MULTI-KERNEL BASED NONLINEAR MODELS FOR CONNECTIVITY IDENTIFICATION OF BRAIN NETWORKS

G. V. Karanikolas[1] *G. B. Giannakis*[1] *K. Slavakis*[2] *and R. M. Leahy*[3]

[1]Univ. of Minnesota, Dept. of Electrical and Comp. Eng., USA; Emails: {karan029,georgios}@umn.edu
[2]Univ. at Buffalo, SUNY, Dept. of Electrical Eng., USA; Email: kslavaki@buffalo.edu
[3]Univ. of Southern California, Signal & Image Processing Institute, USA; Email: leahy@sipi.usc.edu

ABSTRACT

Partial correlations (PCs) of functional magnetic resonance imaging (fMRI) time series play a principal role in revealing connectivity of brain networks. To explore nonlinear behavior of the blood-oxygen-level dependent signal, the present work postulates a kernel-based nonlinear connectivity model based on which it obtains topology revealing PCs. Instead of relying on a single predefined kernel, a data-driven approach is advocated to learn the combination of multiple kernel functions that optimizes the data fit. Synthetically generated data based on both a dynamic causal and a linear model are used to validate the proposed approach in resting-state fMRI scenarios, highlighting the gains in edge detection performance when compared with the popular linear PC method. Tests on real fMRI data demonstrate that connectivity patterns revealed by linear and nonlinear models are different.

Index Terms— fMRI, partial correlation, kernel-based regression, multiple kernel learning.

1. INTRODUCTION

Functional (f)MRI is a neuroimaging procedure for studying brain activity by detecting associated changes in blood oxygenation [10]. Event-related fMRI studies have contributed to our understanding of functional specialization. Although fMRI data analyses focused initially on the role of different brain regions, typically through univariate voxel-wise models of brain activation, much of the recent fMRI literature has focused on efforts to model and characterize the brain as a network [25]. This research has been facilitated by the permeation of network science methodologies to the study of brain connectivity, revealing characteristics such as their small-world structure, as well as the application of machine learning approaches to brain science [20, 4].

This work was supported by MURI AFOSR FA9550-10-1-0567, NIH 1R01GM104975-01, NSF grants 1343860, 1500713, 1514056. This research has been co-financed by the European Union (European Social Fund - ESF), and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the ESF.

Typical steps taken in obtaining brain connectivity graphs from a set of voxel timecourses include: (i) identifying a set of nodes [which may for example correspond to anatomically defined regions of interest (ROIs)]; (ii) assigning a representative timecourse per node; and (iii) inferring edges connecting graph nodes [20]. The last step, which is the focus of this work, typically involves the use of a functional connectivity measure to model dependencies between nodes.

Functional connectivity measures include but are not limited to (partial) correlations [18], coherence, and generalized synchronization [24]. As several studies suggest, the blood-oxygen-level dependent (BOLD) response is a nonlinear function of neuronal input signals [16]. Nonlinear connectivity measures such as mutual information, or, nonlinear variants of linear methods, such as kernel Granger causality (KGC) [14, 17] may therefore be preferable in revealing brain connectivity.

In PC-based identification of brain connectivity graphs, node time series are used to estimate their inverse covariance matrix by maximizing a likelihood criterion regularized with the elastic net [22], or, the ℓ_1-norm [24, 13] to promote sparse graphs. Entries of the estimated inverse covariance matrix correspond to the wanted PC coefficients.

The present contribution adopts a kernel-based nonlinear regression approach to estimate the PC coefficients. The premise is that performing regression with nonlinearly mapped versions of the time-series will offer improved fit of the real data relative to linear models. Intuitively, the proposed approach is motivated by the fact that linear models may be unable to sufficiently capture dependencies. In addition, the problem of choosing the kernel which is critical to the success of any kernel-based method is tackled using a data-driven methodology, namely multi-kernel learning, which learns a combination of kernels, taken from a preselected dictionary of kernel functions, to optimally fit the data. Finally, since PC yields weighted graphs, an edge inference procedure is also presented for identifying those edges which produce a binary graph.

Kernel-based methods have been employed in different fMRI tasks, including KGC and kernel canonical correlation

analysis [8]. Multi-kernel learning techniques have also been applied but in distinct contexts, namely for classification and feature selection [3, 11], or, data fusion from heterogeneous sources [26].

2. NONLINEAR CONNECTIVITY MODEL

Consider a set of nodes $\mathcal{V}$ each representing a collection of voxels belonging to either anatomically defined or data-driven regions [20]. Associated with each node $\nu \in \mathcal{V}$ is a time series represented by a column vector $\mathbf{x}_\nu := [x_\nu[1] \dots x_\nu[T]]^\top$ ($^\top$ stands for transposition), which is a combination of the BOLD fMRI time series of the voxels represented by this node. Based on all $\{\mathbf{x}_\nu\}_{\nu\in\mathcal{V}}$, the goal is to construct a graph $\mathcal{G}$ in which: i) an edge is present only if there is a sufficient level of "similarity" between the time series of the two incident vertices; and, ii) graph edges are indicative of direct influence between vertices rather than indirect influence through an intermediate node.

To this end, partial correlation will be used as a measure of similarity between nodes since it is both intuitively appealing and also has well-documented merits in fMRI-based connectivity studies [24]. To see how PC is indicative of direct influence, in contrast to simple correlation, consider a simple cascade network $\nu_A \to \nu_B \to \nu_C$ of three nodes. All three time series are correlated; hence, ordinary correlation will also suggest a $\nu_A\nu_C$ edge. However, if ν_B is regressed out of ν_A and ν_C, the correlation between ν_A and ν_C disappears, and so does the $\nu_A\nu_C$ edge [24].

Consider data vectors $\mathbf{x}_i, \mathbf{x}_j$ at nodes $i, j \in \mathcal{V}$, and an estimate $\hat{\mathbf{x}}_i$ of $\mathbf{x}_i$ based on $\{\mathbf{x}_k \mid k \in \mathcal{S}\}$, where $\mathcal{S} \subseteq \mathcal{V}\backslash\{i,j\}$, with $|\mathcal{S}| < |\mathcal{V}|$, $|\cdot|$ denoting set cardinality, and $\backslash$ representing set difference. Upon defining $\tilde{\mathbf{x}}_i := \mathbf{x}_i - \hat{\mathbf{x}}_i$, the sample PC of $\mathbf{x}_i, \mathbf{x}_j$ with respect to $\{\mathbf{x}_k\}_{k\in\mathcal{S}}$ is given by

$$\hat{\rho}_{ij|\mathcal{S}} := \frac{(\tilde{\mathbf{x}}_i - \bar{\tilde{\mathbf{x}}}_i)^\top(\tilde{\mathbf{x}}_j - \bar{\tilde{\mathbf{x}}}_j)}{\|\tilde{\mathbf{x}}_i - \bar{\tilde{\mathbf{x}}}_i\|_2 \|\tilde{\mathbf{x}}_j - \bar{\tilde{\mathbf{x}}}_j\|_2} \tag{1}$$

where $\bar{\tilde{\mathbf{x}}}_i := [(1/T)\sum_{t=1}^T \tilde{x}_i[t]]\mathbf{1}$, and $\mathbf{1}$ is the all-ones vector. Notice that $\hat{\rho}_{ij|\mathcal{S}} = \hat{\rho}_{ji|\mathcal{S}}$, and let hereafter $\mathcal{S} := \mathcal{V}\backslash\{i,j\}$. Inferring whether an edge is present or not between i and j entails a hypothesis test, that relies on a statistic expressed in terms of $\hat{\rho}_{ij|\mathcal{S}}$, as it will be demonstrated in Sec. 3. But first, a novel nonlinear approach to finding $\hat{\mathbf{x}}_i$ will be developed.

2.1. Kernel-based nonlinear predictors

Typically, $\hat{\mathbf{x}}_i$ is a linear function of $\{\mathbf{x}_k \mid k \in \mathcal{S}\}$. However, broadening the class $\hat{\mathbf{x}}_i$ belongs to include nonlinear estimators will result in more general connectivity models. To render nonlinear estimators tractable, our approach here relies on kernel-based regression.

Let indices $\{n_{1\backslash ij}, \dots, n_{|V|-2\backslash ij}\}$ enumerate nodes in $\mathcal{S} = \mathcal{V}\backslash\{i,j\}$, and $\boldsymbol{\chi}_{\backslash ij}[t] := [x_{n_{1\backslash ij}}[t], \dots, x_{n_{|V|-2\backslash ij}}[t]]^\top$ denote the snapshot of the whole network at time t, excluding the data observed at nodes i and j. A nonlinear mapping is introduced via $\boldsymbol{\phi} : \mathbb{R}^{|V|-2} \to \mathcal{H}$, that maps vectors $\boldsymbol{\chi}_{\backslash ij}[t]$ to elements of a (potentially infinite-dimensional) Hilbert space $\mathcal{H}$, equipped with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_{\mathcal{H}}$ [23]. The tth entry of $\mathbf{x}_i$ is modeled as $x_i[t] = \langle\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[t]), \boldsymbol{\beta}_i\rangle + \epsilon_i[t]$, where $\boldsymbol{\beta}_i \in \mathcal{H}$ and $\epsilon_i[t]$ captures noise and modeling inaccuracies. Gathering all regressors $\{\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[t])\}_{t=1}^T$ into $\boldsymbol{\Phi}_{\backslash ij} := [\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[1]), \dots, \boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[T])]$, and defining the operation $\boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\beta}_i$ as the $T \times 1$ vector whose tth entry is $\langle\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[t]), \boldsymbol{\beta}_i\rangle$, the previous data generation model reduces to the compact form: $\mathbf{x}_i = \boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i$. Along the lines of ridge regression, the following *infinite-dimensional* optimization task is formulated

$$\begin{aligned}\hat{\boldsymbol{\beta}}_i &:= \arg\min_{\boldsymbol{\beta}_i\in\mathcal{H}} \|\mathbf{x}_i - \boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\beta}_i\|_2^2 + \lambda\|\boldsymbol{\beta}_i\|_{\mathcal{H}}^2 \\ &= \left(\boldsymbol{\Phi}_{\backslash ij}\boldsymbol{\Phi}_{\backslash ij}^\top + \lambda\mathbf{I}_{\mathcal{H}}\right)^{-1}\boldsymbol{\Phi}_{\backslash ij}\mathbf{x}_i \\ &= \frac{1}{\lambda}[\mathbf{I}_{\mathcal{H}} - \boldsymbol{\Phi}_{\backslash ij}(\lambda\mathbf{I}_T + \boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\Phi}_{\backslash ij})^{-1}\boldsymbol{\Phi}_{\backslash ij}^\top]\boldsymbol{\Phi}_{\backslash ij}\mathbf{x}_i \end{aligned} \tag{2}$$

where $\boldsymbol{\Phi}_{\backslash ij}\mathbf{x}_i := \sum_{t=1}^T x_i[t]\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[t])$; product $\boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\Phi}_{\backslash ij}$ denotes the composition of $\boldsymbol{\Phi}_{\backslash ij}^\top$ with $\boldsymbol{\Phi}_{\backslash ij}$; $\mathbf{I}_{\mathcal{H}}$ stands for the identity operator on $\mathcal{H}$; and the last equality relies on the matrix inversion lemma. The (t,τ)th entry of the $T\times T$ Gram matrix $\boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\Phi}_{\backslash ij}$ is given by $\langle\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[t]), \boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[\tau])\rangle$.

Whenever $\mathcal{H}$ is chosen to be a reproducing kernel Hilbert space (RKHS), there exists a unique *reproducing* kernel κ such that $\langle\boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[t]), \boldsymbol{\phi}(\boldsymbol{\chi}_{\backslash ij}[\tau])\rangle = \kappa(\boldsymbol{\chi}_{\backslash ij}[t], \boldsymbol{\chi}_{\backslash ij}[\tau])$ [23]. This way, simple evaluations of κ at the data-vectors $\{\boldsymbol{\chi}_{\backslash ij}[t]\}_{t=1}^T$ generate the entries of the *kernel* matrix $\mathbf{K}_{\backslash ij} = \boldsymbol{\Phi}_{\backslash ij}^\top\boldsymbol{\Phi}_{\backslash ij}$. Popular examples of kernel functions κ, which define uniquely their associated RKHSs $\mathcal{H}$, include the linear kernel $\kappa_l(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) := \boldsymbol{\chi}_1^\top\boldsymbol{\chi}_2$, and the Gaussian or radial basis function (RBF) [23]

$$\kappa_{\text{RBF}}^{(\sigma)}(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2) := e^{\frac{-\|\mathbf{x}_1-\mathbf{x}_2\|_2^2}{2\sigma^2}} \tag{3}$$

where σ^2 denotes the variance of the RBF (here, $\dim(\mathcal{H}) = \infty$). For the linear κ_l, $\boldsymbol{\phi}$ becomes the identity mapping, and (2) boils down to the standard (linear) ridge regression task.

Inserting $\mathbf{K}_{\backslash ij}$ into (2), the estimate $\hat{\mathbf{x}}_i = \boldsymbol{\Phi}_{\backslash ij}^\top\hat{\boldsymbol{\beta}}_i$ is

$$\begin{aligned}\hat{\mathbf{x}}_i &= \frac{\boldsymbol{\Phi}_{\backslash ij}^\top}{\lambda}\left[\mathbf{I}_{\mathcal{H}} - \boldsymbol{\Phi}_{\backslash ij}\left(\lambda\mathbf{I}_T + \mathbf{K}_{\backslash ij}\right)^{-1}\boldsymbol{\Phi}_{\backslash ij}^\top\right]\boldsymbol{\Phi}_{\backslash ij}\mathbf{x}_i \\ &= \mathbf{K}_{\backslash ij}(\mathbf{K}_{\backslash ij} + \lambda\mathbf{I}_T)^{-1}\mathbf{x}_i\,. \end{aligned} \tag{4}$$

So long as $\mathbf{K}_{\backslash ij}$ is available, (4) offers a closed-form expression of the kernel-based estimator $\hat{\mathbf{x}}_i$, which renders it computable via matrix operations even when $\dim(\mathcal{H}) = \infty$.

2.2. Multi-kernel based learning

Choosing a "good" kernel is an application-dependent art. A way to facilitate this selection is multi-kernel learning (MKL)

that deciphers κ from the data [7].

To this end, consider the following reformulation of (2)

$$\min_{\boldsymbol{\beta}_i \in \mathcal{H}} \sum_{t=1}^{T} \xi^2[t] + \lambda \|\boldsymbol{\beta}_i\|_{\mathcal{H}}^2$$
$$\text{s.to} \quad \left\{\xi[t] = x_i[t] - \langle \boldsymbol{\phi}(\boldsymbol{\chi}_{\setminus ij}[t]), \boldsymbol{\beta}_i \rangle\right\}_{t=1}^{T}. \tag{5}$$

Letting $\boldsymbol{\alpha} := [\alpha[1], \ldots, \alpha[T]]^\top$ denote the $T \times 1$ vector of Lagrange multipliers, the dual task of the primal convex program in (5) turns out to be[1]

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^T} -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{x}_i - \boldsymbol{\alpha}^\top \mathbf{K}_{\setminus ij} \boldsymbol{\alpha}. \tag{6}$$

There are several ways of combining multiple kernels [7]. Here, given a number P of user-defined reproducing kernel functions $\{\kappa_p\}_{p=1}^P$, a new kernel κ is obtained via the linear combination $\kappa := \sum_{p=1}^P \theta_p \kappa_p$, where $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_P]^\top \succeq \mathbf{0}$. Since $\boldsymbol{\theta} \succeq \mathbf{0}$, then κ is guaranteed to be reproducing [23]. With regards to kernel matrices, the previous combination translates to $\mathbf{K}_{\setminus ij} = \sum_{p=1}^P \theta_p \mathbf{K}_{p \setminus ij}$. Moreover, to avoid unbounded solutions, $\boldsymbol{\theta}$ is forced to satisfy the sphere constraint $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \Lambda$, for a pre-defined $\Lambda > 0$ and $\boldsymbol{\theta}_0 \in \mathbb{R}^P$. Summarizing, the sought weights belong to $\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^P : \boldsymbol{\theta} \succeq \mathbf{0}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq \Lambda\}$. Plugging now κ in (6), and minimizing the resulting cost w.r.t. $\boldsymbol{\theta}$, yields the following min-max task

$$\min_{\boldsymbol{\theta} \in \Theta} \max_{\boldsymbol{\alpha} \in \mathbb{R}^T} -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{x}_i - \sum_{p=1}^{P} \theta_p \boldsymbol{\alpha}^\top \mathbf{K}_{p \setminus ij} \boldsymbol{\alpha}$$
$$= \max_{\boldsymbol{\alpha} \in \mathbb{R}^T} -\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^\top \mathbf{x}_i + \min_{\boldsymbol{\theta} \in \Theta} -\boldsymbol{\theta}^\top \mathbf{v} \tag{7}$$

where the min-max theorem [9, §4.3] was used to interchange $\min_{\boldsymbol{\theta} \in \Theta}$ with $\max_{\boldsymbol{\alpha} \in \mathbb{R}^T}$, and $\mathbf{v} := [v_1, \ldots, v_P]^\top$ where $v_p := \boldsymbol{\alpha}^\top \mathbf{K}_{p \setminus ij} \boldsymbol{\alpha}$. For an arbitrarily fixed $\boldsymbol{\alpha}$, application of the Karush-Kuhn-Tucker (KKT) conditions on the convex program $\min_{\boldsymbol{\theta} \in \Theta} -\boldsymbol{\theta}^\top \mathbf{v}$ yields the optimal solution $\boldsymbol{\theta}_\star(\boldsymbol{\alpha}) := \boldsymbol{\theta}_0 + \Lambda \mathbf{v}(\boldsymbol{\alpha}) / \|\mathbf{v}(\boldsymbol{\alpha})\|_2$. Finally, plugging $\boldsymbol{\theta}_\star$ into (7) and solving the resultant maximization task w.r.t. $\boldsymbol{\alpha}$ yields the optimal solution $\boldsymbol{\alpha}_\star(\boldsymbol{\theta}_\star) := [\mathbf{K}_{\setminus ij}(\boldsymbol{\theta}_\star) + \lambda \mathbf{I}_T]^{-1} \mathbf{x}_i$. Along the lines of [5], the chain $\boldsymbol{\alpha} \mapsto \boldsymbol{\theta}_\star(\boldsymbol{\alpha}) \mapsto \boldsymbol{\alpha}_\star[\boldsymbol{\theta}_\star(\boldsymbol{\alpha})]$ leads to the iterative Algorithm 1 for solving (7). As output of Alg. 1, the resultant estimates $\{\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j\}_{i,j \in \mathcal{V}}$ are plugged into (1) to obtain the wanted PC coefficients $\hat{\rho}_{ij|\mathcal{S}}$.

3. EDGE INFERENCE

With $\mathcal{E} := \{(i,j) \in \mathcal{V} \times \mathcal{V} \mid \rho_{ij|\mathcal{S}} \neq 0\}$, the hypothesis testing problem for the potential edge (i,j) can be stated as

$$\mathrm{H}_0 : \rho_{ij|\mathcal{S}} = 0; \quad \mathrm{H}_1 : \rho_{ij|\mathcal{S}} \neq 0.$$

Edge inference is performed using as test statistic the Fisher-z transformation $z_{ij|\mathcal{S}} := (1/2) \ln[(1 + \hat{\rho}_{ij|\mathcal{S}})/(1 - \hat{\rho}_{ij|\mathcal{S}})]$

[1]Detailed derivations can be found in the journal version [12].

Algorithm 1 Multi-kernel learning

Require: $\boldsymbol{\theta}_0$, λ, Λ, η, $\{\kappa_p\}_{p=1}^P$.
for $(i,j) \in \mathcal{V} \times \mathcal{V}$, $(i < j)$, **do**
  for $l = 1, 2$ **do**
    $\mathbf{K}_{0 \setminus ij} := \sum_{p=1}^P [\boldsymbol{\theta}_0]_p \mathbf{K}_{p \setminus ij}$.
    $\hat{\boldsymbol{\alpha}}_i := (\mathbf{K}_{0 \setminus ij} + \lambda \mathbf{I}_T)^{-1} \mathbf{x}_i$.
    while $\|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}\|_2 \geq \epsilon$ **do**
      $\boldsymbol{\alpha} := \hat{\boldsymbol{\alpha}}_i$.
      $\mathbf{v} := [\boldsymbol{\alpha}^\top \mathbf{K}_{1 \setminus ij} \boldsymbol{\alpha}, \ldots, \boldsymbol{\alpha}^\top \mathbf{K}_{P \setminus ij} \boldsymbol{\alpha}]^\top$.
      $\boldsymbol{\theta} := \boldsymbol{\theta}_0 + \Lambda \mathbf{v} / \|\mathbf{v}\|_2$.
      $\mathbf{K}_{\setminus ij}^{(i)} := \sum_{p=1}^P \theta_p \mathbf{K}_{p \setminus ij}$.
      $\hat{\boldsymbol{\alpha}}_i := \eta \boldsymbol{\alpha} + (1 - \eta)(\mathbf{K}_{\setminus ij}^{(i)} + \lambda \mathbf{I}_T)^{-1} \mathbf{x}_i$.
    end while
    $i \leftrightarrow j$.
  end for
  $\hat{\mathbf{x}}_i := \mathbf{K}_{\setminus ij}^{(i)} \hat{\boldsymbol{\alpha}}_i$.
  $\hat{\mathbf{x}}_j := \mathbf{K}_{\setminus ij}^{(j)} \hat{\boldsymbol{\alpha}}_j$.
end for

of the estimated PC coefficients $\hat{\rho}_{ij|\mathcal{S}}$; see also [13]. It is assumed that $\{x_\nu[t]\}_{t=1}^T$ are i.i.d normal across time with $x_\nu[t] \sim \mathcal{N}(\mu_\nu, \sigma_\nu^2)$; and possibly correlated across nodes with $\sigma_{\nu\nu'} := \mathbb{E}\{(x_\nu[t] - \mu_\nu)(x_{\nu'}[t'] - \mu_{\nu'})\}$. Under this assumption, it follows that asymptotically (as $T \to \infty$) $z_{ij|\mathcal{S}}$ is zero-mean normal with variance $\sigma_{ij|\mathcal{S}}^2 = 1/[T - (|\mathcal{V}| - 2) - 3]$, $\forall (i,j)$. Since the problem of multiple testing arises here, false discovery rate (FDR) principles, and in particular the method of [1], is utilized to keep FDR $:= \mathbb{E}[\mathrm{FA}/D]$ below a desired threshold, where $\mathbb{E}$ denotes expectation, FA stands for false alarms (rejecting H_0 when it is in effect), and D denotes discoveries (rejecting H_0 regardless of it being true or not).

4. NUMERICAL TESTS

Synthetic fMRI datasets based on the dynamic causal modeling (DCM) fMRI forward model [6], which uses the nonlinear balloon model [2], were generated in order to assess the performance of the proposed method. The setup follows that of [24]. First, vector time series $\mathbf{z}(t) := [z_1(t), \ldots, z_{|\mathcal{V}|}(t)]^\top$, where $z_i(t)$ corresponds to node i, are generated based on the DCM neural network model $\dot{\mathbf{z}}(t) = \delta \mathbf{A} \mathbf{z}(t) + \mathbf{u}(t)$, where $\mathbf{A}$ stands for the network matrix whose entries model the "first-order connectivity among regions" [6], δ is a coefficient that adjusts neural lags (equal to 20 in the following tests), and $\mathbf{u}(t) := [u_1(t), \ldots, u_{|\mathcal{V}|}(t)]^\top$, where $u_i(t)$ denotes the input signal at node i. Each $z_i(t)$ is then fed into the nonlinear balloon model [2] for vascular dynamics. The model is simulated at a 5ms timescale, and each time series is sampled with TR=3s (sampling rate). As a result, the ith node time series $\mathbf{x}_i$, consisting of $T = 200$ time points, is obtained.

Specifically, $u_i(t) = \pi_i(t) + n_i(t)$, where $\pi_i(t)$ denotes a binary pulse train (20% average duty cycle) generated by

a Markov chain, so as to simulate resting state fMRI data [24], and $n_i(t)$ is white Gaussian noise of variance 10^{-2}. A 30×30 upper triangular matrix $\mathbf{A}$ was generated as in [24], with fixed diagonal elements $\mathbf{A}_{ii} = -1$ and randomly placed 100 non-zero entries drawn according to $\mathbf{A}_{ij} \sim \mathcal{U}(0.25, 0.6)$, where $\mathcal{U}(a, b)$ denotes the uniform distribution over $[a, b]$. The balloon model parameters $\{\alpha, \rho, \tau, V_0\}$ were obtained from a fit of experimental BOLD signals obtained under a 3T field to the model, as provided in [19], while $\{\kappa, \gamma\}$ were set according to the DCM priors [6]. A single linear kernel and 19 Gaussian kernels of (3) were used, with variances $\{\sigma_p^2\}_{p=1}^{19}$ taken from the interval $[10^{-6}, 1]$. Moreover, $\boldsymbol{\theta}_0 := \mathbf{1}_P$, $\eta := 0.5$, and the regularization parameters λ, Λ were chosen using k-fold cross-validation separately for each pair of nodes. For each (i, j), two pairs $\{\lambda_i^\star, \Lambda_i^\star\}, \{\lambda_j^\star, \Lambda_j^\star\}$ were obtained, each corresponding to the value of (λ, Λ), in the grid $\{0.1, 1, 10, 100\} \times \{10, 50, 100\}$, that minimizes the cross-validation error for nodes i and j, respectively.

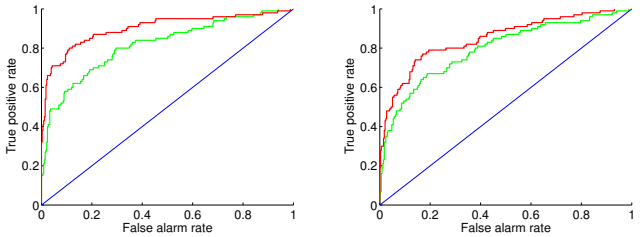
A linear model was also used to generate synthetic data, in order to compare the performance of the proposed method with that of linear PC when the underlying data indeed adhered to a linear model. In particular, time series were generated using a vector auto-regressive model: $\mathbf{z}[n+1] = \mathbf{A}\mathbf{z}[n] + \mathbf{w}[n]$, where $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, along the lines of [21]. Matrix $\mathbf{A}$ was generated using the procedure described previously, except for its diagonal entries which were set to $\mathbf{A}_{ii} = -0.1$. Each time step corresponds to 10ms and the resulting time series of each node is first convolved with a canonical hemodynamic response function (HRF), and then down-sampled to TR=3s. HRF parameters were selected as in [15].

The performance of the proposed approach was compared to that of linear PC on the DCM model through empirical receiver operating characteristic (ROC) curves, obtained using $|\hat{\rho}_{ij|\mathcal{S}}|$ as a test statistic and gradually decreasing the threshold. The improvement offered by the proposed method is evident in Fig. 1a. For example, correctly identifying 70% of the nonzero entries of the ground truth matrix $\mathbf{A}$ results in 13 false alarms (FAs) for the proposed approach, as compared to 69 FAs for the PC method.

With regards to the linear model case, the proposed approach outperformed linear PC (see Fig. 1b) by a smaller margin than in the DCM case. Such results address concerns of linear PC being superior whenever the underlying data are linearly generated. As expected, there was also an increase in the relative contribution of the linear kernel in this case.

Similar results, summarized in Table 1, were obtained in both simulation setups by fixing a maximum FDR level and using the procedure described in Sec. 3.

Tests were also performed on concatenated resting-state portions from the StarPlus fMRI real dataset [27]. Fig. 2 plots $|\hat{\rho}_{ij|\mathcal{S}} - \hat{\rho}_{ij|\mathcal{S}}^{(l)}|$, where $\hat{\rho}_{ij|\mathcal{S}}^{(l)}$ denotes the linear PC coefficient of $\mathbf{x}_i$ and $\mathbf{x}_j$. It is evident that linear and nonlinear models give rise to distinct values of PC coefficients. However, relative merits cannot be assessed in lieu of a ground truth model, which also justifies the role of synthetic data.



(a) DCM (b) Linear model

Fig. 1. ROC curves obtained on DCM and linearly-modeled synthetics. The red curve corresponds to the proposed approach whilst the green one to linear PC.

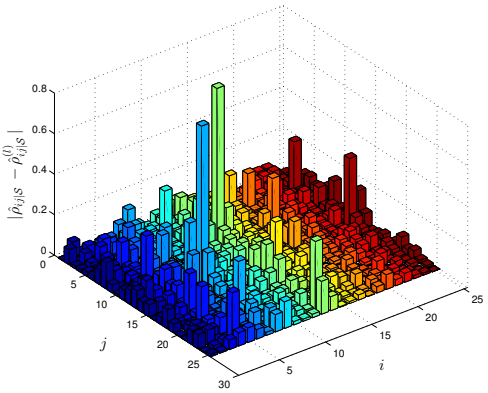


Fig. 2. 3D bar graph of $|\hat{\rho}_{ij|\mathcal{S}} - \hat{\rho}_{ij|\mathcal{S}}^{(l)}|$ obtained on the real data described at the end of Sec. 4.

	DCM		Linear	
	TPR(%)	FDR (%)	TPR(%)	FDR (%)
Proposed approach	65	10.96	45	16.67
Linear PC	49	25.76	37	22.92

Table 1. Testing the procedure of Sec. 3 for desired FDR level of 0.15, both for nonlinear DCM and linear-model-based synthetics. TPR denotes the true positive rate.

5. CONCLUSIONS

In par with the nonlinear generation mechanism of fMRI data, this work presented a novel kernel-based nonlinear connectivity model to infer topology-revealing partial correlations (PCs). A data-driven multi-kernel approach was introduced to learn the model that fits the data optimally. An edge inference procedure was also implemented to map the resultant soft weighted edges to binary-valued ones. Tests on synthetic data, both for linearly and nonlinearly generated data, highlight the superior performance of the proposed method over the popular linear PC alternative. On-going real data tests complementing the preliminary ones here, will be included in the final version and the upcoming journal version [12] of this contribution.

6. REFERENCES

[1] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[2] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: The balloon model," *Magnetic Resonance in Medicine*, vol. 39, no. 6, pp. 855–864, 1998.

[3] E. Castro, V. Gómez-Verdejo, M. Martínez-Ramón, K. A. Kiehl, and V. D. Calhoun, "A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: Application to schizophrenia," *NeuroImage*, vol. 87, pp. 1–17, 2014.

[4] Y.-T. Chang, D. Pantazis, and R. M. Leahy, "To cut or not to cut? Assessing the modular structure of brain networks," *NeuroImage*, vol. 91, pp. 99–108, 2014.

[5] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," in *Proc. Conf. on Uncertainty in Artificial Intelligence*, ser. UAI '09, Arlington, VA, USA, 2009, pp. 109–116.

[6] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[7] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[8] D. R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fMRI data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

[9] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I: Fundamentals*. Berlin: Springer, 1993, vol. 305.

[10] S. A. Huettel, A. W. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*. Sinauer Associates, 2004.

[11] B. Jie, D. Zhang, W. Gao, Q. Wang, C.-Y. Wee, and D. Shen, "Integration of network topological and connectivity properties for neuroimaging classification," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 576–589, 2014.

[12] G. V. Karanikolas, G. B. Giannakis, K. Slavakis, and R. M. Leahy, "Multi-kernel-based nonlinear models for connectivity identification of brain networks," *NeuroImage*, submitted 2016.

[13] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

[14] W. Liao, D. Marinazzo, Z. Pan, Q. Gong, and H. Chen, "Kernel Granger causality mapping effective connectivity on fMRI data," *IEEE Trans. Medical Imaging*, vol. 28, no. 11, pp. 1825–1835, 2009.

[15] M. A. Lindquist, J. M. Loh, L. Y. Atlas, and T. D. Wager, "Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling," *NeuroImage*, vol. 45, no. 1, pp. 187–198, 2009.

[16] N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, "Neurophysiological investigation of the basis of the fMRI signal," *Nature*, vol. 412, no. 6843, pp. 150–157, 2001.

[17] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Physical Review Letters*, vol. 100, no. 14, pp. 144–103, 2008.

[18] G. Marrelec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, J. Doyon, and H. Benali, "Partial correlation for functional brain interactivity investigation in functional MRI," *NeuroImage*, vol. 32, no. 1, pp. 228–237, 2006.

[19] T. Mildner, D. G. Norris, C. Schwarzbauer, and C. J. Wiggins, "A qualitative test of the balloon model for BOLD-based MR signal changes at 3T," *Magnetic Resonance in Medicine*, vol. 46, no. 5, pp. 891–899, 2001.

[20] J. Richiardi, S. Achard, H. Bunke, and D. Van De Ville, "Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience," *IEEE Signal Proc. Mag.*, vol. 30, no. 3, pp. 58–70, 2013.

[21] 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, 2005.

[22] S. Ryali, T. Chen, K. Supekar, and V. Menon, "Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty," *NeuroImage*, vol. 59, no. 4, pp. 3852 – 3861, 2012.

[23] B. Schölkopf and J. A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

[24] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fMRI," *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.

[25] O. Sporns, *Networks of the Brain*. MIT press, 2011.

[26] T. Zhang, D. Zhu, X. Jiang, B. Ge, X. Hu, J. Han, L. Guo, and T. Liu, "Predicting cortical ROIs via joint modeling of anatomical and connectional profiles," *Medical Image Analysis*, vol. 17, no. 6, pp. 601–615, 2013.

[27] Starplus fMRI data. [Online]. Available: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/