \cup \{e_{ij}\} \);
12: end if
13: end if
14: end for
15: end for

At first (Line 1), we consider a training set \( Z_N \) of faulty-free data coming from the sensor network and a user defined confidence level \( \alpha \in (0, 1) \) for the loglikelihood ratio test. A graph \( G = (V, E) \) with \( V = \{s_1, \ldots, s_n\} \) and \( E = \emptyset \) is initially considered (Line 3). By considering a couple of sensors \( s_i \) and \( s_j \) and the dataset \( Z_N \), we are able to compute the test statistics \( F_{ij} \) (Line 11), as in Equation (3). When

\[ F_{ij} \geq F_{\alpha}(p_f - p_r, N - p_f - 1), \]

we have statistical evidence (with confidence \( \frac{\alpha}{n(n-1)} \)) that \( x_j \) Granger causes \( x_i \), conditioned to all the other streams of data, and we add \( e_{ij} = (s_j, s_i) \) to the edge set \( E \) (Line 11). We repeat the test for each couple of datastreams to build the dependency graph (Lines 4-15). Note that we considered confidence level \( \frac{\alpha}{n(n-1)} \) to obtain an overall confidence level of \( \alpha \), thanks to the Bonferroni correction for multiple hypothesis [18].

IV. AN HMM-BASED COGNITIVE FDDS WITH GRANGER-BASED DEPENDENCY GRAPH

As a relevant example, we consider here the use of the proposed Granger-based statistical framework for the Hidden Markov Model(HMM)-based Cognitive FDDS presented in [4]. The bases of this Cognitive FDDS are the use of a HMM Change Detection Test (HMM-CDT) working in the parameter space of linear time-invariant predictive models (modeling the functional relationships in \( G \)) for fault detection and the presence of a cognitive level able to exploit \( G \) to identify and isolate the fault.

The joint use of the Granger-based statistical framework and the HMM-based Cognitive FDDS is provided in Algorithm 2.

Algorithm 2 HMM-CDT algorithm

1: Input: dependency graph \( G = (V, E) \), faulty-free dataset \( Z_N = \{X(t)\}_{t=1}^N \), faulty-free validation set \( Z_O = \{X(t)\}_{t=N+1}^{N+O} \), HMM-CDT parameter \( C \)
2: Output: detection time \( t \), cognitive level information \( I \)
3: for all edge \( e_{ij} \in E \) do
4: Estimate the sequence of model parameters \( \Theta_{ij} \) from \( Z_N^{(ij)} \);
5: Estimate HMM \( \hat{H}_{ij} \) using \( \Theta_{ij} \);
6: Compute the detection threshold \( T_{ij} \) from \( Z_O^{(ij)} \) as in [19];
7: end for
8: while a new vector \( X(t) \) is available do
9: for all edge \( e_{ij} \in E \) do
10: Estimate \( \hat{\theta}_{ij}^{(t)} \) by using \( Z_{ij}^{(t)} \);
11: Build the sequence \( \Theta_{ij}^{(t)} \);
12: Compute \( l_{ij}(t) \) as in [19];
13: if \( l_{ij}(t) \leq T_{ij} \) then
14: Activate the cognitive level based on \( G \) as in [4] to get \( I \);
15: return \( i \leftarrow t, I \);
16: end if
17: end for
18: end while
19: return \( i \leftarrow 0, I \leftarrow 0 \);

After having applied the causal dependency graph algorithm described in Algorithm 1 to learn \( G \), we rely on the faulty-free sequence \( Z_N \) to characterize the nominal concept by estimating a HMM \( \hat{H}_{ij} \) for each edge \( e_{ij} \in E \) (Line 5). More specifically, for each edge \( e_{ij} \in E \) we consider a predictive SISO ARX model on the corresponding dataset \( Z_{ij}^{(t)} = \{(x_{ij}(t), x_{i}(t))\}_{t=1}^N \):

\[ \dot{x}_i(t) = \theta_{ij}^T \begin{pmatrix} x_i(t-1), \ldots, x_i(t-na), \\ x_j(t-1), \ldots, x_j(t-nb) \end{pmatrix} \]

where \( \dot{x}_i(t) \) is the prediction of the datastream \( x_i \) at time instant \( t \), \( \theta_{ij} \) is the parameter vector characterizing the predictive model and \( na, nb \) are the orders of the autoregressive and exogenous part of the model, respectively. By considering overlapping batches of data \( Z_{ij}^{(t)} = \{(x_{ij}(t), x_{i}(t))\}_{t=1}^{h+M-1} \), \( h \in \{1, \ldots, L\}, L = N - M + 1 \), it is possible to extract a sequence of estimated parameter vectors \( \Theta_{ij} = \left( \hat{\theta}_{ij}^{(1)}, \ldots, \hat{\theta}_{ij}^{(L)} \right) \). As pointed out in [19], the use of a HMM ruled by a mixture of Gaussians (GMM) over a sequence of estimated parameter vectors \( \Theta_{ij} \) is the natural solution to model a functional relationship in the parameter space between two generic datastreams. The HMM-CDT for a couple of datastreams \( x_j \) and \( x_i \) relies on the initial parameter vector sequence \( \Theta_{ij} \), used to train a HMM \( \hat{H}_{ij} \), which aims at capturing the statistical behaviour of the estimated parameter vectors sequence \( \Theta_{ij} \) in nominal conditions.

Then, during the operational phase (i.e., \( t > N + O \)), as soon as a new sample \( (x_{ij}(t), x_{i}(t)) \) is available, the statistical
affinity between the newly estimated parameter vector $\hat{\theta}_{ij}^{(t)}$ and $\mathcal{H}_{ij}$ is assessed by looking at the HMM loglikelihood $l_{ij}(t) = l(\mathcal{H}_{ij}, \Theta_{ij}^{(t,k)})$ [19], where $\Theta_{ij}^{(t,k)} = \left(\hat{\theta}_{ij}^{(t-k+1)}, \ldots, \hat{\theta}_{ij}^{(t)}\right)$ is the sequence of the last $k$ estimated parameter vectors. If the relationship does not change over time, the loglikelihood $l_{ij}(t)$ remains above a threshold $T_{ij}$, which is computed on a faulty-free validation set $Z_O^{(ij)} = \{(x_i(t), x_j(t))\}_{t=N+1}^{\infty}$ and which depends on the parameter $C \in \mathbb{R}^+$ (see [19] for a detailed description). Otherwise, in case of a variation in the relationship $f_{(i,j)}$, $l_{ij}(t)$ decreases below the threshold $T_{ij}$, since $\Theta_{ij}^{(t,k)}$ is no more compatible with the statistical model characterized by $\mathcal{H}_{ij}$. It is worth noting that the parameter $C$ is crucial in the HMM-CDT algorithm since it allows to regulate the false positive detections rate: high values for $C$ decrease false positive detection rate, at the expense of an increase of both the detection delay and false negatives. On the contrary, the use of low values of $C$ decreases both false negatives and detection delay, but increases the chance of false positives in detection.

If a detection occurs in a given functional relationship $f_{(i,j)}$, an alarm is raised, the detection time $t = t$ is returned (Line 15) and the cognitive level is activated (Line 14). This algorithm, taking advantage of the learned dependency graph $\mathcal{G}$, is able to distinguish among changes in the environment, model bias and fault (which is also isolated if possible). Further details about the cognitive level can be found in [4].

We emphasize that both the performance and the complexity of the FDSS selected in Algorithm 2 is highly influenced by the learned Granger-based dependency graph. In fact, a HMM-CDT is considered for each edge $e_{ij} \in E$, while the dependency graph is needed to distinguish among faults, change in the environment and model bias.

V. EXPERIMENTAL RESULTS

The aim of this section is to evaluate the performance of the proposed statistical framework both on synthetic (Application D1) and real data coming from a real-world sensor network for rock collapse forecasting (Application D2). In particular, two different scenarios have been considered and evaluated: 1) the ability to correctly identify the meaningful relationships existing in acquired datastreams; 2) the fault detection performance of the HMM-based Cognitive FDSS relying on the Granger-based dependency graph described in Section IV.

A. Figures of merit

As regards the ability to identify the correct relationships in acquired datastreams, we consider the following figures of merit:

- **Recall $R = \frac{|E_\cap E|}{|E_\cap E|}$**, i.e., the fraction of relevant functional relationships that are present in the learned dependency graph;

- **Precision $P = \frac{|E_\cap E|}{|E_\cap E|}$**, i.e., the fraction of functional relationships in the learned dependency graph that are relevant,

where $E_r$ is the set of truly existing relationships among acquired data and $|\cdot|$ is the cardinality operator.

Differently, for the evaluation of the fault detection performance, we consider the following figures of merit:

- **False Negatives rate (FN)**, i.e., the fraction of experiments where the HMM-CDT did not detect any change in the relationship interested by the fault (or environmental change);

- **False Positives rate (FP)**, i.e., the fraction of experiments in which the HMM-CDT detected a change when it was not;

- **Detection Delay (DD)**, i.e., the number of samples necessary to detect a change, if no FP occurred.

B. Comparison

To compare the performance of the proposed Granger-based statistical framework, we considered the crosscorrelation analysis adopted in [4]. In particular, we considered three different values of the crosscorrelation threshold: high crosscorrelation ($\rho_{high} = 0.9$), low crosscorrelation ($\rho_{low} = 0.02$) and best crosscorrelation ($\rho_{best}$ defined as the highest value of correlation that allows to consider all the truly existing relationships in the dependency graph). Obviously, the best crosscorrelation threshold cannot be a priori set (since it requires the knowledge of the true relationships present within the network) and, hence, it represents the correct value for the threshold.

C. Parameters configuration

The confidence on the Granger-based statistical framework is set to $\alpha = 0.05$. The parameters of the HMM-CDT have been assigned as follows: $M = 100$, $k = 10$, the best configuration for the HMMs was selected by considering $s \in \{3, 4, 5, 6\}$ and the number of GMM Gaussian is in $\{1, 2, 4, 8, 16, 32\}$. The orders of the SISO ARX models are chosen through a model selection procedure based on the mean square error on a validation set. We considered the implementation of the Granger causal dependency test provided by the toolbox developed in [8].

D. Application D1: synthetic dataset

a) Experimental setting: Here, we consider datastreams coming from a synthetically generated sensor network with randomly generated relationships among datastreams. At first, a directed acyclic graphs with fixed amount of nodes, i.e., $n = 6$, and fixed number of edges, i.e., $|E| = 6$, is generated. After that, by following the graph topological order, datastreams are generated either by linear $L$ or by sinusoidal $S$ models as:

\[
L: \quad x_i(t) = \theta_{ij}^L [x_i(t-1), x_i(t-2), x_j(t-1), x_j(t-2)] + \eta(t);
\]

\[
S: \quad x_i(t) = \sin \left\{ \theta_{ij}^S [x_i(t-1), x_i(t-2), x_j(t-1), x_j(t-2)] \right\} + \eta(t),
\]

where the input datastream $x_j$ is given either by the computation of streams preceding $x_i$ in topological order or generated from $x_j(t) = \phi x_j(t-1) + \psi(t)$ ($\phi = 0.4$, $\psi(t) \sim \mathcal{N}(0, 0.1)$), $\eta(t) \sim \mathcal{N}(0, \sigma^2)$ is an uncorrelated white noise, $\theta_{ij} \in \mathbb{R}^4$ is a randomly generated vector s.t. the resulting model is stable. A sequence of 7145 samples was generated and the
first $N = 4085$ used for the training phase, while the remaining ones constitute the test set for the detection phase.

In the experiments regarding the ability to correctly identify the meaningful relationships existing in acquired datastreams, we considered 20 equidistant noise levels $\sigma \in [0.05, \ldots, 1]$. Differently, in the experiments on the fault detection performance, an additive constant abrupt fault of magnitude $0.2 \times (\text{max}_t x_i(t) - \text{min}_t x_i(t))$ was injected at time instant 5309 on each sensor one at a time (to model a sensor fault) and on all the sensors at the same time (to model a change in the environment). Results are averaged over the different faults and repeated over 100 runs.

b) Results: The experimental results on the ability to correctly identify the meaningful relationships are presented in Figures 3 to 6. As expected, the choice of the threshold value for the crosscorrelation analysis is critical. In fact, it is possible to see that the choice of a low correlation threshold $\rho_{low}$ implies the selection of $|E| \approx 30$ relationships, leading to a recall $R \approx 1$. As expected, the drawback of this solution is the fact that the precision $P \approx 0.4$, thus considering several non-relevant relationships. Conversely, the main drawback of the high correlation threshold $\rho_{high}$ is the fact that relevant relationships might not be included into the dependency graph. More specifically, when $\sigma > 0.1$, the crosscorrelation method does not select any relationship. As expected, by analysing Figure 7, we can see that the best correlation value $\rho_{best}$ decreases as noise increases. Thus, one should rely on a priori information about the noise level to be able to set the proper threshold for the crosscorrelation analysis. Conversely, the proposed statistical framework based on Granger causality is able to maintain acceptable values for both precision and recall ($P \approx 1$ and $R > 0.7$) without any a priori information about the noise level.

Results about fault detection are shown in Table I. We considered three different noise levels $\sigma \in \{0.05, 0.1, 0.3\}$ and three different values of the parameter $C$ for the HMM-CDT, to allow a fair comparison among the proposed Granger-based framework (identified from now on with $G$), the crosscorrelation algorithm with $\rho_{low}$ and the crosscorrelation algorithm with $\rho_{best}$ (the one with $\rho_{high}$ is not considered here). More specifically, $C_{gran}$ is the smallest parameter that allows to have $FP = 0$ for the HMM-CDT based on the Granger-based dependency graph. Similarly $C_{best}$ and $C_{low}$ are the ones allowing to have $FP = 0$ for the HMM-CDT based on crosscorrelation analysis graph, by using $\rho_{best}$ and $\rho_{low}$, respectively, as thresholds for the crosscorrelation. By looking at the experiments on the linear function $L$ and $\sigma = 0.05$, we can see that, with $C_{gran}$, the Granger-based solution provides $DD$ and $FN$ similar to the ones obtained with $\rho_{best}$ ($DD = 49.7$ and $FN = 0.113$ for $G$, $DD = 49.2$ and $FN = 0.079$ for $\rho_{best}$), and $G$ provides a lower $FP$ ($FP = 0$ for $G$ and $FP = 0.41$ for $\rho_{best}$) without requiring any a priori information about the network or the noise level.

By comparing the Granger-based framework with crosscorrelation with $\rho_{low}$, it presents a higher values for detection delays and false positives rate ($DD = 65$ and $FP = 0.65$) and is able to maintain a lower false negatives rate ($FN = 0.044$). Moreover, the amount of considered edges for $G$ is $|E| = 6$, while both crosscorrelation analysis methods were proposing $|E| > 10$ edges.

Table I highlights in bold those results with threshold $C$ s.t. $FP = 0$, to easily compare $|E|$, $DD$ and $FN$ for
the considered methods. By considering the $G$ performance with $C_{\text{gran}}$ and those obtained with crosscorrelation analysis with $\rho_{\text{best}}$ and $C_{\text{best}}$, we can observe that false negative rates for both methods are similar ($FN \approx 0.11$), while $G$ is characterized by lower detection delay and uses a smaller set of relationships ($|E| = 6$ for $G$ and $|E| = 10.1$ for $\rho_{\text{best}}$). At last, if we compare $G$ with $C_{\text{gran}}$ and $\rho_{\text{low}}$ with $C_{\text{low}}$, we can see that the latter presents a slightly lower number of false negative detections, while $G$ provides lower detection delay ($DD = 49.7$ for $G$ and $DD = 63.4$ for $\rho_{\text{low}}$). This behaviour suggests that, if the only concern is to maintain low $FN$, one should consider all the available relationships at the expense of high $FP$s. Results about the sinusoidal case $S$, as well as those with different noise level, are in line with those of the linear one $L$ with $\sigma = 0.05$.

E. Application D2: Rialba dataset

c) Experimental setting: In this set of experiments, we are considering data coming from a real deployment of a sensor network in the Alpes (Torrioni of Rialba) [20] designed for rock collapse forecasting. In particular, we consider $n = 10$ sensors (i.e., 3 external temperature, 3 in hole temperature...
and 4 accelerometers). A total of 1285 data are considered for each sensor and the first $N = 800$ samples are used for training. Abrupt additive constant faults of magnitude $0.5 \times (\max_i x_i(t) - \min_i x_i(t))$ are injected at time instant 1042 in each of the streams $s$ at a time (to model fault in a sensor) and in all the sensors at the same time (to model a change in the environment).

$d)$ Results: Since we do not have information about the relevant relationships existing among acquired data, we neither can evaluate $R$ and $P$, nor compute $\rho_{\text{next}}$. We report that, by considering the low correlation threshold $\rho_{\text{low}}$, 61 relationship are selected, while, with the high correlation threshold $\rho_{\text{high}}$, 42 relationships are chosen. Differently, the Granger-based statistical framework selects 19 relationships. Interestingly, 12 over 19 of those relationships are in common with the high correlation selection and 15 are in common with the low correlation one. This suggests that, through the correlation analysis method, we are considering an approximate superset of those relationships chosen with the proposed framework.

Results for the detection phase are provided in Table II. The HMM-CDT parameter $C$ is here set to $C_{\text{gran}} = 1.271$ and $C_{\text{high}} = C_{\text{low}} = 3.141$, which are the smallest values s.t. $FP = 0$ by considering the Granger and correlation analysis approaches, respectively. The Granger-based method is able to maintain a $DD$ generally better than the one provided by the crosscorrelation analysis. Is it also possible to see that in the case of fault located in sensor $s_5$, the Granger-based framework does not provide a detection ($FN$), supporting the idea that one should select a high number of relationships to maintain a low $FN$, at the expense of an increase in the $FP$.

### VI. Conclusion

In this paper, we present a statistical framework for automatically learning the dependency graph of a sensor network for fault detection/diagnosis purposes. The framework, which is based on the concept of Granger causality, tests each couple of datastreams for multivariate conditioned Granger causality, to understand the influence of a specific stream of data to the reconstruction of another datastream (conditioned to all the information provided by the other available datastreams). Experimental results obtained by applying the proposed framework to the HMM-CDT Cognitive FDDS provide evidence for the better characterization of the truly existing relationships and consequent better fault detection abilities of the proposed framework over the classical crosscorrelation analysis for dependency graph learning.

### REFERENCES


