# Multivariate time series unsupervised representation on a causality graph for the health monitoring of mechanical systems

Edouard Pineau[†‡⋆], Sébastien Razakarivony[†], Thomas Bonald[‡]

Safran Tech, Signal and Information Technologies[†]
Telecom Paris, Institut Polytechnique de Paris[‡]

**Abstract.** Multivariate time series (MTS) have specific features that complicate their analysis: interactions in space and time between the MTS components, variable length, absence of trivial alignment between samples, and high dimensionality. Hence, finding a representation of MTS from which we can extract meaningful information is a challenging task. In general, specific assumptions are needed to obtain a valuable representation. In our paper, we assume that a dataset of MTS samples has an underlying causal structure that we can exploit to represent samples. Our contribution is a new representation framework that consists of first finding the overall causality graph in a studied dataset and then representing the samples on this graph with a relational neural network. We name this method Sequence-to-Graph (Seq2Graph). We apply Seq2Graph on a health monitoring task, on two MTS datasets coming from mechanical systems, to show the interest of the causality-based representations.

## 1 Introduction

Nowadays, more and more data is packaged as multivariate time series (MTS), for example industrial records, physiological data, and vehicles sensors to name a few. An important preliminary step for data information mining or other downstream machine learning tasks involves finding consistent and meaningful representation of samples. In this paper, we tackle an unsupervised MTS representation problem using a particular feature of MTS: the *causality* between variables.

The standard definition of causality used for time series data is the *Granger causality* [11]. It consists in evaluating, for each couple of variables $(X^{(i)}, X^{(j)})$ of a MTS sample $X$, if the variable $X^{(i)}$ (the cause) is useful to forecast the variable $X^{(j)}$ (the consequence). Hence, Granger causality is a feature that describes the causal dependencies underlying data. Causality can be represented as a graph $\mathcal{G}$, as in Figure 1.

[⋆] Corresponding author: pineau.edouard@gmail.com

When observed samples are generated from a unique mechanical system $\mathcal{S}$, it is common to assume a *unique* causality structure $\mathcal{G}$ for the whole dataset $\mathcal{X}$. For example, $\mathcal{G}$ can be the skeleton of an articulated body or represent the causal relationships (statistical or physical) between sensors arranged in an engine, from which we observe samples. An illustration is given in Figure 2. The causal structure $\mathcal{G}$ is then shared between all observed samples. When the assumption of a unique graphical structure for all samples is valid, like in the aforementioned examples, a causality graph $\mathcal{G}$ can be extracted directly from the studied dataset $\mathcal{X}$, using an appropriate statistical method. $\mathcal{G}$ is then an abstract representation of the system $\mathcal{S}$ (e.g. a mechanical system) that generated all the observed samples.

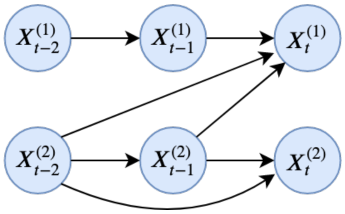

Fig. 1: Graphical representation of the causality underlying a bivariate time series $X$. The edges are weighted with a set of edge's weights $W^X$ that depends on $X$. In this paper, we claim that comparing MTS samples by comparing their causality weights is relevant.

Yet, Granger causality inference only aims at discovering and interpreting causalities between observed variables at the scale of the dataset (i.e. one graph for all samples). In this paper, we propose to relegate the inference of the causal graph $\mathcal{G}$ as a preliminary task preceding the representation inference of individual MTS samples. Our assumption is the following: since $\mathcal{G}$ is the graphical model of the system $\mathcal{S}$ that generated all the observed samples (e.g. a mechanical system), $\mathcal{G}$ is also a natural meaningful latent structure on which each data sample can be represented. In particular, each sample $X$ can have its own set of edge's weights $W^X$ on $\mathcal{G}$, hence its own causality-based representation. We can then compare samples by comparing their causality graph edge's weights. We claim and show in this paper that such a representation is meaningful for mechanical system's state monitoring.

We divide the problem into two subproblems. First, the extraction of the causal structure $\mathcal{G}$ using sparse MTS modeling approach [29]. Second, the choice and training of a relational neural network $F_\phi$ parametrized by $\phi$ that takes MTS samples as input and maps them onto $\mathcal{G}$ [23] to obtain a causality-based representation. Our main contribution is an unsupervised multivariate time series representation framework based on Granger causality, its implementation with neural networks and its application to health monitoring. We name it Sequence-to-Graph (Seq2Graph). Section 2 introduces the problem. Section 3 details the learning procedure. Section 4 gives related work. Section 5 illustrates the interest of Seq2Graph framework with two experiments of unsupervised mechanical systems health monitoring.

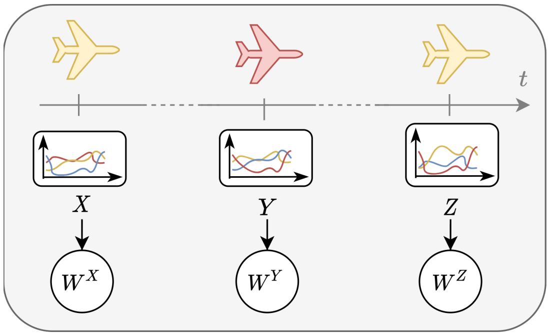

Fig. 2: Under the assumption that a unique causal graph $\mathcal{G}$ represents the causal structure underlying a mechanical system (here a plane), we claim that, if the causality assumption is relevant, the respective set of edge's weights $W^X$, $W^Y$ and $W^Z$ associated to the MTS samples $X$, $Y$ and $Z$ can characterize the *state* of the system. Here, following the color code, $W^Z$ should be closer from $W^X$ than from $W^Y$. We note that a maintenance was effectuated between $Y$ and $Z$ to restore the system's state.

## 2 Seq2Graph

Let $\mathcal{X} \subseteq \mathbb{R}^{d \times T}$ be a finite set of $d$-dimensional multivariate time series (MTS), each indexed over the discrete time range $t = 1, \ldots, T$. We want to represent each sample $X \in \mathcal{X}$ on a causality graph $\mathcal{G}$.

### 2.1 Problem formulation in linear setting

We first assume that for each time series $X$, there exists a tensor $W^X \in \mathbb{R}^{K \times d \times d}$ such that

$$X_t = \sum_{k=1}^{K} W_k^X X_{t-k} \tag{1}$$

Model (1) is a $K$ order linear vector autoregressive model (L-VAR), with sample-wise parameter $W^X$. We consider that $W^X$ represents time series $X$ under linear assumption (1). We therefore can compare different samples $X$ by comparing their parameter $W^X$.

Each parameter $W^X$ can be estimated by maximum likelihood (MLE), in closed form or with a generic likelihood term (for example using a Kalman filter [3]). Yet, when the dataset $\mathcal{X}$ is large, when the samples $X$ are high-dimensional or when $W^X$ has consistency or sparsity constraints, estimating every individual $W^X$ by MLE can be expensive or untractable. As a tractable alternative, we use [23]: a relational neural network (RelNN) [15] $F_\phi$ with parameters $\phi$ is trained to transform time series samples $X$ directly into parameter $W^X$, i.e. $F_\phi(X) = W^X$. Hence, the representation problem becomes an *encoder-decoder* representation learning scheme, with $F_\phi$ the encoder and the VAR model (1) the decoder.

### 2.2 Representation on a common causality graph $\mathcal{G}$

In VAR model, $W^X_{.,i,j} = 0$ means the absence of causality from variable $X^{(i)}$ to variable $X^{(j)}$, for a given sample $X$, [6]. In order to have all samples $X$ represented on the same causality graph $\mathcal{G}$, all $W^X$ should share the same zeros.

In consequence, we assume that each tensor $W^X$ has three underlying components: a dataset-level component $\bar{W} \in \mathbb{R}^{K \times d \times d}$, a sample-level component $P^X$ and dataset-level binary adjacency $A \in \{0,1\}^{d \times d}$, such that:

$$W^X = \bar{A} \odot \left(\bar{W} + P^X\right) \tag{2}$$

.

where $\odot$ is the Hadamard product and $\bar{A}$ the adjacency $A$ extended to match the dimensionality of $\bar{W}$. $\bar{W}$ and $A$ are shared between all samples $X$. The entries of the sparse tensor $A \odot \bar{W}$ are the edge features of the causal graph $\mathcal{G}$. The entries of the tensor $A \odot W^X$ are the adjustment of the graph to match the properties of sample $X$.

*Remark 1.* The decomposition of a regression problem into sample-level $W^X$ and dataset-level $\bar{W}$ is known as *random coefficient regression* (RCR) [19]. The particularity of our approach is the shared sparsity given by $A$.

In practice, we first infer a sparse tensor $\bar{W}$. Then we define a graph adjacency $A$ from $\bar{W}$: $A_{i,j} = \mathbb{1}\{\sum_{k=1}^{K} |\bar{W}_{k,i,j}| > 0\}$. We then build and train a neural network $P_\phi$ with parameters $\phi$ that directly and efficiently outputs the adjustment $P^X$ in the RCR. Hence, the previously introduced neural network $F_\phi$ is defined as $F_\phi(X) := \bar{W} + P^{\mathcal{G}}_\phi(X)$, where $P^{\mathcal{G}}_\phi(X) = \bar{A} \odot P_\phi(X)$. The details of the learning procedure and the implementation are given in Section 3.

The Seq2Graph framework is illustrated in Figure 3.

### 2.3 Generalized L-VAR

To obtain a richer representation learning framework, we extend linear approach to a generalized linear [20] vector autoregressive model (GL-VAR) [29]. We remind that generalized linear models extend linear regression by relating the linear transform and the response variable with a nonlinear link function. Hence, let $g = \{g_{\theta_j}\}_{j=1}^{d}$ be a set of shallow neural networks with parameters $\theta = \{\theta_j\}_{j=1}^{d}$, $g_{\theta_j} : \mathbb{R}^l \to \mathbb{R}$, such that for each sample $X \in \mathcal{X}$ we can find a tensor $W^X \in \mathbb{R}^{K \times d \times d \times l}$ such that $\forall j \in [\![1, d]\!]$:

$$X^{(j)}_t = g_{\theta_j}\left(\sum_{k=1}^{K} W^X_{k,.,.,j} X_{t-k}\right) \tag{3}$$

with $l \in \mathbb{N}^*$. (3) is a neural generalized linear version of a *vector autoregressive* (VAR) model, called *generalized-linear VAR* (GL-VAR) [29].

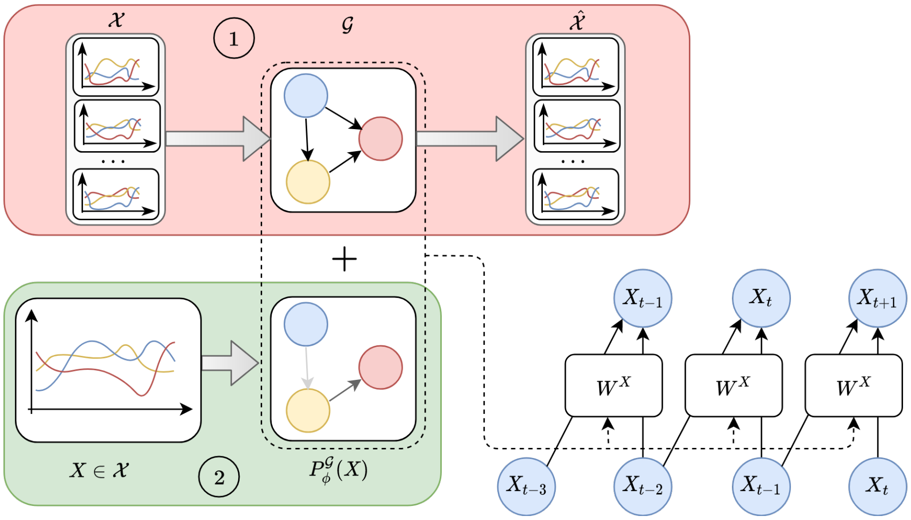

Fig. 3: Seq2Graph. ① is the dataset-level causal graph $\mathcal{G}$ inference. $\hat{\mathcal{X}}$ is the set of predictions for the whole dataset $\mathcal{X}$. $\mathcal{G}$ is built such that it explains the mean dynamical behavior of the dataset. ② is the representation inference. $\hat{X}$ is the prediction of the sample $X$. The adjustments $P_\phi^{\mathcal{G}}(X)$ are built such that they explain the specific dynamical behavior of sample $X$, along the edge of the mean causal graph. Here, $K = 2$.

## 3 Seq2Graph training

### 3.1 Finding $\mathcal{G}$ with group-lasso training

We search a sparse $\bar{W}$ from RCR model (2) and $\theta = \{\theta_j\}_{j=1}^d$ the set of parameters of the neural link functions (in the nonlinear case (3)). We therefore search a solution to the following mean-squared regression with group-lasso regularization:

$$\min_{\bar{W},\theta} \quad \mathbb{E}_{X\sim\mathcal{X}}\left[\sum_{j=1}^{d}\sum_{t=K+1}^{T}\left\|X_t^{(j)} - g_{\theta_j}\left(\sum_{k=1}^{K}\bar{W}_{k,.,j}X_{t-k}\right)\right\|_2^2\right] + \lambda\sum_{i,j=1}^{d}\left\|\bar{W}_{.,i,j}\right\|_F + \gamma\sum_{j=1}^{d}\|\theta_j\| \tag{4}$$

where $\|.\|_F$ is the Frobenius norm. The group-lasso penalty for $W$ associated with coefficient $\lambda$ encourages each $\|\bar{W}_{.,i,j}\|_F$ to be null, meaning that all causal links from $X^{(i)}$ to $X^{(j)}$ would be cut. $\lambda \in \mathbb{R}^+$ controls the speed and intensity of the pruning. Regularization $\|\theta_j\|$ compensates the effect of the group-lasso to avoid the (theoretical) situation where $\bar{W}$ goes to zero while parameters $\theta$ tend to infinite sensitivity.

Problem (4) is first optimized with stochastic gradient descent. Then, we chose proximal gradient descent (PGD) [21] as fine-tuning optimization procedure in order to achieve true zeros in $\bar{W}$ as in [29]. If we had a target sparsity, we could stop the PGD when the level is achieved. In our experiment, we do not have (or assume to not have) the true sparsity level. We propose to monitor the impact of the sparsity on prediction and chose the maximal sparsity that does not degrade the prediction capacity of the model. See experiments for illustration.

We estimate causality adjacency $A$ with $A_{i,j} = \mathbb{1}\{\sum_{k=1}^{K} \|\bar{W}_{k,i,j}\| > 0\}$.

### 3.2 Causality adjustment learning

In this section, we train a neural network that directly infers the causality adjustment of RCR model for a given sample.

We assume that the parameters $\{g, \bar{W}\}$ has been learned on the whole dataset (see Section 3.1). We now build and train a relational neural network $P_\phi$ to infer each sample-wise adjustment. We note $P_\phi^{\mathcal{G}} = \bar{A} \odot P_\phi$ the adjustment constrained to live only on the edges of the inferred causality graph $\mathcal{G}$. $\bar{A} \in \mathbb{R}^{K \times d \times d \times l}$ is the adjacency matrix $A$ of $\mathcal{G}$ expanded to match the dimensions of tensor $P_\phi^{\mathcal{G}}$. Then the problem to solve is:

$$\min_\phi \mathbb{E}_{X \sim \mathcal{X}} \left[ \sum_{j=1}^{d} \sum_{t=K+1}^{T} \left\| X_t^{(j)} - g_{\theta_j} \left( \sum_{k=1}^{K} \left(\bar{W} + P_\phi^{\mathcal{G}}(X)\right)_{k,:,j} X_{t-k} \right) \right\|_2^2 + \eta \Omega \left( P_\phi^{\mathcal{G}}(X) \right) \right] \quad (5)$$

Such learned inference network is therefore both meaningful (since built to represent the dynamics underlying the data) and sparse (since constrained on $\mathcal{G}$).

$\eta$ is a parameter controlling the intensity of the penalty function $\Omega$. A standard penalization $\Omega$ would be the $l^2$ norm on the parameters $\phi$, under the assumption that the parameters of $P^{\mathcal{G}}(X)$ are normally distributed around 0. A more general penalty (yet equivalent when minimized), that we use for Seq2Graph training is $\Omega(P^{\mathcal{G}}(X), P^{reg}) = \|P^{\mathcal{G}}(X) - P^{reg}\|_2 + \|P^{reg}\|_2$. $P^{reg}$ is a parameter trained during the optimization (5). We have found that such penalty helps to obtain more consistent representation. In both cases, the regularization term $\Omega\left(P_\phi^{\mathcal{G}}(X)\right)$ tends to bring the individual representations closer. The idea is to avoid overfitting while improving the consistency between the representations (closer representation for close MTS underlying state). Another interesting regularization would have been to take into account explicitly the temporal relations between samples by forcing the consecutive samples to be closer. Yet, we wanted to show that the causality inductive bias can be sufficient to find temporal consistency when it is relevant, like in certain health monitoring problems.

### 3.3 Implementation details

*Neural network representation inference* For the representation inference function $P_\phi^\mathcal{G}$, we use a relational neural network (RelNN) [24]. The RelNN embeds pair of variables. Its adaptation for time series data is taken from [15], where the RelNN takes a MTS as input and embeds all pairs of variables into a binary space. In [23], an equivalent RelNN embeds pairs of variables as vectors specialized for linear causality, where it proved to be expressive, noise resistant and able to generalize over the notion of linear causality.

*Multi-multivariate time series* There are cases where the $d$ components of a MTS are multidimensional. For example, if the MTS has $d$ variables situated in a 2D space (see Experiment 5.1), hence each variable is represented by a 2D time series. More formally, the problem extends from $\mathcal{X} \subseteq \mathbb{R}^{d \times T}$ to $\mathcal{X} \subseteq \mathbb{R}^{d \times m \times T}$, i.e. $\forall X \in \mathcal{X}\ X_t \in \mathbb{R}^{d \times m}$, with $m$ the dimension of individual time series variables. The approach presented in our paper adapts to this general case by replacing $W^X \in \mathbb{R}^{K \times l \times d}$ by $W^X \in \mathbb{R}^{K \times l \times d \times 1}$. The additional dimension in $W^X$ enables to consider the $m$ time series of each variable as a whole.

## 4 Related work

This section outlines main related work for the two key concepts of our paper: MTS representation and causality inference.

*Time series representation.* Finding interesting and relevant features from time series data has a long-range history. A particularity of time series is that they have no explicit and general features [32]. Hence, MTS representation in an unsupervised manner requires strong assumptions to be relevant. The most common assumption is that similar samples have similar shapes and patterns up to a warping alignment [7] and can be represented closely using bag-of-patterns [26]. This simple assumption gives good results for many time series tasks as long as it is valid, and can be efficiently used [31]. Yet, for multivariate time series, it is limited since it generally does not take into account the interactions between variables.

More recently, the usage representation learning methods based on neural networks have emerged, giving new representation learning models for MTS data. In [17], they propose a method such that the distance between the learned representations is the dynamic time warping (DTW) distance. A powerful family of neural representation learning methods is the autoencoders (AE) and related methods [2]. The principle is to train an encoder and a decoder simultaneously, such that the encoded signal of each sample can be correctly decoded. If so, it means that the encoded signal contains the essential information. The Sequential Autoencoders (SAE) [4,18,34] is the most popular adaptation of AE for time series data. A richer SAE proposed in [8] is based on the joint learning of a discrete variational autoencoder (VAE) [30], a self-organizing map (SOM) latent space [16] and a Markov transition model [10]. Yet, the latter model only treats discrete representation of time series. In [9], an unsupervised scalable time series representation (USTR) is proposed using the notion of triplet loss. The assumption is that a MTS sample is closer to one of its subsamples (*positive* sampling) than to a randomly chosen sample of the dataset (*negative* sampling). This model achieves very good results in classification downstream task. Moreover, it can handle many types of data since the assumptions underlying the model are weak.

Closer to our work, [15] proposes the neural relational inference (NRI), a VAE that transforms time series into a binary relational graph, trained as a variable selection method for neural time series model. Despite high interest and impressive results, NRI is limited to binary representations. [23] extends the binary relationships of NRI to a linear Granger causality latent space, with a model called Seq2VAR. Their model is applied to causality detection in contexts where NRI fails (noisy environment, floating intensity of the causalities). Our approach is built upon Seq2VAR, by adding non-linear causality and dataset-level regularization.

In the current paper, we use SAE, USTR and Seq2VAR as comparative models.

*Causality in time series.* It is common to represent data with graphs [27]. In our paper, we use a particular graph specific to MTS: the Granger causality (GC) graph [6].

Initially, GC is defined with the non zero entries of the parameter of a sparse linear vector autoregressive (L-VAR) model [5,12], obtained with a penalization on L-VAR parameters during training.

For the nonlinear case, finding a useful latent representation is far from obvious. Among standard approaches for nonlinear GC detection in time series, we find kernel methods [1] or information-theory [28,13]. These methods find the existence of causality between variables, i.e., the causality graph. Yet, these methods do not give information on the importance of the causality where it exists. In Seq2Graph, as shown in the following, we require continuous (real) features on the edges of the causality graph. In [33], the authors extend VAR to nonlinear causality with the linear combination of B-spline basis functions. The linear operator, once fitted with appropriate penalization, contains continuous causal information. Another nonlinear extension of L-VAR causality is presented in [29]. The authors build the nonlinearity with neural networks. They find causality by applying a penalization on the input convolutional kernel. We use this last work for our causality-based dataset-level regularization.

## 5 Experiments

We propose two experiments to illustrate the interest of Seq2Graph. Both are based on multivariate time series data generated from mechanical systems. For the first, we use synthetic data with controlled causal structure to show the interest of the regularization with global causality graph $\mathcal{G}$ with a sparse random coefficient model (2). For the second, we use a NASA dataset to assess Seq2Graph on real representation task. In all experiments, we compare Seq2Graph to three time series representation methods: a sequential autoencoder (SAE) [18], to the unsupervised scalable time series representation (USTR) [9] and to the sequence-to-VAR (Seq2VAR) [23].

For all experiments, we use PyTorch [22]. All models are trained with Adam optimizer [14] with learning rate $5.10^{-4}$. For the sample-level representation training, we added a learning rate scheduler with step size 100 and a decay factor = 0.9. All experiments are done on a GPU Nvidia Quadro K4000. The hyperparameters are given in Table 1.

|  | $\lambda$ | $\gamma$ | $\eta$ |
|---|---|---|---|
| Experiment 5.1 | $10^{-3}$ | - | $10^{-5}$ |
| Experiment 5.2 | $5 \times 10^{-3}$ | $5 \times 10^{-3}$ | $5 \times 10^{-4}$ |

Table 1: Hyperparameters for our experiments.

Code to reproduce the experiments can be found at https://github.com/anonym-conf-submission/Seq2Graph.

### 5.1 Interacting Newtonian system

*Dataset* We simulate samples from a 10-ball-springs system, consisting of the simultaneous trajectories of 10 identical balls of unit mass in a 2-dimensional space, each ball being connected to some others by springs (the rate of connection is 56%). This system has a natural bidirectional causal structure: each ball's trajectory acts as a cause for changes in the trajectory of the neighbor balls, and conversely. Using the previously introduced notations, we have $d = 10$ (10 balls) and $m = 2$ (in a 2-dimensional space, see implementation details). System dynamics follows Newton's law of motion. We assume that the system is ageing and is regularly restored. All samples share a common graph graphical structure $\mathcal{G}$ whose adjacency is the interaction matrix formed by the springs.

We simulate a synthetic dataset of 15000 samples (trajectories), 5000 for train, 5000 for validation and 5000 for test. Each trajectory is 49 time-steps-long ($T = 49$). For each batch $b$ of 50 samples, a constant ageing factor $\alpha_b \sim \mathcal{U}([0.9, 1])$ is applied to the system: at each sample $X$ whose index is $s \in [\![0, 50]\!]$ (within the batch $b$ of 50 samples), we randomly choose a spring $(i, j)$ and multiply its rigidity by $\alpha_b^s$, i.e. an exponential ageing coefficient with respect to sample index. Every 50 samples, we *restore* the state of the system and another ageing factor is sampled and applied to the next batch of 50 samples. For some trajectories, $\alpha_b = 1$, i.e. there is no ageing: the initial hidden causality graph has binary adjacency and remains the same along the life of the system. When $\alpha_b < 1$, the initial graph is deteriorating during along the life (observed through 50 samples) of the system, until restoration.

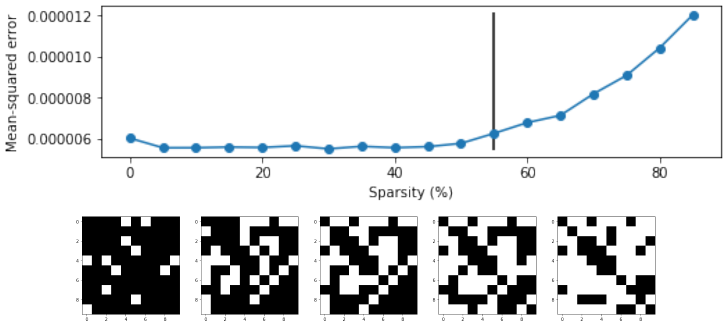

Fig. 4: **Top**: Prediction MSE. Black line is the true sparsity. **Bottom**: Causal graphs for different sparsity levels. The third figure is the inferred causal graph, which matches the ground truth.

*Model* For this first experiment, we assume that the model is linear, i.e. that functions $g_{\theta_j}$ are identities. The objective is to illustrate the impact of representing data on the same causality graph $\mathcal{G}$. We determine by cross-validation that $K = 2$. The level of sparsity is determined by the quality of the prediction for different levels of sparsity of the causality graph, as shown in Figure 4. The prediction is almost invariant until a sparsity of about 56%. We note that we find back the true adjacency with (4).

*Metrics and results* We assess the quality of the representation inference function $P_\phi^\mathcal{G}$. We test if we can represent the ageing of the system with respect to a reference *healthy* sample $X^{ref}$ (first sample of a batch) picked in the validation set. We then build the test ageing curve $\| \sum_{k=1}^{K} (P_\phi^\mathcal{G}(X^{ref}) - P_\phi^\mathcal{G}(X))_k \|_2^2$ for all samples $X \in \mathcal{X}^{test}$. Results are presented in Figure 5 and Table 2.

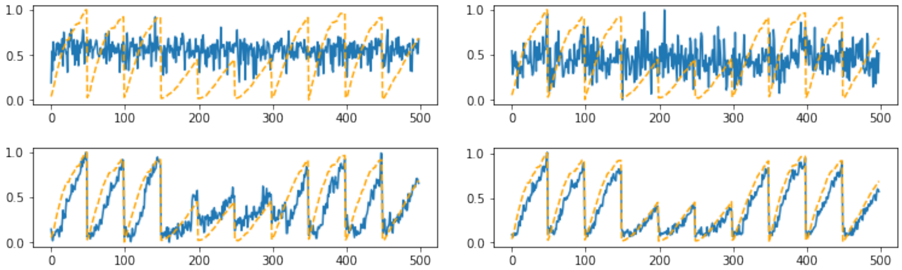

Fig. 5: Unsupervised estimation of ageing curves for the 10 first batches of the test set. **Top-left**: USTR [9], **top-right**: SAE [18], **bottom-left**: Seq2VAR [23], **bottom-right**: Seq2Graph. Orange curve is the ground truth.

| Models | MSE | Ageing score* |
|---|---|---|
| SAE | $2.2 \times 10^{-5}$ | 0.09 |
| USTR | - | 0.05 |
| Seq2VAR | $2.3 \times 10^{-7}$ | 0.62 |
| Seq2Graph | $4.4 \times 10^{-7}$ | **0.97** |

Table 2: Performance of several models plus ours on ageing ball-springs problem. The *Ageing score* is the correlation between estimated and real ageing curve. *Higher is better. MSE stands for mean squared error and serves only as a sanity check (for MSE-based methods).

We see that Seq2Graph outperforms both SAE, USTR and Seq2VAR for unsupervised representation learning, when meaningful information is fully contained in the causality. In particular, SAE and USTR completely miss the consistent ageing information, as expected from pattern-based approaches. Although consistent, Seq2VAR seems to suffer when the causalities become lower. In fact, lowering the causality improves the difficulty to capture them, hence prevents to find the trend hidden in causality. Adding a common causal structure $\bar{W}$ in Seq2Graph naturally helps the identification and the consistency of the sample representations.

### 5.2 NASA turbofan degradation simulation dataset

*Dataset* NASA public Commercial Modular Aero-Propulsion System Simulation dataset (C-MAPSS) is a tool for simulation of realistic large commercial turbofan engine data [25]. An engine degradation simulation was carried out using C-MAPSS, under different conditions and different faults. We use the *FD001* dataset which contains 100 time series recorded at sea level with one fault mode for each (degradation of the high-pressure compressor, a fundamental turbofan piece). The time series are the output of the turbine-engine system that takes a fuel flow as input and outputs 21 variables, whose 13 are not constant (we only keep these 13 variables). Time series are 206 time-steps long on average. Each time series is the recording of a turbine engine going to failure. The engine is operating normally at the start of each time series and develops a fault of unknown initial magnitude in its first moments. We only know that the impact of this fault on the system grows in magnitude until system failure.

For the results of the paper, we split the dataset: the first 60 time series are train set, the 15 next are validation set and the last 25 are test set. We extract from these time series sub-trajectories of length 25, with a rolling window with stride 5 to make our dataset. Hence, as for previous experiment, we have several batches of samples. A batch corresponds to the life of the system from start to failure. At the end of each batch, the engine is *restored* and another batch of samples is recorded.

*Model* Using the previously introduced notations, $d = 13$. All samples share a common (unknown) structure which is the turbine engine mechanics. We assume that this structure can be represented by a sparse causality matrix. For this experiment each link function $g_{\theta_j}$ is a MLP with 2 hidden layers of 4 channels. We determine by cross-validation that $K = 2$. The maximal sparsity is determined by the quality of the prediction for different levels of sparsity, as shown in Figure 6. The prediction is almost invariant until a sparsity of about 75%.

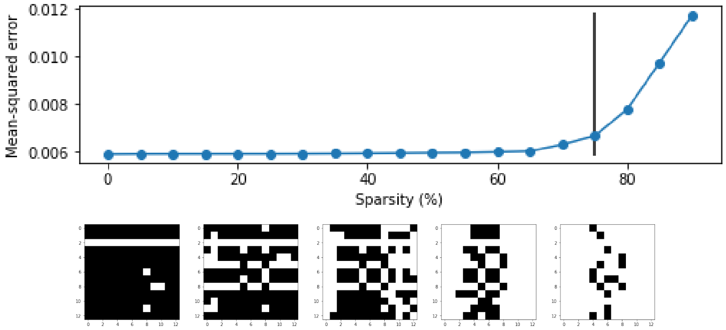

Fig. 6: **Top**: MSE of the prediction. Black line is the sparsity of the inferred causal graph. **Bottom**: Causal graphs for different sparsity levels. Fourth image is the inferred causal graph.

We solve (5) using the previously found causal graph. Contrary to the previous experiment, the system is not *isolated* since the observed variables are the response to an unobserved command (the fuel flow). Yet, the variables can effectively interact with each other and statistical causality still makes sense as a representation assumption.

*Metrics* We assess the quality of the learned $P_\phi^{\mathcal{G}}$ by testing if we can extract *ageing information* from the representations, as for the previous experiment. Yet, we remind that we do not know the importance of the initial fault. What we know is that the 100 engines go to failure and the degradation of the state is monotonic until restoration. We propose a two-step process to predict the imminence of a failure.

First, we build an ageing indicator assuming that is a relative position compare to a healthy sample. We pick a healthy sample $X^{ref}$ (first sample of a batch) in the validation set and build the ageing curve $\| \sum_{k=1}^{K} (P_\phi^{\mathcal{G}}(X^{ref}) - P_\phi^{\mathcal{G}}(X))_k \|_2^2$ for all $X \in \mathcal{X}^{valid}$. We compute a *failure threshold* $\tau^{valid}$ that must indicate when an engine goes to failure. We set $\tau^{valid}$ to the maximal threshold that ensures turbine engine failure detection built from ageing curves for all validation batches, that is formally defined as:

$$\tau^{valid} = \min_{X \in \mathcal{X}^{val,fail}} \left\| \sum_{k=1}^{K} (P_\phi^{\mathcal{G}}(X^{ref}) - P_\phi^{\mathcal{G}}(X))_k \right\|_2^2 \qquad (6)$$

where $\mathcal{X}^{val,fail}$ is the set of validation samples preceding the engine failure. We note that $\tau^{valid}$ has no safety margin, i.e. any threshold above $\tau^{valid}$ misses at least one engine failure in the validation set (under monotony assumption underlying the ageing of a mechanical system). It is possible to add a margin by lowering $\tau^{valid}$.

Second, we build the test ageing curve $\| \sum_{k=1}^{K} (P_\phi^{\mathcal{G}}(X^{ref}) - P_\phi^{\mathcal{G}}(X))_k \|_2^2$ for all $X \in \mathcal{X}^{test}$. We apply the detection test using $\tau^{valid}$ (represented by the horizontal dotted line in Figure 7.

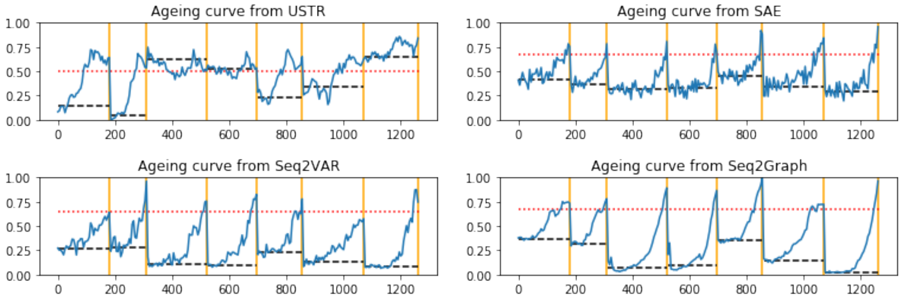

Fig. 7: Unsupervised estimation of C-MAPSS ageing curve with different models on the 7 first test batches. **Top-left**: USTR [9], **top-right**: SAE [18], **bottom-left**: Seq2VAR [23], **bottom-right**: Seq2Graph. Orange picks are engine failures and repair. Long red dotted horizontal line is the threshold $\tau^{valid}$. Black dashed horizontal lines are the estimated initial states of each engine, computed as the mean value of the curve on the 10 first samples of each batch.

*Results* As a first assessment, we see in Figure 7 that the estimated ageing curves built from SAE, Seq2VAR and Seq2Graph are almost monotonic inside each batch (between two vertical orange lines). We recall that monotony is the only ground truth information we have on the ageing of the system. The fact that SAE, Seq2VAR and Seq2Graph unveils monotonic signal means the ageing information is present both in patterns and values (SAE) and in causality (Seq2VAR and Seq2Graph). We do not find consistent representations with USTR. We also observe that the batch's ageing curves do not begin at the same value (dashed horizontal lines in Figure 7), whatever the method. It is partly imputed to the fact that the mechanical faults are located at the beginning of each batch and that they vary in intensity. Hence, the inferred first samples of each batch do not have to be equal.

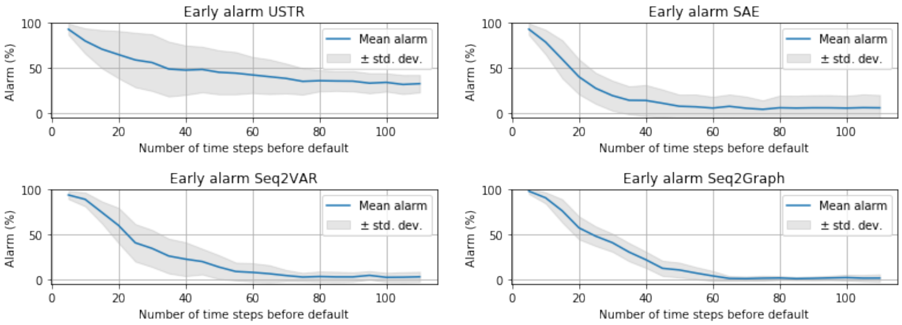

Fig. 8: Early alarm on CMAPSS data using MTS representation models USTR, SAE, Seq2VAR (see related work for details) and Seq2Graph. Means and standard deviations are built using all batch's first samples as $X^{ref}$ and several encoders trained with different seeds.

We now compare the ability of the different MTS representations to relevantly detect failures. In Figure 8, we see the proportion of alarm at different time steps before actual failure happens, built from the estimated ageing curves illustrated in Figure 7. First, we note that all models detect almost 100% of failures before it happens. Second, we want detection of the coming failures to be reasonably early to avoid false alarms. If curves cross threshold too early, the MTS representation is useless. Figure 8 shows that Seq2Graph is the most consistent in early detection with no alarms far from failure, due to the consistency of the extracted monotonic signal. On the contrary, SAE always finds early failures. We note that Seq2Graph also has lower standard deviation, illustrating the interest of the regularizing effect over Seq2VAR.

We have built a representation of the samples that both describes the system dynamics and is consistent with the unknown ageing process since the distance from reference is almost everywhere monotonic before failure, without supervision. We showed an illustration of how to apply our causality-based representation learnig framework.

## 6 Conclusion and future work

In this paper, we have presented a multivariate time series (MTS) representation framework under the assumption that Granger causality contains relevant information about data. We have proposed a two-step approach, based on neural networks. First, the global causal graph is found with a group Lasso penalized neural autoregressive model. Second, a relational neural network is trained to infer the representation of each sample, constrained by the causal structure.

In future work, we intend to include temporal knowledge in Seq2Graph, using the temporal proximity between samples. We will also leverage the interpretability of causality to detect the origin of a degradation in a mechanical system.

## References

1. Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear granger causality of time series. *Physical Review E*, 70(5):056221, 2004.
2. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
3. José Casals, Alfredo Garcia-Hiernaux, Miguel Jerez, Sonia Sotoca, and A. Trindade. *The likelihood of models with varying parameters*, pages 65–82. 09 2018.
4. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
5. Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.
6. Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. *arXiv preprint arXiv:1206.5246*, 2012.
7. Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.
8. Vincent Fortuin, Matthias Hüser, Francesco Locatello, Heiko Strathmann, and Gunnar Rätsch. Som-vae: Interpretable discrete representation learning on time series. *arXiv preprint arXiv:1806.02199*, 2018.
9. Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *LLD workshop, ICLR 2019*, 2019.
10. Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.
11. Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.
12. Fang Han and Han Liu. Transition matrix estimation in high dimensional time series. In *International Conference on Machine Learning*, pages 172–180, 2013.
13. Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.
14. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
15. Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2693–2702, 2018.
16. Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.
17. Qi Lei, Jinfeng Yi, Roman Vaculin, Lingfei Wu, and Inderjit S Dhillon. Similarity preserving representation learning for time series analysis. *arXiv preprint arXiv:1702.03584*, 2017.
18. Pankaj Malhotra, Vishnu TV, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Timenet: Pre-trained deep recurrent neural network for time series classification.

*Proceedings of 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2017.

19. BO Muthén, LK Muthén, and Tihomir Asparouhov. Random coefficient regression, 2015.
20. John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.
21. Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.
22. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
23. Edouard Pineau, Sebastien Razakarivony, and Thomas Bonald. Seq2var: multivariate time series representation with relational neural networks and linear autoregressive model. *Advanced Analytics and Learning on Temporal Data*, 11986, 2019.
24. Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.
25. A Saxena and K Goebel. Turbofan engine degradation simulation data set. *NASA Ames Prognostics Data Repository*, 2008.
26. Pavel Senin and Sergey Malinchik. Sax-vsm: Interpretable time series classification using sax and vector space model. In *2013 IEEE 13th international conference on data mining*, pages 1175–1180. IEEE, 2013.
27. David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.
28. Victor Solo. On causality and mutual information. In *2008 47th IEEE Conference on Decision and Control*, pages 4939–4944. IEEE, 2008.
29. Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality for nonlinear time series. *arXiv preprint arXiv:1802.05842*, 2018.
30. Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.
31. Lingfei Wu, Ian En-Hsu Yen, Jinfeng Yi, Fangli Xu, Qi Lei, and Michael Witbrock. Random warping series: A random features method for time-series embedding. In *International Conference on Artificial Intelligence and Statistics*, pages 793–802, 2018.
32. Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, 12(1):40–48, 2010.
33. Yingxiang Yang, Adams Wei Yu, Zhaoran Wang, and Tuo Zhao. Detecting nonlinear causality in multivariate time series with sparse additive models. *arXiv preprint arXiv:1803.03919*, 2018.
34. Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 27(3):631–644, 2019.