Appendix D

PAPER D

Title: Online Joint Nonlinear Topology Identification and Missing Data Imputation over Dynamic Graphs

Authors: **R. Money**, J. Krishnan, B. Beferull-Lozano

Conference: European Signal Processing Conference 2022

Online Joint Nonlinear Topology Identification and Missing Data Imputation over Dynamic Graphs

R. Money, J. Krishnan, B. Beferull-Lozano

Abstract: Extracting causal graph structures from multivariate time series, termed topology identification, is a fundamental problem in network science with several important applications. Topology identification is a challenging problem in real-world sensor networks, especially when the available time series are partially observed due to faulty communication links or sensor failures. The problem becomes even more challenging when the sensor dependencies are nonlinear and nonstationary. This paper proposes a kernel-based online framework using random feature approximation to jointly estimate nonlinear causal dependencies and missing data from partial observations of streaming graph-connected time series. Exploiting the fact that real-world networks often exhibit sparse topologies, we propose a group lasso-based optimization framework for topology identification, which is solved online using alternating minimization techniques. The ability of the algorithm is illustrated using several numerical experiments conducted using both synthetic and real data.

D.1 Introduction

Data analytics on complex networked systems such as large-scale sensor networks, social networks, brain networks, etc., have gained much research attention in the last decade. Most such complex networks generate data in the form of multivariate time series, which are often interdependent. These dependencies can be represented in the form of a graph. Representing and processing data on graph structures have become increasingly important due to diverse range of applications, such as data compression, denoising, change point detection, etc. Often, such dependencies are not directly observable and must be inferred. Identification of causal graph structure from multivariate time series is termed *topology identification*, which is a challenging task due to the nonstationary and nonlinear nature of the dependencies.

It is essential to have sufficient and good quality data when solving a topology identification problem; however, data might not be fully observable in many real-world situations. Sensor networks, for instance, transmit data captured by sensors

through communication channels to an end-user for processing. These networks are susceptible to data loss due to sensor failures or communication impairments, making it challenging to identify the topology. A practically significant algorithm for topology identification must be *(i)* capable of working online to handle nonstationary dependencies, *(ii)* capable of recognizing nonlinear dependencies, and *(iii)* capable of dealing with noisy and incomplete observations.

Online linear topology identification is fairly well studied in the literature [1, 48]. In [48], an optimization problem is formulated by taking into account the sparse nature of real-world dependencies and solving the problem using composite objective mirror descent (COMID), and in [1], a time-varying convex optimization framework has been used for topology identification. Recently, several works on nonlinear topology identification have been proposed [9–11, 21, 22, 36], among which [21, 22, 36] propose online solutions for nonlinear topology identification problems, whereas [11] and [9] propose batch solutions using kernel and neural networks, respectively.

While the aforementioned works demonstrate promising results in topology estimation, all assume complete data availability with no sensor failures or communication issues. A joint linear topology identification and missing data imputation using block coordinate descent and Kalman smoothing have been recently proposed in [18]. Similarly, [6] proposes a computationally light approach using inexact proximal gradient descent. However, [18] and [6] assume linear causality, which does not make sense for most real-world time series.

In this paper, we propose an online nonlinear topology identification algorithm accounting for missing data by solving a group lasso-based optimization framework. Considering the well-established underlying theory and the ability to carry out online training, kernels are used to model nonlinearity, which are approximated using random features [16] to control the computational complexity. To the best of our knowledge, this is the first attempt to address jointly *(i)* nonlinearity, *(ii)* nonstationarity, and *(iii)* missing data in topology identification.

D.2 Problem formulation

D.2.1 Nonlinear topology identification

A P-th order nonlinear vector autoregressive (VAR) process with N number of nodes can be expressed as

$$y_n[t] = \sum_{n'=1}^{N} \sum_{p=1}^{P} f_{n,n'}^{(p)}(y_{n'}[t-p]) + u_n[t], \tag{D.1}$$

where $y_n[t]$ is the observation of the n-th time series at time t, $f_{n,n'}^{(p)}(.)$ encodes the causal influence of p-th time-lagged value of n'-th time series on n-th time series, and $u_n[t]$ is the observation noise. The nonlinear VAR model is a suitable model owing to the fact that the causal dependencies in the real world are time-lagged in nature. Moreover, the VAR model implies the famous causality hypothesis proposed by Granger [99], under certain assumptions [63].

D.2.1.1 Kernel representation

We assume that the function in (D.1) belongs to a reproducing kernel Hilbert space (RKHS):

$$\mathcal{H}_{n'}^{(p)} := \left\{ f_{n,n'}^{(p)} \mid f_{n,n'}^{(p)}(y) = \sum_{t=p}^{\infty} \beta_{n,n',t}^{(p)} \, \kappa_{n'}^{(p)}(y, y_{n'}[t-p]) \right\}, \tag{D.2}$$

where $\kappa_{n'}^{(p)}(.,.)$ is a positive definite function that measures the similarity between its arguments, and is termed *kernel*. Every positive definite kernel is associated to a RKHS with inner product $\langle \kappa_{n'}^{(p)}(y,x_1), \kappa_{n'}^{(p)}(y,x_2) \rangle := \sum_{t=0}^{\infty} \kappa_{n'}^{(p)}(y[t],x_1)\kappa_{n'}^{(p)}(y[t],x_2)$ and it satisfies the reproducing property $\langle \kappa_{n'}^{(p)}(y,x_1), \kappa_{n'}^{(p)}(y,x_2) \rangle = \kappa_{n'}^{(p)}(x_1,x_2)$, thus thereby inducing the RKHS norm $\|f_{n,n'}^{(p)}\|_{\mathcal{H}_{n'}^{(p)}}^2 = \sum_{t=0}^{\infty}\sum_{t'=0}^{\infty} \beta_{n,n',t}^{(p)} \, \beta_{n,n',t'}^{(p)} \, \kappa_{n'}^{(p)}(y_n[t], y_n[t'])$. As any function in the RKHS can be expressed as an infinite combinations of kernel evaluations, $f_{n,n'}^{(p)}$ can be expressed as (D.2), with $\beta_{n,n',t}^{(p)}$ being the weight associated with each kernel evaluation. A functional optimization problem can be formulated to obtain the required causal dependency for a given node n:

$$\left\{ \hat{f}_{n,n'}^{(p)} \right\}_{n',p} = \underset{\left\{ f_{n,n'}^{(p)} \in \mathcal{H}_{n'}^{(p)} \right\}}{\arg\min} \frac{1}{2} \sum_{\tau=P}^{T-1} \Bigg[y_n[\tau] - \sum_{n'=1}^{N} \sum_{p=1}^{P} f_{n,n'}^{(p)}(y_{n'}[\tau-p]) \Bigg]^2 + \lambda \sum_{n'=1}^{N} \sum_{p=1}^{P} \Omega\left(||f_{n,n'}^{(p)}||_{\mathcal{H}_{n'}^{(p)}} \right), \tag{D.3}$$

where $\sum_{n'=1}^{N}\sum_{p=1}^{P} \Omega\left(||f_{n,n'}^{(p)}||_{\mathcal{H}_{n'}^{(p)}} \right)$ is the regularizer and λ is the hyperparemter associated with it. If $\Omega(.)$ is nondecreasing, the solution of (D.3) can be expressed with a finite number of kernel evaluations using Representer Theorem [38]:

$$\hat{f}_{n,n'}^{(p)}(y_{n'}[\tau-p]) = \sum_{t=p}^{p+T-1} \beta_{n,n',(t-p)}^{(p)} \kappa_{n'}^{(p)}(y_{n'}[\tau-p], y_{n'}[t-p]) \, . \tag{D.4}$$

Here, the number of kernel evaluations required is equal to the number of data samples. As the number of data samples increases, the number of optimization variables increases, which is commonly known as the *curse of dimensionality* in kernel formulations. We use the random feature (RF) approximation to mitigate this problem.

D.2.1.2 RF approximation

RF approximation addresses the curse of dimensionality by restricting the kernel evaluations to an approximate fixed lower-dimensional Fourier space. Furthermore, linear optimization techniques are easier to use in random Fourier space than in infinite-dimensional RKHS. We use shift-invariant kernels to facilitate RF approximation, i.e., $\kappa_{n'}^{(p)}(y_{n'}[\tau], y_{n'}[t]) = \kappa_{n'}^{(p)}(y_{n'}[\tau] - y_{n'}[t])$. According to Bochner's theorem [30], a shift invariant kernel can be represented using an inverse Fourier transform of a probability distribution:

$$\begin{aligned} \kappa_{n'}^{(p)}(y_{n'}[\tau-p], y_{n'}[t-p]) &= \int \pi_{\kappa_{n'}^{(p)}}(v) \; e^{jv(y_{n'}[\tau-p] - y_{n'}[t-p])} dv \\ &= \mathbb{E}_v[e^{jv(y_{n'}[\tau-p] - y_{n'}[t-p])}], \end{aligned} \tag{D.5}$$

where $\mathbb{E}$ is the expectation operator, $\pi_{\kappa_{n'}^{(p)}}(v)$ is the kernel specific probability density function (pdf) and v is the random variable corresponding to the pdf. With sufficient number of i.i.d. samples $\{v_i\}_{i=1}^{D}$, the expectation in (D.5) can be replaced with sample mean:

$$\hat{\kappa}_{n'}^{(p)}\left(y_{n'}[\tau-p], y_{n'}[t-p]\right)=\frac{1}{D}\sum_{i=1}^{D} e^{jv_i(y_{n'}[\tau-p]-y_{n'}[t-p])}. \tag{D.6}$$

Note that (D.6) is an unbiased estimator of the kernel evaluation with a fixed number D of terms [43]. For a Gaussian kernel with variance σ^2, the inverse Fourier transform can be shown to be also a Gaussian with variance σ^{-2}. Using this information, the real part of (D.6), which is also an unbiased estimator of kernel evaluation, can be expressed as

$$\hat{\kappa}_{n'}^{(p)}\left(y_{n'}[\tau-p], y_{n'}[t-p]\right) = \boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[\tau]^{\top}\boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[t], \tag{D.7}$$

$$\begin{aligned} \textit{where, } \boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[\tau] = \frac{1}{\sqrt{D}}\Big[&\sin\left(v_1 y_{n'}[\tau-p]\right),\ldots,\sin\left(v_D y_{n'}[\tau-p]\right),\\ &\cos\left(v_1 y_{n'}[\tau-p]\right),\ldots,\cos\left(v_D y_{n'}[\tau-p]\right)\Big]^{\top}. \end{aligned} \tag{D.8}$$

A fixed dimensional $(2D)$ approximation of the function $\hat{f}_{n,n'}^{(p)}$ is readily obtained by substituting (D.7) in (D.4):

$$\begin{aligned} \tilde{\hat{f}}_{n,n'}^{(p)}\left(y_{n'}[\tau-p]\right) &= \sum_{t=p}^{p+T-1} \beta_{n,n',(t-p)}^{(p)} \boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[\tau]^{\top}\boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[t] \\ &= \boldsymbol{\alpha}_{n,n'}^{(p)\top}\boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[\tau], \end{aligned} \tag{D.9}$$

where $\boldsymbol{\alpha}_{n,n'}^{(p)} = \sum_{t=p}^{p+T-1}\beta_{n,n',(t-p)}^{(p)}\boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[t]$. The following notations are introduced to simplify the formulations:

$$\boldsymbol{\alpha}_{n,n'}^{(p)} = [\alpha_{n,n',1}^{(p)},\ldots,\alpha_{n,n',2D}^{(p)}]^{\top} \in \mathbb{R}^{2D}, \tag{D.10}$$

$$\boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[\tau] = [z_{\boldsymbol{v},n',1}^{(p)}[\tau],\ldots z_{\boldsymbol{v},n',2D}^{(p)}[\tau]]^{\top} \in \mathbb{R}^{2D}, \tag{D.11}$$

$$z_{\boldsymbol{v},n',k}^{(p)}[\tau] = \begin{cases} \sin(v_k y_{n'}[\tau-p]), & \text{if } k \le D \\ \cos(v_{k-D} y_{n'}[\tau-p]), & \text{otherwise.} \end{cases}$$

The functional optimization (D.3) is reformulated as a parametric optimization problem using (D.9):

$$\left\{\widehat{\boldsymbol{\alpha}}_{n,n'}^{(p)}\right\}_{n',p} = \arg\min_{\left\{\boldsymbol{\alpha}_{n,n'}^{(p)}\right\}} \mathcal{L}^n\left(\boldsymbol{\alpha}_{n,n'}^{(p)}\right) + \lambda\sum_{n'=1}^{N}\sum_{p=1}^{P}\Omega(||\boldsymbol{\alpha}_{n,n'}^{(p)}||_2), \tag{D.12}$$

where

$$\mathcal{L}^n\left(\boldsymbol{\alpha}_{n,n'}^{(p)}\right) := \sum_{\tau=P}^{T-1}\frac{1}{2}\left[y_n[\tau] - \sum_{n'=1}^{N}\sum_{p=1}^{P}\boldsymbol{\alpha}_{n,n'}^{(p)\top}\ \boldsymbol{z}_{\boldsymbol{v},n'}^{(p)}[\tau]\right]^2, \tag{D.13}$$

which can be expanded in terms of RF components as

$$\mathcal{L}^n\left(\alpha^{(p)}_{n,n',d}\right) := \sum_{\tau=P}^{T-1}\frac{1}{2}\Big[y_n[\tau] - \sum_{n'=1}^{N}\sum_{p=1}^{P}\sum_{d=1}^{2D}\alpha^{(p)}_{n,n',d}\; z^{(p)}_{\boldsymbol{v},n',d}[\tau]\Big]^2. \tag{D.14}$$

For convenience, the parameters $\{\alpha^{(p)}_{n,n',d}\}$ and $\{z^{(p)}_{\boldsymbol{v},n',d}[\tau]\}$ are stacked in the lexicographic order of the indices p, n', and d to obtain the vectors $\boldsymbol{\alpha}_n \in \mathbb{R}^{2PND}$ and $\boldsymbol{z}_{\boldsymbol{v}}[\tau] \in \mathbb{R}^{2PND}$, respectively, which allows to rewrite the loss function as

$$\mathcal{L}^n(\boldsymbol{\alpha}_n) = \frac{1}{2}\sum_{\tau=P}^{T-1}\Big[y_n[\tau] - \boldsymbol{\alpha}_n^\top \boldsymbol{z}_{\boldsymbol{v}}[\tau]\Big]^2. \tag{D.15}$$

D.2.2 Missing data

To formulate the topology identification problem with missing data and noisy observation, we assume that only a subset of the nodes is observed. The motif of missing data is represented by the masking vector $\mathbf{m}[t] \in R^N$, where $m_n[t], n = 1, ..., N$, are i.i.d Bernoulli random variables. The observed vector signal $\tilde{\mathbf{y}}[t]$ at time t is given by

$$\tilde{\mathbf{y}}[t] = \mathbf{m}[t] \odot (\mathbf{y}[t] + \mathbf{e}[t]), \tag{D.16}$$

where $\mathbf{y}[t] = [y_1[t], ..., y_n[\tau]]\top \in \mathbb{R}^N$ and $\mathbf{e}[t] \in \mathbb{R}^N$ are the original signal and observation noise in vector form and $\odot$ represents the element wise multiplication.

D.2.3 Nonlinear topology identification with missing data

A batch formulation for the joint topology identification and missing data imputation can be formulated similarly to [18] and [6] as follows:

$$\begin{aligned}\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{y}}[\tau]\}_{\tau=P}^{\tau=T-1} = \arg\min_{\boldsymbol{\alpha},\mathbf{y}[\tau]} \sum_{\tau=P}^{T-1}\frac{1}{2}\|\mathbf{y}[\tau] - \boldsymbol{\alpha}^\top \boldsymbol{z}_v[\tau]\|_2^2 \\ + \lambda\sum_{n'=1}^{N}\sum_{d=1}^{2D}\|\boldsymbol{\alpha}_{n,n',d}\|_2 + \sum_{\tau=P}^{T-1}\frac{\nu}{2M_\tau}\|\tilde{\mathbf{y}}[\tau] - \mathbf{m}[\tau] \odot \mathbf{y}[\tau]\|_2^2,\end{aligned} \tag{D.17}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_N^\top] \in \mathbb{R}^{2PND} \times \mathbb{R}^N$, M_τ is cardinality of $\mathbf{m}[\tau]$, and ν is a hyperparameter that regulates the signal reconstruction part.

D.3 Joint online estimation of nonlinear topology and missing data

Note that $\boldsymbol{z}_\nu$ depends on P previous values of all the N time series. Hence the required online estimation strategy should estimate P previous values of the time

series along with the instantaneous values:

$$\{\widehat{\boldsymbol{\alpha}}, \hat{\mathbf{y}}[t], \{\hat{\mathbf{y}}[\tau]\}_{\tau=t-P}^{t-1}\} =$$

$$\underset{\substack{\boldsymbol{\alpha}, \mathbf{y}[t] \\ \{\mathbf{y}[\tau]\}_{\tau=t-P}^{t-1}}}{\arg\min}\, \ell_t\left(\boldsymbol{\alpha}, \mathbf{y}[t], \{\mathbf{y}[\tau]\}_{\tau=t-P}^{t-1}\right) + \lambda \sum_{n'=1}^{N} \sum_{p=1}^{P} \|\boldsymbol{\alpha}_{n,n'}^{(p)}\|_2, \tag{D.18}$$

where the non decreasing function $\Omega(.) = |.|$ and the loss function is defined as

$$\ell_t\left(\boldsymbol{\alpha}, \mathbf{y}[t], \{\mathbf{y}[\tau]\}_{\tau=t-P}^{t-1}\right) =$$
$$\frac{1}{2}\|\mathbf{y}[t] - \boldsymbol{\alpha}^\top \boldsymbol{z}_v[t]\|_2^2 + \frac{\nu}{2M_t}\|\tilde{\mathbf{y}}[t] - \mathbf{m}[t] \odot \mathbf{y}[t]\|_2^2. \tag{D.19}$$

We relax the formulation (D.18) since it is computationally expensive as well as nonconvex. We assume that $\{\hat{\mathbf{y}}[\tau]\}_{\tau=t-P}^{t-1}$ are independent realizations of random variables $\{\mathbf{y}[\tau]\}_{\tau=t-P}^{t-1}$ [6] and obtain a new loss function:

$$\tilde{\ell}_t\left(\boldsymbol{\alpha}, \mathbf{y}[t]\right) = \frac{1}{2}\|\mathbf{y}[t] - \boldsymbol{\alpha}^\top \boldsymbol{z}_v[t]\|_2^2 + \frac{\nu}{2M_t}\|\tilde{\mathbf{y}}[t] - \mathbf{m}[t] \odot \mathbf{y}[t]\|_2^2. \tag{D.20}$$

Now the loss function is convex and separable across n. Hence the optimization problem for a node can be expressed as

$$\{\widehat{\boldsymbol{\alpha}}_n, \hat{y}_n[t]\} = \underset{\boldsymbol{\alpha}_n, y_n[t]}{\arg\min}\, \ell_t^n\left(\boldsymbol{\alpha}_n, y_n[t]\right) + \lambda \sum_{n'=1}^{N} \sum_{p=1}^{P} \|\boldsymbol{\alpha}_{n,n'}^{(p)}\|_2, \tag{D.21}$$

$$where \ell_t^n\left(\boldsymbol{\alpha}_n, y_n[t]\right) = \frac{1}{2}\left[y_n[t] - \boldsymbol{\alpha}_n^\top \boldsymbol{z}_v[t]\right]^2 + \frac{\nu}{2M_t}(\tilde{y}_n[t] - m_n[t] y_n[t])^2. \tag{D.22}$$

We use the alternating minimization method in which (D.21) is solved by alternating between two sub-problems that are convex and have closed-form solutions. Since the optimization problem with respect to $y_n[t]$ (the signal reconstruction problem) is quadratic, a closed-form solution can be obtained. The second optimization problem with respect to $\boldsymbol{\alpha}_n$ (topology identification) is in a form similar to the one discussed in [22], where it is solved in a closed form using composite objective mirror descent (COMID) method.

D.3.1 Signal reconstruction

The signal reconstruction problem can be formulated as

$$\hat{y}_n[t] = \underset{y_n[t]}{\arg\min}\, \ell_t^n\left(\boldsymbol{\alpha}_n, y_n[t]\right). \tag{D.23}$$

The solution of (D.23) is obtained by finding the zero derivative point of the objective function:

$$\hat{y}_n[t] = \frac{\nu m_n[t] \tilde{y}_n[t]}{M_t + \nu m_n[t]} + \frac{k_n[t] M_t}{\nu m_n[t] + M_t}, \tag{D.24}$$

where $k_n[t] = \boldsymbol{\alpha}_n^\top \boldsymbol{z}_v[t]$. Let $\frac{\nu m_n[t]}{M_t + \nu m_n[t]} = q_n[t]$, then,

$$\hat{y}_n[t] = q_n[t] \tilde{y}_n[t] + [1 - q_n[t]] k_n[t]. \tag{D.25}$$

D.3.2 Topology identification

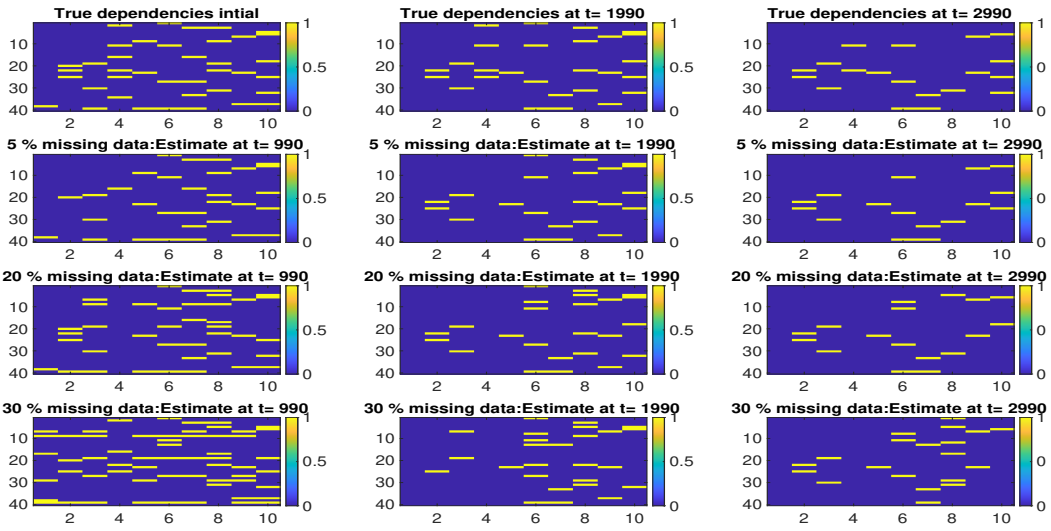

Figure D.1: True edges $(a_{n,n'}^{(p)})$ and estimated weights $(\widehat{b}_{n,n'}^{(p)})$ for various missing data scenarios.

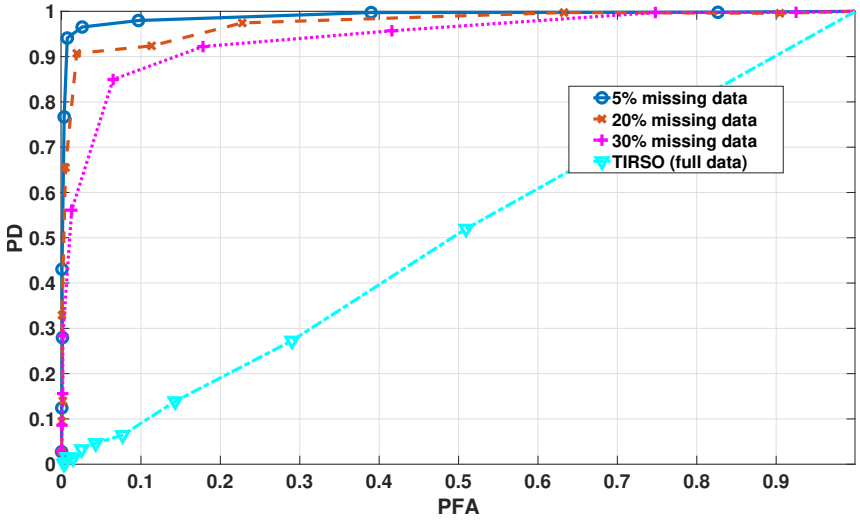

Figure D.2: ROC curve.

Figure D.3: Results: Experiment using synthetic data.

We use the estimates $\{\hat{y}_n[\tau]\}_{\tau=t-P}^{t}$ obtained using (D.25) to find the topology. This sub-problem can be formulated as

$$\widehat{\boldsymbol{\alpha}}_n = \arg \min_{\boldsymbol{\alpha}_n} \ell_t^n(\boldsymbol{\alpha}_n) + \lambda \sum_{n'=1}^{N} \sum_{p=1}^{P} \|\boldsymbol{\alpha}_{n,n'}^{(p)}\|_2. \tag{D.26}$$

where $\ell_t^n(\boldsymbol{\alpha}_n) = \frac{1}{2}[\hat{y}_n[t] - \boldsymbol{\alpha}_n^\top \boldsymbol{z}_{\boldsymbol{v}}[t]]^2$. The convex objective function in (D.26) contains two terms: a smooth loss function and a non-smooth regularizer. Such problems can be solved efficiently using COMID methods [22]. The online COMID update is

given by

$$\boldsymbol{\alpha}_n[t+1] = \arg\min_{\boldsymbol{\alpha}_n} J_t^{(n)}(\boldsymbol{\alpha}_n), \tag{D.27}$$

$$\text{where } J_t^{(n)}(\boldsymbol{\alpha}_n) \triangleq \nabla \ell_t^n(\boldsymbol{\alpha}_n[t])^\top [\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n[t]] + \frac{1}{2\gamma_t}\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n[t]\|_2^2 + \lambda \sum_{n'=1}^{N}\sum_{p=1}^{P} \|\boldsymbol{\alpha}_{n,n'}^{(p)}\|_2. \tag{D.28}$$

In (D.28), $\boldsymbol{\alpha}_n[t] \in \mathbb{R}^{2PND}$ is the estimate of $\boldsymbol{\alpha}_n$ at time t. The objective function $J_t^{(n)}(.)$ consists of three terms: *(i)* gradient of the loss function, *(ii)* Bregman divergence $\|\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n[t]\|_2^2$ chosen such that the optimization problem (D.28) has a closed-form solution (γ_t is the step size associated with the divergence), and *(iii)* a sparsity promoting group lasso regularizer. Note that the Bregman divergence term increases stability of the online algorithm by enforcing the next iterate $\boldsymbol{\alpha}_n[t+1]$ to be closer to current iterate $\boldsymbol{\alpha}_n[t]$. The gradient in (D.28) is evaluated as

$$\mathbf{v}_n[t] := \nabla \ell_t^n(\boldsymbol{\alpha}_n[t]) = \boldsymbol{z}_v[t][\boldsymbol{\alpha}_n^\top \boldsymbol{z}_v[t] - \hat{y}_n[t]]. \tag{D.29}$$

A closed-form solution for (D.27) is obtained via the multidimensional shrinkage-thresholding operator:

$$\boldsymbol{\alpha}_{n,n'}^{(p)}[t+1] = [\boldsymbol{\alpha}_{n,n'}^{(p)}[t] - \gamma_t \mathbf{v}_{n,n'}^{(p)}[t]] \times \left[1 - \frac{\gamma_t \lambda}{\|\boldsymbol{\alpha}_{n,n'}^{(p)}[t] - \gamma_t \mathbf{v}_{n,n'}^{(p)}[t]\|_2}\right]_+, \tag{D.30}$$

where $[x]_+ = \max\{0, x\}$. The above solution is a product of two terms: first term minimizes the loss function $\ell_t^n(\boldsymbol{\alpha}_n)$ and the second term enforces sparsity on the updates. The proposed algorithm for jointly estimating the topology and the missing data is summarized in Algorithm 8.

Algorithm 8:

Result: $\left\{\boldsymbol{\alpha}_{n,n'}^{(p)}[t+1]\right\}_{n,n',p}$, $\hat{\mathbf{y}}[t]$

Initialize $\{\boldsymbol{y}_n[t]\}_{t=1}^P$, $\left\{\boldsymbol{\alpha}_{n,n'}^{(p)}[P]\right\}_{n,n',p}$ as all-ones vector, λ, kernel parameters, γ, D, ν (heuristically chosen)

for $t = P, P+1, \ldots$ **do**

  Get data observation vector $\tilde{\mathbf{y}}_n[t]$ and masking vector $\mathbf{m}[t]$, compute $\boldsymbol{z}_{\boldsymbol{v}}[t]$

  for $n = 1, \ldots, N$ **do**

    compute $\hat{y}_n[t]$ using (D.25)

    compute $\mathbf{v}_n[t]$ using (D.29)

    for $n' = 1, \ldots, N$ **do**

      compute $\boldsymbol{\alpha}_{n,n'}^{(p)}[t+1]$ using (D.30)

    end

  end

end

D.4 Experiment

In this section, we test the capability of our algorithm using both synthetic and real data. We generate graph-connected time series with known topologies and varying levels of missing data for synthetic data experiments, whereas, in the second part, we use real data from Lundin's offshore oil and gas platform[1]. The ℓ_2 norms of the estimated weights ($\widehat{b}^{(p)}_{n,n'}[t] := \|\boldsymbol{\alpha}^{(p)}_{n,n'}[t]\|_2$) are used to visualize the dependencies among the time series. For all the experiments, we used Gaussian reproducing kernel k with variance $\sigma_k^2 = 5$.

D.4.1 Experiments using Synthetic data

The data used in this experiment are generated using nonlinear VAR model described in (D.1) with $N = 10, P = 4$ and random Gaussian noise with mean 0 and variance 0.01. The nonliner function in (D.1) is taken as $f^{(p)}_{n,n'}(x) = a^{(p)}_{n,n'}(x)g(x)$, $\forall n, n', p$, where $g(x) = 0.25\sin(x^2) + 0.25\sin(2x) + 0.5\sin(x)$ and $a^{(p)}_{n,n'}(x) \in \{0, 1\}$. We term $a^{(p)}_{n,n'}$ as *edge* and when $a^{(p)}_{n,n'} = 0$, it disables the dependencies between the nodes n and n' for the time lag p. Furthermore, $a^{(p)}_{n,n'}(x) = 0$, when $g(x) = 0$. The time series are initialized randomly using samples drawn from uniform distribution $\mathcal{U}(0, 1)$. To bring time variance in the topology, 30% of the active edges are made to disappear after every 1000 time stamps, and new equal number of different edges are made active. To simulate various missing data scenarios, we generate different masks $\mathbf{m}[t]\ \forall\ t$, whose samples are drawn from Bernoulli distribution with probabilities $0.95, 0.75, 0.65$, corresponds to $5\%, 25\%, 35\%$ of missing data respectively.

In Fig. D.1, we compare the true edges $a^{(p)}_{n,n'}$ and estimated causal weights $\widehat{b}^{(p)}_{n,n'}$ at three different time instants having different edge patterns. The edges and the estimated weights are arranged in a matrix form of size $N \times N$ for $p = 1, 2, \ldots, P$ and are stacked in Fig. D.1, such that the resulting matrices are of size $NP \times N$. The estimated weights are normalized and hard-thresholded to 0 or 1 to have the best match with the edges. It can be observed in Fig. D.1 that for 5% of missing data, the proposed algorithm estimates most of the edges accurately, and as the number of missing data increases, the estimation accuracy decreases. The ROC curve corresponding to the time stamp $t = 990$ is plotted in Fig. D.2 by computing the probability of detection (PD) and the probability of false alarm (PFA). Figure D.2 shows that the areas under all the three curves are close to 1, indicating the characteristics of a good ROC curve. It can also be observed that the area under the curve deviates more from 1 as the number of missing data increases. Also, the ROC curve for a recent online linear topology estimation algorithm termed TIRSO [48] is included in Fig. D.2. Note that TIRSO's ROC is computed based on full data; even then, its performance significantly lags behind the proposed algorithm. Intuitively, JSTIRSO [6], the extension to TIRSO that accounts for missing data, should also perform inferiorly to the proposed algorithm. These observations illustrate how

[1] https://www.lundin-energy.com/

effectively the proposed algorithm identifies nonlinear topologies compared to its linear counterparts.

D.4.2 Experiments using Real data

We use real data from Lundin's oil and gas plant, consisting of time series recorded from multiple pressure (P), temperature (T), and oil level (L) sensors from $system_{20}$ of the plant. The $system_{20}$ is a plant section where oil, gas, and water are separated from the well extracts. There are 24 sensors in total recording 24 time series, sampled at intervals of $5s$. Below, we examine two different missing data scenarios.

D.4.2.1 Missing data due to limited communication capacity

Assume that only a subset of the sensor values can be transmitted at each timestamp due to the limited capacity of the communication channel. We randomly select 8 out of the 24 sensors ($\sim 33.33\%$) at each time stamp and jointly estimate the topologies and the missing data. The true and observed time series of a sensor, along with the reconstructed values, are shown in Fig. D.4, which shows that the proposed algorithm reconstructs the signal even with a high amount of missing data. Since the ground truth dependencies are unavailable, we compare the dependencies estimated from the partial observations with that from a full observation in Fig. D.5, which shows that the algorithm can estimate most of the dependencies from the partial observations.

D.4.2.2 Missing data due to sensor failure

Here we consider the case where the recording from a particular sensor is missing for a certain period of time due to a sensor failure. In the experiment, time series from sensor-2 are masked from time instant $t = 4000$ to $t = 4200$, which constitutes about 16 minutes of data. Figure D.6 shows that the proposed algorithm reconstructs sensor-2 signals accurately during the missing data interval without having access to any information from sensor-2. This clearly showcases the advantage of exploiting causal dependencies in missing data imputations.

Conclusion

This paper presents a novel algorithm for joint nonlinear topology identification and missing data imputation. The nonlinear causal dependencies are modeled using a computationally light kernel formulation based on random feature approximations. Experiments on real and synthetic data have demonstrated the effectiveness of the proposed algorithm under various missing data scenarios.

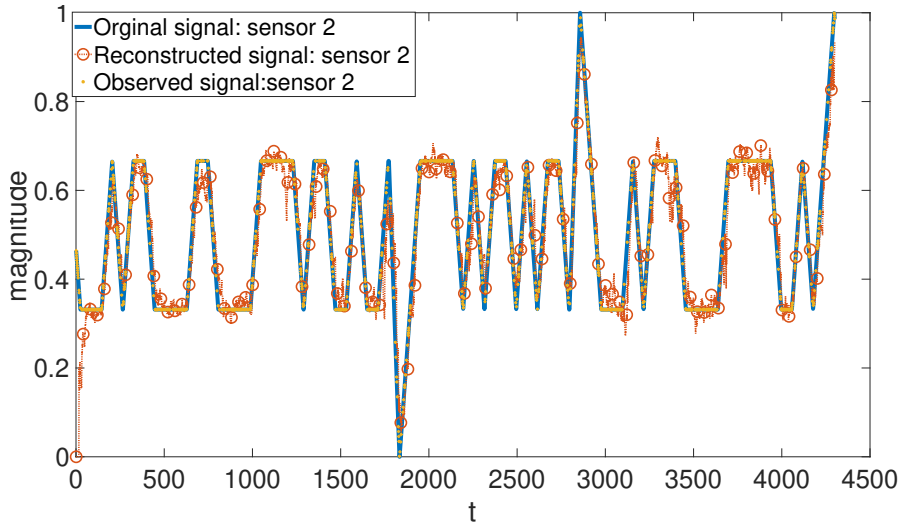

Figure D.4: Original and reconstructed signal when only 33.33% of data is observable.

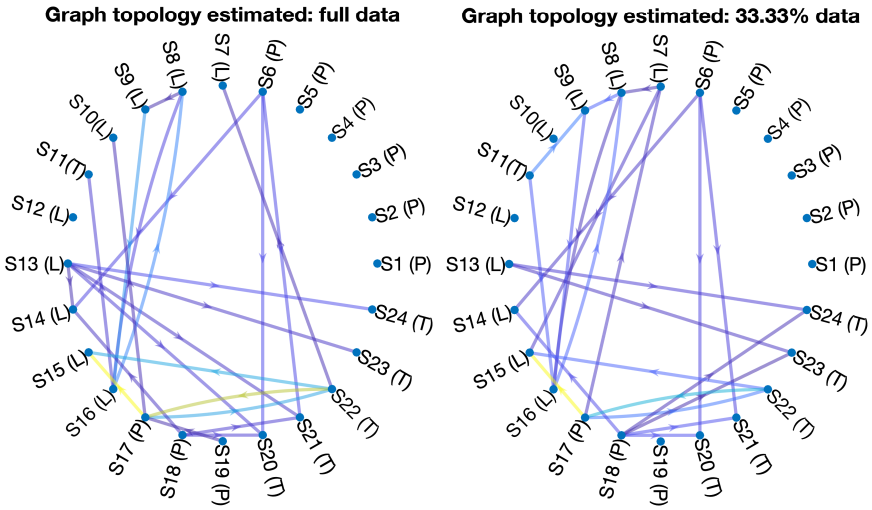

Figure D.5: Causality graph estimated for oil and gas platform (Only the significant edges are shown).

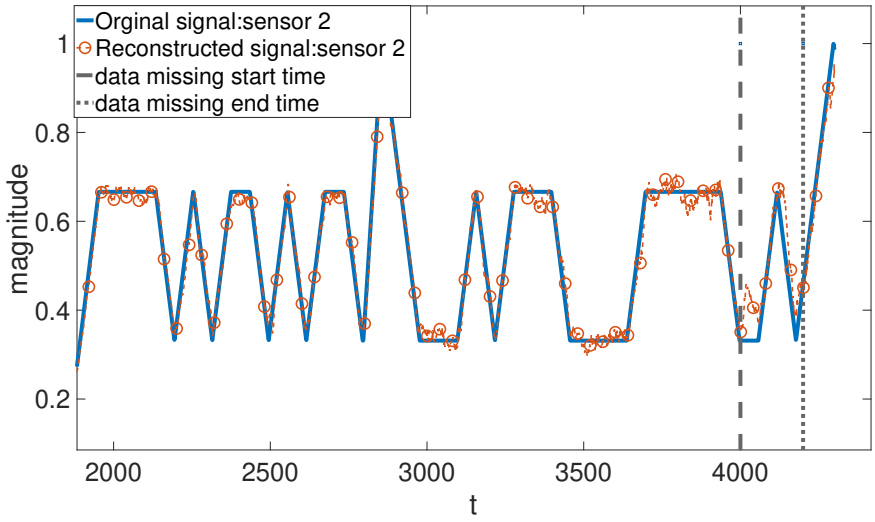

Figure D.6: Comparison of real and reconstructed signal when an interval of data is missing for a sensor.

Figure D.7: Results: Experiment using real data.