TOPOLOGY INFERENCE OF DIRECTED GRAPHS USING NONLINEAR STRUCTURAL VECTOR AUTOREGRESSIVE MODELS[†]

Yanning Shen, Brian Baingana, and Georgios B. Giannakis

Dept. of ECE and DTC, University of Minnesota, Minneapolis, USA

ABSTRACT

Linear structural vector autoregressive models constitute a generalization of structural equation models (SEMs) and vector autoregressive (VAR) models, two popular approaches for topology inference of directed graphs. Although simple and tractable, linear SVARMs seldom capture nonlinearities that are inherent to complex systems, such as the human brain. To this end, the present paper advocates kernel-based nonlinear SVARMs, and develops an efficient sparsity-promoting least-squares estimator to learn the hidden topology. Numerical tests on real electrocorticographic (ECoG) data from an Epilepsy study corroborate the efficacy of the novel approach.

***Index Terms*—** Network topology inference, structural vector autoregressive models, nonlinear

1. INTRODUCTION

Graph topology inference plays a crucial role in network studies, since edges are often not directly observable, but network processes may be measurable at nodes. For example, one may have access to functional magnetic resonance (fMRI) time series per brain region, yet the dependency structure between them is hidden. Granger causality or vector autoregressive (VAR) models [6], and structural equation models (SEMs) [11] are widely adopted to infer directed network topologies. VAR models postulate that causal relationships are captured through time-lagged dependencies between nodal time series, while SEMs are based on instantaneous interactions.

Interestingly, structural vector autoregressive models (SVARMs) [3] offer a unified view, postulating that nodal time series result from both instantaneous and time-lagged interactions with other time series. Indeed, SVARMs are more flexible and explanatory than VARs and SEMs used independently, at the expense of increased model complexity.

Contemporary SVARMs are generally based on linear models, due to their simplicity and tractability. However, resorting to linear SVARMs is limiting, since interactions within complex systems (e.g., the human brain) are generally best captured by nonlinear models. In fact, several variants of nonlinear SEMs and VARs have been put forth in several recent works; see e.g., [7, 9, 10, 12, 13, 15, 18, 20] and references therein.

The present paper extends the merits of these prior works, and puts forth a more general *nonlinear* SVARM for inference of *directed* network topologies. Unlike prior nonlinear modeling efforts for SEMs and VARs, a novel additive nonlinear model that assumes no knowledge of the link structure is advocated. The developed estimator leverages kernels as an encompassing framework for nonlinear learning tasks, and exploits edge sparsity, which is inherent to most real networks. Preliminary tests conducted on real brain signals from an Epilepsy study demonstrate that estimated brain networks contain new directed edges that were previously overlooked by linear SVARMs.

2. PRELIMINARIES

Consider an N-node directed graph, whose topology is unknown, but a time-series $\{y_{it}\}_{t=1}^{T}$ is observed per node i, over T time intervals. Traditional linear SVARMs postulate that each y_{jt} is a linear combination of instantaneous measurements at other nodes $\{y_{it}\}_{i\neq j}$, and their time-lagged versions $\{\{y_{i(t-\ell)}\}_{i=1}^{N}\}_{\ell=1}^{L}$; see e.g., [3]. Specifically, y_{jt} obeys the following linear time-lagged model

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

with a_{ij}^{ℓ} capturing the causal influence of node i upon node j over a lag of ℓ time points, while a_{ij}^{0} encodes the corresponding instantaneous causal relationship between them. The coefficients encode the causal structure of the network, that is, a causal link exists between nodes i and j only if there exists $a_{ij}^{\ell} \neq 0$ for $\ell = 0, \ldots, L$. On the other hand, if $a_{ij}^{0} = 0 \;\forall i, j$, then (1) reduces to classical Granger causality (a.k.a., vector autoregression); see also [16]. On the other hand, setting $a_{ij}^{\ell} = 0$ for $\ell = 1, \ldots, L$, reduces (1) to a special instance of a linear SEM with no exogenous inputs [11].

Let $\mathbf{A}^{\ell} \in \mathbb{R}^{N\times N}$ denote the "time-lagged" adjacency matrix, with $\left[\mathbf{A}^{\ell}\right]_{ij} = a_{ij}^{\ell}$. Given the multivariate time series $\{\mathbf{y}_t\}_{t=1}^{T}$, where $\mathbf{y}_t := [y_{1t}, \ldots, y_{Nt}]^{\top}$, the goal is to esti-

[†] Work in this paper was supported by grants NSF 1500713, 1514056 and NIH 1R01GM104975-01.

mate the coefficient matrices $\{\mathbf{A}^\ell\}_{\ell=1}^L$, and consequently unveil the hidden network topology. Although generally known, one can readily deduce L via standard model order selection tools e.g., the Bayesian information criterion [4], or Akaike's Information Criterion (AIC) [2].

With the topology known, several approaches have been put forth to learn the model coefficients. Examples are based upon ordinary least-squares [3], hypothesis tests developed to detect presence or absence of pairwise causal links under prescribed false alarm rates [16], and proximal gradient iterations exploiting known network structural properties [15]. Albeit conceptually simple and computationally tractable, the linear SVARM is incapable of capturing nonlinear dependencies inherent to complex networks such as the human brain. The present paper generalizes the linear SVARM in (1) to a nonlinear kernel-based SVARM.

3. NONLINEAR SVARM

Aiming for enhanced flexibility and accuracy, we transcend the linearity assumption, and resort to more general nonlinear models which encompass their linear counterparts as special cases. Consistent with the instantaneous and time-lagged structure inherent to the linear SVARM, we postulate that [cf. (1)]

$$y_{jt} = \psi_j^0\left(\mathbf{y}_{-jt}\right) + \sum_{\ell=1}^{L} \psi_j^\ell\left(\mathbf{y}_{t-\ell}\right) + e_{jt} \tag{2}$$

where $\mathbf{y}_{-jt} := [y_{1t}, \ldots, y_{(j-1)t}, y_{(j+1)t}, \ldots, y_{Nt}]^\top$ collects present observations on all nodes but the j-th one, at time slot t, and $\psi_j^\ell(.)$ denotes a nonlinear function. Inspired by *generalized additive models* (GAMs) [8, Ch. 9], we further posit that

$$\psi_j^0\left(\mathbf{y}_{-jt}\right) = \sum_{i \neq j} a_{ij}^0 \psi_{ij}^0\left(y_{it}\right) \tag{3}$$

$$\psi_j^\ell(\mathbf{y}_{t-\ell}) = \sum_{i=1}^{N} a_{ij}^\ell \psi_{ij}^\ell(y_{i(t-\ell)}),\ \ell = 1, \ldots, L \tag{4}$$

with each $\psi_{ij}^\ell(.)$ denoting a nonlinear univariate function. Note that (3) and (4) deviate from traditional GAMs by inclusion of the explicit unknown coefficients $\{a_{ij}^\ell\}$, which will facilitate identification of the directed graph edges. Furthermore, suppose that $\{\psi_{ij}^\ell(y_{it})\}$ are expressed as

$$\psi_{ij}^\ell(y_{it}) = \sum_{p=1}^{P} c_{ijp}^\ell \phi_p^\ell(y_{i(t-\ell)}) \tag{5}$$

where $\{c_{ijp}^\ell\}_{p=1}^P$ are unknown coefficients, but $\{\phi_p^\ell(.)\}_{p=1}^P$ are (possibly) known functions, and P can take on any positive integer or even infinity. Letting $\mathbf{c}_{ij}^\ell := [c_{ij1}^\ell, \ldots, c_{ijP}^\ell]^\top$, and defining $\boldsymbol{\phi}^\ell(y_{it}) := [\phi_1^\ell(y_{it}), \ldots, \phi_P^\ell(y_{it})]^\top$, one obtains the following nonlinear SVARM [cf. (2) –(5)]

$$y_{jt} = \sum_{i \neq j} a_{ij}^0 (\boldsymbol{\phi}^0)^\top (y_{it}) \mathbf{c}_{ij}^0 + \sum_{i=1}^{N} \sum_{\ell=1}^{L} a_{ij}^\ell (\boldsymbol{\phi}^\ell)^\top (y_{it}) \mathbf{c}_{ij}^\ell + e_{jt} \tag{6}$$

for $j = 1, \ldots, N$, and $t = 1, \ldots, T$. Let $\tilde{\mathbf{y}}_j := [y_{j1}, \ldots, y_{jT}]^\top$, $\mathbf{e}_j := [e_{j1}, \ldots, e_{jt}]^\top$, $\mathbf{Y} := [\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_N]$, and define $\boldsymbol{\Phi}_i^\ell := [\boldsymbol{\phi}^\ell(y_{i1}), \ldots, \boldsymbol{\phi}^\ell(y_{iT})]^\top$, and $\boldsymbol{\Phi}^\ell := [\boldsymbol{\Phi}_1^\ell, \ldots, \boldsymbol{\Phi}_N^\ell]$. One then obtains the nonlinear SVARM in matrix form

$$\mathbf{Y} = \sum_{\ell=0}^{L} \boldsymbol{\Phi}^\ell \mathbf{W}^\ell + \mathbf{E} \tag{7}$$

where

$$\mathbf{W}^\ell := \begin{bmatrix} a_{11}^\ell \mathbf{c}_{11}^\ell & \cdots & a_{1N}^\ell \mathbf{c}_{1N}^\ell \\ \vdots & \ddots & \vdots \\ a_{N1}^\ell \mathbf{c}_{N1}^\ell & \cdots & a_{NN}^\ell \mathbf{c}_{NN}^\ell \end{bmatrix} \tag{8}$$

is an $NP \times N$ block matrix, whose structure is modulated by the entries of $\mathbf{A}^\ell$. For instance, if $a_{ij}^\ell = 0$, then $a_{ij}^\ell \mathbf{c}_{ij}^\ell$ is an all-zero block regardless of the values of entries in $\mathbf{c}_{ij}^\ell$.

It is worth observing that real-world networks often exhibit edge sparsity, that is, $\mathbf{A}^\ell$ has only a few nonzero entries, and more efficient estimators can be realized by capitalizing on this prior knowledge. Note that entries $\{a_{ij}^\ell\}$ determine whether certain blocks are all-zero or not [cf. (8)], naturally leading $\mathbf{W}^\ell$ to exhibit group sparsity. The rest of the paper will leverage this group sparsity, and put forth a novel kernel-based estimator for inference of the unknown network topology. The problem can now be formally stated as follows.

Problem statement. Given nodal measurements $\mathbf{Y}$ in (7), the goal is to estimate the edge-modulated matrix $\mathbf{W}^\ell$, and correspondingly the unknown adjacency matrix $\mathbf{A}^\ell$. Whether $\boldsymbol{\Phi}$ is (un)known will be clarified in the ensuing section, which puts forth a novel estimator.

4. KERNEL-BASED TOPOLOGY ESTIMATION

To estimate the unknowns in (7) with no prior knowledge about additive noise statistics, the present paper advocates a regularized least-squares (LS) estimator, namely,

$$\underset{\mathbf{W}^0 \in \mathcal{W},\ \{\mathbf{W}^\ell\}_{\ell=1}^L}{\arg\min} \ (1/2) \left\| \mathbf{Y} - \textstyle\sum_{\ell=1}^{L} \boldsymbol{\Phi}^\ell \mathbf{W}^\ell \right\|_F^2 + \textstyle\sum_{\ell=1}^{L} \lambda_\ell \|\mathbf{W}^\ell\|_{\mathcal{G},1} \tag{9}$$

where $\mathcal{W} := \{\mathbf{W}^0 \in \mathbb{R}^{NP \times N} : \mathbf{w}_{ii}^0 = \mathbf{0},\ \ i = 1, \ldots, N\}$, and with $\mathbf{w}_{ij}^\ell := a_{ij}^\ell \mathbf{c}_{ij}^\ell$ denoting the (i, j)-th block of $\mathbf{W}^\ell$. The constraint set $\mathcal{W}$ restricts solutions to network topologies free of self-loops; that is, $a_{ii}^0 = 0 \iff \mathbf{w}_{ii}^0 = \mathbf{0}$. Furthermore, the penalty term $\|\mathbf{W}^\ell\|_{\mathcal{G},1} := \sum_{i,j} \|\mathbf{w}_{ij}^\ell\|_2$ is a well-known regularizer that has been shown to promote group sparsity [22], while the regularization parameters $\{\lambda_l \geq 0\}$ allow the estimator to trade off group sparsity for the LS fit, and can take different values. For simplicity, it will be assumed in the sequel that $\lambda_1 = \ldots = \lambda_L = \lambda$.

With $\{\phi^\ell\}_{\ell=0}^L$ available, the regularized cost in (9) is jointly convex with respect to (w.r.t.) $\{\mathbf{W}^\ell\}$, and in principle the problem can be solved to global optimality. The solutions $\{\widehat{\mathbf{W}}^\ell\}_{\ell=0}^L$ contain all the information necessary to recover the graph topology captured by $\{\mathbf{A}^\ell\}$, by simply identifying nonzero blocks. But $\{\mathbf{\Phi}^\ell\}$ may not be available, since one may not know the functions $\{\phi_p^\ell(.)\}_{p=1}^P$ explicitly. Moreover, even when $\{\phi_p^\ell(.)\}_{p=1}^P$ are known, it may be impossible to efficiently solve (9), as $P \to \infty$. In lieu of these challenges, the present paper advocates kernel-based approaches that have well-appreciated merits in nonlinear modeling, often circumventing the need for explicit knowledge of the nonlinear mappings. The kernel-based estimator developed next will rely on the following result, which is a variant of the *representer* theorem [21], and its proof can be found in [18].

Proposition 1: *If $\{\widehat{\mathbf{W}}^\ell\}_{\ell=0}^L$ is the optimal solution of the regularized LS estimator in* (9)*, with $\hat{\mathbf{w}}_{ij}^\ell$ denoting the (i,j)-th block entry of $\widehat{\mathbf{W}}^\ell$, then $\hat{\mathbf{w}}_{ij}^\ell$ can be written as*

$$\hat{\mathbf{w}}_{ij}^\ell = \sum_{t=1}^{T} \alpha_{ijt}^\ell \boldsymbol{\phi}^\ell(y_{it}) = (\mathbf{\Phi}_i^\ell)^\top \boldsymbol{\alpha}_{ij}^\ell, \quad \forall i,j,\ell \tag{10}$$

where $\boldsymbol{\alpha}_{ij}^\ell := [\alpha_{ij1}^\ell, \ldots, \alpha_{ijT}^\ell]^\top$ is a $T \times 1$ coefficient vector.

Acknowledging Proposition 1, and substituting (10) into (9), the LS term can be written as

$$\begin{aligned}
&(1/2)\|\mathbf{Y} - \sum_{\ell=0}^{L} \mathbf{\Phi}^\ell \mathbf{W}^\ell\|_F^2 \\
&= \sum_{j=1}^{N} (1/2) \left\| \tilde{\mathbf{y}}_j - \sum_{i \neq j} \mathbf{\Phi}_i^0 (\mathbf{\Phi}_i^0)^\top \boldsymbol{\alpha}_{ij}^0 - \sum_{\ell=1}^{L}\sum_{i=1}^{N} \mathbf{\Phi}_i^\ell (\mathbf{\Phi}_i^\ell)^\top \boldsymbol{\alpha}_{ij}^\ell \right\|_2^2 \\
&\qquad + \lambda \sum_{j=1}^{N}\sum_{i=1}^{N} \sqrt{\boldsymbol{\alpha}_{ij}^\top \mathbf{\Phi}_i^\ell (\mathbf{\Phi}_i^\ell)^\top \boldsymbol{\alpha}_{ij}^\ell}.
\end{aligned} \tag{11}$$

Clearly, each entry of $\mathbf{\Phi}_i^\ell (\mathbf{\Phi}_i^\ell)^\top$ constitutes an inner product in the "lifted" space; i.e., $[\mathbf{\Phi}_i^\ell (\mathbf{\Phi}_i^\ell)^\top]_{k,l} = (\boldsymbol{\phi}^\ell(y_{ik}))^\top \boldsymbol{\phi}^\ell(y_{il})$. Defining the set of kernel matrices $\{\mathbf{K}_i^\ell \in \mathbb{R}^{M \times M}\}_{i=1}^N$, with $\mathbf{K}_i^\ell := \mathbf{\Phi}_i^\ell (\mathbf{\Phi}_i^\ell)^\top$, it is possible to recast the estimator (9), so that all dependencies on the functions $\{\phi^\ell(.)\}$ are captured through entries of $\mathbf{K}_i^\ell$, for $i = 1, \ldots, N$. Accordingly, one obtains the following estimator of the coefficient vector in (10) for each $\ell = 1, \ldots, L$

$$\begin{aligned}
\{\hat{\boldsymbol{\alpha}}_{ij}^\ell\} = \underset{\hat{\boldsymbol{\alpha}}_{ii}^0 = \mathbf{0}, \{\boldsymbol{\alpha}_{ij}^\ell\}}{\arg\min} \; & (1/2)\|\mathbf{Y} - \sum_{\ell=1}^{L} \tilde{\mathbf{K}}^\ell \mathbf{W}_\alpha^\ell\|_F^2 \\
& + \lambda \sum_{\ell=0}^{L}\sum_{j=1}^{N}\sum_{i=1}^{N} \|(\mathbf{K}_i^\ell)^{1/2} \boldsymbol{\alpha}_{ij}^\ell\|_2
\end{aligned} \tag{12}$$

where $\tilde{\mathbf{K}}^\ell := [\mathbf{K}_1^\ell, \ldots, \mathbf{K}_N^\ell]$, $\mathbf{W}_\alpha^\ell := [\boldsymbol{\alpha}_0^\ell, \ldots, \boldsymbol{\alpha}_N^\ell]$, and $\boldsymbol{\alpha}_j^\ell := [(\boldsymbol{\alpha}_{1j}^\ell)^\top, \ldots, (\boldsymbol{\alpha}_{Nj}^\ell)^\top]^\top$. Examination of (12) reveals that $\mathbf{W}_\alpha$ inherits the block-sparse structure of $\mathbf{W}$; that is, $\mathbf{w}_{ij}^\ell = \mathbf{0} \Leftrightarrow \boldsymbol{\alpha}_{ij}^\ell = \mathbf{0}$.

For ease of exposition, let the equality constraints ($\boldsymbol{\alpha}_{jj}^\ell = \mathbf{0}$) temporarily remain implicit. Introducing the change of variables $\boldsymbol{\gamma}_{ij}^\ell = (\mathbf{K}_{ij}^\ell)^{1/2} \boldsymbol{\alpha}_{ij}$, problem (12) can be equivalently recast as

$$\begin{aligned}
\underset{\{\boldsymbol{\alpha}_{ij}^\ell\}}{\arg\min} \quad & (1/2)\|\mathbf{Y} - \textstyle\sum_{\ell=0}^{L} \tilde{\mathbf{K}}^\ell \mathbf{W}_\alpha^\ell\|_F^2 + \textstyle\sum_{\ell=0}^{L} g(\mathbf{\Gamma}^\ell) \\
\text{s.t.} \quad & \boldsymbol{\gamma}_{ij}^\ell - (\mathbf{K}_{ij}^\ell)^{1/2} \boldsymbol{\alpha}_{ij}^\ell = \mathbf{0} \quad \forall i,j,\ell
\end{aligned} \tag{13}$$

where $\mathbf{\Gamma}^\ell := [\boldsymbol{\gamma}_1^\ell, \ldots, \boldsymbol{\gamma}_N^\ell]$, $\boldsymbol{\gamma}_j^\ell := [(\boldsymbol{\gamma}_{1j}^\ell)^\top, \ldots, (\boldsymbol{\gamma}_N^\ell)^\top]^\top$, and $g(\mathbf{\Gamma}^\ell) := \lambda \sum_{i=1}^{N}\sum_{j=1}^{N} \|\boldsymbol{\gamma}_{ij}^\ell\|_2$ is the nonsmooth regularizer.

Given matrices $\mathbf{Y}$ and $\tilde{\mathbf{K}}^\ell$, the next section capitalizes on convexity to develop efficient topology inference algorithms to estimate $\{\boldsymbol{\alpha}_{ij}^\ell\}$. Proximal-splitting approaches have been shown useful for convex optimization when the cost function comprises both smooth and nonsmooth components [5]. Prominent among these is the alternating direction method of multipliers (ADMM), see e.g., [17] for an early application of ADMM to distributed estimation. In the next section, a novel approach leveraging ADMM iterations will be introduced to solve (13), and henceforth identify the graph topology.

4.1. ADMM solver

Let $\mathbf{D}^\ell$:=Bdiag$(\mathbf{K}_1^\ell, \ldots, \mathbf{K}_N^\ell)$, and $\mathbf{D}$:=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 (13) as

$$\begin{aligned}
\mathcal{L}_\rho(\mathbf{W}_\alpha, \mathbf{B}, \mathbf{\Gamma}, \mathbf{\Xi}) &= (1/2)\|\mathbf{Y} - \tilde{\mathbf{K}}\mathbf{W}_\alpha\|_F^2 + g(\mathbf{\Gamma}) \\
&+ \langle \mathbf{\Xi}, \mathbf{D}^{1/2}\mathbf{W}_\alpha - \mathbf{\Gamma} \rangle + (\rho/2)\|\mathbf{\Gamma} - \mathbf{D}^{1/2}\mathbf{W}_\alpha\|_F^2
\end{aligned} \tag{14}$$

where $\tilde{\mathbf{K}} := [\tilde{\mathbf{K}}^0, \ldots, \tilde{\mathbf{K}}^L]$, $\mathbf{W}_\alpha := [(\mathbf{W}_\alpha^0)^\top, \ldots, (\mathbf{W}_\alpha^L)^\top]^\top$, and $\mathbf{\Gamma} := [(\mathbf{\Gamma}^0,)^\top \ldots, (\mathbf{\Gamma}^L)^\top]^\top$. Note that $\mathbf{\Xi}$ is a matrix of dual variables that collects Lagrange multipliers corresponding to the equality constraints introduced in (13), $\langle \mathbf{P}, \mathbf{Q} \rangle$ denotes the inner product between $\mathbf{P}$ and $\mathbf{Q}$, while $\rho > 0$ is prescribed a priori as a penalty parameter. ADMM boils down to a sequence of alternating minimization (AM) iterations to minimize $\mathcal{L}_\rho(\mathbf{W}_\alpha, \mathbf{B}, \mathbf{\Gamma}, \mathbf{\Xi})$ over the primal variables $\mathbf{W}_\alpha, \mathbf{B}$, and $\mathbf{\Gamma}$, followed by a gradient ascent step over the dual variables $\mathbf{\Xi}$; see also [1, 17]. Per iteration $k+1$, this entails the following provably-convergent steps, see e.g. [17]

$$\mathbf{W}_\alpha[k+1] = \underset{\mathbf{W}_\alpha}{\arg\min} \;\; \mathcal{L}_\rho(\mathbf{W}_\alpha, \mathbf{\Gamma}[k], \mathbf{\Xi}[k]) \tag{15a}$$

$$\boldsymbol{\Gamma}[k+1] = \arg\min_{\boldsymbol{\Gamma}} \ \mathcal{L}_\rho(\mathbf{W}_\alpha[k+1], \boldsymbol{\Gamma}, \boldsymbol{\Xi}[k]) \tag{15b}$$

$$\boldsymbol{\Xi}[k+1] = \boldsymbol{\Xi}[k] + \rho(\mathbf{D}^{1/2}\mathbf{W}_\alpha[k+1] - \boldsymbol{\Gamma}[k+1]). \tag{15c}$$

Focusing on $\mathbf{W}_\alpha[k+1]$, note that (15a) decouples across columns of $\mathbf{W}_\alpha$, and admits closed-form, parallelizable solutions. Incorporating the structural constraint $\boldsymbol{\alpha}_{jj}^0 = \mathbf{0}$, one obtains the following decoupled subproblem per column j

$$\tilde{\boldsymbol{\alpha}}_j[k+1] = \arg\min_{\tilde{\boldsymbol{\alpha}}_j} \ (1/2)\tilde{\boldsymbol{\alpha}}_j^\top \left(\tilde{\mathbf{K}}_j^\top \tilde{\mathbf{K}}_j + \rho\mathbf{D}_j\right)\tilde{\boldsymbol{\alpha}}_j - \tilde{\boldsymbol{\alpha}}_j^\top \mathbf{q}_j[k] \tag{16}$$

where $\tilde{\boldsymbol{\alpha}}_j$ denotes the $(NL-1)T \times 1$ vector obtained by removing entries of the j-th column of $\mathbf{W}_\alpha$ indexed by $\mathcal{I}_j := \{(j-1)T+1, \ldots, jT\}$. Similarly, $\tilde{\mathbf{K}}_j$ collects columns of $\tilde{\mathbf{K}}$ excluding the columns indexed by $\mathcal{I}_j$, the block-diagonal matrix $\mathbf{D}_j$ is obtained by eliminating rows and columns of $\mathbf{D}$ indexed by $\mathcal{I}_j$, while $\mathbf{q}_j[k]$ is constructed by removal of entries indexed by $\mathcal{I}_j$ from $\rho\mathbf{D}^{1/2}\boldsymbol{\gamma}_j[k] + \tilde{\mathbf{K}}^\top\tilde{\mathbf{y}}_j - \mathbf{D}^{1/2}\boldsymbol{\xi}_j[k]$, with $\boldsymbol{\xi}_j[k]$ denoting the j-th column of $\boldsymbol{\Xi}[k]$. Assuming $\left(\tilde{\mathbf{K}}_j^\top\tilde{\mathbf{K}}_j + \rho\mathbf{D}_j\right)$ is invertible, the per-column subproblem (16) admits the following closed-form solution per j

$$\tilde{\boldsymbol{\alpha}}_j[k+1] = \left(\tilde{\mathbf{K}}_j^\top\tilde{\mathbf{K}}_j + \rho\mathbf{D}_j\right)^{-1}\mathbf{q}_j[k]. \tag{17}$$

On the other hand, (15b) can be solved per component vector $\boldsymbol{\gamma}_{ij}^\ell$, and a closed-form solution can be obtained via the so-termed *group shrinkage* operator for each i and j, namely,

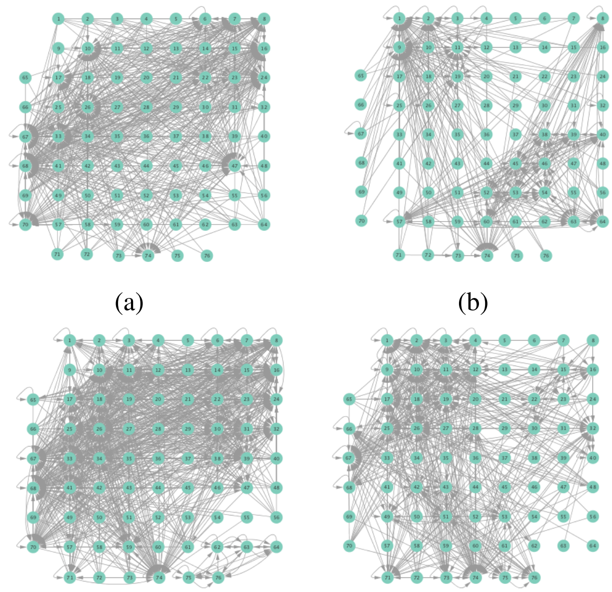
$$\boldsymbol{\gamma}_{ij}^\ell[k] = \mathcal{P}_{\lambda/\rho}\left((\mathbf{K}_i^\ell)^{1/2}\boldsymbol{\alpha}_{ij}^\ell[k+1] + \boldsymbol{\xi}_{ij}^\ell[k]/\rho\right) \tag{18}$$

where $\mathcal{P}_\lambda(\mathbf{z}) := (\mathbf{z}/\|\mathbf{z}\|_2)\max(\|\mathbf{z}\|_2 - \lambda, 0)$. 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 $\ell = 1, \ldots, L$.

5. NUMERICAL TESTS

Dataset description. Nodal time-series were obtained from an Epilepsy study reported in [14]. A 76-electrode grid was implanted into the subject's brain, and electrocorticographic (ECoG) time series spanning 0.5s were recorded per electrode (node). The goal of the experiment was to assess whether modeling nonlinearities, and adopting the novel algorithm would yield significant insights pertaining to *effective* dependencies between brain regions, that linear versions would otherwise fail to capture.

Results. Figure 1 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. Each node in the network is



(a) (b) (c) (d)

Fig. 1: Visualizations of networks inferred from ECoG data: (a) linear SVARM with $L = 1$ on preictal time series; (b) linear SVARM with $L = 1$ on ictal time series; (c) K-SVARM with $L = 1$ on preictal time series, using a polynomial kernel of order 2; (d) the same K-SVARM on ictal time series.

representative of an electrode. A cursory inspection of the visual maps reveals significant variations in connectivity patterns between ictal and preictal intervals. Specifically, networks inferred via the K-SVARM, 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. Clearly, acknowledging that interactions among brain regions may be driven by nonlinear dynamics, the novel nonlinear modeling framework facilitates discovery of global change in connectivity pattern that may not be captured by linear SVARMs.Due to space constraints, more extensive numerical tests can be found in an extended version of the paper [19].

6. CONCLUSIONS

This paper put forth a novel nonlinear SVARM framework for inference of sparse directed network topologies. Adopting a regularized LS estimator, and leveraging kernels, ADMM iterations were developed to estimate the network topology. Tests on real data from an Epilepsy study demonstrated that the novel approach is capable of identifying structural differences in the brain network that could not be captured by the linear model. Future directions include broadening the scope of the approach to dynamic networks, and distributed implementations that are suitable for large-scale networks.

7. REFERENCES

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

[2] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, Sep. 1987.

[3] G. Chen, D. R. Glen, Z. S. Saad, J. P. Hamilton, M. E. Thomason, I. H. Gotlib, and R. W. Cox, "Vector autoregression, structural equation modeling, and their synthesis in neuroimaging data analysis," *Computers in Biology and Medicine*, vol. 41, no. 12, pp. 1142–1155, Dec. 2011.

[4] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 8, Virginia, USA, 1998, pp. 127–132.

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

[6] C. W. Granger, "Some recent development in a concept of causality," *Journal of Econometrics*, vol. 39, no. 1, pp. 199–211, Sep. 1988.

[7] J. R. Harring, B. A. Weiss, and J.-C. Hsu, "A comparison of methods for estimating quadratic effects in nonlinear structural equation models." *Psychological Methods*, vol. 17, no. 2, pp. 193–214, Jun. 2012.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001, vol. 1.

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

[10] K. G. Jöreskog, F. Yang, G. Marcoulides, and R. Schumacker, "Nonlinear structural equation models: The Kenny-Judd model with interaction effects," *Advanced Structural Equation Modeling: Issues and Techniques*, pp. 57–88, Jan. 1996.

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

[12] G. V. Karanikolas, G. B. Giannakis, K. Slavakis, and R. M. Leahy, "Multi-kernel based nonlinear models for connectivity identification of brain networks," in *Proc. of Intl. Conf. on Acoust., Speech, and Signal Processing*, Shanghai, China, Mar. 2016, pp. 6315–6319.

[13] A. Kelava, B. Nagengast, and H. Brandt, "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, Jun. 2014.

[14] M. A. Kramer, E. D. Kolaczyk, and H. E. Kirsch, "Emergent network topology at seizure onset in humans," *Epilepsy Research*, vol. 79, no. 2, pp. 173–186, May 2008.

[15] N. Lim, F. d'Alché Buc, C. Auliac, and G. Michailidis, "Operator-valued kernel-based vector autoregressive models for network inference," *Machine learning*, vol. 99, no. 3, pp. 489–513, Jun. 2015.

[16] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using Granger causality and fMRI," *Neuroimage*, vol. 25, no. 1, pp. 230–242, Mar. 2005.

[17] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc WSNs with noisy links -Part I: Distributed estimation of deterministic signals," *IEEE Trans. Sig. Proc.*, vol. 56, pp. 350–364, Jan. 2008.

[18] Y. Shen, B. Baingana, and G. B. Giannakis, "Kernel-based structural equation models for topology identification of directed networks," 2016. [Online]. Available: http://arxiv.org/abs/1605.03122

[19] ——, "Nonlinear structural vector autoregressive models for inferring effective brain network connectivity," 2016. [Online]. Available: https://arxiv.org/abs/1610.06551

[20] X. Sun, "Assessing nonlinear granger causality from multivariate time series," in *Proc. Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, Antwerp, Belgium, Sep. 2008, pp. 440–455.

[21] G. Wahba, *Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1990.

[22] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc.*, vol. 68, no. 1, pp. 49–67, Feb. 2006.