IEEE Open Journal of
Signal Processing

Received 26 October 2022; revised 12 January 2023; accepted 17 January 2023. Date of publication 1 February 2023;
date of current version 16 February 2023. The review of this article was arranged by Associate Editor Vincenzo Matta.

Digital Object Identifier 10.1109/OJSP.2023.3241580

Scalable and Privacy-Aware Online Learning of Nonlinear Structural Equation Models

ROHAN MONEY [1], JOSHIN KRISHNAN [2] (Member, IEEE), BALTASAR BEFERULL-LOZANO [1,2] (Senior Member, IEEE), AND ELVIN ISUFI [3] (Member, IEEE)

[1]WISENET Center, Department of ICT, University of Agder, Grimstad 4879, Norway
[2]SIGIPRO Department, Simula Metropolitan Center for Digital Engineering, Oslo 0167, Norway
[3]Faculty of EEMCS, TU Delft, 2628 Delft CD, The Netherlands

CORRESPONDING AUTHOR: ROHAN MONEY (e-mail: rohantm@uia.no).

This work was supported in part by the IKTPLUSS INDURB through the Research Council of Norway under Grant 270730/O70 and in part by TU Delft AI labs programme.

ABSTRACT An online topology estimation algorithm for nonlinear structural equation models (SEM) is proposed in this paper, addressing the nonlinearity and the non-stationarity of real-world systems. The nonlinearity is modeled using kernel formulations, and the curse of dimensionality associated with the kernels is mitigated using random feature approximation. The online learning strategy uses a group-lasso-based optimization framework with a prediction-corrections technique that accounts for the model evolution. The proposed approach has three properties of interest. First, it enjoys node-separable learning, which allows for scalability in large networks. Second, it offers privacy in SEM learning by replacing the actual data with node-specific random features. Third, its performance can be characterized theoretically via a dynamic regret analysis, showing that it is possible to obtain a linear dynamic regret bound under mild assumptions. Numerical results with synthetic and real data corroborate our findings and show competitive performance w.r.t. state-of-the-art alternatives.

INDEX TERMS Network topology inference, time-varying graph learning, structural equation models, random feature approximation.

I. INTRODUCTION

Structural Equation Models (SEM) are a prevalent tool to model interactions in real-world networks due to their simplicity and ability to express instantaneous directed relationships between interacting entities [1], [2], [3]. The advantages of SEM over simple correlation-based models lie in leveraging the directionality, which is key to many applications, such as modeling the functional connectivity between brain regions [4] and interactions in financial networks [5], to name a few. SEM modeling and its topology estimation are challenging because real-life networks are large, dynamic, and comprise nonlinear interactions, as well as leveraging directly node-specific data may raise privacy concerns [1].

Although SEM-based topology estimation has been explored in literature, most of the approaches are developed for stationary linear systems and provide offline (batch-based) solutions [6], [7]. Modeling time-varying systems call for online optimization strategies, which can be classified into *time-unstructured* and *time-structured* methods [8], [9]. The former update the model only when a new data sample arrives [10], whereas the latter first predict the model based on its evolution and then correct the prediction when the new data sample is available [11]. The time-structured algorithms are expected to perform better since they take advantage of the prior related to the model evolution but typically have a slightly higher computational cost. A SEM-based online topology estimation has been proposed in [12], but it adopts the time-unstructured strategy and fails to exploit the model evolution; hence, suboptimal. On the other hand, [9] and [13] propose time-structured online SEM learning strategies, but the models are restricted to linear interactions. Moreover, the node operations of these algorithms are computationally expensive, and they assume symmetric interactions of the network data, which destroys SEM's directionality features.

Aiming to overcome the above challenges, we propose an online topology learning algorithm for nonlinear and directed SEM using a time-structured optimization framework. The nonlinearity is tackled using kernel methods, and the curse of dimensionality of kernels is mitigated through random feature (RF) approximation. Kernel techniques are conventionally used for nonlinear topology estimation [14], [15], [16] and help transform the problem into an amenable form. Instead, RF is typically used to reduce the complexity of nonlinear models as well as to ensure that connectivity is inferred without revealing nodal attributes [17], [18], [19], [20], [21], [22]. Through a series of design choices and theoretical derivations, we show how kernels and RFs can be incorporated into the online nonlinear SEM model and show that the proposed algorithm has the following four properties:

i) *Sparse model evolution:* The proposed SEM learning strategy uses a prediction-correction approach to model the SEM evolution. Exploiting the fact that real-world networks exhibit sparse directed interactions, we introduce a group-lasso-based regularizer to learn sparse models.
ii) *Scalability:* The proposed algorithm is separable across the nodes with a fixed computational complexity per iteration, thereby facilitating scalability in large graphs.
iii) *Privacy:* The node separability and the random features avoid sharing the true data, thus, ensuring node privacy.
iv) *Convergence Guarantee:* A dynamic regret analysis of the proposed algorithm is conducted, guaranteeing convergence, and showing the role played by the different components of the proposed method.

Numerical experiments on synthetic data and real data from neuroscience and finance corroborate the above contributions and show superior performance to competing alternatives.

The rest of the paper is organized as follows. Section II presents the nonlinear SEM, kernel formulation, and random feature approximation. Section III develops an online strategy for learning the nonlinear SEM using a prediction-correction algorithm. The dynamic regret analysis of the proposed algorithm is performed in Section IV, and the numerical results are provided in Section V. Section VI concludes the paper. All proofs are collected in the Appendix.

II. PROBLEM FORMULATION

Consider N interdependent time series, and let $y_n[t]$ be the value of the n-th time series at time t. A nonlinear SEM with no exogenous variables models the dependencies among these time series as

$$y_n[t] = \sum_{n'=1, n' \neq n}^{N} f_{n,n'}(y_{n'}[t]) + u_n[t],\ n = 1, \ldots, N, \tag{1}$$

where $f_{n,n'}(\cdot)$ encodes the nonlinear influence of n'-th time series on n-th time series, and $u_n[t]$ is the observation noise [23]. For example, in the context of brain networks, $y_n[t]$ represents the electroencephalogram (EEG) recorded at the n-th node (sensor) at time t, and $f_{n,n'}(\cdot)$ encodes the functional connectivity between the nodes n and n'.

Kernel representation: The nonlinear structure in (1) allows modeling a broader range of problems, but at the same time makes it more difficult to analyse and model the time series interactions. A typical way to approach these challenges is to consider the nonlinear function in (1) belonging to a reproducing kernel Hilbert space (RKHS):

$$\mathcal{H}_{n'} := \Big\{ f_{n,n'}(\cdot) \,|\, f_{n,n'}\left(y_n[t']\right) = \sum_{t=0}^{\infty} \beta_{n,n,'t}\, \kappa_{n'}\left(y_n[t'], y_{n'}[t]\right) \Big\}, \tag{2}$$

where $\kappa_{n'}(\cdot,\cdot)$ is a positive definite kernel function, measuring the similarity between its arguments. Every positive definite kernel has an associated RKHS characterized by the inner product: $\langle \kappa_{n'}(y, x_1), \kappa_{n'}(y, x_2)\rangle := \sum_{t=0}^{\infty} \kappa_{n'}(y[t], x_1)\kappa_{n'}(y[t], x_2)$. RKHS kernels satisfy the reproducing property $\left\langle \kappa_{n'}^{(p)}(y, x_1), \kappa_{n'}(y, x_2)\right\rangle = \kappa_{n'}(x_1, x_2)$, and induces a norm $\|f_{n,n'}\|^2_{\mathcal{H}_{n'}} = \sum_{t=0}^{\infty}\sum_{t'=0}^{\infty} \beta_{n,n,'t}\, \beta_{n,n,'t'}\, \kappa_{n'}(y_n[t], y_n[t'])$. It is possible to express any function in RKHS as an infinite sum of kernel evaluations weighted by $\beta_{n,n,'t}$ [15].

For a node n, the functional dependency can be obtained by solving

$$\left\{\hat{f}_{n,n'}\right\}_{n'} = \underset{\{f_{n,n'} \in \mathcal{H}_{n'}\}}{\operatorname{argmin}} \frac{1}{2} \sum_{\tau=0}^{T-1} \left[y_n[\tau] - \sum_{n'=1, n' \neq n}^{N} f_{n,n'}(y_{n'}[\tau]) \right]^2 + \lambda \sum_{n'=1, n' \neq n}^{N} \Omega\left(||f_{n,n'}||_{\mathcal{H}_{n'}}\right), \tag{3}$$

where $\Omega(\cdot)$ is a regularizing function with the hyperparameter $\lambda > 0$. We consider $\Omega(x) = |x|$ to induce a sparse SEM model. In (3), the function $f_{n,n'}(\cdot)$ belongs to the RKHS, which is an infinite dimensional space [cf. (2)]. However, for non-decreasing regularizing functions such as $\Omega(x) = |x|$, $x \geq 0$, the solution of (3) can be expressed with a finite number of parameters using the Representer Theorem [24]:

$$\hat{f}_{n,n'}\left(y_{n'}[\tau]\right) = \sum_{t=0}^{T-1} \beta_{n,n,'t}\, \kappa_{n'}\left(y_{n'}[\tau], y_{n'}[t]\right). \tag{4}$$

As the number of data samples increases, the number of kernel evaluations in (4) and the parameters required to express the function also increases. We overcome this curse of dimensionality by using random feature (RF) approximation.

RF approximation: RF approximation limits the kernel evaluations to a fixed lower-dimensional Fourier space for kernels with a shift-invariant property, i.e., $\kappa_{n'}(y_{n'}[\tau], y_{n'}[t]) = \kappa_{n'}(y_{n'}[\tau] - y_{n'}[t])$; thus, preventing the dimensionality growth. According to Bochner's theorem [25], an inverse Fourier transform of a probability distribution can represent

a shift-invariant kernel:

$$\kappa_{n'}\left(y_{n'}[\tau], y_{n'}[t]\right) = \int_{\mathbb{R}} \pi_{\kappa_{n'}}(v)\, e^{jv\left(y_{n'}[\tau]-y_{n'}[t]\right)} dv$$
$$= \mathbb{E}_v\left[e^{jv\left(y_{n'}[\tau]-y_{n'}[t]\right)}\right], \quad (5)$$

where $\pi_{\kappa_{n'}}(v)$ is the kernel-specific probability density function (pdf), v is the random variable associated to the pdf, and $\mathbb{E}[\cdot]$ is the expectation operator. Given a sufficient number D of i.i.d. samples $\{v_i\}_{i=1}^{D}$ drawn from distribution $\pi_{\kappa_{n'}}(v)$, the expectation is estimated by the sample mean:

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

Finding the probability distribution which is the inverse Fourier transform of a kernel is a difficult task in general. However, choosing a Gaussian kernel with variance σ^2 avoids this difficulty since its Fourier transform is also a Gaussian with variance σ^{-2}. This allows writing the real part of (6) as

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

where

$$\boldsymbol{z}_{\boldsymbol{v},n'}[\tau] = \frac{1}{\sqrt{D}}\big[\sin\left(v_1 y_{n'}\tau]\right), \ldots, \sin\left(v_D y_{n'}\tau]\right),$$
$$\cos\left(v_1 y_{n'}\tau]\right), \ldots, \cos\left(v_D y_{n'}\tau]\right)\big]^{\top}. \quad (8)$$

A fixed dimensional ($2D$) representation of the function $\hat{f}_{n,n'}(\cdot)$ is obtained by substituting (7) into (4):

$$\hat{\tilde{f}}_{n,n'}\left(y_{n'}\tau]\right) = \sum_{t=0}^{T-1} \beta_{n,n',t}\boldsymbol{z}_{\boldsymbol{v},n'}[\tau]^{\top}\boldsymbol{z}_{\boldsymbol{v},n'}[t]$$
$$= \boldsymbol{\alpha}_{n,n'}{}^{\top}\boldsymbol{z}_{\boldsymbol{v},n'}[\tau], \quad (9)$$

where $\boldsymbol{\alpha}_{n,n'} = \sum_{t=0}^{T-1}\beta_{n,n',t}\boldsymbol{z}_{\boldsymbol{v},n'}[t]$. Using (9), we can reformulate the non-parametric problem (3) into a parametric optimization problem:

$$\left\{\widehat{\boldsymbol{\alpha}}_{n,n'}\right\}_{n'} = \underset{\{\boldsymbol{\alpha}_{n,n'}\}}{\arg\min}\, \frac{1}{2}\sum_{\tau=0}^{T-1}\left[y_n[\tau] - \sum_{n'=1,n'\neq n}^{N}\boldsymbol{\alpha}_{n,n'}^{\top}\boldsymbol{z}_{\boldsymbol{v},n'}[\tau]\right]^2$$
$$+\lambda\sum_{n'=1,n'\neq n}^{N}||\boldsymbol{\alpha}_{n,n'}||_2, \quad (10)$$

The regularizer in (10) is a group-lasso regularizer to enforce sparsity in the RF coefficient $\boldsymbol{\alpha}_{n,n'} \in \mathbb{R}^{2D}$. For brevity, we stack the vectors $\boldsymbol{\alpha}_{n,n'}$ and $\boldsymbol{z}_{\boldsymbol{v},n'}[t]$ along the index $n' = 1, \ldots, N$, $n' \neq n$ to form $\boldsymbol{\alpha}_n \in \mathbb{R}^{2(N-1)D}$ and $\boldsymbol{z}_n[t] \in \mathbb{R}^{2(N-1)D}$, and compactly write (10) as

$$\widehat{\boldsymbol{\alpha}}_n = \arg\min_{\boldsymbol{\alpha}_n} \mathcal{L}^n(\boldsymbol{\alpha}_n) + \lambda\sum_{n'=1,n'\neq n}^{N}||\boldsymbol{\alpha}_{n,n'}||_2, \quad (11)$$

where

$$\mathcal{L}^n(\boldsymbol{\alpha}_n) = \frac{1}{2}\sum_{\tau=0}^{T-1}\left[y_n[\tau] - \boldsymbol{\alpha}_n^{\top}\boldsymbol{z}_n[\tau]\right]^2. \quad (12)$$

Solving problem (11) requires access to all the batch of time series $\{y_n[\tau]\}_{\tau=0}^{T-1}$ which may be practically infeasible as they evolve over time and, at the same time, it is computationally demanding. Targeting real-world nonstationary systems with streaming data, we develop an online strategy enhanced by prediction correction mechanisms [11] that exploit the nonlinear SEM evolution. However, the group-lasso regularizer, required to enforce sparse dependencies is non-differentiable, making the deployment of prediction-correction methods not straightforward.

III. TIME-VARYING SOLUTION

A. ONLINE LOSS FUNCTION

Following online optimization, we replace the batch loss in (12) with a recursive least square loss (RLS) using an exponential window:

$$\tilde{\ell}_t^n(\boldsymbol{\alpha}_n) = \mu\sum_{\tau=0}^{t}\gamma^{t-\tau}\ell_\tau^n(\boldsymbol{\alpha}_n). \quad (13)$$

where $\ell_\tau^n(\boldsymbol{\alpha}_n) = \frac{1}{2}[y_n[\tau] - \boldsymbol{\alpha}_n^{\top}\boldsymbol{z}_n[\tau]]^2$ is the instantaneous loss function, $\gamma \in (0,1)$ is the forgetting factor of the window, and $\mu = 1-\gamma$ normalizes the window. The RLS loss function can be expanded as

$$\tilde{\ell}_t^n(\boldsymbol{\alpha}_n) = \frac{1}{2}\mu\sum_{\tau=0}^{t}\gamma^{t-\tau}\left(y_n^2[\tau] + \boldsymbol{\alpha}_n^{\top}\boldsymbol{z}_n[\tau]\boldsymbol{z}_n[\tau]^{\top}\boldsymbol{\alpha}_n\right.$$
$$\left.-2y_n[\tau]\boldsymbol{z}_n[\tau]^{\top}\boldsymbol{\alpha}_n\right)$$
$$= \frac{1}{2}\mu\sum_{\tau=0}^{t}\gamma^{t-\tau}y_n^2[\tau] + \frac{1}{2}\boldsymbol{\alpha}_n^{\top}\boldsymbol{\Phi}_n[t]\boldsymbol{\alpha}_n - \boldsymbol{r}_n^{\top}\boldsymbol{\alpha}_n, \quad (14)$$

where

$$\boldsymbol{\Phi}_n[t] = \mu\sum_{\tau=0}^{t}\gamma^{t-\tau}\boldsymbol{z}_n[\tau]\boldsymbol{z}_n[\tau]^{\top}, \quad (15)$$

$$\boldsymbol{r}_n[t] = \mu\sum_{\tau=0}^{t}\gamma^{t-\tau}y_n[\tau]\boldsymbol{z}_n[\tau]. \quad (16)$$

The new optimization problem using the RLS loss becomes

$$\arg\min_{\boldsymbol{\alpha}_n}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n) + \lambda\sum_{n'=1,n'\neq n}^{N}\|\boldsymbol{\alpha}_{n,n'}\|_2. \quad (17)$$

The objective function in (17) has a differentiable loss but a non-differentiable regularizer. We solve it using composite objective mirror descent (COMID) [26] with the online updates:

$$\boldsymbol{\alpha}_n^{(1)}[t+1] = \arg\min_{\boldsymbol{\alpha}_n}\Big[\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])^{\top}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_n[t])$$

$$+\frac{1}{2\nu_t}\|\boldsymbol{\alpha}_n-\boldsymbol{\alpha}_n[t]\|_2^2+\lambda\sum_{n'=1,n'\neq n}^{N}\|\boldsymbol{\alpha}_{n,n'}\|_2\Bigg], \quad (18)$$

where $\boldsymbol{\alpha}_n^{(1)}[t+1]$ denotes the one-step COMID descent of $\boldsymbol{\alpha}_n[t]$, ν_t the step size, and $\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])$ the gradient of $\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])$ w.r.t. $\boldsymbol{\alpha}_n$, which can be computed from (14) as

$$\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])=\boldsymbol{\Phi}_n[t]\boldsymbol{\alpha}_n-\boldsymbol{r}_n[t]. \quad (19)$$

In an online setting, the parameters $\boldsymbol{\Phi}_n[t]$ and $\boldsymbol{r}_n[t]$ can be estimated recursively as $\boldsymbol{\Phi}_n[t]=\gamma\boldsymbol{\Phi}_n[t-1]+\mu\boldsymbol{z}_v[t]\boldsymbol{z}_n[t]^\top$ and $\boldsymbol{r}_n[t]=\gamma\boldsymbol{r}_n[t-1]+\mu y_n[t]\boldsymbol{z}_n[t]$ [cf. (15) and (16)].

The COMID update (18) can be solved in closed-form for each lasso group $\boldsymbol{\alpha}_{n,n'}\in\boldsymbol{\alpha}_n$ [cf. (10)] using the multidimensional shrinkage thresholding operator (MSTO) [27]:

$$\boldsymbol{\alpha}_{n,n'}^{(1)}[t+1]=\left(\boldsymbol{\alpha}_{n,n'}[t]-\nu_t\mathbf{v}_{n,n'}\right)\times\left[1-\frac{\nu_t\lambda}{\|\boldsymbol{\alpha}_{n,n'}[t]-\nu_t\mathbf{v}_{n,n'}\|_2}\right]_+, \quad (20)$$

where $\left[\mathbf{v}_{n,1}^\top,\mathbf{v}_{n,2}^\top,\ldots,\mathbf{v}_{n,N}^\top\right]^\top\triangleq\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])$ and $[x]_+=\max\{0,x\}$. The MSTO solution (20) involves a one-step COMID update. For brevity of the succeeding formulation, we represent the K-step version of (20) as

$$\boldsymbol{\alpha}_n^{(K)}[t+1]=\text{MSTO}^{(K)}\left(\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t]),\nu_t,\lambda\right), \quad (21)$$

which computes the K-step descent update of $\boldsymbol{\alpha}_{n,n'}[t]$ as in (20), for $n'=1,\ldots,N$, $n'\neq n$, for the loss function $\tilde{\ell}_t^n(\cdot)$ with the parameters ν_t and λ, and stacks them to form $\boldsymbol{\alpha}_n^{(K)}[t+1]$.

B. PREDICTION-CORRECTION ALGORITHM

Although we can follow a time-unstructured learning strategy by directly using (21), such an approach discards the model evolution and leads to a suboptimal solution. Problem (17) features a strongly convex time-varying loss function and a properly convex regularizer, and such an optimization problem can be solved online using time-structured optimization methods that account for the model evolution. We follow the prediction-correction strategy as proposed in [11].

Prediction: The first step is to predict at time t, the yet unobserved loss function $\tilde{\ell}_{t+1}^n(\boldsymbol{\alpha}_n)$ using Taylor series expansion:

$$\tilde{\ell}_{t+1}^{n,pr}(\boldsymbol{\alpha}_n)=\boldsymbol{\alpha}_n{}^\top\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n)\boldsymbol{\alpha}_n+\Big[\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])+\nabla_{t\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])-\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])\boldsymbol{\alpha}_n[t]\Big]^\top\boldsymbol{\alpha}_n \quad (22)$$

In addition to the gradient computed in (19), prediction (22) requires computing the Hessian $\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])$ and the partial derivative of $\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])$ w.r.t. time $\nabla_{t\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])$ which have the forms

$$\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])=\boldsymbol{\Phi}_n[t], \quad (23)$$

$$\nabla_{t\boldsymbol{\alpha}}\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])=(\boldsymbol{\Phi}_n[t]-\boldsymbol{\Phi}_n[t-1])\boldsymbol{\alpha}-(\boldsymbol{r}_n[t]-\boldsymbol{r}_n[t-1]). \quad (24)$$

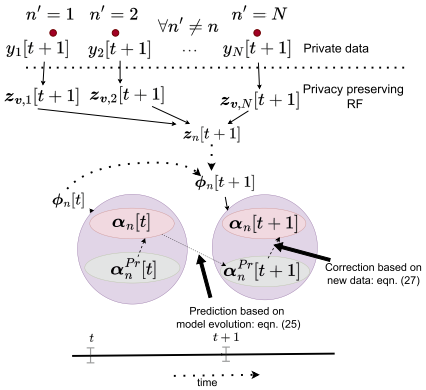

FIGURE 1. Schematic of the proposed method.

The group-lasso regularizer is a time-invariant function and just performs the thresholding operation in (20), irrespective of the time indices. Hence, it does not require prediction. Using the predicted loss (22) in place of (17), we predict the RF coefficients as

$$\boldsymbol{\alpha}_n^{pr}[t+1]=\text{MSTO}^{(P)}\left(\tilde{\ell}_{t+1}^{n,pr}(\boldsymbol{\alpha}_n[t]),\nu_t,\lambda\right), \quad (25)$$

where $\boldsymbol{\alpha}_n^{pr}[t+1]$ denotes the P-step COMID descent of $\boldsymbol{\alpha}_n[t]$ under the predicted loss. The gradient of the predicted loss involved in the MSTO operation (25) can be obtained from (22) as

$$\nabla_{\boldsymbol{\alpha}}\tilde{\ell}_{t+1}^{n,pr}(\boldsymbol{\alpha}_n[t])=(2\boldsymbol{\Phi}_n[t-1]-\boldsymbol{\Phi}_n[t-2])\boldsymbol{\alpha}_n+2\boldsymbol{r}_n[t-1]-\boldsymbol{r}_n[t-2]. \quad (26)$$

Correction: At time $t+1$, the loss $\tilde{\ell}_{t+1}^n(\cdot)$ [cf. the one appearing in (17)] becomes available, and the predicted RF coefficients $\boldsymbol{\alpha}_n^{pr}[t+1]$ are corrected via C-step COMID descents:

$$\boldsymbol{\alpha}_n[t+1]=\text{MSTO}^{(C)}\left(\tilde{\ell}_{t+1}^n\left(\boldsymbol{\alpha}_n^{pr}[t+1]\right),\nu_t,\lambda\right), \quad (27)$$

A high-level system model of the proposed method is presented in Fig. 1. At each time instant, the algorithm computes two estimates of the model parameters. For instance, at time $t+1$, we have two estimates: i) the predicted coefficient $\boldsymbol{\alpha}_n^{pr}[t+1]$, which is predicted based on evolution of the model and ii) the corrected coefficient $\boldsymbol{\alpha}_n[t+1]$, which is obtained by correcting the predicted coefficient when a new data sample is available. Note that in the proposed framework, nodes do not share the actual data $\{y_n[t+1]\}_{n=1,n\neq n'}^N$ with each other; instead the random features $\{z_{v,n}[t+1]\}_{n=1,n\neq n'}^N$ are shared, which ensures the privacy. A pseudocode of the proposed prediction-correction algorithm is provided in Algorithm 1. The computational complexity of the proposed algorithm is mainly contributed by the gradient evaluation steps (26) and (19); and it is of order $\mathcal{O}(N^2D^2)$ per node.

IV. DYNAMIC REGRET

To characterize the performance of the proposed online algorithm, we analyse its dynamic regret [28], which characterizes

Algorithm 1: Proposed Algorithm.

Result: $\{\boldsymbol{\alpha}_{n,n'}\}_{n,n'}$
Initialize $\lambda > 0$, $\nu_t > 0$, D, σ_n, P and C
for $t = 1, 2, \ldots$ **do**
  Get data samples $y_n[t],\ \forall n$ and compute $\boldsymbol{z_n}[t],\ \forall n$
  for $n = 1, \ldots, N$ **do**
    $\boldsymbol{\Phi}_n[t] = \gamma \boldsymbol{\Phi}_n[t-1] + \mu \boldsymbol{z}_n[t]\boldsymbol{z}_n[t]^\top$
    $\boldsymbol{r}_n[t] = \gamma \boldsymbol{r}_n[t-1] + \mu y_n[t]\boldsymbol{z}_n[t]$
    compute $\tilde{\ell}_{t+1}^{n,pr}(\boldsymbol{\alpha}_n)$ using (22)
    compute $\boldsymbol{\alpha}_n^{pr}[t+1]$ using (25)
    compute $\boldsymbol{\alpha}_n[t+1]$ using (27)
  end
end

the distance of the online loss function from the optimal counterpart in each time instant. The regret analysis is derived under the following mild assumptions:

A1) *Bounded time series:* there exists $B_y > 0$ such that $\{|y_n[t]|^2\}_{n,t} \leq B_y \leq \infty$,
A2) *Shift-invariant kernels:* the kernels are shift-invariant, i.e., $k(x_i, x_j) = k(x_i - x_j)$.
A3) *Bounded minimum eigenvalue of* $\boldsymbol{\Phi}_n[t]$*:* There exists $\rho_l > 0$ such that $\Lambda_{\min}(\boldsymbol{\Phi}_n[t]) \geq \rho_l,\ \forall t$, where $\Lambda_{\min}(\cdot)$ is the minimum eigenvalue operator.
A4) *Bounded maximum eigenvalue:* there exists $L > 0$ such that $2\Lambda_{\max}(\boldsymbol{\Phi}_n[t]) < L < \infty,\ \forall t$, where $\Lambda_{\max}(\cdot)$ is the maximum eigenvalue operator.

Dynamic regret is defined as the sum of differences between the online estimated cost function and optimal cost function:

$$R_n[T] = \sum_{t=0}^{T-1} \left[h_t^n(\boldsymbol{\alpha}_n[t], z_n[t]) - h_t^n\left(\boldsymbol{\beta}_n^*[t], \boldsymbol{\kappa}_n[t]\right)\right], \quad (28)$$

where $\boldsymbol{\alpha}_n[t]$ collects the estimated RF coefficients [cf. (27)] and $z_n[t]$ is the RF features; and $\boldsymbol{\beta}_n^*[t] \in \mathbb{R}^{(N-1)t}$ and $\boldsymbol{\kappa}_n[t]$ are the optimal coefficients and the kernel-based features in RKHS, respectively. The function $h_t^n(\cdot,\cdot)$ is defined as

$$h_t^n(\mathbf{w}, \mathbf{x}) = \frac{1}{2}\left[y_n[t] - \mathbf{w}^\top \mathbf{x}\right]^2 + \lambda \sum_{n'=1}^{N} \|\mathbf{w}_{n,n'}\|_2, \quad (29)$$

which is related to (11) by replacing the cumulative loss by an instantaneous loss. We also define the optimal RF coefficients $\boldsymbol{\alpha}_n^*[t]$ as

$$\boldsymbol{\alpha}_n^*[t] = \arg\min_{\boldsymbol{\alpha}_n} h_t^n(\boldsymbol{\alpha}_n, z_n[t]). \quad (30)$$

Adding and subtracting $h_t^n(\boldsymbol{\alpha}_n^*[t], z_n[t])$ in (28) gives

$$R_n[T] = \underbrace{\sum_{t=0}^{T-1} \left(h_t^n(\boldsymbol{\alpha}_n[t], z_n[t]) - h_t^n\left(\boldsymbol{\alpha}_n^*[t], z_n[t]\right)\right)}_{R_n^{\text{rf}}[T]}$$
$$+ \underbrace{\sum_{t=0}^{T-1} \left(h_t^n(\boldsymbol{\alpha}_n^*[t], z_n[t]) - h_t^n\left(\boldsymbol{\beta}_n^*[t], \boldsymbol{\kappa}_n[t]\right)\right)}_{\xi_n[T]}, \quad (31)$$

where $R_n^{\text{rf}}[T]$ is the regret w.r.t. optimal cost in RF space and $\xi_n[T]$ is the cumulative error in RF approximation. Dynamic regret can be bounded by bounding $R_n^{\text{rf}}[T]$ and $\xi_n[T]$.

Theorem 1: Under assumptions A1, A2, A3, and A4, the dynamic regret $R_n(T)$ satisfies

$$R_n(T) \leq \left(\left(1 + \frac{L}{2\rho_l}\right)\sqrt{2(N-1)DB_y} + \lambda\sqrt{N-1}\right)$$
$$\times\, T\left(q^{(P+C)}\|\boldsymbol{\alpha}_n^*[0]\|_2 + q^{(P+C)}d + q^{(P+C+1)}l\right) + \epsilon\eta L_h T,$$

where $\eta > 0$ is a constant, L_h is the Lipschitz continuity parameter of function $h_t^n(\cdot,\cdot)$, d is the maximum temporal variation in the optimal solution $\|\boldsymbol{\alpha}_n^*[t] - \boldsymbol{\alpha}_n^*[t-1]\|_2$, and l is the maximum error in the optimal prediction $\|\boldsymbol{\alpha}_n^*[t] - \boldsymbol{\alpha}_n^{pr*}[t]\|_2$ with $\boldsymbol{\alpha}_n^{pr*}[t]$ the optimum prediction at time t. The quantity $q \in (0, 1)$ is the contraction coefficient, and its value for various optimization techniques is provided in [29].

Proof: See Appendix.

The dynamic regret bound in Theorem 1 is linear in time, which implies that $\lim_{t\to\infty} R_n(T)/T = constant$, where *constant* is the steady state error, which depends on $l = \|\boldsymbol{\alpha}_n^*[t] - \boldsymbol{\alpha}_n^{pr*}[t]\|_2$, $d = \|\boldsymbol{\alpha}_n^*[t] - \boldsymbol{\alpha}_n^*[t-1]\|_2$, and the constant $\epsilon \geq 0$. Having a linear variation of the dynamic regret is a favourable attribute of the proposed online algorithm, since it implies that asymptotically, the solution will converge to the optimal online solution, but with steady state errors. Further, if d and l are low (slowly varying systems), it is possible to have a very low bound for the asymptotic $R_n(T)/T$ by controlling ϵ. The constant ϵ is inversely related to the number of RF features [30], meaning that setting ϵ to zero requires an infinite number of RF features. Hence, ϵ is controlled in the expense of model complexity.

V. NUMERICAL EXPERIMENTS

This section compares the proposed algorithm with competing alternatives using both synthetic data from Erdös-Rényi graph models and real data from epileptic seizure and financial time series. We compare the proposed approach with the following alternatives:

- *Pro-SEM:* the time-unstructured linear time-varying SEM from [12], based on a proximal online gradient framework;
- *TV-SEM:* the time-structured linear time-varying SEM from [31];
- *MSTO:* A nonlinear SEM by merely performing a one-step multidimensional shrinkage thresholding [cf. (21)] without any prediction-correction steps.

The first two alternatives are considered as baselines as they have also shown superior performance to other online learning strategies in the respective papers. Instead, the third

alternative is considered to highlight the importance of the proposed time-structured strategy.

In all experiments, the proposed algorithm has one-step prediction ($P = 1$) and one-step correction ($C = 1$). Wherever the SEM topologies are plotted for visualization, we use the normalized ℓ_2 norms of the RF coefficients as the topology estimates, defined as $\widehat{b}_{n,n'}[t] := \|\boldsymbol{\alpha}_{n,n'}[t]\|_2/(\max_m \|\boldsymbol{\alpha}_{n,m}[t]\|_2)$. The per node computational complexity of Pro-SEM, TV-SEM, MSTO, and the proposed method are in the order of $\mathcal{O}(N^2)$, $\mathcal{O}(N^3)$, $\mathcal{O}(N^2D^2)$, and $\mathcal{O}(N^2D^2)$, respectively. Compared to the proposed algorithm, MSTO has the same computational complexity, whereas the linear Pro-SEM is computationally lighter. TV-SEM's computational complexity increases cubically, while that of the proposed method increases quadratically, which is favourable when scaling to large networks.

A. SYNTHETIC DATA

In this experiment, we consider simulated data from a slowly-varying SEM model. We generate graph-connected time series using the following nonlinear SEM model:

$$\mathbf{y}[t] = 0.1(\mathbf{I} - \mathbf{W}[t])^{-1}\mathbf{u}[t] + 0.1\sin((\mathbf{I} - \mathbf{W}[t])^{-1}\mathbf{u}[t]), \tag{32}$$

where $\mathbf{y}[t] \in \mathbb{R}^5$ is the signal at time t, $\mathbf{u}[t] \sim \mathcal{N}(0, 0.1)$, $\mathbf{I} \in \mathbb{R}^{5\times5}$ is the identity matrix, and the operator $\sin(\cdot)$ acts element-wise to introduce non-linearities. The matrix $\mathbf{W}[t] \in \mathbb{R}^{5\times5}$ is constructed such that it attributes slowly-evolving model dynamics to (32), and is of the form:

$$\mathbf{W}[t+1] = \mathbf{W}[t] + 0.001\sin(0.01t)\mathbf{W}[t], \tag{33}$$

where $\mathbf{W}[0] \in \mathbb{R}^{5\times5}$ is constructed using an Erdős-Rényi random graph with diagonal entries zero.[1]

Our synthetic data set consists of 100 multi-variate time series, generated using (32), each having $T = 5000$ signal samples. Out of the 100 multi-variate time series, 20 are used to tune the hyperparameter of all the algorithms based on a grid search for the best model fitness. The model fitness is measured via *Mean Squared Error* (MSE), defined as

$$\text{MSE}[T] = \frac{\sum_{t=0}^{T-1} \|\mathbf{y}[t] - \hat{\mathbf{y}}[t]\|_2^2}{NT}, \tag{34}$$

where $\hat{\mathbf{y}}[t] \in \mathbb{R}^5$ is the signal estimated using the learned SEM model. The hyperparameter values of the proposed algorithm are $(\sigma_n, \lambda, \gamma, \nu_t) = (5, 0.0009, 0.98, 2/\max\{\Lambda_{\max}(\boldsymbol{\Phi}_n[t])\}_n)$ and the RF count is $D = 5$. The MSEs averaged across the remaining 80 multi-variate time series are plotted in Fig. 2, which shows that the proposed method outperforms all alternatives. This is because the alternatives do not exploit the evolution of the model or cannot learn non-linearities, whereas the proposed algorithm features both.

Dynamic Regret: In Fig. 3, we plot the rate of change of the dynamic regret w.r.t. optimal cost function in RF space $R_n^{\text{rf}}[T]/T$. The convergence of $R^{\text{rf}}[T]/T$ is evident from Fig. 3, which supports our theoretical analysis in Theorem 1. We wish to note that a numerical evaluation of the second component of the dynamic regret $\xi_n[T]$ is a daunting, complex process since it involves finding the optimal parameters in a high dimensional RKHS. However, $\xi_n[T]/T$ is upper bounded by the value $\epsilon\eta L_h$ [cf. Lemma 2], where ϵ is a user-controlled parameter. By setting ϵ to be very small, the rate of change of the overall dynamic regret $R_n[T]/T$ can be made closer to $R^{\text{rf}}[T]/T$, when $T \to \infty$.

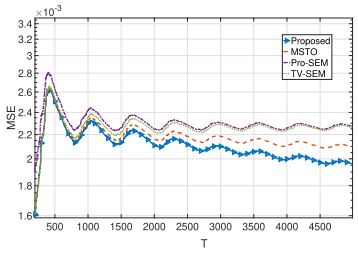

FIGURE 2. MSE comparison on the synthetic data set.

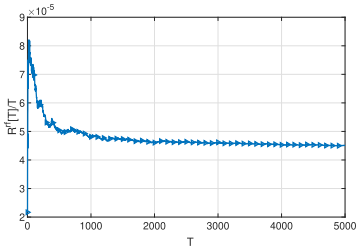

FIGURE 3. Dynamic regret in RF space.

B. REAL DATA: EPILEPTIC SEIZURE

In this experiment, we examine the functional connectivities among different brain regions via learned SEM topologies using an EEG dataset. Our goal is to distinguish between the normal and epileptic dynamics in the brain networks. We use an EEG dataset of children with intractable seizures collected from the Children's Hospital, Boston [32]. The data set consists of multivariate time series of potential differences between electrodes inserted in the brain. There are a total of 23 times series measuring EEG activities in different brain regions. We fit this data using different algorithms and test their capability to distinguish the pre-seizure and the seizure events. We measure the performance via the *Maximum Mean Discrepancy* (MMD) of the distribution of nodal degrees, which is a standard approach used to measure the distance between two graphs [33], [34]. The MMD is defined as

$$\text{MMD}^2(p_1||p_2) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_1}\left[k(\mathbf{x},\mathbf{y})\right] + \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_2}\left[k(\mathbf{x},\mathbf{y})\right] - 2\mathbb{E}_{\mathbf{x}\sim p_1,\mathbf{y}\sim p_2}\left[k(\mathbf{x},\mathbf{y})\right] \tag{35}$$

where $k(\mathbf{x}, \mathbf{y})$ is the radial basis kernel function computing the distance between $\mathbf{x}$ and $\mathbf{y}$, and MMD^2 measures the distance between distributions p_1 and p_2. In this experiment, p_1 and p_2

[1] We choose a small Erdős-Rényi graph of size 5 to corroborate the dynamic regret, which involves high computational complexity.

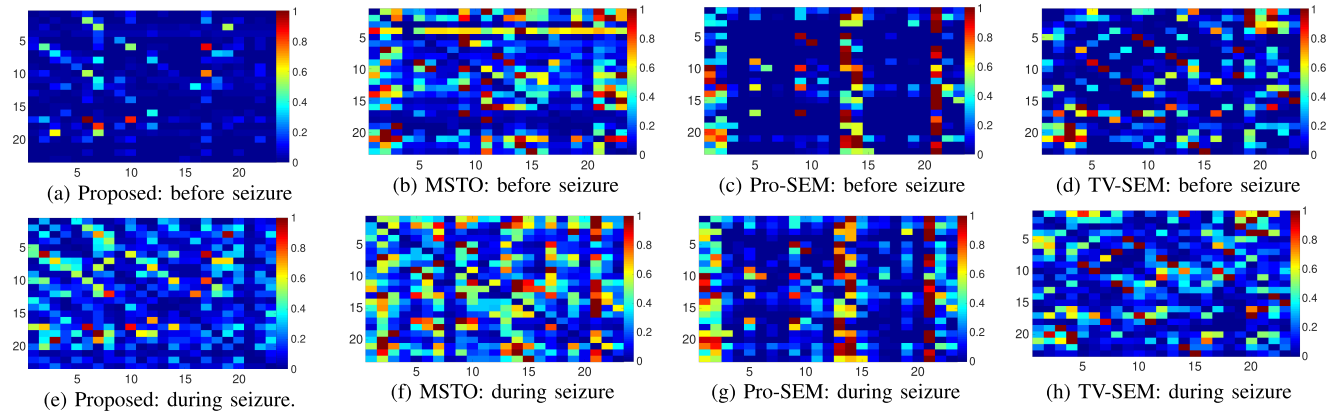

FIGURE 4. Snapshots of estimated topologies.

TABLE 1. Maximum Mean Discrepancy for Node Degree on EEG Data

MMD	S_1	S_2
Proposed	**0.0532**	**0.0550**
TV-SEM	0.0037	0.0038
Pro-SEM	0.0032	0.0013
MSTO	0.0067	0.0070

The bold value signifies our outperforms the competitors.

correspond to the distributions of nodal degrees for the pre-seizure and seizure events, respectively.

We used the proposed method with the RF count $D = 5$ along with the hyperparameters $(\sigma_n, \lambda, \gamma, \nu_t) = (1, 0.1, .98, 2/\max\{\Lambda_{\max}(\mathbf{\Phi}_n[t])\})$, obtained using a grid search for the best MMD. The hyperparameters of other algorithms are also tuned using the same strategy.

Table 1 compares the MMD of the different algorithms using the seizure data from two subjects, S_1 and S_2. The MMD of the proposed algorithm is an order-one magnitude higher compared to alternatives, which highlights that the proposed algorithm distinguishes the seizure and the pre-seizure events better. This is due to the fact that the functional connectivities in brain are highly nonlinear [15], and all alternatives, except MSTO, discard the nonlinear components in the connectivity. MSTO, on the other hand, can accommodate the non-linearities; however, it does not take advantage of the brain connectivity evolution, and is at the second place in the comparison.

A snapshot of the estimated graph topology before seizure and after seizure is shown in Fig. 4(a) and (e), respectively. Before the seizure, the connections are concentrated across certain regions, and during the seizure, they get more disrupted, which agrees with the observations in [35]. The reason for the disrupted topology is the increase in pathogenic neural discharge during seizure [36].

We further compare the per-node computational complexity of the proposed method and the time-structured benchmark

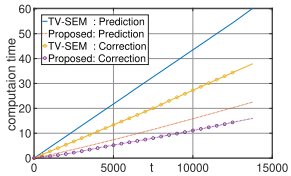

FIGURE 5. Comparison of cumulative computational time on epileptic data.

TABLE 2. Categorized List of Financial Times Series

Groups	Stocks
Group-1	Delta Air Lines (DAL), Air Canada (AC), Air France (AF), (Airlines) United Airlines (UAL), American Airlines (AAL).
Group-2 (Oil)	British Petroleum (BP), ConocoPhillips (COP), Chevron (CVX), Shell (SHEL), ExxonMobil (XOM).
Group-3 (Crypto)	Bitcoin (BTC), Dogecoin (DOGE), Ripple (XRP), Cardano (ADA), Ethereum (ETH).

TV-SEM. The experiment is conducted in a machine with specifications: 2.4 GHz 8-core Intel Core *i*9 and 16 *GB* 2667 *MHz DDR*4 *RAM*. In Fig. 5, we plot the cumulative computation time of the prediction and the correction steps, where it can be observed that the proposed model performs the prediction and the correction much faster. The shorter computation time stems from the node separability feature, which the TV-SEM does not have. The other alternatives are not considered in Fig. 5 since they are time-unstructured algorithms that do not take advantage of the model evolution, and hence, are faster than the time-structured methods.

C. FINANCIAL TIME SERIES

We consider financial time series belonging to three categories: airline industry, oil industry, and cryptocurrency, which are listed in Table 2. The data set includes 15 time series of 879 samples each, which are the closing price values

TABLE 3. Clustering Coefficients of Stock Groups Under COVID and post-COVID Market Dynamics, Computed Using (36)

	Algorithm	*Airlines*	*Oil*	*Crypto*
COVID	Proposed	0.45	**0.54**	**0.54**
	TV-SEM	0.45	0.44	0.45
	Pro-SEM	0.23	0.33	0.00
	MSTO	0.38	0.45	0.40
post-COVID	Proposed	**0.81**	**0.80**	1.00
	TV-SEM	0.60	0.40	0.44
	Pro-SEM	0.20	0.42	1.00
	MSTO	0.63	0.54	1.00

The bold value signifies our algorithm outperforms the competitors.

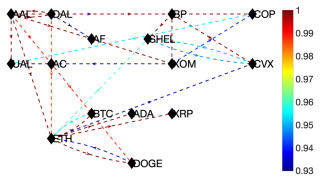

FIGURE 6. Estimated SEM topology on 05-05-2020 (during COVID).

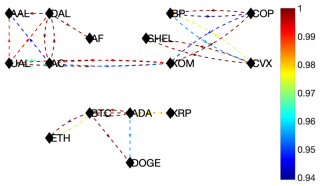

FIGURE 7. Estimated SEM topology on 08-12-2021 (after COVID).

of the stocks from 01-06-2019 to 14-10-2022, including the COVID-19 outbreak. The pandemic had a serious impact on world economy, affecting the natural dynamics of the stock market. A high dip in the S & P 500 index was observed around 25-02-2020 to 25-06-2020, which we mark as the pandemic period. Our goal in this experiment is to identify clusters in the data using the learned SEM topologies and examine the variations in the clusters during and after the pandemic. Since the stock groups in Table 2 are formed by selecting the stocks from similar industries, they are expected to show stronger intra-group dependencies than intergroup dependencies, under the normal market conditions [37].

Let $\mathcal{V}_i = 1, 2, 3$, denote the set of nodes corresponding to the stocks in each group. We measure the performance via the clustering coefficient ρ_i that computes the ratio of the number of edges within group-i to the total number of edges connected to group-i members:

$$\rho_i = \frac{\sum_{n\in\mathcal{V}_i} \mathbb{1}(b_{n,n'} > \delta | n' \in \mathcal{V}_i)}{\sum_{n\in\mathcal{V}_i} \mathbb{1}(b_{n,n'} > \delta) + \sum_{n'\in\mathcal{V}_i} \mathbb{1}(b_{n,n'} > \delta)}, \tag{36}$$

where δ is a threshold selected to consider the strongest $2N$ edges for clustering; and $\mathbb{1}(\cdot)$ is an indicator function defined as $\mathbb{1}(x) = 1$, when x is *true*, and 0, otherwise. A high value of ρ_i indicates that intra-group interactions in group-i are stronger compared to its intergroup interactions. The first 20% of the data samples are used to tune the hyperparameter for the lowest MSE resulting in $(\sigma_n, \lambda, \gamma, \nu_t) = (1, 1, .98, 2/\max\{\Lambda_{\max}(\mathbf{\Phi}_n[t])\})$ and RF count for the experiment is $D = 10$.

Table 3 lists the clustering coefficients of the three groups, averaged across 80 days, randomly sampled from the COVID and post-COVID intervals. As expected, the clustering is more predominant with post-COVID market dynamics than with the COVID market dynamics. The proposed method identifies better such clusters compared with the alternatives. The MSTO algorithm is next in the comparison. This observation is supported by the fact that the interactions among the financial time series are complex [38], which cannot be effectively modeled using the linear Pro-SEM and TV-SEM. It is further interesting to note here as the crypto cluster is much easier identified in the post-COVID period. This follows the intuition that the airline and oil sectors have more financial transactions between them, whereas cryptocurrencies are exchanged only with each other.

Further, the SEM topologies estimated using the proposed algorithm for a COVID-affected market day and a post-COVID day are shown in Figs. 6 and 7, respectively. In line with the expectation, more intra-group market interactions can be observed in Fig. 7, whereas these interactions get disrupted in Fig. 6.

VI. CONCLUSION

This paper proposed an online algorithm to learn the nonlinear structural equation model (SEM), targeting the streaming data from real-world systems with nonlinear dynamics. The proposed method leverages the kernel formulation with random feature approximation to obtain a low-dimensional representation of the nonlinear dynamics. The algorithm uses a prediction-correction strategy equipped with a group-lasso-based optimization framework, solved via composite object mirror descent. Unlike the state-of-the-art algorithms, the proposed method offers data privacy at the network node through node separability and random features. In addition, the proposed online problem is separable across nodes, improving scalability in large graphs. A dynamic regret analysis has been derived to ensure the theoretical guarantee of the algorithm. Using synthetic, epileptic, and financial data, we demonstrated that the SEM topology learned using the proposed model fits the data better and can distinguish between the changes in the system dynamics with less computational complexity compared to the state-of-the-art alternatives. Our future research will involve extending the algorithm by incorporating vector autoregressive (VAR) or structural VAR models, considering time-lagged interactions among the nodes, which are inherent to many real-world networks.

APPENDIX
PROOF OF THEOREM 1

Theorem 1 provides an upper bound for the dynamic regret $R_n(T) = R_n^{\text{rf}}(T) + \xi_n(T)$. We prove the theorem by bounding $R_n^{\text{rf}}(T)$ and $\xi_n(T)$ using the following two lemmas.

Lemma 1: Under assumptions A1, A3, and A4, and letting $\nu_t = \frac{2}{L}$, the dynamic regret w.r.t. the optimal cost function in the RF space is upper bounded by

$$R_n^{\text{rf}}(T) \leq \left(\left(1 + \frac{L}{2\rho_l}\right)\sqrt{2(N-1)DB_y} + \lambda\sqrt{N-1}\right) \times T\left(\|\boldsymbol{\alpha}_n^*[0]\|_2 + q^{(P+C)}d + q^{(P+C+1)}l\right).$$

Proof: The Cauchy-Schwarz inequality allows us to bound $R_n^{\text{rf}}[T]$ by bounding the cumulative optimality gap $\sum_{t=0}^{T-1}\|\boldsymbol{\alpha}_n[t] - \boldsymbol{\alpha}_n^*[t]\|_2$ and the gradient of the loss function $\|\nabla\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])\|_2$ [39].

The bound for optimality gap is given in [9, Prop. 1]:

$$\|\boldsymbol{\alpha}_n[t] - \boldsymbol{\alpha}_n^*[t]\|_2 \leq q^C\left(q^P\|\boldsymbol{\alpha}_n[t-1] - \boldsymbol{\alpha}_n^*[t-1]\|_2 + q^P d + \left(1+q^P\right)l\right) \tag{37}$$

Since $q < 1$, we can express cumulative error in terms of the initial optimal solution $\boldsymbol{\alpha}_n^*[0]$. Setting $\boldsymbol{\alpha}_n[0] = 0$, we bound the cumulative optimality gap as

$$\sum_{t=0}^{T-1}\|\boldsymbol{\alpha}_n[t] - \boldsymbol{\alpha}_n^*[t]\|_2 \leq Tq^{(P+C)}\|\boldsymbol{\alpha}_n^*[0]\|_2 + Tq^{(P+C)}d + Tq^{(P+C+1)}l \tag{38}$$

The gradient of the loss is bounded by following Lemma 1.2 in [19]:

$$\|\nabla\tilde{\ell}_t^n(\boldsymbol{\alpha}_n[t])\|_2 \leq \left(\left(1+\frac{L}{2\rho_l}\right)\sqrt{2(N-1)DB_y} + \lambda\sqrt{N-1}\right) \tag{39}$$

The claim can be then proved by adding (38) and (39). □

Lemma 2: Under assumptions A1 and A2, there exists a constant $\epsilon \geq 0$ such that the cumulative approximation error $\xi_n[T]$ satisfies

$$\xi_n(T) \leq \epsilon\eta L_h T.$$

Proof: The proof follows from [19, Th, 2]. □

REFERENCES

[1] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. IEEE*, vol. 106, no. 5, pp. 787–807, May 2018.

[2] G. Mateos, S. Segarra, A. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019.

[3] A. G. Marques, S. Segarra, and G. Mateos, "Signal processing on directed graphs: The role of edge directionality when processing and learning from network data," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 99–116, Nov. 2020.

[4] A. McLntosh and F. Gonzalez-Lima, "Structural equation modeling and its application to network analysis in functional brain imaging," *Hum. Brain Mapping*, vol. 2, pp. 2–22, 1994.

[5] M. Maxim and A. Ashif, "A new method of measuring stock market manipulation through structural equation modeling (sem)" MPRA Paper 82891, Univ. Library of Munich, Germany, 2017.

[6] Y. Shen, X. Fu, G. B. Giannakis, and N. D. Sidiropoulos, "Topology identification of directed graphs via joint diagonalization of correlation matrices," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 271–283, 2020.

[7] B. Baingana and G. B. Giannakis, "Switched dynamic structural equation models for tracking social network topologies," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2015, pp. 682–686.

[8] A. Simonetto, E. Dall'Anese, S. Paternain, G. Leus, and G. B. Giannakis, "Time-varying convex optimization: Time-structured algorithms and applications," *Proc. IEEE*, vol. 108, no. 11, pp. 2032–2048, Nov. 2020.

[9] A. Natali, E. Isufi, M. Coutino, and G. Leus, "Online graph learning from time-varying structural equation models," in *Proc. IEEE 55th Asilomar Conf. Signals, Syst., Comput.*, 2021, pp. 1579–1585.

[10] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations TrendsR Mach. Learn.*, vol. 4, no. 2, pp. 107–194, 2012.

[11] A. Simonetto and E. Dall'Anese, "Prediction-correction algorithms for time-varying constrained optimization," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5481–5494, Oct. 2017.

[12] B. Zaman, L. M. L. Ramos, and B. Beferull-Lozano, "Dynamic regret analysis for online tracking of time-varying structural equation model topologies," in *Proc. IEEE 15th Conf. Ind. Electron. Appl.*, 2020, pp. 939–944.

[13] A. Natali, E. Isufi, M. Coutino, and G. Leus, "Learning time-varying graphs from online data," *IEEE Open J. Signal Process.*, vol. 3, pp. 212–228, 2022.

[14] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel-Granger causality and the analysis of dynamical networks," *Phys. Rev. E*, vol. 77, 2008, Art. no. 056215.

[15] Y. Shen, G. Giannakis, and B. Baingana, "Nonlinear structural vector autoregressive models with application to directed brain networks," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5325–5339, Oct. 2019.

[16] R. Money, J. Krishnan, and B. Beferull-Lozano, "Online non-linear topology identification from graph-connected time series," in *Proc. IEEE Data Sci. Learn. Workshop*. 2021, pp. 1–6.

[17] Y. Shen, G. Leus, and G. Giannakis, "Online graph-adaptive learning with scalability and privacy," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2471–2483, May 2019.

[18] R. Money, J. Krishnan, and B. Beferull-Lozano, "Random feature approximation for online nonlinear graph topology identification," in *Proc. IEEE 31st Int. Workshop Mach. Learn. Signal Process.*, 2021, pp. 1–6.

[19] R. Money, J. Krishnan, and B. Beferull-Lozano, "Sparse online learning with kernels using random features for estimating nonlinear dynamic graphs," 2022, *arXiv:10.36227.19210092*.

[20] J. Lu, S. Hoi, J. Wang, P. Zhao, and Z. Liu, "Large scale online kernel learning," *J. Mach. Learn. Res.*, vol. 17, no. 47, pp. 1–43, 2016. [Online]. Available: http://jmlr.org/papers/v17/14-148.html

[21] T. Nguyen, T. Le, H. Bui, and D. Phung, "Large-scale online kernel learning with random feature reparameterization," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2543–2549.

[22] Y. Shen, T. Chen, and G. Giannakis, "Random feature-based online multi-kernel learning in environments with unknown dynamics," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 773–808, Jan. 2019.

[23] Y. Shen, B. Baingana, and G. B. Giannakis, "Kernel-based structural equation models for topology identification of directed networks," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2503–2516, May 2017.

[24] B. Olkopf, R. Herbrich, A. Smola, and R. Williamson, "A generalized representer theorem," *Comput. Learn. Theory*, vol. 42, no. 6, pp. 416–426, 2000.

[25] S. Bochner, *Lectures on Fourier Integrals*. vol. 42. Princeton, NJ, USA: Princeton Univ. Press, 1959.

[26] J. Duchi, S. Shwartz, and A. Tewari, "Composite objective mirror descent," in *Proc. COLT*, 2010, pp. 14–26.

[27] A. T. Puig, A. Wiesel, and A. O. Hero, "A multidimensional shrinkage-thresholding operator," in *Proc. IEEE/SP 15th Workshop Stat. Signal Process.*, 2009, pp. 113–116.

[28] L. Zhang, T. Yang, R. Jin, and Z.-H. Zhou, "Dynamic regret of strongly adaptive methods," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5882–5891.

[29] A. Beck, *First-Order Methods in Optimization*. Philadelphia, PA, USA: Soc. Ind. Appl. Math., 2017.
[30] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1177–1184.
[31] A. Natali, M. Coutino, E. Isufi, and G. Leus, "Online time-varying topology identification via prediction-correction algorithms," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 5400–5404.
[32] A. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment." Ph.D. dissertation, Inst. Technol., Cambridge, MA, USA, 2009.
[33] K. Martinkus, A. Loukas, N. Perraudin, and R. Wattenhofer, "Spectre: Spectral conditioning helps to overcome the expressivity limits of one-shot graph generators," in *Proc. Int. Conf. Mach. Learn.* 2022, pp. 15159–15179. .
[34] R. Liao et al., "Efficient graph generation with graph recurrent attention networks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4255–4265.
[35] Y. Hu, Q. Zhang, R. Li, T. Potter, and Y. Zhang, "Graph-based brain network analysis in epilepsy: An EEG study," in *Proc. IEEE/EMBS 9th Int. Conf. Neural Eng.*, 2019, pp. 130–133.
[36] F. Pittau, F. Fahoum, R. Zelmann, F. Dubeau, and J. Gotman, "Negative BOLD response to interictal epileptic discharges in focal epilepsy," *Brain Topogr*, vol. 26, pp. 627–640, 2013.
[37] M. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton University Press, 2008.
[38] F. Jawadi and W. Barnett, *Nonlinear Model. of Econ. and Financial Time-Ser.* (International Symposia in Economic Theory and Econometrics Series), vol. 20 . Bingley, U.K.: Emerald Group Publishing Ltd., 2020.
[39] R. Dixit, A. Bedi, R. Tripathi, and K. Rajawat, "Online learning with inexact proximal online gradient descent algorithms," *IEEE Trans. Signal Process.*, vol. 67, no. 5, pp. 1338–1352, Mar. 2019.