Title

Nonlinear Structural Vector Autoregressive Models With Application to Directed Brain Networks

Journal

IEEE Transactions on Signal Processing, 67(20)

ISSN

1053-587X

Authors

Shen, Yanning
Giannakis, Georgios B
Baingana, Brian

Publication Date

2019-10-15

DOI

10.1109/tsp.2019.2940122

Peer reviewed

HHS Public Access

Author manuscript

IEEE Trans Signal Process. Author manuscript; available in PMC 2020 October 15.

Published in final edited form as:

IEEE Trans Signal Process. 2019 October 15; 67(20): 5325–5339. doi:10.1109/TSP.2019.2940122.

Nonlinear Structural Vector Autoregressive Models with Application to Directed Brain Networks

Yanning Shen [Member, IEEE],
Dept. of EECS and the Center for Pervasive Communications and Computing at the University of California, Irvine, CA 92697

Georgios B. Giannakis [Fellow, IEEE],
Dept. of ECE and the Digital Technology Center, University of Minnesota

Brian Baingana [Member, IEEE]
NextEra Analytics

Abstract

Structural equation models (SEMs) and vector autoregressive models (VARMs) are two broad families of approaches that have been shown useful in effective brain connectivity studies. While VARMs postulate that a given region of interest in the brain is directionally connected to another one by virtue of time-lagged influences, SEMs assert that directed dependencies arise due to instantaneous effects, and may even be adopted when nodal measurements are not necessarily multivariate time series. To unify these complementary perspectives, linear structural vector autoregressive models (SVARMs) that leverage both instantaneous and time-lagged nodal data have recently been put forth. Albeit simple and tractable, linear SVARMs are quite limited since they are incapable of modeling nonlinear dependencies between neuronal time series. To this end, the overarching goal of the present paper is to considerably broaden the span of linear SVARMs by capturing nonlinearities through kernels, which have recently emerged as a powerful nonlinear modeling framework in canonical machine learning tasks, e.g., regression, classification, and dimensionality reduction. The merits of kernel-based methods are extended here to the task of learning the effective brain connectivity, and an efficient regularized estimator is put forth to leverage the edge sparsity inherent to real-world complex networks. Judicious kernel choice from a preselected dictionary of kernels is also addressed using a data-driven approach. Numerical tests on ECoG data captured through a study on epileptic seizures demonstrate that it is possible to unveil previously unknown directed links between brain regions of interest.

Keywords

Network topology inference; structural vector autoregressive models; nonlinear models

yannings@uci.edu.

This work was done when the authors were with the Dept. of ECE and the Digital Technology Center, University of Minnesota, Minneapolis, MN 55455.

I. Introduction

Several contemporary studies in the neurosciences have converged on the well-accepted view that information processing capabilities of the brain are facilitated by the existence of a complex underlying network; see e.g., [39] for a comprehensive review. The general hope is that understanding the behavior of the brain through the lens of network science will reveal important insights, with an enduring impact on applications in both clinical and cognitive neuroscience.

However, brain networks are not directly observable, and must be inferred from processes observed or measured at nodes. To this end, functional magnetic resonance imaging (fMRI) has emerged as a powerful tool, capable of revealing varying blood oxygenation patterns modulated by brain activity [37]. Other related brain imaging modalities include positron emission tomography (PET), electroencephalography (EEG), and electrocorticography (ECoG), to name just a few. Most state-of-the-art tools for inference of brain connectivity leverage variants of causal and correlational analysis methods, applied to time-series obtained from the imaging modalities [7], [12], [13], [17], [18].

Contemporary brain connectivity analyses fall under two broad categories, namely, *functional connectivity* which pertains to discovery of non-directional pairwise correlations between regions of interest (ROIs), and *effective connectivity* which instead focuses on inference of directional dependencies between them [14]. Granger causality [38], vector autoregressive models (VARMs) [18], structural equation models (SEMs) [34], [36], and dynamic causal modeling (DCM) [15] constitute widely used approaches for effective connectivity studies. VARMs postulate that connected ROIs exert time-lagged dependencies among one another, while SEMs assume instantaneous directed interactions among them. Interestingly, these points of view are unified through the so-termed structural vector autoregressive model (SVARM) [9], which postulates that the spatio-temporal behavior observed in brain imaging data results from both instantaneous and time-lagged interactions between ROIs. It has been shown that SVARMs lead to markedly more flexibility and explanatory power than VARMs and SEMs treated separately, at the expense of increased model complexity [9].

The fundamental appeal of the aforementioned effective connectivity approaches stems from their inherent simplicity, since they adopt linear models. However, this is an oversimplification that is highly motivated by the need for tractability, even though consideration of nonlinear models for directed dependence may lead to more accurate approaches for inference of brain connectivity. In fact, recognizing the limitations associated with linear models, several variants of nonlinear SEMs have been put forth in a number of recent works; see e.g., [16], [19], [23], [24], [26], [29], [51].

For example, [29] and [30] advocate SEMs in which nonlinear dependencies only appear among the so-termed *exogenous* variables. Furthermore, [23] puts forth a hierarchical Bayesian nonlinear modeling approach in which unknown random parameters capture the strength and directions of directed links among variables. Several other studies adopt polynomial SEMs, which offer an immediate extension to classical linear SEMs; see e.g.,

[19], [24], [26], [42], [51]. In all these contemporary approaches, it is assumed that the network connectivity structure is *known a priori,* and developed algorithms only estimate the unknown edge weights. The Bayesian network method proposed in [22] also relies on probabilistic assumptions that are also related to prior information of the structure of the network. However, this is a rather major limitation since such prior information may not be available in practice, especially when dealing with potentially massive networks, e.g., the brain.

Similarly, several variants of nonlinear VARMs have been shown useful in unveiling links that often remain undiscovered by traditional linear models; see e.g., [31]–[33], [47]. More recently, [31] proposed a kernel-based VARM, with nonlinear dependencies among nodes encoded by unknown functions belonging to a reproducing kernel Hilbert space.

Building upon these prior works, the present paper puts forth a novel additive nonlinear VARM to capture dependencies between observed ROI-based time-series, without explicit knowledge of the edge structure. Similar to [43], [44], kernels are advocated as an encompassing framework for nonlinear learning tasks. Note that SVARMs admit an interesting interpretation as SEMs, with instantaneous terms viewed as *endogenous* variables, and time-lagged terms as exogenous variables. Since numerical measurement of external brain stimuli is often impractical, or extremely challenging in conventional experiments, adoption of such a fully-fledged SEM (with both endo- and exogenous inputs) is often impossible with traditional imaging modalities.

A key feature of the novel approach is the premise that edges in the unknown network are sparse, that is, each ROI is linked to only a small subset of all potential ROIs that would constitute a maximally-connected power graph. This sparse edge connectivity has recently motivated the development of efficient regularized estimators, promoting the inference of sparse network adjacency matrices; see e.g., [1], [3], [21], [31], [48], [49] and references therein. Based on these prior works, this paper develops a sparse-regularized kernel-based nonlinear SVARM to estimate the effective brain connectivity from per-ROI time series. Compared with [31], the novel approach incorporates instantaneous variables, turns out to be more computationally efficient, and facilitates a data-driven approach for kernel selection.

The rest of this paper is organized as follows. Section II introduces the conventional SVARM, while Section III puts forth its novel nonlinear variant. Section IV advocates a sparsity-promoting regularized least-squares estimator for topology inference from the nonlinear SVARM, while Section V deals with an approach to learn the kernel that ‘best’ matches the data. Results of extensive numerical tests based on EEG data from an Epilepsy study are presented in Section VI, and pertinent comparisons with linear variants demonstrate the efficacy of the novel approach. Finally, Section VII concludes the paper, and highlights several potential future research directions opened up by this work.

Notation. Bold uppercase (lowercase) letters will denote matrices (column vectors), while operators $(\cdot)^{\mathsf{T}}$, and diag$(\cdot)$ will stand for matrix transposition and diagonal matrices, respectively. The identity matrix will be represented by $\mathbf{I}$, while $\mathbf{0}$ will denote the all-zero

matrix, and their dimensions will be clear from the context. Finally, ℓ_p and Frobenius norms will be denoted by $\|\cdot\|_p$, and $\|\cdot\|_F$, respectively.

II. Preliminaries on Linear SVARMs

Consider a directed network whose topology is unknown, comprising N nodes, each associated with an observable time series $\{y_{it}\}_{t=1}^{T}$ measured over T time-slots, for $i = 1, \ldots, N$. Note that y_{it} denotes the t-th sample of the time series measured at node i. In the context of the brain, each node could represent a ROI, while the per-ROI time series are obtainable from standard imaging modalities, e.g., EEG or fMRI time courses. The network topology or edge structure will be captured by the weighted graph adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, whose (i,j)-th entry a_{ij} is nonzero only if a directed effect exists from region i to region j.

In order to unveil the hidden directed network topology, traditional linear SVARMs postulate that each y_{jt} can be represented as a linear combination of instantaneous measurements at other nodes $\{y_{it}\}_{i \neq j}$, and their time-lagged versions $\{\{y_{i(t-\ell)}\}_{i=1}^{N}\}_{\ell=1}^{L}$ [9]. Specifically, y_{jt} admits the following linear instantaneous plus time-lagged model

$$y_{jt} = \sum_{i \neq j} a_{ji}^{0} y_{it} + \sum_{i=1}^{N} \sum_{\ell=1}^{L} a_{ji}^{\ell} y_{i(t-\ell)} + e_{jt} \quad (1)$$

with a_{ij}^{ℓ} capturing the directed influence of region i upon region j over a lag of ℓ time points, while a_{ij}^{0} encodes the corresponding instantaneous directed relationship between them. The coefficients encode the directed structure of the network, that is, a directed link exists from nodes i to j, i.e. $a_{ij} \neq 0$, only if $a_{ij}^{0} \neq 0$, or if there exists $a_{ij}^{\ell} \neq 0$ for $\ell = 1, \ldots, L$. If $a_{ij}^{0} = 0 \ \forall i,j$, then (1) reduces to classical Granger causality [38]. Similarly, setting $a_{ij}^{\ell} = 0 \ \forall i,j \neq 0$ reduces (1) to a linear SEM with no exogenous inputs [25]. Defining $\mathbf{y}_t := [y_{1t}, \ldots, y_{Nt}]^{\mathsf{T}}$, $\mathbf{e}_t := [e_{1t}, \ldots, e_{Nt}]^{\mathsf{T}}$, and the time-lagged adjacency matrix $\mathbf{A}^{\ell} \in \mathbb{R}^{N \times N}$ with the (i,j)-th entry $[\mathbf{A}^{\ell}]_{ij} := a_{ij}^{\ell}$ one can write (1) in vector form as

$$\mathbf{y}_t = \mathbf{A}^{0}\mathbf{y}_t + \sum_{\ell=1}^{L} \mathbf{A}^{\ell}\mathbf{y}_{t-\ell} + \mathbf{e}_t \quad (2)$$

where $\mathbf{A}^{0}$ has zero diagonal entries $a_{ii}^{0} = 0$ for $i = 1, \ldots, N$.

Given the multivariate time series $\{\mathbf{y}_t\}_{t=1}^{T}$ the goal is to estimate matrices $\{\mathbf{A}^{\ell}\}_{\ell=0}^{L}$, and consequently unveil the hidden network topology. Admittedly, overfitting is a potential risk since L is assumed prescribed. Nevertheless, this can be mitigated via standard order

selection methods that control model complexity, e.g., the Bayesian information criterion [10], or Akaike's information criterion [6].

Knowing which entries of $\mathbf{A}^0$ are nonzero, several approaches have been put forth to estimate their values. Examples are based upon ordinary least-squares [9], and hypothesis tests developed to detect presence or absence of pairwise directed links under prescribed false-alarm rates [38]. Albeit conceptually simple and computationally tractable, the linear SVARM is incapable of capturing nonlinear dependencies inherent to complex networks such as the human brain. To this end, the present paper generalizes the *linear* SVARM in (1) to a *nonlinear* kernel-based SVARM.

It is also worth noting that most real world networks (including the brain) exhibit edge sparsity, the tendency for each node to link with only a few other nodes compared to the maximal $\mathcal{O}(N)$ set of potential connections per node. This means that per j, only a few coefficients $\{a_{ij}^{\ell}\}$ are nonzero. In fact, several recent approaches exploiting edge sparsity have been advocated, leading to more efficient topology estimation; see e.g., [1], [3], [31].

III. From linear to nonlinear SVARMs

To enhance flexibility and accuracy, this section generalizes (1) so that nonlinear directed dependencies can be captured. The most general nonlinear model with both instantaneous (spatial) and time-lagged (temporal) dependencies can be written in multivariate form as $\mathbf{y}_t = \bar{\mathbf{f}}(\mathbf{y}_t, \{\mathbf{y}_{t-\ell}\}_{\ell=1}^{L}) + \mathbf{e}_t$ or, entry-wise as

$$y_{jt} = \bar{f}_j(\mathbf{y}_{-jt}, \{\mathbf{y}_{t-\ell}\}_{\ell=1}^{L}) + e_{jt}, \quad j = 1, \ldots, N \qquad (3)$$

where $\mathbf{y}_{-jt} := [y_{1t}, \ldots, y_{(j-1)t}, y_{(j+1)t}, \ldots, y_{Nt}]^{\mathsf{T}}$ collects all but the j-th nodal observation at time t, $\mathbf{y}_{t-\ell} := [y_{1(t-\ell)}, \ldots, y_{N(t-\ell)}]^{\mathsf{T}}$ and $\bar{f}_j(\cdot)$ denotes a nonlinear function of its multivariate argument. With limited (NT) data available, $\bar{f}_j$ in (3) entails $(L+1)N-1$ variables. This fact motivates simpler model functions to cope with the emerging 'curse of dimensionality' in estimating $\{\bar{f}_j\}_{j=1}^{N}$. A simplified form of (3) has been studied in [31] with $L = 1$, and without instantaneous influences $\mathbf{y}_{-jt}$, which have been shown of importance in applications such as brain connectivity [34] and gene regulatory networks [8]. Such a model is simplified compared with (3) because the number of variables of $\bar{f}_j$ reduces to N. Nevertheless, estimating such an N-variate functional model still suffers from the curse of dimensionality, especially when the size of typical networks scales up.

To circumvent this challenge, we further posit that the multivariate function in (3) is separable with respect to each of its $(L+1)N-1$ variables. Such a simplification of (3) amounts to adopting a generalized additive model (GAM) [20, Ch. 9]. In the present context, the GAM adopted is $\bar{f}_j(\mathbf{y}_{-jt}, \{\mathbf{y}_{t-\ell}\}_{\ell=1}^{L}) = \sum_{i\neq j} \bar{f}_{ij}^{0}(y_{it}) + \sum_{i=1}^{N}\sum_{\ell=1}^{L} \bar{f}_{ij}^{\ell}(y_{i(t-\ell)})$, where the nonlinear functions $\{\bar{f}_{ij}^{\ell}\}$ will be specified in the next section. Defining

$\bar{f}_{ij}^{\ell}(y) := a_{ij}^{\ell} f_{ij}^{\ell}(y)$, the node j observation at time t is a result of both instantaneous and multi-lag effects; that is [cf. (1)]

$$y_{jt} = \sum_{i \neq j} a_{ij}^{0} f_{ij}^{0}(y_{it}) + \sum_{i=1}^{N} \sum_{\ell=1}^{L} a_{ij}^{\ell} f_{ij}^{\ell}(y_{i(t-\ell)}) + e_{jt} \quad (4)$$

where similar to (1), $\{a_{ij}^{\ell}\}$ define the matrices $\{\mathbf{A}^{\ell}\}_{\ell=0}^{L}$. In order to avoid scaling ambiguity, we $a_{ij}^{\ell} \in \{0, 1\}$ as binary variables indicating the presence of edge, meaning a directed edge from node j to node i exists if the corresponding $a_{ij}^{\ell} \neq 0$ for any $\ell = 0,1, \ldots, L$. Instead of having to estimate an $[(L+1)N-1]$-variate function in (3) or an N-variate function in [31], (4) requires estimating $(L+1)N-1$ *univariate* functions. Note that conventional linear SVARMs in (1) assume that the functions $\{f_{ij}^{\ell}\}$ in (4) are linear, a limitation that the ensuing Section IV will address by resorting to a reproducing kernel Hilbert space (RKHS) formulation to model $\{f_{ij}^{\ell}\}$.

Problem statement.

Given $\{\mathbf{y}_t \in \mathbb{R}^N\}_{t=1}^{T}$, the goal now becomes to find the nonlinear functions $\{f_{ij}^{\ell}\}$, as well as the adjacency matrices $\{\mathbf{A}^{\ell}\}_{\ell=0}^{L}$ in (4).

IV. Kernel-based Sparse SVARMs

Suppose that each univariate function $f_{ij}^{\ell}(\cdot)$ in (4) belongs to the RKHS

$$\mathcal{H}_i^{\ell} := \{f_{ij}^{\ell} | f_{ij}^{\ell}(y) = \sum_{t=1}^{\infty} \beta_{ijt}^{\ell} \kappa_i^{\ell}(y, y_{i(t-\ell)})\} \quad (5)$$

where $\kappa_i^{\ell}(y, \psi): \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a preselected basis (so-termed kernel) function that measures the similarity between y and ψ. Different choices of κ_i^{ℓ} specify their own basis expansion spaces, and the linear functions can be regarded as a special case associated with the linear kernel $\kappa_i^{\ell}(y, \psi) = y\psi$. An alternative popular kernel is the Gaussian one that is given by $\kappa_i^{\ell}(y, \psi) := \exp[-(y-\psi)^2/(2\sigma^2)]$. Defining the inner product as $\left\langle \kappa_i^{\ell}(y, \psi_1), \kappa_i^{\ell}(y, \psi_2) \right\rangle := \sum_{\tau} \kappa_i^{\ell}(y_{\tau}, \psi_1) \kappa_i^{\ell}(y_{\tau}, \psi_2)$, a kernel is reproducing if it satisfies $\left\langle \kappa_i^{\ell}(y, \psi_1), \kappa_i^{\ell}(y, \psi_2) \right\rangle = \kappa_i^{\ell}(\psi_1, \psi_2)$,, which induces the RKHS norm $||f_{ij}^{\ell}||_{\mathcal{H}_i^{\ell}}^2 = \sum_{\tau} \sum_{\tau'} \beta_{ij\tau}^{\ell} \beta_{ij\tau'}^{\ell} \kappa_i^{\ell}(y_{i\tau}, y_{i\tau'})$[50].

Considering the measurements per node j, with functions $f_{ij}^{\ell} \in \mathscr{H}_i^l$, for $i = 1, N$ and $\ell = 0, 1, \ldots, L$, the present paper advocates the following regularized least-squares (LS) estimates of the aforementioned functions obtained as

$$\{\hat{f}_{ij}^{\ell}\} = \arg \min_{\{f_{ij}^{\ell} \in \mathscr{H}_i^{\ell}\}} \frac{1}{2} \sum_{t=1}^{T} \left[y_{jt} - \sum_{i \neq j} a_{ij}^0 f_{ij}^0(y_{it}) - \sum_{i=1}^{N} \sum_{\ell=1}^{L} a_{ij}^{\ell} f_{ij}^{\ell}(y_{it}) \right]^2 + \quad (6)$$
$$\lambda \sum_{i=1}^{N} \sum_{\ell=0}^{L} \Omega \left(\| a_{ij}^{\ell} f_{ij}^{\ell} \|_{\mathscr{H}^{\ell}} \right)$$

where $\Omega(.)$ denotes a regularizing function, which will be specified later. An important result that will be used in the following is the representer theorem [20, p. 169], according to which the optimal solution for each f_{ij}^{ℓ} in (6) is given by

$$\hat{f}_{ij}^{\ell}(y) = \sum_{t=1}^{T} \beta_{ijt}^{\ell} \kappa_i^{\ell}(y, y_{i(t-\ell)}). \quad (7)$$

Although the function spaces in (5) include infinite basis expansions, since the given data are finite, namely T per node, the optimal solution in (7) entails a finite basis expansion. Substituting (7) into (6), and letting $\boldsymbol{\beta}_{ij}^{\ell} := [\beta_{ij1}^{\ell}, \ldots, \beta_{ijT}^{\ell}]^{\mathsf{T}}$, and $\boldsymbol{\alpha}_{ij}^{\ell} := a_{ij}^{\ell} \boldsymbol{\beta}_{ij}^{\ell}$, the functional minimization in (6) boils down to optimizing over vectors $\{\boldsymbol{\alpha}_{ij}^{\ell}\}$ Specifically, (6) can be equivalently written for each j in vector form as

$$\{\hat{\boldsymbol{\alpha}}_{ij}^{\ell}\} = \arg \min_{\{\boldsymbol{\alpha}_{ij}^{\ell}\}} \frac{1}{2} \left\| \mathbf{y}_j - \sum_{i \neq j} \mathbf{K}_i^0 \boldsymbol{\alpha}_{ij}^0 - \sum_{i=1}^{N} \sum_{\ell=1}^{L} \mathbf{K}_i^{\ell} \boldsymbol{\alpha}_{ij}^{\ell} \right\|_2^2 \quad (8)$$
$$+ \lambda \sum_{i=1}^{N} \sum_{\ell=0}^{L} \Omega \left(\sqrt{(\boldsymbol{\alpha}_{ij}^{\ell})^{\mathsf{T}} \mathbf{K}_i^{\ell} \boldsymbol{\alpha}_{ij}^{\ell}} \right)$$

where $\mathbf{y}_j := [y_{j1}, \ldots, y_{jT}]^{\mathsf{T}}$, and the $T \times T$ matrices $\{\mathbf{K}_i^{\ell}\}$ have entries $[\mathbf{K}_i^{\ell}]_{t,\tau} = \kappa_i^{\ell}(y_{it}, y_{i(\tau-\ell)})$. Furthermore, collecting all the observations at different nodes in $\mathbf{Y} := [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{T \times N}$ and letting $\bar{\mathbf{K}}^{\ell} := [\mathbf{K}_1^{\ell} \ldots \mathbf{K}_N^{\ell}]$, (8) can be written as

$$\{\boldsymbol{\alpha}_{ij}^{\ell}\} = \arg \min_{\boldsymbol{\alpha}_{ii}^{0}=0,\{\boldsymbol{\alpha}_{ij}^{\ell}\}} \frac{1}{2}\left\|\mathbf{Y} - \sum_{l=1}^{L} \overline{\mathbf{K}}^{\ell}\mathbf{W}_{\boldsymbol{\alpha}}^{\ell}\right\|_F^2 \quad (9)$$
$$+ \lambda \sum_{j=1}^{N}\sum_{i=1}^{N}\sum_{\ell=0}^{L} \Omega\left(\sqrt{(\boldsymbol{\alpha}_{ij}^{\ell})^{\mathsf{T}}\mathbf{K}_i^{\ell}\boldsymbol{\alpha}_{ij}^{\ell}}\right)$$

where the $NT \times N$ block matrix

$$\mathbf{W}_{\alpha}^{\ell} := \begin{bmatrix} \boldsymbol{\alpha}_{11}^{\ell} & \cdots & \boldsymbol{\alpha}_{1N}^{\ell} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\alpha}_{N1}^{\ell} & \cdots & \boldsymbol{\alpha}_{NN}^{\ell} \end{bmatrix} \quad (10)$$

exhibits a structure 'modulated' by the entries of $\mathbf{A}^{\ell}$. For instance, if $a_{ij}^{\ell} = 0$, then $\boldsymbol{\alpha}_{ij:}^{\ell} = a_{ij}^{\ell}\boldsymbol{\beta}_{ij}^{\ell}$ is an all-zero block, irrespective of the values taken by $\boldsymbol{\beta}_{ij}^{\ell}$.

Instead of the LS cost used in (6) and (9), alternative loss functions could be employed to promote robustness using the ϵ-insensitive, or, the ℓ_1-error norm; see e.g., [20, Ch. 12]. Regarding the regularizing function $\Omega(\cdot)$, typical choices are $\Omega(z) = |z|$, or, $\Omega(z) = z^2$. The former is known to promote sparsity of edges, which is prevalent to most networks; see e.g., [39]. In principle, leveraging such prior knowledge naturally leads to more efficient topology estimators, since $\{\mathbf{A}^{\ell}\}$ are promoted to have only a few nonzero entries. The sparse nature of $\mathbf{A}^{\ell}$manifests itself as *block sparsity* in $\mathbf{W}_{\alpha}^{\ell}$. Specifically, using $\Omega(z) = |z|$, one obtains the following estimator of the coefficient vectors $\{\boldsymbol{\alpha}_{ij}^{\ell}\}$

$$\{\hat{\boldsymbol{\alpha}}_{ij}^{\ell}\} = \arg \min_{\hat{\boldsymbol{\alpha}}_{ii}^{0}=0,\{\boldsymbol{\alpha}_{ij}^{\ell}\}} \frac{1}{2}\left\|\mathbf{Y} - \sum_{l=1}^{L} \overline{\mathbf{K}}^{\ell}\mathbf{W}_{\boldsymbol{\alpha}}^{\ell}\right\|_F^2 \quad (11)$$
$$+ \lambda \sum_{\ell=0}^{L}\sum_{j=1}^{N}\sum_{i=1}^{N} \sqrt{(\boldsymbol{\alpha}_{ij}^{\ell})^{\mathsf{T}}\mathbf{K}_i^{\ell}\boldsymbol{\alpha}_{ij}^{\ell}}.$$

Recognizing that summands in the regularization term of (11) can be written as $\sqrt{(\boldsymbol{\alpha}_{ij}^{\ell})^{\mathsf{T}}\mathbf{K}_i^{\ell}\boldsymbol{\alpha}_{ij}^{\ell}} = \|(\mathbf{K}_i^{\ell})^{1/2}\boldsymbol{\alpha}_{ij}^{\ell}\|_2$, which is the weighted ℓ_2-norm of $\boldsymbol{\alpha}_{i,j}$ the entire regularizer can henceforth be regarded as the weighted $\ell_{2,1}$-norm of $\mathbf{W}_{\alpha}^{\ell}$, that is known to be useful for promoting block sparsity. It is clear that (11) is a strongly convex problem, which admits a globally optimal solution. In fact, the problem structure of (11) lends itself naturally to efficient iterative *proximal optimization* methods e.g., proximal gradient descent iterations [5, Ch. 7], or, the alternating direction method of multipliers (ADMM) [41].

For a more detailed description of algorithmic approaches adopted to unveil the hidden topology by solving (11), the reader is referred to Appendix A. All in all, Algorithm 1 is a summary of the novel iterative solver of (11) derived based on ADMM iterations. Per iteration, the complexity of ADMM is in the order of $\mathcal{O}(T^2NL)$ which is linear in the network size N. A couple of remarks are now in order.

Remark 1: Selecting $\Omega(z) = z^2$ is known to control model complexity, and thus prevent overfitting [20, Ch. 3]. Let $\mathbf{D}^\ell := \text{Bdiag}(\mathbf{K}_1^\ell \ldots \mathbf{K}_N^\ell)$, and $\mathbf{D}$:=Bdiag($\mathbf{D}^0 \ldots \mathbf{D}^L$), where Bdiag$(\cdot)$ is a block diagonal of its matrix arguments. Substituting $\Omega(z) = z^2$ into (9), one obtains

$$\{\hat{\boldsymbol{\alpha}}_{ij}^\ell\} = \arg \min_{\boldsymbol{\alpha}_{ii}^0 = 0, \{\boldsymbol{\alpha}_{ij}^\ell\}} \frac{1}{2}\left\|\mathbf{Y} - \overline{\mathbf{K}}\mathbf{W}_\alpha\right\|_F^2 + \lambda \ \text{trace}(\mathbf{W}_\alpha^\mathsf{T}\mathbf{D}\mathbf{W}_\alpha) \quad (12)$$

where $\overline{\mathbf{K}} := [\overline{\mathbf{K}}^0 \ldots \overline{\mathbf{K}}^L]$, and $\mathbf{W}_\alpha := [(\mathbf{W}_\alpha^0)^\mathsf{T} \ldots (\mathbf{W}_\alpha^L)^\mathsf{T}]^\mathsf{T}$. Problem (12) is convex and can be solved in closed form as

$$\hat{\bar{\boldsymbol{\alpha}}}_j = (\overline{\mathbf{K}}_j^\mathsf{T}\overline{\mathbf{K}}_j + 2\mathbf{D}_j)^{-1}\overline{\mathbf{K}}_j^\mathsf{T}\mathbf{y}_j \quad (13)$$

where $\bar{\boldsymbol{\alpha}}_j$ denotes the $(NL-1)T \times 1$ vector obtained after removing entries of the j-th column of $\mathbf{W}_\alpha$ indexed by $\mathcal{J}_j := \{(j-1)T+1, \ldots, jT\}$; $\overline{\mathbf{K}}_j$ collects columns of $\overline{\mathbf{K}}$ excluding the columns indexed by $\mathcal{J}_j$; and the block-diagonal matrix $\mathbf{D}_j$ is obtained after eliminating rows and columns of $\mathbf{D}$ indexed by $\mathcal{J}_j$. Using the matrix inversion lemma, the complexity of solving (13) is in the order of $\mathcal{O}(T^3NL)$.

Remark 2: Relying on an operator kernel (OK), the approach in [31] offers a more general nonlinear VARM (but not SVARM) than the one adopted here. However, [31] did not account for instantaneous or the multiple-lagged effects. Meanwhile, estimating $\bar{\mathrm{f}}(\mathbf{y}_{t-1})$ in [31] does not scale well as the size of the network (N) increases. Also OK-VARM is approximated in [31] using the Jacobian, which again adds to the complexity of the algorithm, and may degrade the generality of the proposed model. Finally, the model in [31] is limited in its ability to incorporate the structure of the network (e.g., edge sparsity). In order to incorporate prior information on the model structure, [31] ends up solving a nonconvex problem, which might experience local minima, and the flexibility in choosing kernel functions will also be sacrificed. In contrast, our approach entails a natural extension to a data-driven kernel selection, which will be outlined in the next section.

Remark 3: Estimation of a nonlinear function generally requires a large number of samples (T), consequently incurring increased complexity. Clearly, the proposed method is prone to

scalability issues when dealing with even moderately-sized networks. This motivates solving the kernel-based optimization problem "on the fly,"; see also [40], [46] for recent examples. However, deriving such efficient and real-time solvers is beyond the scope of the present paper, whose focus is the novel nonlinear modeling framework. Only batch algorithms have been presented, but pursuit of more efficient online algorithms constitutes an important future direction.

Remark 4: Note that the sparsity prior in the present work is considered separately for each $\mathbf{A}^{\ell}$. This results in the block sparsity of $\mathbf{W}_{\alpha}^{\ell}$, see (10). This can be easily adapt to the case where more prior information is available with respect to the structure of the network. For example, in case when $\{\mathbf{A}^{\ell}\}$ share the same sparsity pattern. One could stack the weight matrices in (10), and impose group sparse structure on the corresponding blocks of $\{\mathbf{W}_{\alpha}^{\ell}\}$. Such adaptation will not influence the convexity of the problem in (11), henceforth the resulting problem can still be solved to the global optimum.

Remark 5: Stability of the generative KSVARM depends on the matrices $\{\mathbf{A}^{\ell}\}$, the noise level, the initial condition, as well as the choice of the kernel functions. It is an interesting future research direction, but goes beyond the scope of the present paper.

V. Data-driven kernel selection

Choice of the kernel function determines the associated Hilbert space, and it is therefore of significant importance in estimating the nonlinear functions $\{f_{ij}^{\ell}\}$. Although Section IV assumed that the kernels $\{\kappa_i^{\ell}\}$ are available, this is not the case in general, and this section advocates a data-driven strategy for selecting them. Given a dictionary of reproducing kernels $\{\kappa_p\}_{p=1}^{P}$ it has been shown that any function in the convex hull $\mathscr{K} := \{\kappa | \kappa = \sum_{p=1}^{P} \theta_p \kappa_p, \theta_p \geq 0, \sum_{p=1}^{P} \theta_p = 1\}$ is a reproducing kernel [35], [45]. Therefore, the goal of the present section is to select a kernel from $\mathscr{K}$ that best fits the data. For ease of exposition, consider $\kappa_i^{\ell} = \kappa \in \mathscr{K}$ for all $\ell = 0,1, \ldots, L$ and $i = 1, \ldots, N$ in (6), therefore $\mathscr{H}_i^{\ell} = \mathscr{H}^{(\kappa)}$. Note that the formulation can be readily extended to settings when $\{\kappa_i^{\ell}\}$ are different. Incorporating κ as a variable function in (6) yields

$$\{\hat{f}_{ij}^{\ell}\} = \arg \min_{\kappa \in \mathscr{K}, \{f_{ij}^{\ell} \in \mathscr{H}^{(\kappa)}\}} \frac{1}{2} \sum_{t=1}^{T} \left[y_{jt} - \sum_{i \neq j} a_{ij}^{0} f_{ij}^{0}(y_{it}) - \sum_{i=1}^{N} \sum_{\ell=1}^{L} a_{ij}^{\ell} f_{ij}^{\ell}(y_{it}) \right]^2 + \lambda \sum_{i=1}^{N} \sum_{\ell=0}^{L} \Omega \left(\| a_{ij}^{\ell} f_{ij}^{\ell} \|_{\mathscr{H}^{(\kappa)}} \right) \quad (14)$$

where $\mathscr{H}^{(\kappa)}$ denotes the Hilbert space associated with kernel function κ. With $\mathscr{H}_p$ denoting the RKHS induced by κ_p, it has been shown in [4] and [35] that the optimal $\{\hat{f}_{ij}^{\ell}\}$ in (14) is expressible in a separable form as

$$\hat{f}_{ij}^{\ell}(y) := \sum_{p=1}^{P} f_{ij}^{\ell;p}(y) \quad (15)$$

where $f_{ij}^{\ell;p}$ belongs to RKHS $\mathscr{H}_p$, for $p = 1, \ldots, P$. Substituting (15) into (14), one obtains

$$\{\hat{f}_{ij}^{\ell}\} = \arg \min_{\{f_{i,j}^{\ell;p} \in \mathscr{H}_p\}} \frac{1}{2} \sum_{t=1}^{T} \left[y_{jt} - \sum_{i \neq j} \sum_{p=1}^{P} a_{ij}^{0} f_{ij}^{0,p}(y_{it}) - \sum_{i=1}^{N} \sum_{\ell=1}^{L} \sum_{p=1}^{P} a_{ij}^{\ell} f_{ij}^{\ell;p}(y_{it}) \right]^2 + \lambda \sum_{i=1}^{N} \sum_{\ell=0}^{L} \sum_{p=1}^{P} \Omega\left(\| a_{ij}^{\ell} f_{ij}^{\ell;p} \|_{\mathscr{H}_p} \right). \quad (16)$$

Note that (16) and (6) have similar structure, and their only difference pertains to an extra summation over P candidate kernels. Hence, (16) can be solved in an efficient manner along the lines of the iterative solver of (6) listed under Algorithm 1 [cf. the discussion in Section IV]. Further details of the solution are omitted due to space limitations.

Remark 6: Note that $\{\theta_p\}$ does not show up in the optimization problem in (16), since all the coefficients can be readily absorbed into the nonlinear functions without affecting optimality. This is a consequence of the property that given any function belonging to a prescribed RKHS, all its scaled versions belong to the same RKHS [c.f. (5)]; see also [35] for a detailed proof.

VI. Numerical Tests

This section presents results from numerical tests conducted on both synthetic and real data to corroborate the effectiveness of the proposed approach. Simulated data were generated via a different model in order to assess the impact of the presence of nonlinear dependencies. Tests on real data were based on seizure experiments captured from a number of subjects.

A. Synthetic data tests

Data generation.—Setting $L = 1$, synthetic data were generated via a random 20-node ($N = 20$) Erdös-Rényi graph, with edge probability 0.4. The resulting graph was encoded as a binary 20×20 adjacency matrix. Using this graph, simulated data were generated via both linear and nonlinear models. After drawing vectors $y_{n,1} \sim \mathcal{N}(0, 1)$ and matrices $\{\mathbf{A}^l\}$, the

output data $\{y_t\}_{t=2}^{T}$ were generated recursively. Furthermore, matrices $\{\mathbf{K}_m^{\ell}\}$ were generated using prescribed kernels, that is, entry (i, j) of $\mathbf{K}_m$ was set to $[\mathbf{K}_l^{\ell}]_{ij} = \kappa(y_{it}, y_{jt})$ where the kernel function $\kappa(\cdot, \cdot)$ is known a priori. Entries of coefficient vectors $\boldsymbol{\alpha}_{ij} \in \mathbb{R}^T$ were drawn independently from $\mathcal{N}(0, 1)$, while noise terms were generated i.i.d. as $e_{i1} \sim \mathcal{N}(0, \sigma_e^2)$.

Experiments were run for different values of T, with the edge detection threshold τ selected in each setting to obtain the lowest edge identification error rate (EIER), defined as

$$\text{EIER} := \frac{\|\mathbf{A} - \widehat{\mathbf{A}}\|_0}{N(N-1)} \times 100\% . \quad (17)$$

The operator $\|\cdot\|_0$ denotes the number of nonzero entries of its argument. For all experiments, error plots were generated with values of EIER averaged over 100 independent runs.

Test results.—Figures 2 and 3 depict EIER values plotted against the measurement ratio (T/N) under varying signal-to-noise ratios (SNR := 10 log(P)$_{\text{signal}}$/P_{noise}), where P_{signal} denotes the variance of y_{j1}) for polynomial and Gaussian kernels, respectively. The synthetic graph was generated with edge probability $p = 0.3$. Figure 2 plots the EIER when data are generated by (4), using a polynomial kernel of order $P = 2$, and Figure 3 plots the error performance realized with data generated via a Gaussian kernel with bandwidth $\sigma^2 = 1$. It is clear that adoption of nonlinear SVARMs yields markedly better performance than topology inference approaches based on linear SVARMs, which corroborates the effectiveness of the proposed algorithm in identifying the network topology when the dependencies among nodes are nonlinear. It is also worth observing that as SNR decreases, the performance of the proposed algorithm deteriorates slightly, but can still yield much better performance than the linear approach.

Figure 5 plots heatmaps of actual and inferred adjacency matrices, under varying modeling assumptions. Plots of inferred adjacency matrices are based on a single realization of $T = 40$ samples, with white entries representing presence of an edge, that is, $a_{ij}^{\ell} \neq 0$. As shown by the plots, accounting for nonlinearities yields more accurate recovery of the unknown network topology. Figure 4 plots EIER values against the measurement ratio (T/N) when the underlying Erdös-Rényi graphs are generated using several values of the edge probability $p \in \{0.1, 0.5, 0.9\}$, with λ and η optimally chosen for each scenario. For each value of p, the nonlinear approach outperforms the linear variant, since nonlinear dependencies are accounted for.

In order to assess edge detection performance, receiver operating characteristic (ROC) curves are plotted under different modeling assumptions in Figure 6. With P_D denoting the probability of detection, and P_{FA} the probability of false alarms, each point on the ROC corresponds to a pair (P_{FA}, P_D) for a prescribed threshold. Figure 6 (a) results from tests run on data generated by Gaussian kernels with $\sigma^2 = 1$, while Figure 6 (b) corresponds to

polynomial kernels of order $P = 2$. Using the *area under the curve* (AUC) as the edge-detection performance criterion, Figures 6 (a) and (b) clearly emphasize the benefits of accounting for nonlinearities. In both plots, kernel-based approaches result in the higher AUC metrics than approaches resorting to linear SVARMs. Moreover, the kernel based SVARM outperforms the kernel based VARM which does not take into account the instantaneous effects.

Figure 6 (c) plots ROC curves based on linear and kernel based SVARMs, with simulated data actually generated using a linear SVARM. The curves are parameterized by the sparsity-control parameter λ. Not surprisingly, kernel-based SVARMs adopting polynomial kernels underperform the linear SVARM, due to the inherent model mismatch. However, the kernel SVARM endowed with a multi-kernel learning scheme (MK-SVARM) is shown to attain comparable performance to the linear SVARM when the prescribed dictionary comprises both linear and polynomial kernels. Figure 7 plots the ROC curves based on linear and kernel based nonlinear SVARMs with $L = 2$. The trend exhibited by these curves is similar to those in Figure 6.

B. Real data tests

This section presents test results on seizure data, captured through experiments conducted in an *epilepsy* study [28]. Epilepsy refers to a chronic neurological condition characterized by recurrent seizures, globally afflicting over 20 million people, and often associated with abnormal neuronal activity within the brain. Diagnosis of the condition sometimes involves comparing EEG or ECoG time series obtained from a patient's brain before and after onset of a seizure. Recent studies have shown increasing interest in analysis of connectivity networks inferred from the neuronal time series, in order to gain more insights about the unknown physiological mechanisms underlying epileptic seizures. In this section, connectivity networks are inferred from the seizure data using the novel approach, and a number of comparative measures are computed from the identified network topologies.

C. Seizure data description

Seizure data were obtained for a 39-year-old female subject with a case of intractable epilepsy at the *University of California, San Francisco (UCSF) Epilepsy Center;* see also [28]. An 8 × 8 subdural electrode grid was implanted into the cortical surface of the subject's brain, and two accompanying electrode strips, each comprising six electrodes (a.k.a., depth electrodes) were implanted deeper into the brain. Over a period of five days, the combined electrode network recorded 76 ECoG time series, consisting of voltage levels measured in a region within close proximity of each electrode.

ECoG epochs containing eight seizures were extracted from the record and analyzed by a specialist. The time series at each electrode were first passed through a bandpass filter, with cut-off frequencies of 1 and 50 Hz, and the so-termed *ictal* onset of each seizure was identified as follows. A board-certified neurophysiologist identified the initial manifestation of rhythmic high-frequency, low-voltage focal activity, which characterizes the onset of a seizure. Samples of data before and after this seizure onset were then extracted from the ECoG time series. The per-electrode time series were then divided into 1s windows, with

0.5s overlaps between consecutive windows, and the average spectral power between 5Hz and 15Hz was computed per window. Finally, power spectra over all electrodes were averaged, and the ictal onset was identified by visual inspection of a dramatic increase (by at least an order of magnitude) in the average power. Two temporal intervals of interest were picked for further analysis, namely, the *preictal* and ictal intervals. The preictal interval is defined as a 10s interval preceding seizure onset, while the ictal interval comprises the 10s immediately afterwards. Further details about data acquisition and pre-processing are provided in [28].

The goal here was to assess whether modeling nonlinearities, and adopting the novel kernel-based approach would yield significant insights pertaining to causal/effective dependencies between brain regions, that linear variants would otherwise fail to capture. Toward this goal, several standard network analysis measures were adopted to characterize the structural properties of the inferred networks.

D. Inferred networks

Prior to running the developed algorithm, 10s intervals were chosen from the preprocessed ECoG data with the sampling rate set to 400Hz, and then divided into 20 successive segments, each comprising 200 data samples over a 0.5s horizon. To illustrate this, suppose the 10s interval starts from $t = 0$s and ends at $t = 10s$, then the first segment comprises samples taken over the interval $[0s, 0.5s]$, the second one would be $[0.5s, 1s]$, and so on. After this segmentation of the time series, directed network topologies were inferred using Algorithm 1 with $L = 1$, based on the 0.5s segments, instead of the entire signal, to ensure that the signal is approximately *stationary* per experiment run. A directed link from electrode i to j was drawn if at least one of the estimates of a_{ij}^{ℓ} turned out to be nonzero.

Upon inference of the 20 networks the data pertaining to each seizure, presence/absence of an edge is established via a *t-test* with $P_{FA} = 0.1$. Inference results were not averaged since different seizures may originate from disparate parts of the brain, leading to non-trivial differences between connectivity patterns among electrodes.

Networks inferred from the preictal and ictal intervals were compared using linear, the kernel-based (K-)SVARMs, and K-SVARM with data-driven kernel selection. The lag lengths were set to $L = 1$ for all cases. For K-SVARM, a Gaussian kernel with $\sigma = 1$ was selected, and with $\rho = 10$, Algorithm 1 with regularization parameter Λ was selected via cross-validation. For the data-driven kernel selection scheme, two candidate kernels were employed, namely, a linear kernel, and a polynomial kernel of order 2.

Figure 8 depicts networks inferred from different algorithms for both preictal and ictal intervals of the time series. The figure illustrates results obtained by the linear SVARM, and the K-SVARM approach with and without kernel selection. Each node in the network is representative of an electrode, and it is depicted as a circle, while the node arrangement is forced to remain consistent across the six visual representations. A cursory inspection of the visual maps reveals significant variations in connectivity patterns between ictal and preictal intervals for both models. Specifically, networks inferred via the K-SVARMs, reveal a global decrease in the number of links emanating from each node, while those inferred via the

linear model depict increases and decreases in links connected to different nodes. Interestingly, the K-SVARM with kernel selection recovered most of the edges inferred by the linear and the K-SVARM using a Gaussian kernel, which implies that both linear and nonlinear interactions may exist in brain networks. Moreover, network topologies inferred via K-SVARM with kernel selection are more similar to those obtained using a single kernel than a linear SVARM, implying that the simple linear model is insufficient to capture the network topology in complex brain networks. However, one is unlikely to gain much insight only by visual inspection of the network topologies. Moreover, to further analyze differences between inferred networks from both models, and to assess the potential benefits gained by adopting the novel scheme, several network topology metrics are computed and compared in the next subsection.

E. Comparison of network metrics

First, in- and out-degree was computed for nodes in each of the inferred networks. Note that the in-degree of a node counts its number of incoming edges, while the out-degree counts the number of out-going edges. The total degree per node sums the in- and out-degrees, and is indicative of how well-connected a given node is. Figure 9 depicts nodes in the network and their total degrees encoded by the radii of circles associated with the nodes. As expected from the previous subsection, Figures 9 (a) and (b) demonstrate that the linear SVARM yields both increases and deceases in the inferred node degree. On the other hand, the nonlinear SVARM leads to a more spatially consistent observation with most nodes exhibiting a smaller degree after the onset of a seizure (see Figures 9 (c) and (d)), which may imply that directed dependencies thin out between regions of the brain once a seizure starts, the same trend is also revealed by the K-SVARM with kernel selection (see Figures 9 (e) and (f)); see also Table I for the number of disappearing and appearing edges, as well as the number of nodes for which the degree increases or decreases.

In order to assess the reachability of brain regions before and after seizure onset, comparisons of the so-termed *average shortest path* lengths were done. Average shortest path of a node computes the average length of shortest paths between the given node and all other nodes; see e.g., [27] for more details. The per-node average shortest path length for each inferred network is depicted in Figure 10, with node radii similarly encoding the computed values. Little variation between preictal and ictal average shortest path length is seen for the linear model (Figures 10 (a) and (b)), while variations are more marked for the K-SVARM, see Figures 10 (c–f). It can be seen that modeling nonlinearities reveals subtle changes in reachability of nodes between preictal and ictal phases.

Figure 11 depicts the *closeness centrality* computed per node in the inferred networks. Closeness centrality measures how reachable a node is from all other nodes, and is generally defined as the reciprocal of the sum of geodesic distances of the node from all other nodes in the network; see also [27]. Once again, Figure 11 depicts a more general decrease in closeness centralities after seizure onset in networks inferred by the nonlinear SVARM, as compared to the linear variant. This empirical result indicates a change in reachability between regions of the brain during an epileptic seizure.

In addition to the local metrics, a number of global measures were computed over entire inferred networks, and pertinent comparisons were drawn between the two phases; see Table II for a summary of the average global measures of the inferred networks from 8 different seizures. Several global metrics were considered, e.g., network density, global clustering coefficient, network diameter, and average number of neighbors.These metrics are obtained by averaging the metrics of networks inferred from 8 seizures.

Network density refers to the number of actual edges divided by the number of potential edges, while the global clustering coefficient is the fraction of connected triplets that form triangles, adjusted by a factor of three to compensate for double counting. On the other hand, network diameter is the length of the longest geodesic, excluding infinity. Table II shows that networks inferred via the K-SVARMs and MKL-SVARMs exhibit lower network cohesion after seizure onset, as captured by network density, global clustering coefficient, and average number of neighbors, while the network diameter increases. These changes provide empirical evidence that the brain network becomes less connected, and diffusion of information is inhibited after the onset of an epileptic seizure.

In addition, note that we do not have access to the ground truth topologies for the brain networks to evaluate the effectiveness of the nonlinear methods. In order to verify the benefit of the nonlinear method, we further carried out classification on the obtained adjacency matrix as well as the nodal feature vectors, including degree, average shortest path length, as well as clustering coefficients of the network nodes. We adopted SVM with 5-fold cross-validation. It can been observed from Table III that extracting nonlinear dependencies can improve the accuracy of classification between ictal and pre-ictal data. Therefore, the novel nonlinear model can help improve the accuracy of disease diagnosis.

VII. Conclusions

This paper put forth a novel nonlinear SVARM framework that leverages kernels to infer effective connectivity networks in the brain. Postulating a generalized additive model with unknown functions to capture the hidden network structure, a novel regularized LS estimator that promotes sparse solutions was advocated. In order to solve the ensuing convex optimization problem, an efficient algorithm that resorts to ADMM iterations was developed, and a data-driven approach was introduced to select the appropriate kernel. Extensive numerical tests were conducted on ECoG seizure data from a study on epilepsy.

In order to assess the utility of the novel approach, several local and global metrics were adopted and computed on networks inferred before and after the onset of a seizure. By observing changes in network behavior that are revealed by standard metrics before and after seizure onset, it is possible identify key structural differences that may be critical to explain the mysteries of epileptic seizures. With this in mind, the paper focused on identifying structural differences in the brain network that could not be captured by the simpler linear model. Interestingly, empirical results support adoption of a nonlinear modeling perspective when analyzing differences in effective brain connectivity for epilepsy patients. Specifically, adopting the novel kernel-based approach revealed more significant differences between the preictal and ictal phases of ECoG time series. For instance, it turned out that some regions

exhibited fewer dependencies, reduced reachability, and weakened information-routing capabilities after the onset of a seizure. Since the kernel-based model includes the linear SVARM as an instance, the conducted experiments suggest that one may gain more insights by adopting the nonlinear model, a conclusion that may yield informative benefits to studies of epilepsy that leverage network science.

This work paves the way for a number of exciting research directions in analysis of brain networks. Although it has been assumed that inferred networks are static, overwhelming evidence suggests that topologies of brain networks are dynamic, and may change over rather short time horizons. Future studies will extend this work to facilitate tracking of dynamic brain networks. Furthermore, the novel approach will be empirically tested on a wider range of neurological illnesses and disorders, and pertinent comparisons will be done to assess the merits of adopting the advocated nonlinear modeling approach.

Acknowledgments

Work in this paper was supported by grants NSF 1901134, 1711471, 1500713 and NIH 1R01GM104975-01.

Biographies

Yanning Shen joined the Department of Electrical Engineering and Computer Science as an assistant professor in August 2019. Her research interests span the areas of machine learning, network science, data science, and statistical-signal processing.

She received her B.Sc. and M.Sc. degrees in Electrical Engineering from the University of Electronic Science and Technology of China (UESTC), in 2011 and 2014, respectively, and her Ph.D degree from the University of Minnesota (UMN) in 2019. She received the UESTC distinguished B.Sc. thesis Award in 2011, and distinguished M.Sc. Thesis Award in 2014. She was a Best Student Paper Award finalist of the 2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, and the 2017 Asilomar Conference on Signals, Systems, and Computers. She was selected as Rising Stars in EECS by Stanford University in 2017, and received the UMN Doctoral Dissertation Fellowship in 2018.

Brian Baingana (M’16) received the B.Sc. degree in Electrical Engineering from Makerere University, Kampala, Uganda, in 2007, and the M.Sc. and Ph.D. degrees in Electrical

Engineering from the University of Minnesota, Minneapolis, MN, USA, in 2013 and 2016, respectively. His broader research interests include data and network science, and statistical learning, with applications in social networks, brain connectivity studies, and power grids. His recent interests have focused on developing real-time online algorithms for inference and learning over large scale, time-varying networks.

Before his graduate studies, he was a Software Systems Engineer with MTN Uganda, a telecommunications operator. He currently works with WindLogics, a subsidiary of NextEra Energy Resources, as a Senior Data Scientist. He received the prestigious university-wide Graduate School Fellowship for the academic years 2009-2012, and a Paper Award at the 2015 Global Conference on Signal and Information Processing.

Georgios B. Giannakis (Fellow'97) received his Diploma in Electrical Engr. from the Ntl. Tech. Univ. of Athens, Greece, 1981. From 1982 to 1986 he was with the Univ. of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engr., 1986. He was with the University of Virginia from 1987 to 1998, and since 1999 he has been a professor with the Univ. of Minnesota, where he holds an Endowed Chair in Wireless Telecommunications, a University of Minnesota McKnight Presidential Chair in ECE, and serves as director of the Digital Technology Center.

His general interests span the areas of communications, networking and statistical signal processing - subjects on which he has published more than 450 journal papers, 750 conference papers, 25 book chapters, two edited books and two research monographs (h-index 124). Current research focuses on learning from Big Data, wireless cognitive radios, and network science with applications to social, brain, and power networks with renewables. He is the (co-) inventor of 32 patents issued, and the (co-) recipient of 8 best paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, the G. W. Taylor Award for Distinguished Research from the University of Minnesota, and the IEEE Fourier Technical Field Award (2015). He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.

Appendix

A. Topology Inference via ADMM

Given matrices $\mathbf{Y}$ and $\overline{\mathbf{K}} := [\overline{\mathbf{K}}^0 \ldots \overline{\mathbf{K}}^L]$, this section capitalizes on convexity, and the nature of the additive terms in (11) to develop an efficient topology inference algorithm. Proximal

optimization approaches have recently been shown useful for convex optimization when the cost function comprises the sum of smooth and nonsmooth terms; see e.g., [11]. Prominent among these approaches is the alternating direction method of multipliers (ADMM), upon which the novel algorithm is based; see e.g., [41] for an early application of ADMM to distributed estimation.

Algorithm 1

ADMM for network topology identification

Input: $\mathbf{Y}, \{\{\mathbf{K}_i^{\ell}\}_{i=1}^{N}\}_{\ell=1}^{L}, \tau_{\alpha}, \lambda, \rho$

Initialize: $\mathbf{\Gamma}[0] = 0_{NT \times N}, \mathbf{\Xi}[0] = \mathbf{0}_{NT \times N}, \kappa = 0$

for $\ell = 1, \ldots, L$ **do**

$\mathbf{D}^{\ell} = \text{Bdiag}(\mathbf{K}_1^{\ell}, \ldots, \mathbf{K}_N^{\ell})$

$\bar{\mathbf{K}}^{\ell} = [\mathbf{K}_1^{\ell} \ldots \mathbf{K}_N^{\ell}]$

end for

$\bar{\mathbf{K}} := [\bar{\mathbf{K}}^0 \ldots \bar{\mathbf{K}}^L], \mathbf{D} = \text{Bdiag}(\mathbf{D}^0 \ldots \mathbf{D}^L)$

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

$\mathcal{I}_j := \{(j-1)T + 1, \ldots, jT\}$

$\overline{\mathcal{I}}_j := \{(\kappa, l) | \kappa \notin \mathcal{I}_j, \text{or } l \notin \mathcal{I}_j\}$

$\widetilde{\mathcal{I}}_j := \{(\kappa, l) | l \notin \mathcal{I}_j\}$

$\mathbf{D}_j = [\mathbf{D}]_{\overline{\mathcal{I}}_j}, \overline{\mathbf{K}}_j = [\overline{\mathbf{K}}]_{\widetilde{\mathcal{I}}_j}$

end for

while not converged **do**

for $j = 1, \ldots, N$ (in parallel) **do**

$\mathbf{q}_j[\kappa] = \rho \mathbf{D}_j^{1/2} \gamma_j[\kappa] + \bar{\mathbf{K}}_j^{\mathsf{T}} \mathbf{y}_j - \mathbf{D}_j^{1/2} \boldsymbol{\xi}_j[\kappa]$

$\bar{\boldsymbol{\alpha}}_j[\kappa + 1] = (\overline{\mathbf{K}}_j^{\mathsf{T}} \overline{\mathbf{K}}_j + \rho \mathbf{D}_j)^{-1} \mathbf{q}_j[\kappa]$

$\gamma_{ij}^{\ell}[\kappa] = \mathcal{P}_{\lambda/\rho}((\mathbf{K}_i^{\ell})^{1/2} \boldsymbol{\alpha}_{ij}^{\ell}[\kappa + 1] + \boldsymbol{\xi}_{ij}^{\ell}[\kappa]/\rho),$

for $i = 1, \ldots N, \ell = 0, \ldots L$

end for

$\mathbf{W}_{\alpha}[\kappa + 1] := [(\mathbf{W}_{\alpha}^0)^{\mathsf{T}}[\kappa + 1], \ldots, (\mathbf{W}_{\alpha}^L)^{\mathsf{T}}[\kappa + 1]]^{\mathsf{T}}$

$\mathbf{\Gamma}[\kappa + 1] := [(\mathbf{\Gamma}^0[\kappa + 1])^{\mathsf{T}}, \ldots, (\mathbf{\Gamma}^L[\kappa + 1])^{\mathsf{T}}]^{\mathsf{T}}$

$\mathbf{\Xi}[\kappa + 1] = \mathbf{\Xi}[\kappa] + \rho(\mathbf{D}^{1/2} \mathbf{W}_{\alpha}[\kappa + 1] - \mathbf{\Gamma}[\kappa + 1])$

$\kappa = \kappa + 1$

end while

26: **Edge identification:** (after converging to $\widehat{\boldsymbol{\alpha}}_{ij}^{*}$)

27: $\widehat{a}_{ij}^{*} \neq 0$ if $||\widehat{\boldsymbol{\alpha}}_{ij}^{*}|| \geq \tau_{\alpha,}$ else $\widehat{a}_{ij}^{*} = 0$, $\forall (i, j, \ell)$

28: **return** $\{\widehat{\mathbf{A}}^{\ell}\}_{\ell=0}^{L}$

For ease of exposition, let the equality constraints ($\boldsymbol{\alpha}_{jj}^{\ell} = \mathbf{0}$) temporarily remain implicit. Introducing the change of variables $\gamma_{ij}^{\ell} := (\mathbf{K}_i^{\ell})^{1/2}\boldsymbol{\alpha}_{ij}^{\ell}$, problem (11) can be recast as

$$\arg\min_{\{\boldsymbol{\alpha}_{ij}^{\ell}\}} (1/2)\left\|\mathbf{Y} - \sum_{\ell=0}^{L} \overline{\mathbf{K}}^{\ell}\mathbf{W}_{\alpha}^{\ell}\right\|_F^2 + \sum_{\ell=0}^{L} g(\boldsymbol{\Gamma}^{\ell}) \quad (18)$$

$$\text{s.t. } \gamma_{ij}^{\ell} - (\mathbf{K}_i^{\ell})^{1/2}\boldsymbol{\alpha}_{ij}^{\ell} = \mathbf{0} \ \forall i, j, \ell$$

where $\boldsymbol{\Gamma}^{\ell} := [\gamma_1^{\ell} \ldots \gamma_N^{\ell}], \gamma_j^{\ell} := [(\gamma_{1j}^{\ell})^{\mathsf{T}} \ldots (\gamma_{Nj}^{\ell})^{\mathsf{T}}]^{\mathsf{T}}$ and $g(\boldsymbol{\Gamma}^{\ell}) := \lambda \sum_{i=1}^{N}\sum_{j=1}^{N} ||\gamma_{ij}^{\ell}||_2$ is the nonsmooth regularizer. Let $\mathbf{D}^{\ell} := \text{Bdiag}(\mathbf{K}_1^{\ell} \ldots \mathbf{K}_N^{\ell})$, and $\mathbf{D} := \text{Bdiag}(\mathbf{D}^0 \ldots \mathbf{D}^{\ell})$, where Bdiag(.) is a block diagonal of its matrix arguments. One can then write the augmented Lagrangian of (18) as

$$\mathscr{L}_{\rho}(\mathbf{W}_{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Xi}) = (1/2)\left\|\mathbf{Y} - \overline{\mathbf{K}}\mathbf{W}_{\alpha}\right\|_F^2 + g(\boldsymbol{\Gamma}) \quad (19)$$
$$+ \left\langle \boldsymbol{\Xi}, \mathbf{D}^{1/2}\mathbf{W}_{\alpha} - \boldsymbol{\Gamma} \right\rangle + (\rho/2)\left\|\boldsymbol{\Gamma} - \mathbf{D}^{1/2}\mathbf{W}_{\alpha}\right\|_F^2$$

where $\mathbf{W}_{\alpha} := [(\mathbf{W}_{\alpha}^{0})^{\mathsf{T}} \ldots (\mathbf{W}_{\alpha}^{L})^{\mathsf{T}}]^{\mathsf{T}}$ and $\boldsymbol{\Gamma} := [(\boldsymbol{\Gamma}^{0})^{\mathsf{T}} \ldots (\boldsymbol{\Gamma}^{L})^{\mathsf{T}}]^{\mathsf{T}}$. Note that $\boldsymbol{\Xi}$ is a matrix of dual variables that collects Lagrange multipliers corresponding to the equality constraints in (18), $\langle \mathbf{P}, \mathbf{Q} \rangle$ denotes the inner product between $\mathbf{P}$ and $\mathbf{Q}$, while $\rho > 0$ a prescribed penalty parameter. ADMM boils down to a sequence of alternating minimization iterations to minimize $\mathscr{L}_{\rho}(\mathbf{W}_{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{\Xi})$ over the primal variables $\mathbf{W}_{\alpha}$, and $\boldsymbol{\Gamma}$, followed by a gradient ascent step over the dual variables $\boldsymbol{\Xi}$; see also [2], [41]. Per iteration $\kappa + 1$, this entails the following provably-convergent steps, see e.g. [41]

$$\mathbf{W}_{\alpha}[\kappa + 1] = \arg\min_{\mathbf{W}_{\alpha}} \mathscr{L}_{\rho}(\mathbf{W}_{\alpha}, \boldsymbol{\Gamma}[\kappa], \boldsymbol{\Xi}[\kappa]) \quad (20a)$$

$$\boldsymbol{\Gamma}[\kappa + 1] = \arg\min_{\boldsymbol{\Gamma}} \mathscr{L}_{\rho}(\mathbf{W}_{\alpha}[\kappa + 1], \boldsymbol{\Gamma}, \boldsymbol{\Xi}[\kappa]) \quad (20b)$$

$$\boldsymbol{\Xi}[\kappa+1] = \boldsymbol{\Xi}[\kappa] + \rho(\mathbf{D}^{1/2}\mathbf{W}_{\alpha}[\kappa+1] - \boldsymbol{\Gamma}[\kappa+1]) . \quad (20c)$$

Focusing on $\mathbf{W}_{\alpha}[\kappa+1]$, note that (20a) decouples across columns of $\mathbf{W}_{\alpha}$, and admits closed-form, parallelizable solutions. Incorporating the structural constraint $\alpha_{ij}^{0} = 0$, one obtains the following decoupled subproblem per column j

$$\bar{\boldsymbol{\alpha}}_j[\kappa+1] = \quad (21)$$

$$\arg\min_{\bar{\boldsymbol{\alpha}}_j} (1/2)\bar{\boldsymbol{\alpha}}_j^{\mathsf{T}}(\overline{\mathbf{K}}_j^{\mathsf{T}}\overline{\mathbf{K}}_j + \rho\mathbf{D}_j)\bar{\boldsymbol{\alpha}}_j - \bar{\boldsymbol{\alpha}}_j^{\mathsf{T}}\mathbf{q}_j[\kappa]$$

where $\mathbf{q}_j[\kappa]$ is constructed by removal of entries indexed by $\mathcal{I}_j$ from $\rho\mathbf{D}^{1/2}\gamma_j[\kappa] + \overline{\mathbf{K}}^{\mathsf{T}}y_j - \mathbf{D}^{1/2}\boldsymbol{\xi}_j[\kappa]$, with $\boldsymbol{\xi}_j[\kappa]$, denoting the j-th column of $\boldsymbol{\Xi}[\kappa]$. Assuming $(\mathbf{K}_j^{\mathsf{T}}\overline{\mathbf{K}}_j + \rho\mathbf{D}_j)$ is invertible, the per-column subproblem (21) admits the following closed-form solution per j

$$\bar{\boldsymbol{\alpha}}_j[\kappa+1] = (\overline{\mathbf{K}}_j^{\mathsf{T}}\overline{\mathbf{K}}_j + \rho\mathbf{D}_j)^{-1}\mathbf{q}_j[\kappa] . \quad (22)$$

On the other hand, (20b) can be solved per component vector γ_{ij}^{ℓ}, and a closed-form solution can be obtained via the so-termed *block shrinkage* operator for each i and j, namely,

$$\gamma_{ij}^{\ell}[\kappa] = \mathcal{P}_{\lambda/\rho}((\mathbf{K}_i^{\ell})^{1/2}\boldsymbol{\alpha}_{ij}^{\ell}[\kappa+1] + \boldsymbol{\xi}_{ij}^{\ell}[\kappa]/\rho) \quad (23)$$

where $\mathcal{P}_{\lambda}(\mathbf{z}) := (\mathbf{z}/||\mathbf{z}||_2)\max(||z||_2 - \lambda, 0)$. Upon convergence, $\{a_{ij}^{\ell}\}$ can be determined by thresholding $\hat{\boldsymbol{\alpha}}_{ij}^{\ell}$ and declaring an edge present from i to j, if there exists any $\hat{\boldsymbol{\alpha}}_{ij}^{\ell} \neq \mathbf{0}$ for $\ell = 1, \ldots, L$.

References

[1]. Angelosante D and Giannakis GB, "Sparse graphical modeling of piecewise-stationary time series," in Proc. Int. Conf. Acoust. Speech Signal Process., Prague, Czech Republic, 5 2011, pp. 1960–1963.

[2]. Baingana B, Mateos G, and Giannakis GB, "Dynamic structural equation models for tracking topologies of social networks," in Proc. of Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Saint Martin, 12 2013, pp. 292–295.

[3]. —, "Proximal-gradient algorithms for tracking cascades over social networks," IEEE J. Sel. Topics Sig. Proc, vol. 8, no. 4, pp. 563–575, 8 2014.

[4]. Bazerque JA and Giannakis GB, “Nonparametric basis pursuit via sparse kernel-based learning: A unifying view with advances in blind methods,” IEEE Signal Processing Magazine, vol. 30, no. 4, pp. 112–125, 7 2013.

[5]. Bertsekas DP, Nonlinear Programming, 2nd ed Athena-Scientific, 1999.

[6]. Bozdogan H, “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions,” Psychometrika, vol. 52, no. 3, pp. 345–370, 9 1987.

[7]. Büchel C and Friston KJ, “Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI.” Cerebral Cortex, vol. 7, no. 8, pp. 768–778, 12 1997. [PubMed: 9408041]

[8]. Cai X, Bazerque JA, and Giannakis GB, “Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations,” PLoS Comp. Biol, vol. 9, no. 5, p. e1003068, 5 2013.

[9]. Chen G, Glen DR, Saad ZS, Hamilton JP, Thomason ME, Gotlib IH, and Cox RW, “Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis,” Computers in Biology and Medicine, vol. 41, no. 12, pp. 1142–1155, 12 2011. [PubMed: 21975109]

[10]. Chen S and Gopalakrishnan P, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” in Proc. DARPA Broadcast News Transcription and Understanding Workshop, Virginia, USA, 8 1998, pp. 127–132.

[11]. Daubechies I, Defrise M, and De Mol C, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” Comm. Pure Appl. Math, vol. 57, no. 11, pp. 1413–1457, 8 2004.

[12]. Deshpande G, Hu X, Stilla R, and Sathian K, “Effective connectivity during haptic perception: A study using Granger causality analysis of functional magnetic resonance imaging data,” NeuroImage, vol. 40, no. 4, pp. 1807–1814, 5 2008. [PubMed: 18329290]

[13]. Friston KJ, Frith CD, and Frackowiak RS, “Time-dependent changes in effective connectivity measured with PET,” Human Brain Mapping, vol. 1, no. 1, pp. 69–79, 1 1993.

[14]. Friston KJ, “Functional and effective connectivity in neuroimaging: A synthesis,” Human Brain Map, vol. 2, no. 1–2, pp. 56–78, 1 1994.

[15]. Friston KJ, Harrison L, and Penny W, “Dynamic causal modelling,” NeuroImage, vol. 19, no. 4, pp. 1273–1302, 8 2003. [PubMed: 12948688]

[16]. Giannakis GB, Shen Y, and V Karanikolas G, “Topology identification and learning over graphs: Accounting for nonlinearities and dynamics,” Proc. of the IEEE, vol. 106, no. 5, pp. 787–807, 5 2018.

[17]. Gitelman DR, Penny WD, Ashburner J, and Friston KJ, “Modeling regional and psychophysiologic interactions in fMRI: The importance of hemodynamic deconvolution,” NeuroImage, vol. 19, no. 1, pp. 200–207, 5 2003. [PubMed: 12781739]

[18]. Goebel R, Roebroeck A, Kim D-S, and Formisano E, “Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping,” Magnetic Resonance Imaging, vol. 21, no. 10, pp. 1251–1261, 12 2003. [PubMed: 14725933]

[19]. Harring JR, Weiss BA, and Hsu J-C, “A comparison of methods for estimating quadratic effects in nonlinear structural equation models.” Psychological Methods, vol. 17, no. 2, pp. 193–214, 6 2012. [PubMed: 22429193]

[20]. Hastie T, Tibshirani R, and Friedman J, The Elements of Statistical Learning. Springer, 2009.

[21]. Haufe S, Tomioka R, Nolte G, Muller K-R, and Kawanabe M, “Modeling sparse connectivity between underlying brain sources for eeg/meg,” Trans. on Bio. Engineer, vol. 57, no. 8, pp. 1954–1963, 5 2010.

[22]. Itani S, Ohannessian M, Sachs K, Nolan GP, and Dahleh MA, “Structure learning in causal cyclic networks,” in Causality: Objectives and assessment, 2010, pp. 165–176.

[23]. Jiang X, Mahadevan S, and Urbina A, “Bayesian nonlinear structural equation modeling for hierarchical validation of dynamical systems,” Mechanical Systems and Signal Processing, vol. 24, no. 4, pp. 957–975, 4 2010.

[24]. Jöreskog KG, Yang F, Marcoulides G, and Schumacker R, “Non-linear structural equation models: The Kenny-Judd model with inter-action effects,” Advanced Structural Equation Modeling: Issues and Techniques, pp. 57–88, 1 1996.

[25]. Kaplan D, Structural Equation Modeling: Foundations and Extensions. Sage, 2009.

[26]. Kelava A, Nagengast B, and Brandt H, “A nonlinear structural equation mixture modeling approach for nonnormally distributed latent predictor variables,” Structural Equation Modeling: A Multidisciplinary Journal, vol. 21, no. 3, pp. 468–481, 6 2014.

[27]. Kolaczyk ED, Statistical Analysis of Network Data: Methods and Models. Springer, 2009.

[28]. Kramer MA, Kolaczyk ED, and Kirsch HE, “Emergent network topology at seizure onset in humans,” Epilepsy Research, vol. 79, no. 2, pp. 173–186, 5 2008. [PubMed: 18359200]

[29]. Lee S-Y and Song X-Y, “Model comparison of nonlinear structural equation models with fixed covariates,” Psychometrika, vol. 68, no. 1, pp. 27–47, 3 2003.

[30]. Lee S-Y, Song X-Y, and Lee JC, “Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data,” J. of Educ. and Behavioral Stat, vol. 28, no. 2, pp. 111–134, 6 2003.

[31]. Lim N, d’Alché Buc F, Auliac C, and Michailidis G, “Operator-valued kernel-based vector autoregressive models for network inference,” Ma-chine Learning, vol. 99, no. 3, pp. 489–513, 6 2015.

[32]. Marinazzo D, Pellicoro M, and Stramaglia S, “Kernel-Granger causal-ity and the analysis of dynamical networks,” Physical Review E, vol. 77, no. 5, p. 056215, 5 2008.

[33]. —, “Kernel method for nonlinear Granger causality,” Physical Review Letters, vol. 100, no. 14, p. 144103, 4 2008. [PubMed: 18518037]

[34]. Mclntosh A and Gonzalez-Lima F, “Structural equation modeling and its application to network analysis in functional brain imaging,” Human Brain Mapping, vol. 2, no. 1–2, pp. 2–22, 10 1994.

[35]. C. A. Micchelli and M. Pontil, “Learning the kernel function via regularization,” Journal of Machine Learning Research, vol. 6, no. Jul, pp. 1099–1125, 7 2005.

[36]. Peters J, Janzing D, and Sch’olkopf B, “Causal inference on time series using restricted structural equation models,” in Advances in Neural Information Processing Systems, Lake Tahoe, 2013, pp. 154–162.

[37]. Roche R and Commins S, Pioneering Studies in Cognitive Neuro-science. McGraw-Hill Education, 2009.

[38]. Roebroeck A, Formisano E, and Goebel R, “Mapping directed influ-ence over the brain using Granger causality and fMRI,” NeuroImage, vol. 25, no. 1, pp. 230–242, 3 2005. [PubMed: 15734358]

[39]. Rubinov M and Sporns O, “Complex network measures of brain connectivity: Uses and interpretations,” NeuroImage, vol. 52, no. 3, pp. 1059–1069, 9 2010. [PubMed: 19819337]

[40]. Sahoo D, Hoi SC, and Li B, “Online multiple kernel regression,” in Intl. Conf. on Know. Disc. and Data Mining, New York, 8 2014, pp. 293–302.

[41]. Schizas ID, Ribeiro A, and Giannakis GB, “Consensus in ad hoc WSNs with noisy links -Part I: Distributed estimation of deterministic signals,” IEEE Trans. Sig. Proc, vol. 56, pp. 350–364, 1 2008.

[42]. Shen Y, Baingana B, and Giannakis GB, “Nonlinear structural equation models for network topology inference,” in Proc. of Conf. on Info. Sciences and Systems, Princeton, NJ, 3 2016.

[43]. —, “Kernel-based structural equation models for topology identification of directed networks,” IEEE Trans. Signal Processing, vol. 65, no. 10, pp. 2503–2516, 5 2017.

[44]. —, “Topology inference of directed graphs using nonlinear structural vector autoregressive models,” in Proc. ofIntl. Conf. on Acoustic Speech and Signal Process., New Orleans, USA, 2017, pp. 6513–6517.

[45]. Shen Y, Chen T, and Giannakis GB, “Random feature-based online multi-kernel learning in environments with unknown dynamics,” Journal of Machine Learning Research, vol. 20, no. 22, pp. 1–36, 2019.

[46]. Shen Y and Giannakis GB, “Online identification of directional graph topologies capturing dynamic and nonlinear dependencies,” in IEEE Data Science Workshop, Lausanne, Switzerland, 2018, pp. 195–199.

[47]. Sun X, “Assessing nonlinear Granger causality from multivariate time series,” in Proc. Eur. Conf. Mach. Learn. Knowl. Disc. Databases, Antwerp, Belgium, 9 2008, pp. 440–455.

[48]. Valdés-Sosa PA, Sánchez-Bornot JM, Lage-Castellanos A, Vega-Hernández M, Bosch-Bayard J, Melie-García L, and Canales-Rodríguez E, “Estimating brain functional connectivity with sparse multi-variate autoregression,” Philosophical Trans. of the Royal Soc. of London B: Bio. Sci, vol. 360, no. 1457, pp. 969–981, 5 2005.

[49]. Varoquaux G, Gramfort A, Poline J-B, and Thirion B, “Brain covari-ance selection: better individual functional connectivity models using population prior,” in Advances in Neural Info. Process. Sys, Vancouver, Canada, 2010, pp. 2334–2342.

[50]. Wahba G, Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics).

[51]. Wall M and Amemiya Y, “Estimation for polynomial structural equation models,” J. American Stat. Assoc, vol. 95, no. 451, pp. 929–940, 10 2000.

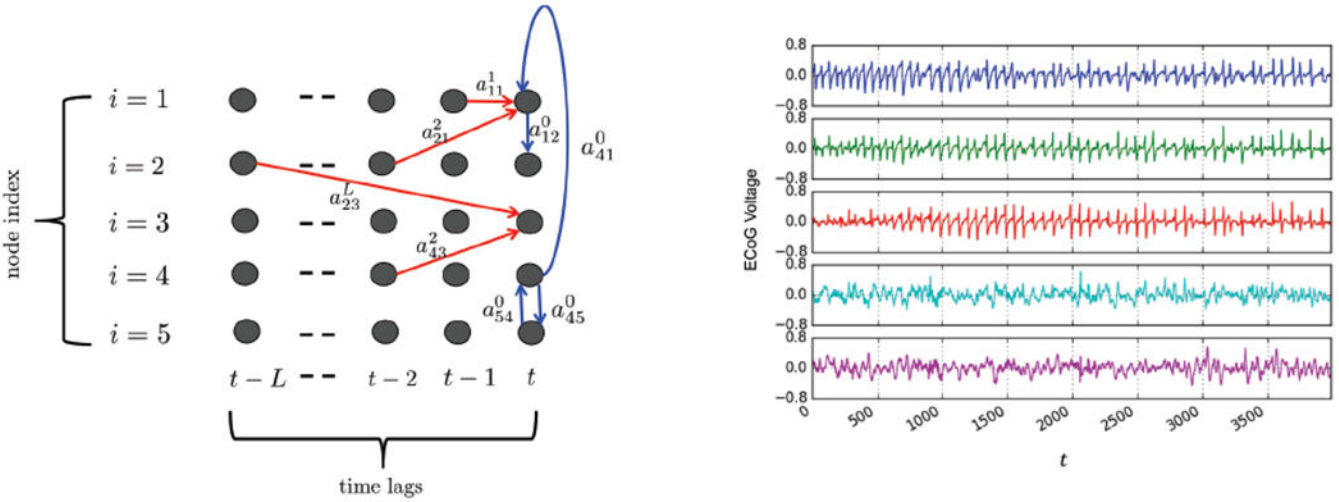

Fig. 1:

(left) A simple illustration of a 5-node brain network; and (right) a set of five neuronal time series (e.g., ECoG voltage) each associated with a node. Per interval *t*, SVARMs postulate that directed dependencies between the 5 nodal time series may be due to both the instantaneous effects (blue links), and/or time-lagged effects (red links). Estimating the values of the unknown coefficients amounts to learning the directed (link) structure of the network.

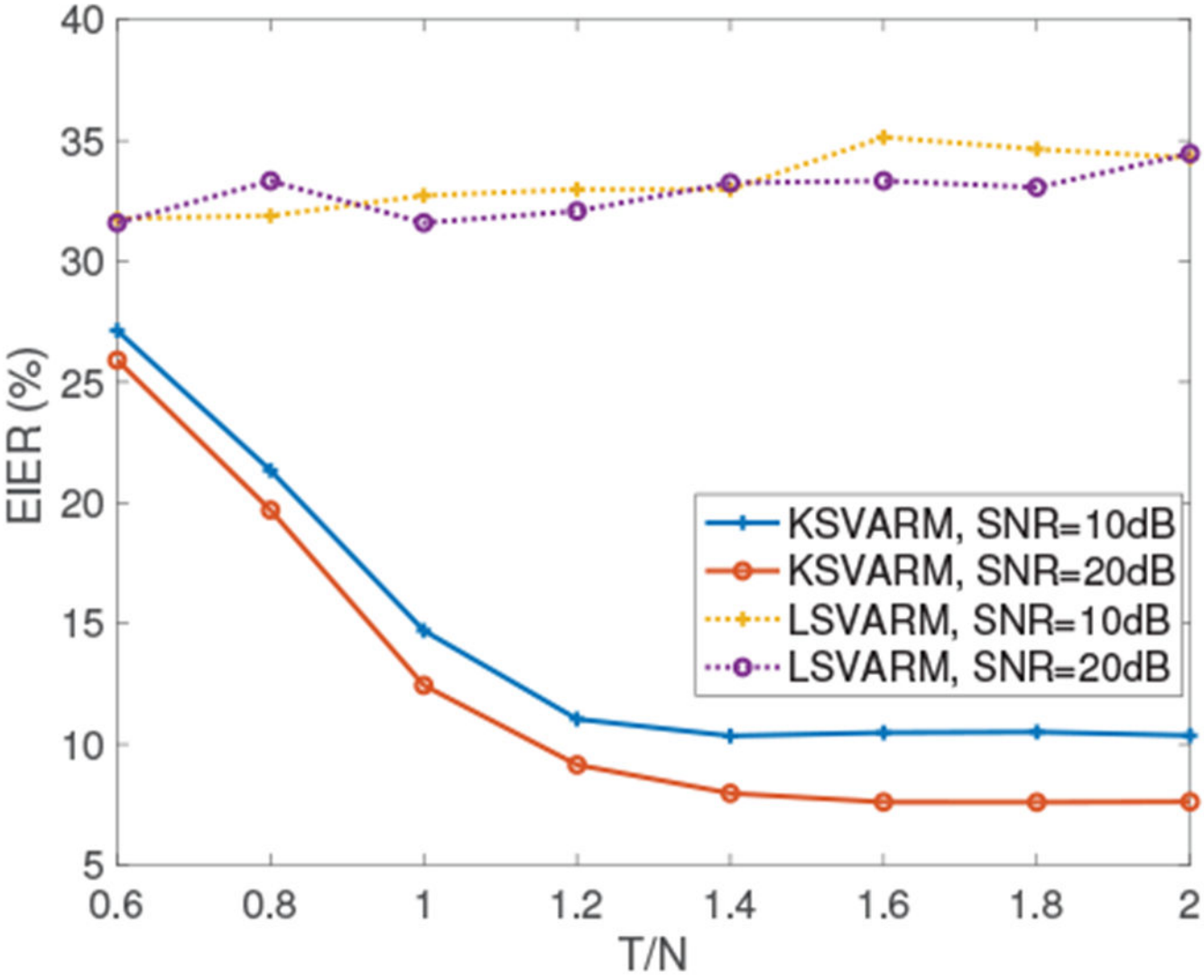

Fig. 2:

Plot of EIER vs. measurement ratio (T/N), with simulated data generated via a polynomial kernel of order $P = 2$. Note that K-SVARMs consistently outperform LSVARMs.

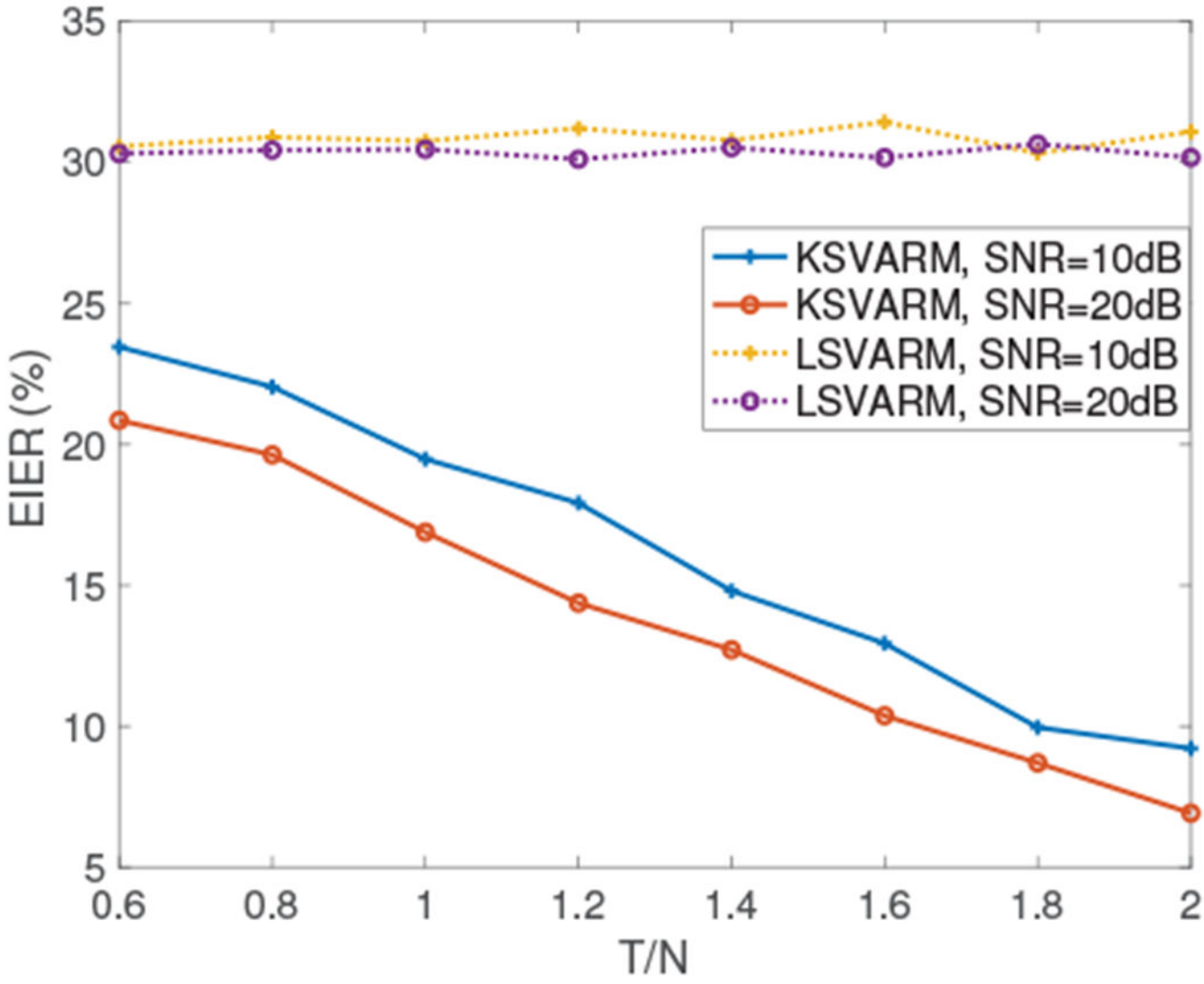

Fig. 3:

Plot of EIER vs. (*T/N*) with data generated using a Gaussian kernel with $\sigma^2 = 1$; K-SVARMs uniformly lead to lower errors than linear LSVARMs over varying SNR levels, based on empirical observations.

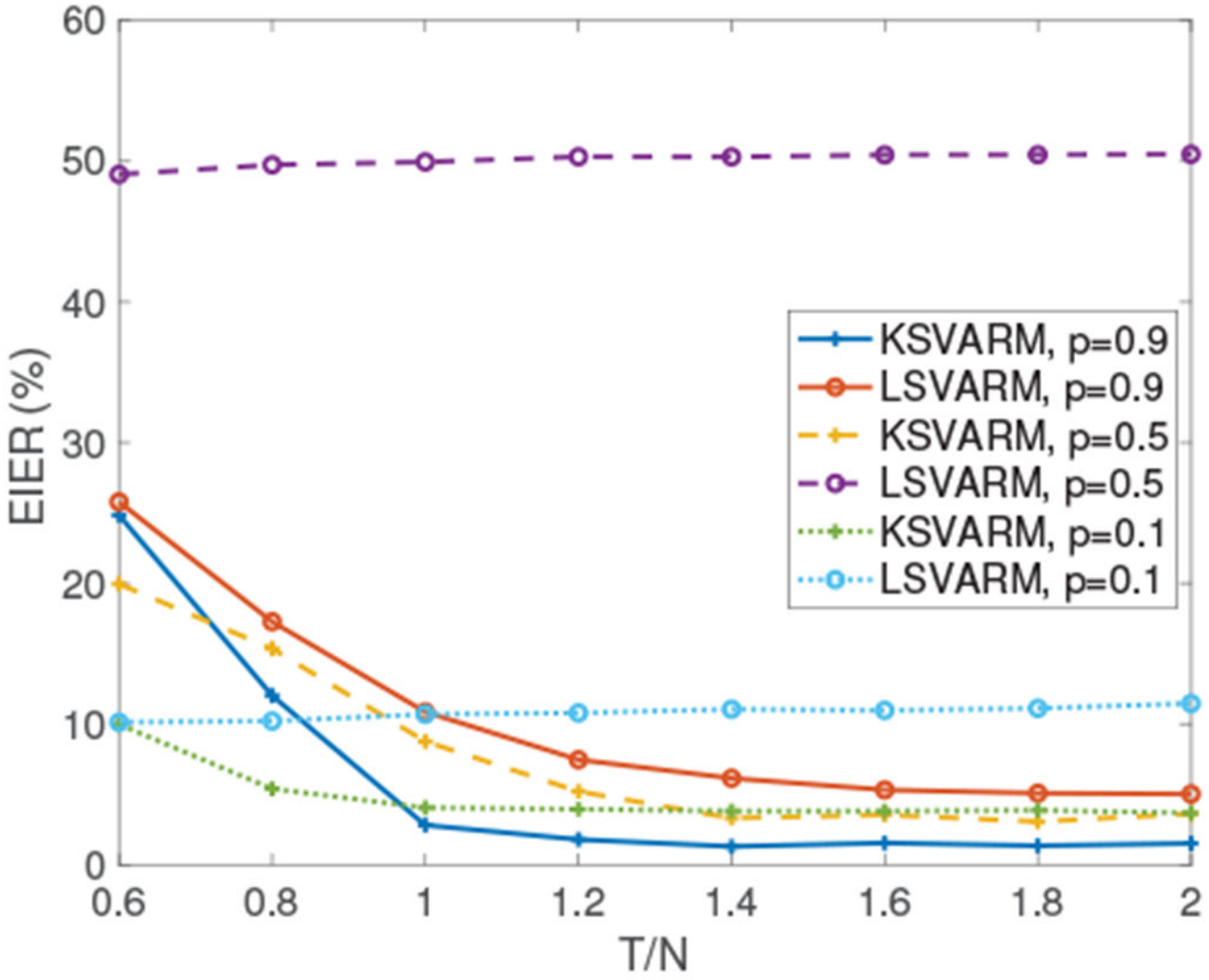

Fig. 4:

Plot of EIER vs. (T/N) with simulated data generated via a polynomial kernel of order $P = 2$; it can be empirically observed that K-SVARMs uniformly outperform LSVARMs across varying edge densities.

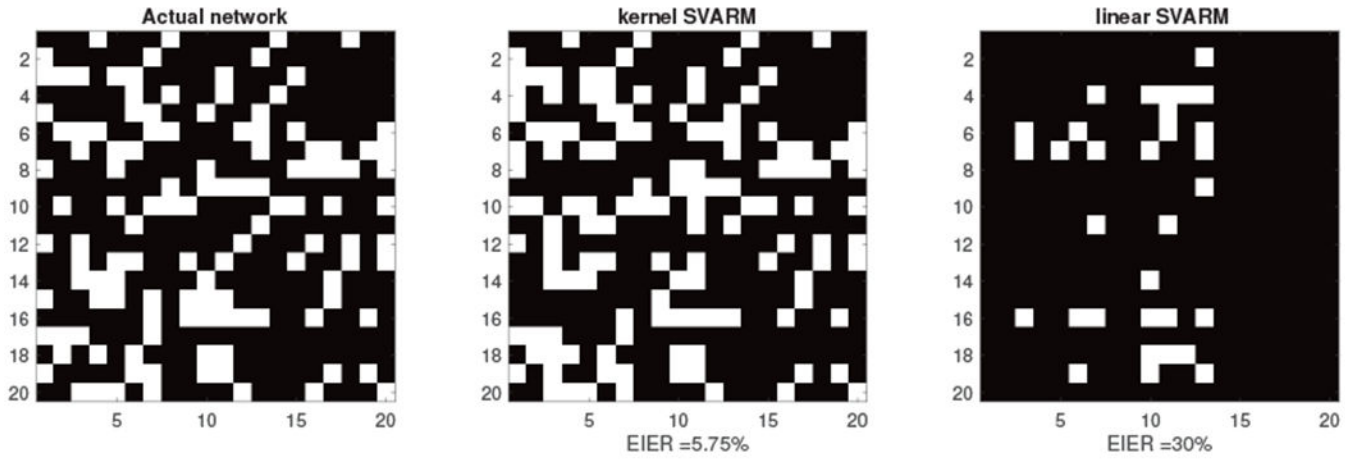

Fig. 5:

Plots of actual and inferred adjacency matrices resulting from adopting linear and nonlinear SVARM with $T = 40$.

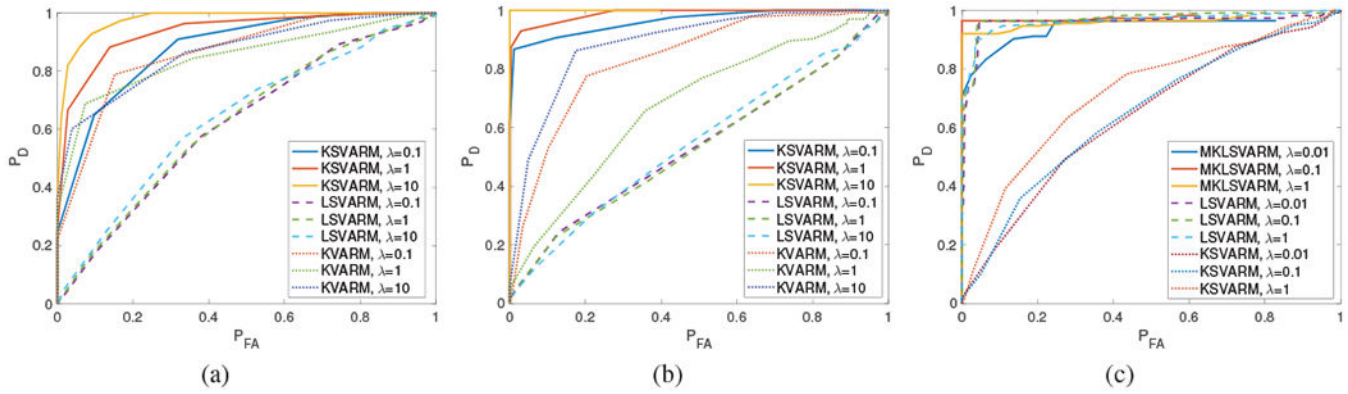

Fig. 6:

ROC curves generated under different modeling assumptions: a) K-SVARM based on a Gaussian kernel with $\sigma^2 = 1$; b) K-SVARM based on polynomial kernel of order $P = 2$; and c) Linear SVARM.

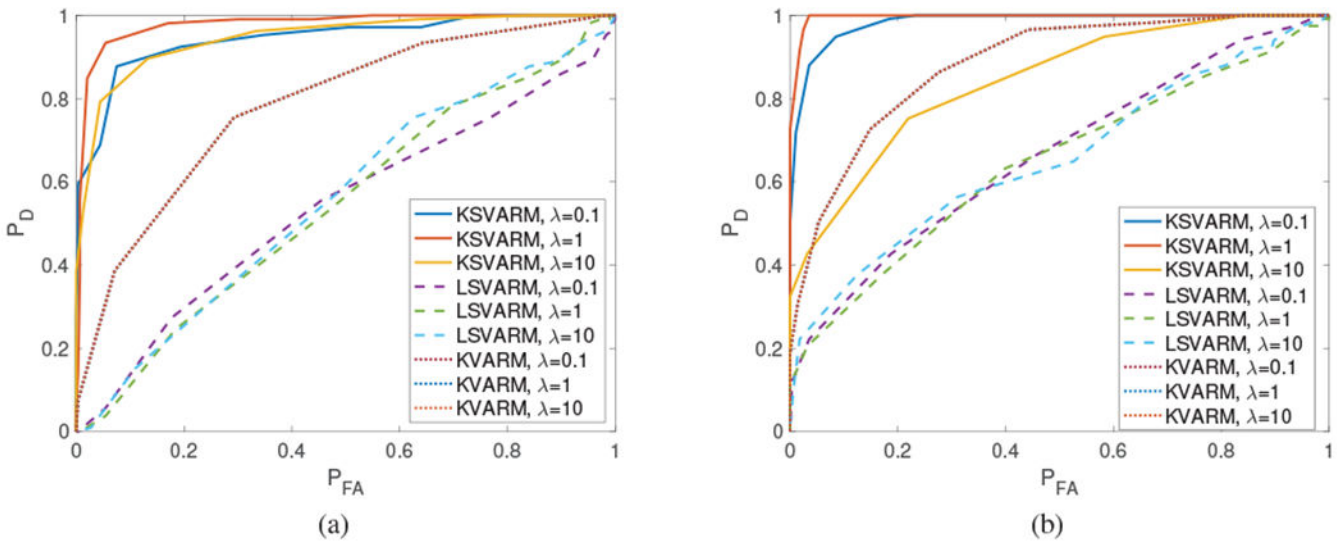

Fig. 7:

ROC curves generated under different modeling assumptions: a) K-SVARM based on a Gaussian kernel with $\sigma^2 = 1$, $L = 2$; and b) K-SVARM based on a polynomial kernel of order $P = 2$, $L = 2$.

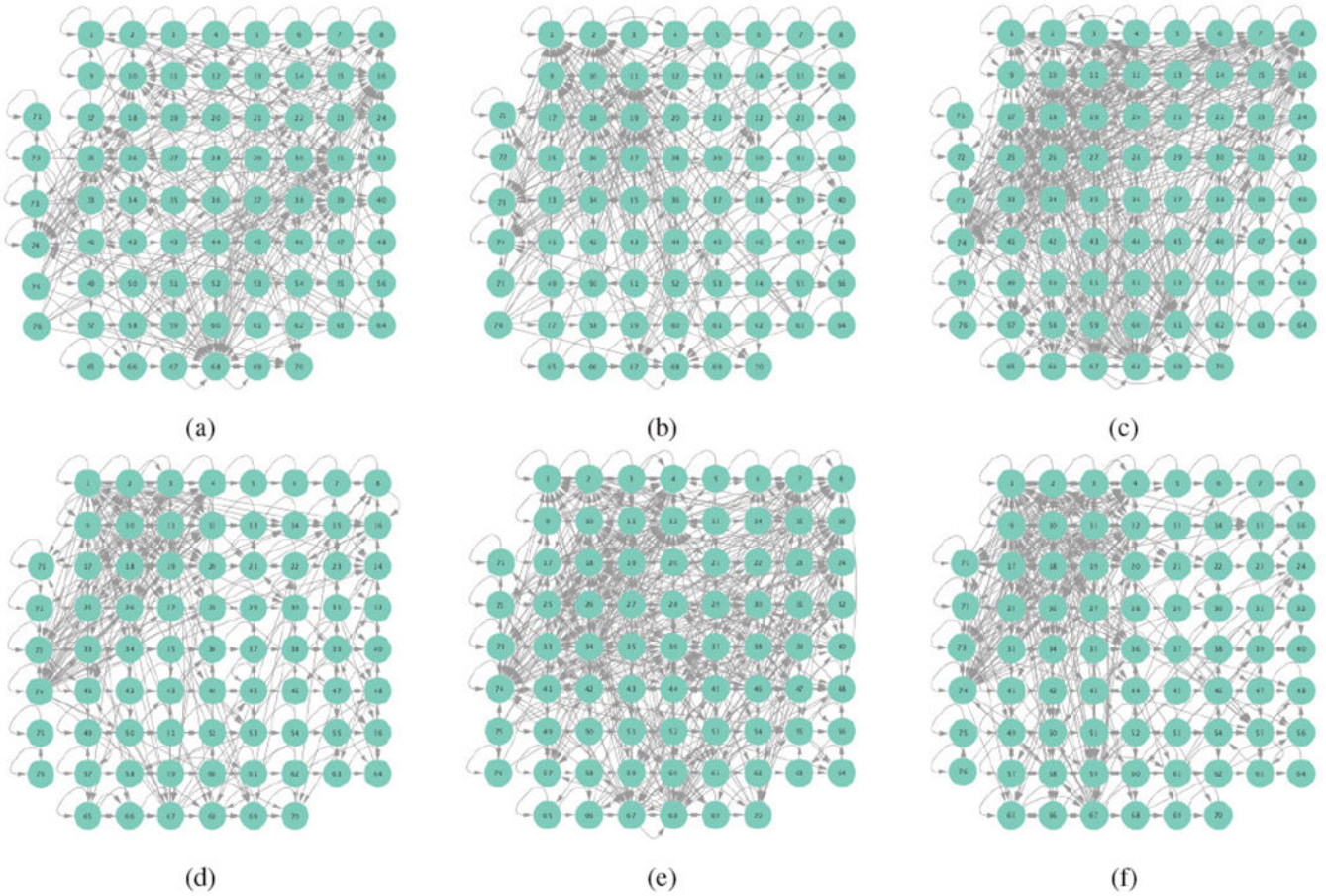

Fig. 8:

Visualizations of 76-electrode networks inferred from ECoG data: (a) linear SVARM with $L = 1$ on preictal time series; (b) linear SVARM on ictal time series; (c) K-SVARM on preictal time series, using Gaussian kernel with $\sigma = 1$; (d) the same K-SVARM on ictal time series; (e) K-SVARM with kernel selection on preictal time series; and finally (f) K-SVARM with kernel selection on ictal time series.

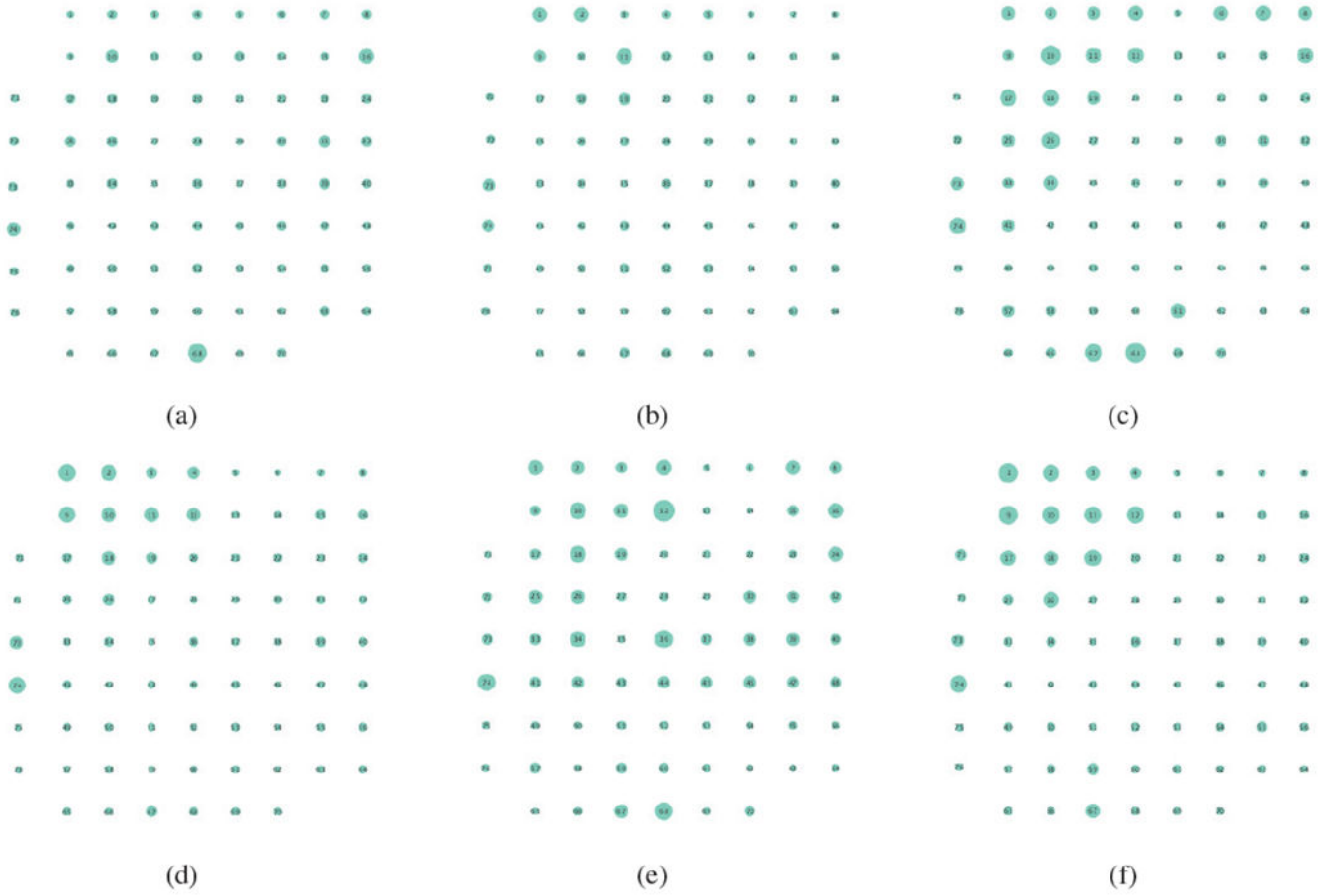

Fig. 9:

Node degrees of networks inferred from ECoG data encoded by circle radii: (a) linear SVARM on preictal data; (b) linear SVARM on ictal data; (c) K-SVARM on preictal time series; (d) K-SVARM on ictal data; (e) MKL-SVARM on preictal time series; (f) MKL-SVARM on ictal time series.

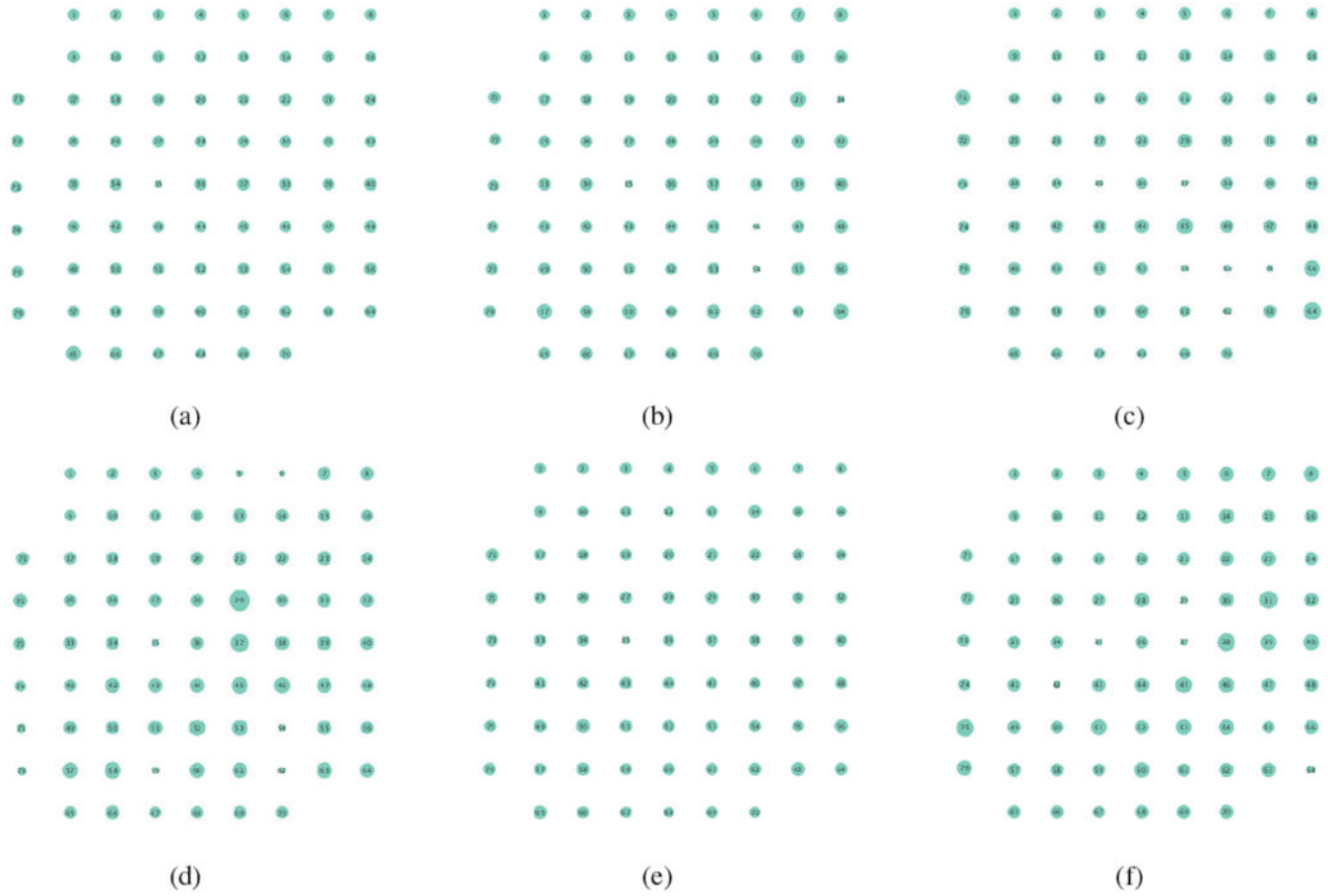

Fig. 10:
Same as in Figure 9 for comparison based on average shortest path length of inferred graphs.

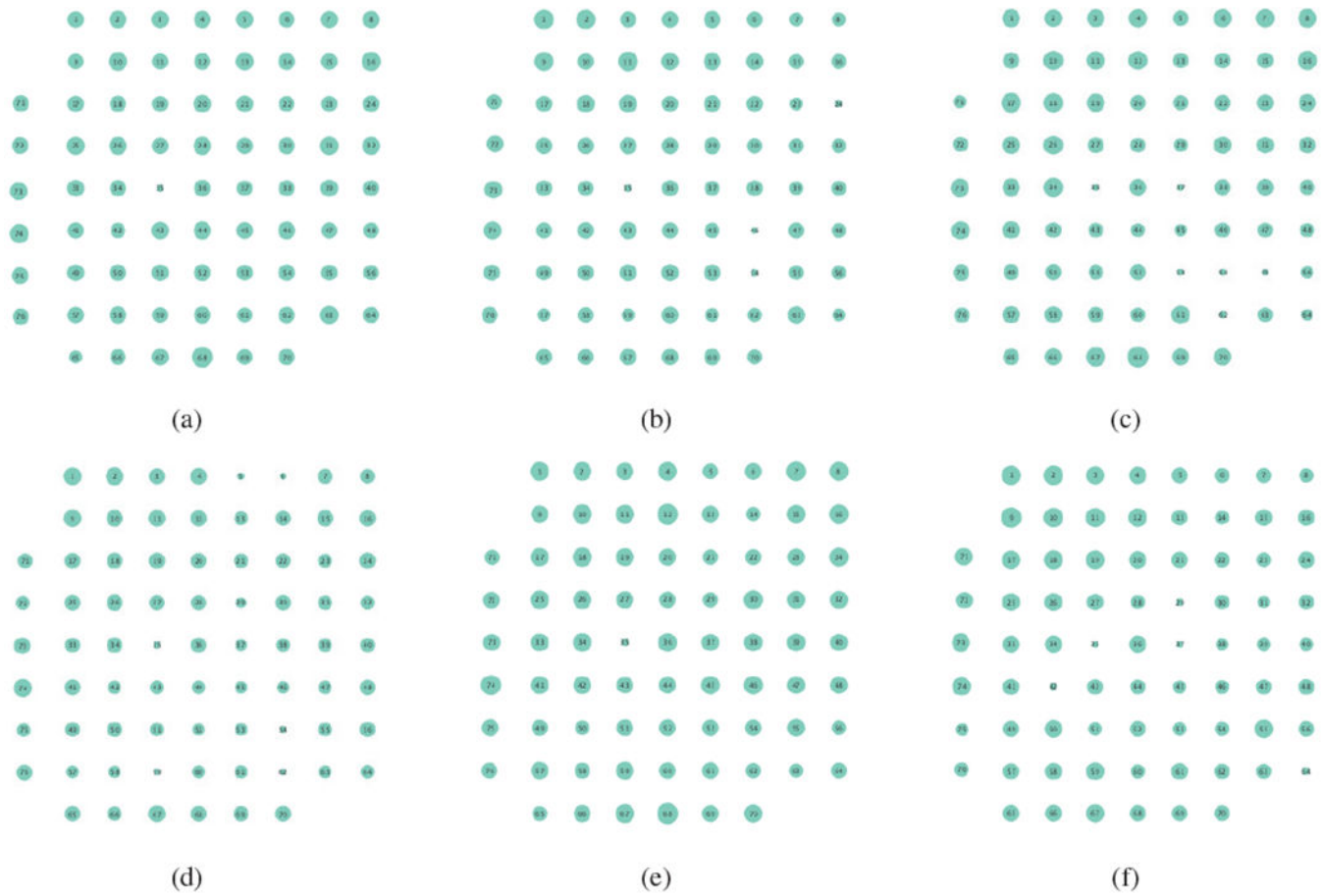

Fig. 11:
Same as in Figure 9 for comparison based on closeness centrality of inferred graphs.

TABLE I:

Changes of edges and nodal degrees from preictal to ictal stages using networks inferred from ECoG seizure data using the linear, K-SVARM, and K-SVARM with kernel selection schemes ↑: appear/increase, ↓: disappear/decrease.

	Linear SVARM		**K-SVARM**		**MKL-SVARM**	
	↑	↓	↑	↓	↑	↓
edge	166	184	103	252	233	421
degree	24	31	23	40	22	48

TABLE II:

Comparison of global metrics associated with networks inferred from ECoG seizure data using the linear, K-SVARM, and K-SVARM with kernel selection scheme. Major differences between the computed metrics indicate that one may gain insights from network topologies inferred via models that capture nonlinear dependencies.

	Linear SVARM		**K-SVARM**		**MKL-SVARM**	
	Preictal	**Ictal**	**Preictal**	**Ictal**	**Preictal**	**Ictal**
Network density	0.103	0.095	0.136	0.089	0.148	0.101
Glob. clustering coeff.	0.246	0.220	0.372	0.356	0.391	0.343
No. of connect. comp.	8	7	2	5	2	6
Network diameter	5	5	7	9	4	7
Avg. no. of neighbors	7.73	6.89	10.23	6.71	11.11	7.55

TABLE III:

Comparison of classification accuracy (%) using networks inferred from ECoG seizure data using the linear, K-SVARM, and K-SVARM with kernel selection scheme.

	L-SVARM	K-SVARM	MKL-SVARM
Degree	65.4	68.3	70.8
Clust. coeff.	56.1	61.1	61.2
Centrality	59.4	65.2	65.3