Chaos

Entropic regression with neurologically motivated applications

Cite as: Chaos **31**, 113105 (2021); https://doi.org/10.1063/5.0039333
Submitted: 02 December 2020 • Accepted: 14 October 2021 • Published Online: 08 November 2021

Jeremie Fish, Alexander DeWitt, Abd AlRahman R. AlMomani, et al.

Chaos ARTICLE scitation.org/journal/cha

Entropic regression with neurologically motivated applications

Cite as: Chaos **31**, 113105 (2021); doi: 10.1063/5.0039333
Submitted: 2 December 2020 · Accepted: 14 October 2021 ·
Published Online: 8 November 2021

Jeremie Fish,[1,2,a] Alexander DeWitt,[1,2] Abd AlRahman R. AlMomani,[1,2] Paul J. Laurienti,[3] and Erik Bollt[1,2]

AFFILIATIONS

[1]Department of Electrical and Computer Engineering, Clarkson University, Potsdam, New York 13699, USA
[2]Clarkson Center for Complex Systems Science (C^3S^2), Potsdam, New York 13699, USA
[3]Department of Radiology, Wake Forest School of Medicine, Winston-Salem, North Carolina 27101, USA

[a)]**Author to whom correspondence should be addressed:** fishja@clarkson.edu

ABSTRACT

The ultimate goal of cognitive neuroscience is to understand the mechanistic neural processes underlying the functional organization of the brain. The key to this study is understanding the structure of both the structural and functional connectivity between anatomical regions. In this paper, we use an information theoretic approach, which defines direct information flow in terms of causation entropy, to improve upon the accuracy of the recovery of the true network structure over popularly used methods for this task such as correlation and least absolute shrinkage and selection operator regression. The method outlined above is tested on synthetic data, which is produced by following previous work in which a simple dynamical model of the brain is used, simulated on top of a real network of anatomical brain regions reconstructed from diffusion tensor imaging. We demonstrate the effectiveness of the method of AlMomani *et al.* [Chaos **30**, 013107 (2020)] when applied to data simulated on the realistic diffusion tensor imaging network, as well as on randomly generated small-world and Erdös–Rényi networks.

The field of cognitive neuroscience seeks to understand the function of the brain, as related to the physical brain structure. Knowledge of the connectivity between functional regions is central to this comprehension, spurring the emergence of network neuroscience as a sub-field. However, these relationships are unknown *a priori*. Much of the work in this area, often called connectomics,[9] assesses so-called functional connectivity using well known methods for deduction from time series, including correlation and Least Absolute Shrinkage and Selection Operator (LASSO) regression. Effective connectivity methods, such as dynamic causal modeling or DCM,[10,11] are also popular and identify causal links but the methods do not scale well to whole brain analyses that are common in network neuroscience. Methods such as LASSO can lead to the recovery of a majority of true, causal connections with sufficient sample size. However, as we show, LASSO also infers a large fraction of false connections, which do not exist in the true network. To circumvent this issue, we utilize our recently developed entropic regression. Entropic regression is an information theoretic technique that is especially well suited for this problem, as it allows for optimal selection of basis functions as related to the underlying information flow of the dynamical system. We show that entropic regression yields a high recovery of true edges, while simultaneously limiting the number of falsely inferred connections, thus associated with excellent ROC (receiver operating characteristic curve) performance.

I. INTRODUCTION

Complex networks are all around us, from structural networks such as roads and flights,[1,2] to social networks such as Facebook and Twitter[3,4] to chemical networks, such as Belousov–Zhabotinsky oscillators,[5,6] biological systems,[7,8] and network neuroscience,[9,16,17] which we will be studying extensively here. We will consider a description of the brain as partitioned into 83 anatomical regions,[29] whose function of these interacting elements is to be described as a complex network structure.

Knowledge of the structure of the complex networks helps us to determine how they will respond to stresses, such as dynamical perturbations,[18,19] or to structural network perturbations.[20,21] However, frequently, the underlying network structure of a system is unknown. What may be available to us is time series data collected at each node. In the case of functional magnetic imaging (fMRI) data, time varying intensity is associated with each of the voxels, as

a three-dimensional movie. The associated metabolism of an active region of the brain becomes apparent by increased oxygen levels, resulting in the blood oxygenation level dependent (BOLD) signal that can be inferred by magnetic resonance.[22] These perturbations and interacting variations carry information about the underlying network connectivity of the brain. However, the inverse problem of inferring the underlying network connectivity from observations of the state is inherently an ill-posed problem that is sometimes called network tomography, and it remains a difficult problem today. This is true particularly for connectomics studies that attempt to infer the entire brain network using 100s or 1000s of network nodes.

In this work, we utilize a recently developed method named entropic regression[23] to recover the underlying network structure of synthetic data that are generated from a network Kuramoto model from three different types of networks from synthetic models to a real structural brain network: an Erdös–Rényi[26] (ER) network as a classical random graph model, a small-world network following[25] with designed structure, and the experimentally observed diffusion tensor imaging (DTI) network by methods as discussed below. Associated with Ref. 23, we soon thereafter made a general purpose code that we have previously made available described in Ref. 24 available as MATLAB .mlx style live-script self-explanatory codes at Github, https://github.com/almomaa/ERFit-Package, and also we produced a live-script specialized to producing our solutions associated with this current manuscript, found at Github https://github.com/almomaa/ERFit_Neurological_Applications.

On these network models of successive challenge and realism, we will simulate a popular synthetic dynamic to challenge the accuracy entropic regression. We simulate the first order network Kuramoto model, which has popularly been used including for the generation of synthetic fMRI data.[25] With the simulated data, we demonstrate that entropic regression accurately outperforms for network recovery as compared to other leading methods, including correlation and LASSO regression utilizing the Bayesian information criterion (BIC).

II. MATERIALS AND METHODS

A. Complex networks

For this work, we will use the terms graph and complex network interchangeably. A graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ is a set of nodes $\mathscr{V}$ and edges $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$. In context here, where we will associate each anatomical region of the brain as a node, the goal is to infer interactions as information flow between these regions. The unweighted structure of a graph of n nodes can be encoded by an $n \times n$ adjacency as follows:

$$[A]_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathscr{E}\ \forall (i \neq j), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

A *weighted* directed graph $\mathscr{G} = (\mathscr{V}, \mathscr{E}, \mathscr{W})$ has a weight for each edge, and it can be encoded by a weight matrix, $[W]_{ij} = w_{ij} sign([A]_{ij})$, with weights w_{ij}. A simple adjacency matrix is defined by a graph where no self-edges occur. A *complete* network is defined by the presence of all possible edges. We will generally allow for weighted directed graphs, which are well defined by a (possibly asymmetric) weight matrix W, which will be our goal to find from node-level observations only.

The degree distribution is often used to classify complex networks, even though it does not uniquely define its generative process. Recall that the *degree* of a node d_i is defined by $d_i = \sum_{j=1}^{n} [A]_{ij}$, and the degree distribution is the discrete probability function of d_i sampled across i. The Erdös–Rényi[26] (ER) graph, for example, is the classical random graph, wherein edges are assigned at random with probability p. It has been shown[27] in the ER graph, the degree distribution is Poisson in the limit of a large number of nodes. It became clear over time that real world networks tend to be highly clustered and yet have a small shortest path distance between any pair of nodes[28] and this led to the development of the Watts–Strogatz (WS) "community" graph. ER graphs do not exhibit the clustering or short path length that the community graphs do.

B. Structural brain networks

Simulated time series data were generated using an actual structural brain network but a synthetic model; thus, we call the data semi-synthetic, in addition to the synthetic complex networks with synthetic model dynamics, which we call fully synthetic data. The true brain networks were generated experimentally in previously published work[29] using diffusion tensor imaging (DTI) data. The networks generated in this prior work are publicly available and "Subject1" through "Subject20" were used in the current work. Details of the image processing and network generation can be found in that manuscript but are briefly described below. The Diffusion Toolkit (FDT) in the FSL software package[30] was utilized for image preprocessing and for diffusion tensor estimation. The cerebral gray matter was parcellated into 83 regions of interest (ROIs) based on the Lausanne anatomical atlas using Freesurfer software.[31] There were 41 ROIs in each hemisphere and a single brainstem region. The regions were warped to the native space of each study participant based on the anatomical image and then transformed to the DTI space using FSL. Deterministic tractography was performed using the Diffusion Toolkit software. All white matter voxels were seeded and resulting fibers were assessed to determine if they connected two of the ROIs. The ROIs served as the nodes in the adjacency matrix. An edge was present between two nodes if there was at least one white matter fiber connecting the nodes. A weighted adjacency matrix was generated by counting the number of fibers connecting two nodes and normalizing this value using the length of the fibers and the area of the ROIs as previously proposed.[32] The absence of a fiber connecting two ROIs was indicated in the adjacency matrix with a zero (0).

C. Correlation with thresholding

Among the oldest methods of network inference is correlation with thresholding. This method is still widely used in many fields, including in neuroscience.[12–14] Though it has been known since at least the time of Granger[15] that correlation is inadequate for discovery of causal relationships, we review it here due to its relative popularity in the neuroscience field.

Let $x_i \in \mathbb{R}^q$ represent length q time series and $\bar{x}_i \in \mathbb{R}$ represent the arithmetic mean of x_i. Then, the Pearson (linear) correlation

between x_i and x_j is

$$\mathrm{corr}(x_i, x_j) = \frac{\sum_{k=1}^{q}(x_i^{(k)} - \bar{x}_i)(x_j^{(k)} - \bar{x}_j)}{\left[\sum_{k=1}^{q}(x_i^{(k)} - \bar{x}_i)^2\right]^{\frac{1}{2}}\left[\sum_{k=1}^{q}(x_j^{(k)} - \bar{x}_j)^2\right]^{\frac{1}{2}}}. \quad (2)$$

The corresponding correlation matrix C will be

$$C_{ij} = \mathrm{corr}(x_i, x_j). \quad (3)$$

Network inference from the correlation matrix often proceeds by choosing a threshold value τ and declaring a relationship between the pair (i, j) to be causal if $C_{ij} > \tau$ or if $\mathrm{abs}(C_{ij}) > \tau$, where $\mathrm{abs}(\cdot)$ is the absolute value. We note that this formulation is incapable of finding directional causal relationships between variables. Choice of the threshold in the neurological context is discussed in Ref. 12.

D. Regression and regularization

Suppose a parametric model $y = x\beta$ is assumed for scalar real variable $y \in \mathbb{R}$ and $x \in \mathbb{R}^n$. Then, q samples may be stated by

$$Y = X\beta + \varepsilon, \quad (4)$$

with data matrices $X \in \mathbb{R}^{q\times n}$, $Y \in \mathbb{R}^q$ and here noise is assumed to be normal $\varepsilon \sim \mathcal{N}(0, \eta)$, gives $\beta \in \mathbb{R}^n$ unknown regressors. We assume enough samples so that the problem is over-determined, $q > n$. The goal in ordinary least squares (OLS) is to find the closest parametric fit to the data in the sense of square norm residual, so the minimizer of the loss function,[33,34]

$$\min_{\beta\in\mathbb{R}^n} ||X\beta - Y||_2. \quad (5)$$

A flaw of OLS, however, as noted in Ref. 35 is that while the OLS estimator has a low bias, it typically has a large variance.[35] It tends to suffer from over-fitting if model complexity is not well chosen.

Tikhonov regularization is a strategy to mitigate overfitting of OLS. The regularization term introduces a penalty term to Eq. (5). Let

$$\min_{\beta\in\mathbb{R}^n} ||X\beta - Y||_2^2 + R(\beta), \quad (6)$$

where $R(\beta)$ is a regularization term on β. OLS paired with common types of regularization are often referred to as ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO). In ridge regression, we choose

$$R(\beta) = \lambda||\beta||_2, \quad (7)$$

emphasizing reduced variance. In LASSO regression, we have

$$R(\beta) = \lambda||\beta||_1, \quad (8)$$

which emphasizes sparsity. Tikhonov regularization, in general, defines the selection of a favored solution among infinitely many ill-posed problems. In the ridge regression scenario, this can be thought of as filtering noise that is related to the small but nonzero singular values of the data matrix.[48]

The choice of the value of regularity parameter λ is an important practical aspect of regularization, as a poor choice for this value can lead to poor estimation results. There exist several methods for its choice, but here we choose the Bayesian information criterion (BIC),[49] which is one of the popular choices for selecting the value of λ. The BIC solution is considered to be similar to the maximum likelihood method for asserting the dimension of a model. The BIC estimation relies on the optimizing,[50]

$$\mathrm{BIC}(\beta, \lambda|X) = \kappa \ln(q) - 2\ln(\mathscr{L}(\beta, \lambda|X)), \quad (9)$$

where κ is the number of nonzero model parameters and $\mathscr{L}$ is the maximum likelihood solution of the given model with respect to the parameters β for a given fixed value of the parameter λ and given the data X. Then, the optimal value (λ_{opt}) can be found by minimizing the BIC with respect to λ for each β, that is,

$$\lambda_{opt} = \min_{(\beta,\lambda)} \mathrm{BIC}(\beta, \lambda). \quad (10)$$

This formulation introduces a penalty term on making the dimension of the model too large as well as a penalty for having a model with too many parameters, which generally provides popularly pleasing estimates of the parameter λ_{opt} when using LASSO.[35,49,50]

E. Entropic regression

Causal network inference from time series is an important and challenging task. Numerous methods exist for this problem, including Granger causality, convergent cross mapping, and information theoretic methods such as partial mutual information from mixed embedding.[36–38] We note that the performance of the best information theoretic methods have been shown to be comparable.[38] For this section, we will focus on entropic regression,[23] a method that has been adapted from optimal causation entropy.[43] It will be shown that entropic regression is capable of outperforming causation entropy on the task of network inference as will be shown below.

Here, we review entropic regression as developed in Ref. 23 specializing in the network discovery algorithm of optimal causation entropy,[42,43] which is an efficient and accurate method to infer a parametrically defined model, as a system identification problem but based on an information theoretic criterion. In entropic regression,[23] we use the conditional mutual information of time delayed measurements of the time series as to how they interact with basis functions, for an information theoretic criterion to iteratively select relevant basis functions. This approach is based on our prior work in causation entropy,[42,43] which is an information flow estimator that is capable of distinguishing direct vs indirect influence and in this way, it is a generalization of the concept of transfer entropy[44] that is meant only for component dynamics. The system identification problem can be stated in matrix form,

$$\dot{Z} = \Phi(Z)\Gamma, \quad (11)$$

where $Z \in \mathbb{R}^{q\times n}$ as before and is the measured state variables of the n-dimensional system with q observations, $\dot{Z}$ is the vector field estimated from Z, $\Phi : \mathbb{R}^{q\times n} \mapsto \mathbb{R}^{t\times K}$ is a function that maps the state variables Z, to the expanded set of candidate functions (not necessarily linear), and $\Gamma \in \mathbb{R}^{K\times n}$ is the parameters matrix. Given a basis set of functions $\Phi = \{\phi_i(Z)\}_{i=1}^K$, where a row vector Φ_j is given by $\Phi_j = \{\phi_1(Z_j), \phi_2(Z_j), \ldots, \phi_K(Z_j)\}$, and $\phi_i(Z_j) : \mathbb{R}^n \mapsto \mathbb{R}$, is a candidate function on the n-dimensional observation Z_j, that has a high flexibility of choice.

We illustrate the idea of entropic regression using a simple example. The most relevant (basis) functions of the observable variables, where most relevant refers to the most informative variable, of a most informative variable, which is measured by a mutual information objective and is discovered via a greedy search. For example, in this form, we can write Lorenz's equations, with its three-dimensional state vector $Z = (z_1, z_2, z_3)$, as

$$\dot{z}_1 = \eta(z_2 - z_1), \dot{z}_2 = z_1(\rho - z_3) - z_2, \dot{z}_3 = z_1 z_2 - b z_1. \quad (12)$$

Even though the Lorenz equation is nonlinear, it is a linear combination of nonlinear functions. Therefore, using the second-order power polynomial functions, we can see that

$$\Phi(Z = (z_1, z_2, z_3)) = \{1, z_1, z_2, z_3, z_1^2, z_1 z_2, z_1 z_3, z_2^2, z_2 z_3, z_3^2\}, \quad (13)$$

and we have ten candidate functions that contain linear and nonlinear terms (functions). However, the underlying dynamics of the Lorenz system is a linear combination of these candidate functions, and we see, for example, that $\dot{z}_1$ in the Lorenz system can be found as

$$\dot{z}_1 = \Phi(Z)[0, -\eta, \eta, 0, 0, 0, 0, 0, 0, 0]^T. \quad (14)$$

A key idea of system identification as stated through Eq. (11) is that linear combinations of basis functions may well define even a nonlinear system. The idea of entropic regression is to infer the exact interaction between the state variables Z, as measured through the basis functions, ϕ_i to the rates of change $\dot{Z}$ and this is associated with information flow that we define in Sec. III by a mutual information optimization process. The same applies for $\dot{z}_2$ and $\dot{z}_3$. This kind of linearization is generally known as Carleman linearization.[57] The reconstructed vector field using the least squares solution may be written as

$$V(\dot{Z}, \Phi) = \Phi \Phi^{\dagger} \dot{Z} = \Phi L(\dot{Z}, \Phi), \quad (15)$$

where $\Phi^{\dagger}$ is the pseudoinverse of the matrix Φ. The problem of regression for differential equation models Γ has recently become especially relevant for data-driven science[58–60] and recently popularized by sparse methods such as[61] optimization by LASSO regression, and by information theoretic methods.[23,42,43] For simplicity, we will write $\Phi(Z)$ as Φ in the following discussion. Our entropic regression has proved competitive[23] in comparison with other methods.

a. Entropic regression algorithm. The entropic regression involves an optimization method to associate data to a most informative and sparse set of basis functions as illustrated in Fig. 1. Most informative is interpreted in the sense of mutual information between projected basis function in the least squares fit sense [Eq. (15)] and the measured data. For simplicity, we will refer to the projected basis function in the least squares fit sense as PBFs (Projected Basis Functions). The underlying optimization proceeds in two stages: forward greedy search (selection) and (backward) possible elimination of PBF, and these are based on the conditional mutual information amongst competing observations through the selected PBF. In the forward stage, our objective is to select the subset $s \subset \mathscr{S} = \{1, 2, \ldots, K\}$, which represent strong candidate functions. Starting from empty set $s_0 = \emptyset$, the forward selection stage can be written as

$$u_k = \underset{i \in \mathscr{S}, i \notin s_{k-1}}{\arg\max}\, I(\dot{Z}; V(\dot{Z}, \Phi_i) | V(\dot{Z}, \Phi_{s_{k-1}})), \quad s_k = s_{k-1} \cup u_k, \quad (16)$$

where $k = 1, \ldots$ is the iteration index, u_k is the index with the maximum objective function value. Note that $s_0 = \emptyset \implies V(\dot{Z}, \Phi_{s_0}) = \emptyset$, which reduces the conditional $I(\cdot;\cdot|\cdot)$ to the mutual information $I(\cdot;\cdot)$. During the forward stage, at each iteration k and given the information $(V(\dot{Z}, \Phi_{s_{k-1}}))$ we already have from the set

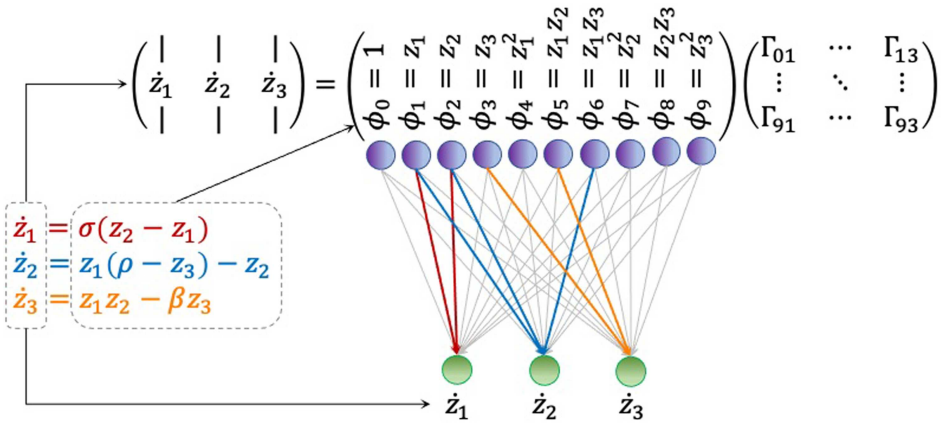

FIG. 1. The goal of entropic regression is to deduce which basis functions of the observational variables interact with the observed rates of changes of the variables. In the case of the Lorenz systems equation (12), the basis set Eq. (12) implies an interaction structure equation (14), which can be reconstructed equation (15). To find these generally sparse interactions, we developed a maximal mutual information principle stated in Eq. (16) but shown in summary here by the solid arrows between the rates $\dot{z}_1, \dot{z}_2, \dot{z}_3$, and the maximally informative functions. As it turns out in this example, this reconstructed set is in keeping with the true Lorenz equations.

s_{k-1}, we are looking for the function that maximally adds extra information to the model. The process terminates when either all PBF are exhausted (with maximum number of iterations equal K), or the reward function $I(\dot{Z}; V(\dot{Z}, \Phi_i)|V(\dot{Z}, \Phi_{s_{k-1}})) = 0$ (or thresholded to $\leq \varepsilon$), indicating that none of the remaining PBF are *relevant*, in an information theoretic sense. In other words, the process terminates when the strongest candidate has no further information beyond what we already have. Note that entropy, mutual information, and conditional mutual information can be estimated using any valid estimator. We choose the K-nearest neighbors estimator[39–41] for its accuracy, especially with relatively small sample sizes.

After the forward entropic regression, we have the set s that has the indices of the strong candidate functions. Eventually, s may have a few non-relevant functions that are selected due to a high degree of uncertainty and the rounding error at the end of forward entropic regression. Since we have reduced set ($|s| \ll K$), it will be computationally inexpensive to validate the accuracy of the model. This backward stage is an elimination stage, where the functions indexed by s are re-examined for their information theoretic relevance so that redundancy may be removed. In particular, we label the set s as initial set $s_0 = s$ for the backward stage, and we perform the following computations and updates:

$$u_k = \arg\min_{i \in s_{k-1}} I(\dot{Z}; \mathscr{V}(\dot{Z}, \Phi_i)|V(\dot{Z}, \Phi_{\{s_{k-1}-i\}})),$$
$$s_k = s_{k-1}/u_k. \tag{17}$$

In particular, note the differences of this stage regarding the index set, which is different between Eqs. (16) and (17). The backward stage includes a loss function, where at each iteration k, we examine what information will be lost if we remove the index i from the set s_{k-1} and the process terminates when $I(\dot{Z}; V(\dot{Z}, \Phi_i)|V(\dot{Z}, \Phi_{\{s_{k-1}-i\}})) > 0$. The result of the backward entropic regression is a set of indices s. We emphasize a practical strength that the forward entropic regression stage can substantially reduce the computational complexity of the backward stage, by limiting the elimination search space to a few candidate functions. However, the backward elimination is key in that in the special case of a low-dimension system, or we have efficient computational resources, we can plausibly skip the forward stage, to rely only on the backward stage directly, with initial set $s = \{1, 2, \ldots, K\}$. Parameters $\boldsymbol{\beta} \in \mathbb{R}^K$ can be found by updating the vector of zeros $\boldsymbol{\beta} = \mathbf{0}_K$ such that

$$\boldsymbol{\beta}_s = \mathscr{L}(\dot{X}, \Phi_s), \tag{18}$$

where $\boldsymbol{\beta}_s$ are the entries of $\boldsymbol{\beta}$ indexed by the elements of s. Thus, $\|\boldsymbol{\beta}\|_0 = |s|$. The primary role of entropic regression is to find the minimally optimal informative set of PBF. Once that set is identified, the parameter values themselves are easily found, for example, by ordinary least squares (OLS).

In general statistical estimators for information theoretic quantities must be dealt with to make a practical method. In theory, mutual information $I(x; y|z)$ is always non-negative and $I(x; y|z) = 0$ if and only if x and y are statistically independent given z. In practice, however, due to finite sample size and estimation inaccuracies, the estimated mutual information may be nonzero even when x and y are independent, and even worse, some otherwise useful favorite estimators may yield negative numbers.[40,45] We require a way to deduce with statistical confidence that they differ from zero to confidently decide whether x and y should be deemed to be independent given the estimated value of $I(x; y|z)$. In 46 and 43, a shuffle test with a "confidence" parameter $\alpha \in [0, 1]$ for tolerance estimation was developed. The shuffle test involves random shuffling of one of the variables, repeating n_s times, to build a test statistic. In particular, for the ith random shuffle, a random permutation $\pi^{(i)} : [T] \to [T]$ is generated to shuffle one of the variables, say y, which produces a new variable $(\tilde{y}^{(i)})$ where $\tilde{y}^{(i)} = y_{\pi^{(i)}}$; x and z are kept the same. Then, we estimate the mutual information $I(x; \tilde{y}^{(i)}|z)$ using the (partially) permuted variable $(x, \tilde{y}^{(i)}, z)$, for each $i = 1, \ldots, n_s$. For given α, we then compute a threshold value $I_\alpha(x; y|z)$ as the α-percentile from the values of $I(x; \tilde{y}^{(i)}|z)$. If $I(x; y|z) > I_\alpha(x; y|z)$, we determine x and y as dependent given z, otherwise independent. This threshold (tolerance) of mutual independence is adopted in the forward and backward stages as the termination condition. Hence, the tolerance describes the minimum effective quantity of information. In this sense, in forward entropic regression, we are selecting the functions, which add a significant quantity of information to the model, while in backward entropic regression, we are discarding functions which are determined to be negligible.

F. Kuramoto oscillators

The Kuromoto model is a popular model of network coupled phase oscillators.[51] While in the standard Kuromoto model the network is assumed to be a complete network,[51] but this is easily generalized by[47,51]

$$\dot{\theta}_i = \omega_i + \eta \sum_{j=1}^{n} a_{ij} \sin(\theta_j - \theta_i), \quad i = 1, \ldots n, \tag{19}$$

where $\theta_i(T)$ represents the phase angle of the i^{th} oscillator at time q, and ω_i represents the natural frequency and $a_{ij} \in A$. The ω_i are drawn from some distribution $g(\omega)$ that is usually assumed to be unimodal and symmetric about the mean value $\bar{\omega}$.[51] Let $\boldsymbol{\theta}^{(i)}(T_j) = (\theta_i(T_1), \theta_i(T_2), \ldots, \theta_i(T_q))^T$, with $(i = 1, \ldots, n)$, and time T_j for $(j = 1, \ldots, q)$ equally spaced times. The resulting trajectories data $Z = (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^n)$ is a $q \times n$, matrix of vector states as columns. The first order Kuramoto model will eventually result in a synchronous state almost certainly[47] if the value of coupling strength η is large enough. So, if $\eta > \eta_c$, the result will most likely synchronize for any initial condition. In Fig. 2, we show three different scenarios around the critical coupling strength (η_c). As $\eta \to \infty$,[47] the system tends to synchronize faster making the system indistinguishable sooner.

III. RESULTS

To demonstrate the effectiveness of entropic regression, we perform analysis on synthetic and semi-synthetic time series data. We know the ground truth for the underlying structural networks, so we can compare the effectiveness of entropic regression to other standard leaders based on the correlation method and LASSO regression.

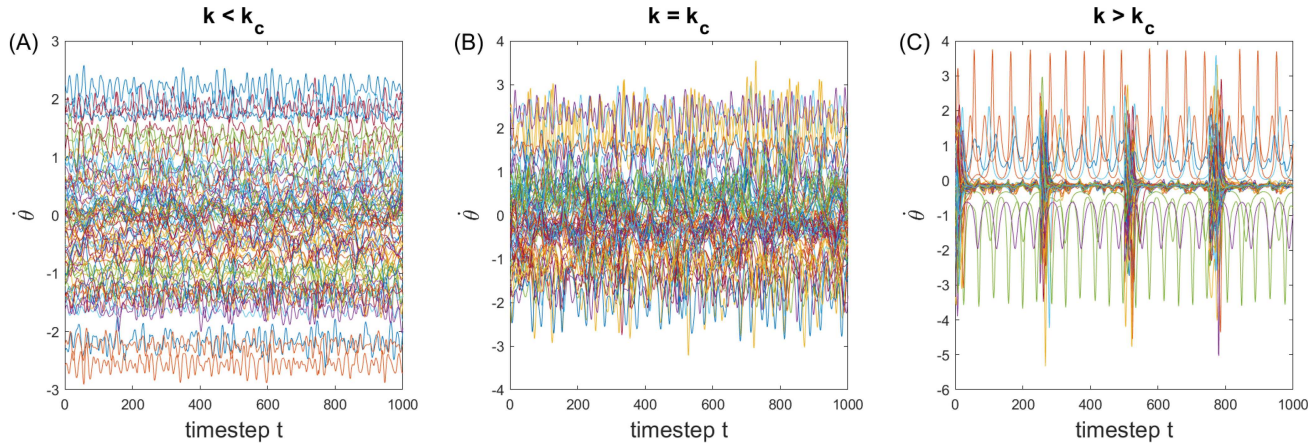

FIG. 2. The angular frequency $\dot{\theta}$ vs time from a Kuramoto network model coupled with the DTI adjacency matrix shown in Fig. 3(c). For values of the coupling strength η_c that are below the critical value η_c as in (a), the natural frequencies of the oscillators ω that are drawn from a normal distribution $\mathcal{N}(0, 1)$ have more variance than the mean field of the model. This results in loosely interacting individual oscillators. If the coupling η is greater than the critical value η_c, then the oscillators quickly fall into a periodic stable state. (b) shows the coupling that was chosen for our model. In this mode, the oscillators can be shown to be just inside the basin of attraction for the stable synchronous state, which is shown in (c); however, in practice, even without the re-sampling, this transient state is long lived when η is chosen to be η_c. Note that when $\eta > \eta_c$, the oscillators are rapidly attracted to the synchronous basin.

A. Synthetic data

For the testing purposes, we simulate data according to the first order network Kuramoto model as described in Eq. (19). We simulate the Kuramoto model different types of networks with approximately 80 nodes (the DTI network is 83 nodes, the other two types are set to 80 to be close to the same number of nodes), including the directed ER network, a special type of the community network and the DTI network, examples of which are shown in Fig. 3. The community network is a simple representation of the brain as seen in Ref. 25 and the DTI network as a realistic representation. We note that all methods perform poorly if the coupling coefficient η is set too high as a result of near immediate synchronization, thus rendering the nodes indistinguishable. Thus, the coupling strength is chosen to be nearby to its critical value, that is, $\eta \approx \eta_c$.

We ran 50 trials for each of the randomly generated types of adjacency matrix. The parameters of the Kuramoto model where held fixed throughout all trials. The critical coupling of the system is defined as $\eta_c = \frac{2}{\pi g(0)} \frac{1}{\lambda_1} N$.[51] For all trials, we set $\eta = \eta_c$, calculated at runtime. The function $g(\omega)$ denotes the pdf of ω across the network of oscillators, and λ_1 is the largest eigenvalue of the

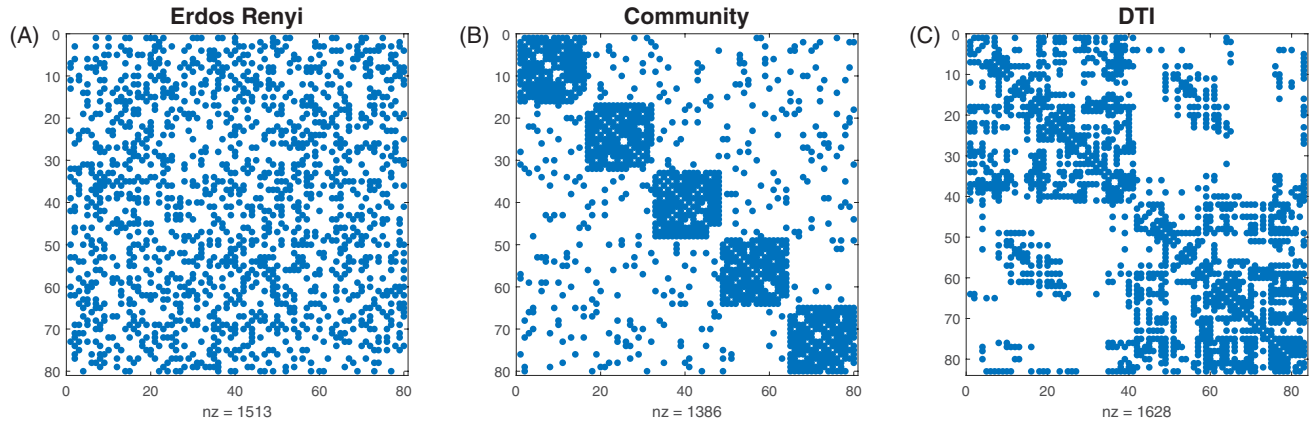

FIG. 3. We show an example spy plot of the adjacency matrices of (a) an Erdös–Rényi network, (b) a community network, and (c) a DTI network. The blue dots represent nonzero values of the matrix. As can be seen, the ER networks have edges assigned in a random manner, whereas community networks are highly structured. The DTI network exists somewhere in between these two network types.

corresponding adjacency matrix A.[47] The initial conditions $\boldsymbol{\theta_1}$ are randomly drawn from $\mathscr{U}[0, 2\pi]$, and the natural frequencies ω are drawn from $\mathscr{N}(0, 1)$. Each sample was integrated by an adaptive stepsize for target precision RK45 numerical solver, and solutions at 1001 equally spaced time steps were computed in the time range $0 \leq t \leq 1001$. Finally, θ is returned as modulated such that $0 \leq \theta < 2\pi$. The derivative $\dot{\theta}_i$ is estimated by finite differences, yielding 1000 estimated differences. For the Erdös–Rényi graph[26] [Fig. 3(a)], the sparsity was chosen such that it was as close as possible to the DTI graph (0.2363). These, along with choosing a coupling coefficient based upon the resultant structure, are experimental controls available to eliminate sparsity as a contributing factor in the outcome of the estimation. The community graph [Fig. 3(b)] is created with by adding edges randomly with equal weight as well; however, groups are chosen such that each node has a community. We chose 80 nodes with 5 communities of 16 nodes each. Each node has some number of intra-community connections and inter-community connections. We chose 13 intra-community connections and 5 inter-community connections. 13 and 5 were chosen so that the community graph has a maximum sparsity of 0.225 as close as possible to the DTI graph. The community graph is undirected, which increases the odds of not finding a configuration with the maximum number of edges, and as such, they tend to be sparser than the Erdös–Rényi graphs.

In Fig. 4, we show receiver operator characteristic (ROC) curves for a single realization of the network coupled dynamics discussed above, in which true positive and false positive rates are displayed together. It is desirable in an ROC curve for a method to come as close as possible to the upper left hand corner, representing almost no false positives and almost all true positives. In the three examples, networks from the Erdös–Rényi example, the community example shown in Fig. 3, and the DTI derived example, we compare the accuracy of standard methods such as correlation and LASSO to that of our entropic regression for these three scenarios. For the purposes of entropic regression, we let $Y^{(i)} = \theta^{(i)}(T_k)$ and $X^{(i)}_{kj} = \sin[\theta^{(j)}(T_k) - \theta^{(i)}(T_k)]$ $(\forall j \neq i)$, yielding different input data $Z^{(i)} = [Y^{(i)}, X^{(i)}] \in \mathbb{R}^{q\times n}$, for each variable $i = 1, \ldots, n$. All methods were tested on the same input data for each i. For entropic regression, we used the non-parametric KNN method to estimate conditional mutual information with $K = 1$. Clearly, the entropic regression outperforms correlation method. For both Gaussian and K-nearest neighbors (KNN) versions of oCSE we chose $\alpha = 0.01$, and for the KNN version of oCSE, $k = 10$ was chosen as this gave favorable performance over other tested values of k in the set $\{1, 2, \ldots, 20\}$. In the case of LASSO, the ROC curves are calculated across of a range of values of the regularization parameter λ. We see that for some λ that entropic regression clearly outperforms LASSO but for some λ, it is close to a tie, with Lasso slightly outperforming entropic regression at least for the Erdös–Rényi example. However, without a method of model selection, or otherwise knowing the answer *a priori*, it is not possible to know what is the better λ, the two leading methods being based on either cross validation or a Aikake information criterion (AIC)/BIC formalism already described, and we choose BIC optimum value λ_{opt} as discussed above by Eqs. (9) and (10). Assuming errors are Gaussian and i.i.d.; therefore, the term $\ln(\mathscr{L}(\beta, \lambda|X))$ can be approximated as

$$\ln(\mathscr{L}(\beta, \lambda|X)) = -\frac{q}{2}\ln(2q\pi\,\mathrm{MSE}), \tag{20}$$

where MSE denotes the mean squared error. We arrive at this approximation by noting that in the Gaussian case, we have $\mathrm{MSE} = \frac{\sigma^2}{q}$.[52] We combine this with the well known Gaussian log-likelihood[53] in the i.i.d. case and ignoring a constant term that does not affect the optimization of BIC. This yields

$$\mathrm{BIC} = \kappa\ln(q) + q\ln(2q\pi\,\mathrm{MSE}). \tag{21}$$

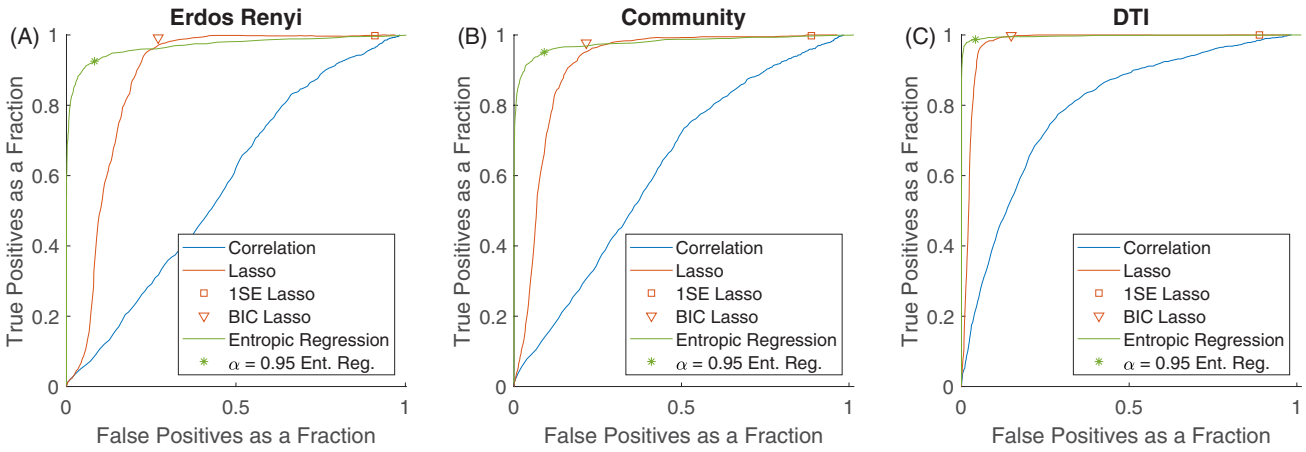

FIG. 4. Here, we show the ROC curves comparing the performance of LASSO, correlation, and entropic regression on (a) an Erdös–Rényi networks, (b) a community network, and (c) a DTI network. The solid lines represent the ROC curves and the star represents the parameter chosen for the entropic regression method, and the triangle represents the parameter λ chosen by the BIC method for LASSO. In all three types of network including the community network as discussed in Ref. 25, it is clear that entropic regression outperforms both correlation and LASSO. Furthermore, both entropic regression and LASSO offer significant performance benefits over correlation. As a result, the performance of entropic regression is much better than that of both LASSO and correlation.

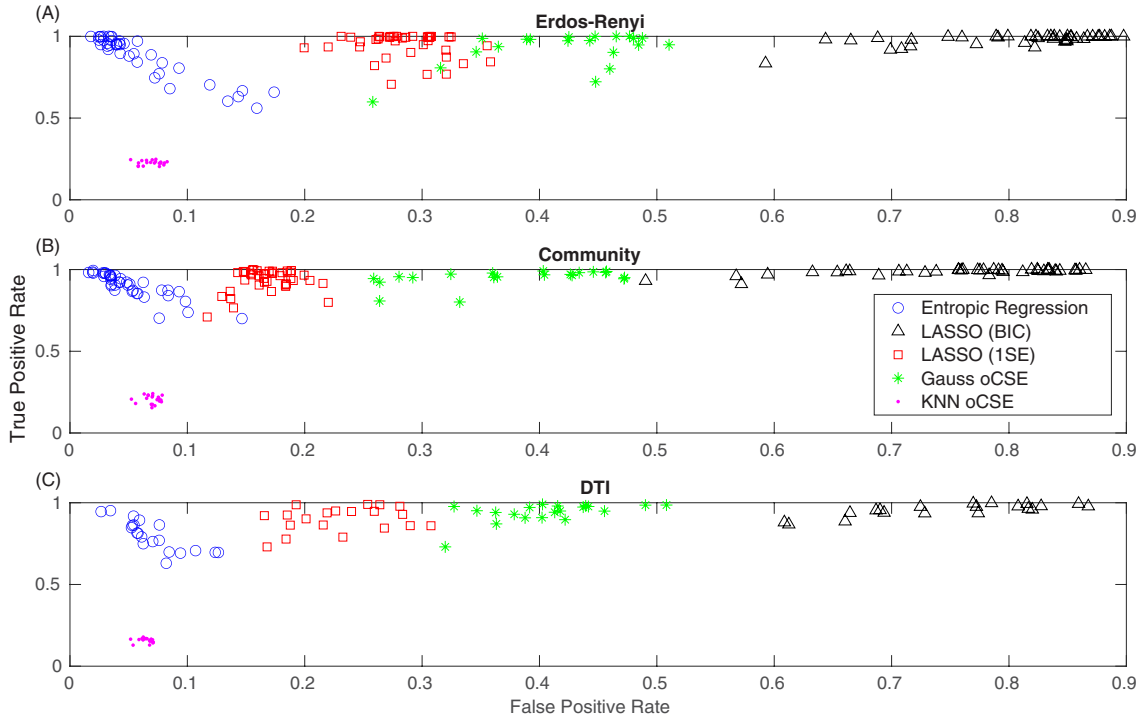

FIG. 5. Twenty runs on a different network for each run are shown above. In (a), there are 20 different random Erdös–Rényi networks with Kuramoto dynamics. The true positive rates and false positive rates are shown for entropic regression and LASSO choosing the appropriate value of λ either with BIC or using 1 standard error. (b) Random community networks; here, we additionally compare performance with two versions of oCSE as well, the Gaussian version and a non-parametric KNN version. As can be seen, neither version of oCSE outperforms either entropic regression or LASSO in this setting, KNN because it requires unrealistic amounts of data and the Gaussian version because the data are clearly not normally distributed. (c) DTI networks are used in the place of ER networks from (a). The DTI networks are reconstructed from 20 different patients in this case. It is clear that entropic regression clusters closest to the upper left hand corner, showing its improved performance over the other methods. These results are summarized in Table I.

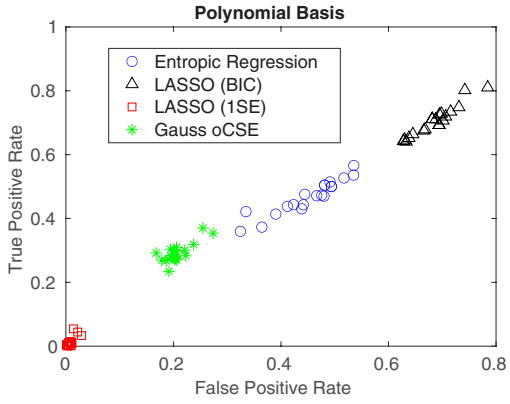

FIG. 6. Network estimation using a polynomial basis. As can be seen, all of the methods perform poorly for network estimation when using a polynomial basis, rather than choosing the appropriate basis of the coupling. In this figure, we show only a polynomial basis of a single term, as we found that including higher order terms led to an even larger FPR and TPR, as all models discovered more edges.

Estimates of the MSE follow a LASSO solution and using tenfold cross validation and optimization estimated for Eq. (21).

We can see that entropic regression exceeds the performance of the other methods as the method identifies parameters much closer to the upper left hand corner of the ROC curve than the other methods. For LASSO, we show the BIC estimation of the parameter λ_{opt}, which is marked by a triangle in Fig. 4. A common choice of confidence level is $\alpha = 95\%$, which is the value we chose, and is marked by a star. It is clear that entropic regression significantly outperforms the other methods.

Figure 5 shows the true positive rate (TPR) and false positive rate (FPR) over 20 runs on different networks of the types shown in Fig. 4. 20 ER and community networks were generated at random, while the DTI networks were reconstructed from 20 different patients. Only the point chosen by $\alpha = 0.95$ for entropic regression and λ chosen by BIC or one standard error (1SE) in LASSO are shown, rather than the full ROC. Entropic regression in this case clusters closest to the top left corner in all network types, making it preferable to all of the other methods including LASSO and oCSE. Across the 20 runs, the mean TPRs of LASSO (BIC) and entropic regression are similar, slightly favoring LASSO. Entropic

TABLE I. True and false positive rates (TPR and FPR) of entropic regression and LASSO averaged over 20 synthetic network realizations of either Erdös–Rényi or Community Network types, and 20 DTI networks constructed from different patients. On average, both LASSO and entropic regression return similar TPR; however, entropic regression returns a much lower FPR than LASSO. In the parenthesis, the standard deviation is reported.

	ER network	Community network	DTI network
Mean TPR Ent. Reg.	0.8938 (0.1035)	0.8858 (0.0892)	0.7983 (0.0940)
Mean TPR LASSO (BIC)	0.9538 (0.0571)	0.9150 (0.086)	0.8994 (0.0743)
Mean FPR Ent. Reg.	0.0535 (0.0301)	0.0551 (0.0305)	0.0718 (0.0268)
Mean FPR LASSO (BIC)	0.2844 (0.0410)	0.1647 (0.0247)	0.2350 (0.0440)

regression significantly outperforms LASSO (BIC) in FPR, similar to above.

In Fig. 6, we show the performance of the top three reviewed algorithms assuming a polynomial basis. As can be seen, the performance of all algorithms degrades in this scenario. However, we note that the brain is often modeled in this context with a sinusoidal coupling function[17,25,54–56] as it captures some of the observable features of the brain dynamics, which we believe justifies our choice of PBF shown in Fig. 5.

As seen in Fig. 3, graphs produced by DTI of real brains lie somewhere between ER and community in terms of ordered structure. For this reason, the synthetic brain activity data are more likely to match the dynamics exhibited by the brain. In all 20 examples of simulated dynamics from different DTI networks, entropic regression had a lower false positive rate than any of the networks produced by LASSO. This allows for increased confidence in the edges, which are inferred by entropic regression in the context of the brain. This highlights the utility of our approach over other existing methods in neurological applications.

LASSO in all cases averaged more than triple the FPR of entropic regression over the 20 networks examined. This could have severe implications for the network tomography. For example, a future treatment relying on accurate knowledge of connectivity between the ROIs would suffer greatly from having a significant number of false inferred edges.

In Table I, we average across 20 sample simulations and for each with a new sampled randomly generated network and randomly selected initial condition. The average TPRs and FPRs as well as the standard deviation of the two best methods are reported. As can be seen, on average, entropic regression and LASSO perform similarly in true positive rate (TPR); however, entropic regression significantly outperforms, especially in terms of false positive rate (FPR). Across the different k Overall, entropic regression is capable of recovering the majority of the true networks while only generating very few false edges, whereas LASSO also recovers the majority of the true network but introduces many more false edges.

IV. CONCLUSIONS

In this work, we have presented a simple model for brain dynamics using Kuramoto oscillators running on synthetic ER and DTI networks. Having access to accurate network structure is essential to understanding the dynamics of any network coupled system. Unfortunately, in the brain, the ground truth of the network structure is unknown and thus must be inferred. Commonly used methods for inference of functional connectivity-based networks from data include correlation and LASSO. We additionally compare performance to a recently developed information based inference method known as oCSE. We show that a new method, entropic regression, offers improvement upon the above listed methods in terms of accuracy. Specifically, entropic regression offers a similar true positive rate with a much improved false negative rate over the other methods.

ACKNOWLEDGMENTS

E.B., A.A.R.A., and A.D. were supported by the Army Research Office (N68164-EG), and J.F. and E.B. were supported by the DARPA.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts of interest to declare.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]B. F. Zhan and C. E. Noon, "Shortest path algorithms: An evaluation using real road networks," Transp. Sci. **32**, 65–73 (1998).
[2]P. Morrell and C. Lu, "The environmental cost implication of hub–hub versus hub by-pass flight networks," Transp. Res. Part D: Transp. Environ. **12**, 143–157 (2007).
[3]K. Lewis *et al.*, "Tastes, ties, and time: A new social network dataset using Facebook.com," Soc. Networks **30**, 330–342 (2008).
[4]B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network," in *IEEE Second International Conference on Social Computing* (IEEE, 2010), pp. 177–184.
[5]N. Tompkins, M. C. Cambria, A. L. Wang, M. Heymann, and S. Fraden, "Creation and perturbation of planar networks of chemical oscillators," Chaos **25**, 064611 (2015).
[6]K. Torbensen, F. Rossi, S. Ristori, and A. Abou-Hassan, "Chemical communication and dynamics of droplet emulsions in networks of Belousov-Zhabotinsky micro-oscillators produced by microfluidics," Lab Chip **17**, 1179–1189 (2017).
[7]E. de Silva and M. P. H. Stumpf, "Complex networks and simple models in biology," J. R. Soc. Interface **2.5**, 419–430 (2005).
[8]O. Mason and M. Verwoerd, "Graph theory and networks in biology," IET Syst. Biol. **1.2**, 89–119 (2007).
[9]O. Sporns, "Brain connectivity," Scholarpedia **2**, 4695 (2007).

[10]K. J. Friston, "Functional and effective connectivity in neuroimaging: A synthesis," Hum. Brain Mapp. **2**, 56–78 (1994).
[11]K. J. Friston, "Functional and effective connectivity: A review," Brain Connect. **1**, 13–36 (2011).
[12]K. Garrison *et al.*, "The (in)stability of functional brain network measures across thresholds," Neuroimage **118**, 651–661 (2015).
[13]D. Vidaurre, S. Smith, and M. W. Woolrich, "Brain network dynamics are hierarchically organized in time," Proceedings of the National Acadamies of Sciences (PNAS) **114**, 12827–12832 (2017).
[14]H. Chen *et al.*, "Alterations of brain network topology and structural-functional connectivity coupling in capsular versus pontine stroke," Eur. J. Neurol. **28**, 1967–1976 (2021).
[15]C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometr. J. Econ. Soc. **37**, 424–438 (1969).
[16]D. Bassett and E. D. Bullmore, "Small-world brain networks," Neuroscientist **12**, 512–523 (2006).
[17]D. Bassett *et al.*, "Dynamic reconfiguration of human brain networks during learning," Proc. Natl. Acad. Sci. U.S.A. **108**, 7641–7646 (2011).
[18]B. Barzel and A. Barabási, "Universality in network dynamics," Nat. Phys. **9**, 673–681 (2013).
[19]L. Wang *et al.*, "A geometrical approach to control and controllability of nonlinear dynamical networks," Nat. Commun. **7**, 1–11 (2016).
[20]E. I. Pas and S. L. Principio, "Braess' paradox: Some new insights," Transport. Res. B: Methodol. **31**, 265–276 (1997).
[21]Y. A. Korilis *et al.*, "Avoiding the Braess paradox in non-cooperative networks," J. Appl. Probab. **36**, 211–222 (1999).
[22]S. Ogawa *et al.*, "Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model," Biophys. J. **64**, 803–812 (1993).
[23]A. A. R. AlMomani, J. Sun, and E. Bollt, "How entropic regression beats the outliers problem in nonlinear system identification," Chaos **30**, 013107 (2020).
[24]A. A. AlMomani and E. Bollt, "ERFit: Entropic regression fit MATLAB package, for data-driven system identification of underlying dynamic equations," arXiv:2010.02411 (2020).
[25]B. J. Stoltz, H. A. Harrington, and M. A. Porter, "Persistent homology of time-dependent functional networks constructed from coupled time series," Chaos **27**, 047410 (2017).
[26]P. Erdös and A. Rényi, "On the evolution of random graphs," Publ. Math. Inst. Hung. Acad. Sci. **5**, 17–60 (1960).
[27]M. E. J. Newman, "Random graphs as models of networks," *Handbook of Graphs and Networks* (Wiley, 2003), Vol. 1, pp. 36–68.
[28]D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," Nature **393**, 440–442 (1998).
[29]L. Bonilha *et al.*, "Reproducibility of the structural brain connectome derived from diffusion tensor imaging," PLos One **10**, e0135247 (2015).
[30]T. E. Behrens *et al.*, "Characterization and propagation of uncertainty in diffusion-weighted MR imaging," Magn. Reson. Med. **50**, 1077–1088 (2003).
[31]B. Fischl *et al.*, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," Neuron **33**, 341–355 (2002).
[32]P. Hagmann *et al.*, "Mapping the structural core of human cerebral cortex," PLoS Biol. **6**, e159 (2008).
[33]A. E. Raftery, D. Madigan, and J. A. Hoeting, "Bayesian model averaging for linear regression models," J. Am. Stat. Assoc. **92**, 179–191 (1997).
[34]G. H. Golub and C. F. Van Loan, *Matrix Computations* (Johns Hopkins University Press, 2013).
[35]R. Tibshirani, "Regression shrinkage and selection via the LASSO," J. Roy. Stat. Soc. B **58**, 267–288 (1996).
[36]G. Sugihara *et al.*, "Detecting causality in complex ecosystems," Am. Assoc. Advance. Sci. **338**, 496–500 (2012).
[37]J. Runge *et al.*, "Inferring causation from time series in earth system sciences," Nat. Commun. **10**, 1–13 (2019).
[38]J. Kovřenek and J. Hlinka, "Causal network discovery by iterative conditioning: Comparison of algorithms," Chaos **30**, 013117 (2020).
[39]L. F. Kazachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," Problem. Pered. Inform. **23**, 9–16 (1987).
[40]A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," Phys. Rev. E **69**, 066138 (2004).
[41]M. Vejmelka and M. Paluš, "Inferring the directionality of coupling with conditional mutual information," Phys. Rev. E **77**, 026214 (2008).
[42]J. Sun, C. Cafaro, and E. M. Bollt, "Identifying the coupling structure in complex systems through the optimal causation entropy principle," Entropy **16**, 3416–3433 (2014).
[43]J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," SIAM J. Dyn. Syst. **14**, 73–106 (2015).
[44]T. Schreiber, "Measuring information transfer," Phys. Rev. Lett. **85**, 461 (2000).
[45]W. M. Lord, J. Sun, and E. M. Bollt, "Geometric k-nearest neighbor estimation of entropy and mutual information," Chaos **28**, 033114 (2018).
[46]J. Runge, J. Heitzig, N. Marwan, and J. Kurths, "Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy," Phys. Rev. E **86**, 061121 (2012).
[47]F. A. Rodrigues, T. K. D. M. Peron, P. Ji, and J. Kurths, "The Kuramoto model in complex networks," Phys. Rep. **610**, 1–98 (2016).
[48]G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," SIAM J. Matrix Anal. Appl. **21**, 185–194 (1999).
[49]G. Schwarz, "Estimating the dimension of a model," Ann. Stat. **6**, 461–464 (1978).
[50]E. Wit, E. Heuvel, and J. W. Romeijn, "'All models are wrong. . .': An introduction to model uncertainty," Stat. Neerl. **66**, 217–236 (2012).
[51]A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhao, "Synchronization in complex networks," Phys. Rep. **469**, 93–153 (2008).
[52]M. H. DeGroot, *Probability and Statistics* (Wesley Publishing Company, 1986).
[53]P. Zwiernik, C. Uhler, and D. Richards, "Maximum likelihood estimation for linear Gaussian covariance models," arXiv:1408.5604 (2014).
[54]J. L. P. Velazquez, "Brain research: A perspective from the coupled oscillators field," NeuroQuantology **4**, 155–165 (2006).
[55]M. Breakspear, S. Heitmann, and A. Daffertshofer, "Generative models of cortical oscillations: Neurobiological implications of the Kuramoto model," Front. Human Neurosci. **4**, 190 (2010).
[56]R. G. Andrzejak, C. Rummel, F. Mormann, and K. Schindler, "All together now: Analogies between chimera state collapses and epileptic seizures," Sci. Rep. **6**, 1–10 (2016).
[57]T. Carleman, "Application de la théorie des équations intégrales linéaires aux systèmes d'équations différentielles non linéaires," Acta Math. **59**, 63–87 (1932).
[58]J. P. Crutchfield and B. S. McNamara, "Equations of motion from a data series," Complex Syst. **1**, 417–452 (1987).
[59]C. Yao and E. M. Bollt, "Modeling and nonlinear parameter estimation with Kronecker product representation for coupled oscillators and spatiotemporal systems," Physica D **227**(1), 78–99 (2007).
[60]W.-X. Wang, Y.-C. Lai, and C. Grebogi, "Data based identification and prediction of nonlinear and complex dynamical systems," Phys. Rep. **644**, 1–76 (2016).
[61]S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proc. Natl. Acad. Sci. U.S.A. **113**(15), 3932–3937 (2016).