Echo State Network models for nonlinear Granger causality

Andrea Duggento*, Maria Guerrisi*, Nicola Toschi†,
*Department of Biomedicine and Prevention, University of Rome Tor Vergata, Rome, Italy
†Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Boston, MA, USA

***Abstract*—While Granger Causality (GC) has been employed in a large number of problems in network neuroscience, the vast majority of GC applications are based on linear multivariate autoregressive (MVAR) models. However, it is well known that real-life systems (and biological networks in particular) exhibit notable nonlinear behavior, hence undermining that validity of MVAR-based GC (MVAR-GC) approaches. Current nonlinear approaches to GC estimation only cater for additive nonlinearities or, alternatively, are based on recurrent neural networks (RNN) or Long short-term memory (LSTM) networks, which present considerable training difficulties and/or the need to be carefully tailored to specific applications. We define a novel approach to estimating nonlinear, directed within-network interactions based on a specific class of RNNs termed echo-state networks (ESN), where training is replaced by random initialization and the internal basis is built through orthonormal matrices. We reformulate the classical GC framework in terms of ESN-based models for arbitrarily complex networks, and characterize the ability of our ESN-based Granger Causality (ES-GC) estimator to capture nonlinear causal relations in a network of noisy Duffing oscillators, showing a net advantage of ES-GC in detecting nonlinear, causal links. We then explore the structure of ES-GC networks in the human brain employing functional MRI data from 1003 healthy subjects scanned at rest at 3T within the human connectome project, demonstrating the existence of previously unknown directed within-brain interactions. In summary, ES-GC performs better than commonly used and recently developed GC detection approaches, making it a valuable tool for the analysis of e.g. multivariate biological networks.**

I. Introduction

Multivariate Granger causality [1], [2] estimates how much the forecast of a timeseries can be improved by including information from the past of another timeseries, while accounting for additional, mutually interacting signals. It is defined in terms of conditional dependencies in the time or frequency domains [3], and can be considered an estimator for directed information flow between pairs of nodes (possibly) belonging to complex networks [4]. Granger Causality(GC)-based approaches, including the recent State Space (SS) (SS-GC) reformulation [5], have been employed in a vast number of problems which can be assimilated to network science, and the majority of CG applications are based on linear multivariate autoregressive (MVAR) models [2]. However, it is well known that real-life systems in general (and biological networks in particular) exhibit notable nonlinear behavior, hence undermining the validity of MVAR-based approaches in estimating GC (MVAR-GC) [6]. A typical case study is the analysis of brain networks from functional MRI (fMRI) signals, which result from convolving neural activity with a locally hemodynamic response function (HRF) [7], [8]. Here, a linear MVAR approach is not suitable for reconstructing neither the nonlinear components of neural coupling, nor the multiple nonlinearities and time-scales which concur to generating the signals. In this context, nonlinear extensions of MVAR models have been proposed, mostly based on kernel learning [9]. However, such models only cater for additive nonlinearities. In contrast, neural network (NN) models can be used to flexibly account for multiscale nonlinear dynamics and interactions [10]. Neural networks like multi-layer perceptions [11] or neural networks with non-uniform embeddings [12] have been used to introduce nonlinear estimation capabilities which also include "extended" GC [13] and wavelet-based approaches [14]. Also, recent preliminary work has employed deep learning to estimate bivariate GC interactions [15], convolutional neural networks to reconstruct temporal causal graphs [16] or Recurrent NN (RNN) with a sparsity-inducing penalty term to improve parameter interpretability [17], [18]. While RNNs provide flexibility and a generally vast modelling capability, RNN training can prove complex and their employment in real-world data, where data paucity is often an issue, may prove impractical and/or unstable. In this respect, a subclass of RNN, termed long-short term memory (LSTM) models have been designed to explicitly include a "forgetting element" [19] which facilitates training (see [20] and references therein for a general discussion of LSTM in various learning tasks), and one paper also employed LSTM models in brain connectivity estimation [21]. Still, successful design and training of both RNN and LSTM models requires memory-bandwidth-bound computation, involves in-depth tailoring to a specific application, and the final architecture is often defined through trial and error procedures.

In this paper, we introduce a novel approach to estimating nonlinear, directed within-network interactions while retaining ease of training and a good degree of generality. Our framework is based on a specific class of RNN termed echo-state networks (ESN) [22]. The peculiarity of ESN is that, contrary to the general RNN model, ESN weights are not trained but rather randomly initialized, after which a linear mixing matrix is employed to map internal states to predicted outputs. The main hypothesis is that a fixed but randomly connected RNN can provide output with a state space rich enough to provide flexible fitting capabilities while eliminating the training issues

Manuscript received May 15, 2019.
Corresponding author: A. Duggento (email: duggento@med.uniroma2.it).

common in RNNs. In addition, we modify and optimize the current ESN formulation to simultaneously model nonlinear, multivariate signal coupling while decoupling internal model representations into separate orthonormal weight matrices. We then reformulate the classical GC framework in terms of ESN-based models for multivariate signals generated by arbitrarily complex networks, and characterize the ability of our ESN-based Granger Causality (ES-GC) estimator to capture nonlinear causal relations by simulating multivariate coupling in a network of interacting, noisy Duffing oscillators. Synthetic validation shows a net advantage of ES-GC over other estimators in detecting nonlinear, causal links. We then explore the structure of EC-GC networks in the human brain employed functional MRI data from 1003 healthy subjects scanned at rest at 3T withing the human connectome project (HCP), demonstrating the existence of previously unknown directed within-brain interactions.

II. Methods

A. Granger causality

GC was introduced [23], [24] under the assumptions that a) the cause happens prior to its effect, and b) the cause contains unique information about the future of the effect. Under these assumptions, given a time-evolving system $\mathbf{u}(t)$ with L components $\{u_1, u_2, \dots u_L\}$, component u_j is said to be causal on component u_i ($u_j \to u_i$) if:

$$\mathcal{P}[u_i(t+1) \mid \mathcal{I}(\mathbf{u}(t))] \neq \mathcal{P}[u_i(t+1) \mid \mathcal{I}(\mathbf{u}^{(-j)}(t))] \quad (1)$$

where $\mathcal{P}$ is a probability density function, while $\mathcal{I}(\mathbf{u}(t))$ and $\mathcal{I}(\mathbf{u}^{(-j)}(t))$ denote (with loose notation) all information provided by $\mathbf{u}$ up to time t including or excluding component j, respectively.

A common simplification is that $\mathcal{I}(\mathbf{u}(t))$ can be represented by a MVAR process defined over $\mathbf{u}$. The inequality (1) is then replaced by a test of equality between the estimated variances of the two distributions, i.e. $\mathrm{Var}(\varepsilon') \neq \mathrm{Var}(\varepsilon)$, where ε and ε' are the prediction errors derived from the so called restricted model (RM) and an unrestricted model (UM), respectively:

$$\begin{aligned} \varepsilon'_i(t) &= u_i(t) - \left(\sum_{\tau=1}^{p} \mathbf{k'}_{i,\tau} \mathcal{L}^{\tau} \right) \mathbf{u}^{(-j)}(t) \\ \varepsilon_i(t) &= u_i(t) - \left(\sum_{\tau=1}^{p} \mathbf{k}_{i,\tau} \mathcal{L}^{\tau} \right) \mathbf{u}(t) \end{aligned} \quad (2)$$

where $\mathcal{L}^{\tau}\mathbf{u}(t) = \mathbf{u}(t-\tau)$ is the lag operator, the autoregressive order p is a suitably chosen parameter, and $\mathbf{k'}_{i,\tau}$ and $\mathbf{k}_{i,\tau}$ are to be estimated from data. Further, it is common practice to use the logarithm of the ratio of average squared residuals ($\epsilon_i = \langle \varepsilon_i^2 \rangle$) as a measure of MVAR-GC strength as follows: $s_{j\to i} = \log(\epsilon_i^{(-j)}/\epsilon_i)$ [1]. This measure also has a natural interpretation as the rate of "information transfer" between i and j and has been shown to be equal to the transfer entropy between i and j in the case of Gaussian variables [25]–[27]. Also, recently [5] the MVAR approach has been refined through a latent state-space (SS) model, where the inference of SS-GC is done over observables which are a linear mixture $\mathbf{v}(t)$ of the state variables $\mathbf{u}(t)$ with added white Gaussian noise: $\mathbf{v}(t) = A\mathbf{u}(t) + \xi(t)$ (where A is a mixing matrix). SS-GC has been extended in [28] to define multiscale causality, and has been shown to augment performance when classical MVAR methods fail [29].

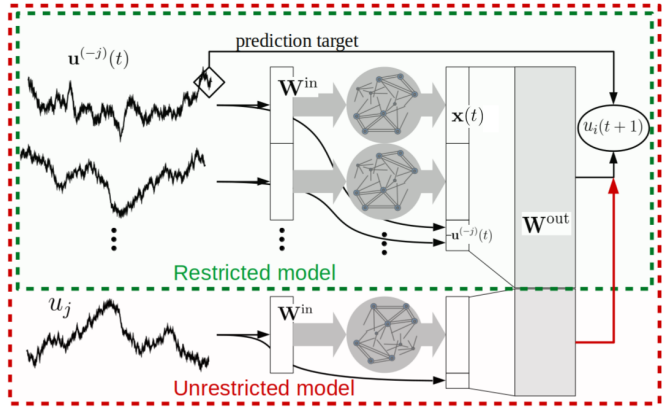

Figure 1. Schematic representation of ES-GC.

B. Echo State Network based causality

ESNs were introduced as a specific type of RNN which can be associated with an architecture and a supervised learning principle. In an ESN, a random, large and fixed RNN is fed with input signals eliciting a nonlinear response in each neuron within the network's "reservoir". The output is then derived as a linear combination of these nonlinear responses. In this way, the information $\mathcal{I}(\mathbf{u}(t))$ can be encoded through an M-dimensional array (the "reservoir") of dynamical states $\mathbf{x}(t)$ which evolve in time as a function of the input $\mathbf{u}$ and of previous states $\mathbf{x}(t-1)$. The reservoir is often [30] modeled with an exponentially decaying memory and an innovation terms $\tilde{\mathbf{x}}$ whose relative contribution is linearly weighted by the so called leak-rate α:

$$\mathbf{x}(t) = (1-\alpha)\mathbf{x}(t-1) + \alpha\tilde{\mathbf{x}}(t)\,. \quad (3)$$

Additionally, the innovation term is a nonlinear function of the contribution of two other terms: a linear combination of the input states $\mathbf{W}^{\text{in}}\mathbf{u}(t)$, and a linear combination of the previous reservoir states $\mathbf{W}\mathbf{x}(t-1)$, yielding $\tilde{\mathbf{x}}(t) = f\left(\mathbf{W}^{\text{in}}[1;\mathbf{u}(t)] + \mathbf{W}\mathbf{x}(t-1)\right)$. $\mathbf{W} \in \mathbb{R}^{M\times M}$ is a mixing matrix between reservoir states, and $\mathbf{W}^{\text{in}} \in \mathbb{R}^{M\times L}$ is a mixing matrix between input states. Typically, both $\mathbf{W}$ and $\mathbf{W}^{\text{in}}$ are constant and randomly initialized, and $\mathbf{W}$ is usually a sparse matrix whose initialization is controlled by its largest eigenvalue ρ (the so-called spectral radius) and its density. Also, a typical choice for the function $f : \mathbb{R}^M \to \mathbb{R}^M$ is an element-wise sigmoid function (e.g. hyperbolic tangent) which is symmetrical around the origin, approximates identity for "small" inputs, and is asymptotically bounded. The choices of α, $\mathbf{W}^{\text{in}}$ and $\mathbf{W}$ are crucial for forecasting accuracy [22].

C. Redefining Causality trough ESNs

For a suitable choice of reservoir size M, leak parameter α, spectral radius ρ, matrix $\mathbf{W}$ and matrix $\mathbf{W}^{\text{in}}$, we can define an

"extended" state $\mathbf{z}(t) = [1; \mathbf{x}(t), \mathbf{u}(t)]$ which, arguably, contains (most of) $\mathcal{I}(\mathbf{u}(t))$. Then, under a linear approximation, we can assume that the expected value of $u_i(t+1) \mid \mathcal{I}(\mathbf{u}(t))$ can be written as a linear combination of an "optimal" matrix $\mathbf{W}^{\text{out}}$ (to be estimated numerically) and $\mathbf{z}(t)$:

$$\mathrm{E}[u_i(t+1) \mid \mathcal{I}(\mathbf{u}(t))] = \mathbf{W}^{\text{out}}\mathbf{z}(t) \ ; \tag{4}$$

given a realization of the system, $\mathbf{W}^{\text{out}} \in \mathbb{R}^{1\times(1+M+L)}$ is found by minimizing the sum of the squared residuals generated when using the next time-point of the i-th component as the 'influenced' variable. Then, the RM and UM in equation (2) can be reformulated as:

$$\begin{aligned} \varepsilon_i'(t) &= u_i(t+1) - \mathbf{W}'^{\text{out}}\mathbf{z}^{(-j)}(t) \quad \text{(RM)} \\ \varepsilon_i(t) &= u_i(t+1) - \mathbf{W}^{\text{out}}\mathbf{z}(t) \qquad \text{(UM)} \end{aligned} \tag{5}$$

and, just like in the classical definition of GC, $s_{j\to i} = \log(\epsilon_i^{(-j)}/\epsilon_i)$ is the estimate of ES-GC strength.

In this paper, under the assumption that each component of $\mathbf{u}$ interacts weakly (as compared to its own dynamics) with other components, we introduce the choice of $\mathbf{W}$ as a block diagonal matrix $\mathbf{W} = \text{diag}(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L)$. Equivalently, we assume that, for each component i, the expected value of $u_i(t+1) \mid \mathcal{I}(\mathbf{u}(t))$ is linearly separable in terms of all echo states including its own. Figure 1 shows a pictorial representation of this model, which can also be thought of as a larger ESN composed of several separable ESNs. As a further improvement, in this paper we introduce the use of orthonormal matrices (as opposed to sparse, randomly initialized matrices) as the block diagonal matrices $\mathbf{W}_i$, obtained through random initialization followed by orthonormalization. This is heuristically motivated by the idea of providing the network with a "maximally orthogonal" basis for signal representation. Experimentally, we found that this choice i) consistently yields superior forecasting performance in terms of residual sum of squares of univariate models, and ii) renders performance largely insensitive to the choice of parameters $(\rho, \mathbf{W}^{\text{in}})$ within a wide range of values (data not shown).

D. Synthetic validation of ES-GC and comparison to other estimators

1) Network generation: In order to compare the performances of ES-GC, SS-GC and MVAR-GC in detecting true causal connections within complex directed networks, we generate data from a family of 10-node ground-truth random networks derived by the Erdös-Rényi model [31], [32]. This entails randomly sampling from a uniform graph distribution, i.e. a graph is constructed by connecting nodes randomly or, equivalently, each edge is included in the graph with constant probability independent from every other edge. Specifically, starting with L disconnected nodes, "edges" (i.e. connections) between two not already connected nodes are successively and randomly assigned up to the required density. Bidirectional connections as well as loops are explicitly allowed. The total number of edges n_{e} depends on the network density d_{n} which, for a network with L nodes, is defined as $n_{\text{e}}/(L(L-1))$. Here, we generated graph families at 9 different densities, where values are

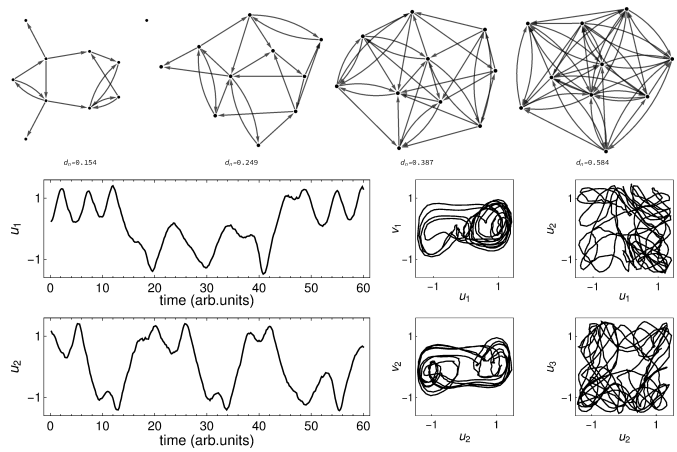

Figure 2. Top: example networks used to generate synthetic data at densities $d_{\text{n}} = 0.1544, 0.2485, 0.387, 0.5837$. Bottom: signals from a network of forced, weakly coupled Duffing oscillators (eq.(6)). Parameters were chosen so that: i) each oscillator would exhibit chaotic behaviour even without coupling; ii) none of any two oscillators would be synchronized regardless of network density or coupling strength. In this paper, $\gamma = 0.5, \beta = -1, \alpha = 1, \delta = 0.3$; ω_i is randomly chosen from the interval $[1.19; 1.21]$ and ϕ_i is randomly chosen from the interval $[0; 2\pi]$; $\Sigma = \sigma^{\text{T}}\sigma$ where $\sigma = \mathbf{m}^{\text{T}}(0.05\ I)\mathbf{m}$ and $\mathbf{m}$ is a matrix whose elements are randomly sampled from a uniform distribution in the [-1;1] interval.

chosen so that the corresponding densities are approximately equidistant on a logarithmic scale between 0.01 and 1: $d_{\text{n}} = \{0.022, 0.044, 0.067, 0.1, 0.154, 0.249, 0.387, 0.584, 0.822\}$. For each value of d_{n} we generate 30 different networks to account for fluctuations with respect to network topology. Each network is described by a binary, zero-diagonal, asymmetric adjacency matrix $\mathbf{A}$, whose elements A_{ij} represent the direct influence of node j on note i. Examples of the generated networks at different densities are shown in Fig. 2.

2) Node-wise Duffing oscillators: For each ground-truth network, a set of forced, noisy, weakly coupled, Duffing oscillators $\mathbf{u} = \{u_1, \ldots, u_L\}$ are generated and assigned to network nodes as follows:

$$\begin{aligned} \dot{u}_i &= v_i + c_{ij}(u_j - u_i) + \xi_i \\ \dot{v}_i &= -\delta v_i - \beta u_i - \alpha {u_i}^3 + \gamma\cos(\omega_i t + \phi_i) \end{aligned} \tag{6}$$

where $\xi_i(t)$ is a spatially correlated white noise process: $\langle \xi_i(t), \xi_j(\tau)\rangle = \delta(t-\tau)\Sigma_{ij}$. The coupling coefficient c_{ij} is defined by a global coupling strength w and a ground-truth matrix $\mathbf{A}$ that defines the topology of the network. Specifically, for the i-th node, if $A_{ij} = 0$ then $c_{ij} = 0$; otherwise c_{ij} is equal to w normalized by the number of incoming connections $c_{ij} = w/\sum_i A_{ij}$. Here, nine values (approximately equidistant on a logarithmic scale) for w were employed ($w = 0.02 - 0.5$). Each networks was evolved for a total of 10000 timepoints ($\Delta t = 0.5$, undersampled from a signal generated with a stochastic second-order Runge-Kutta numerical integration scheme with integration step $\Delta t' = 0.1$). Example synthetic signals are shown in Fig. 2 along with details about parameter choices.

3) Causality estimation in ground truth networks: For each set of synthetics signals (30 networks $\times$ 9 density values $\times$

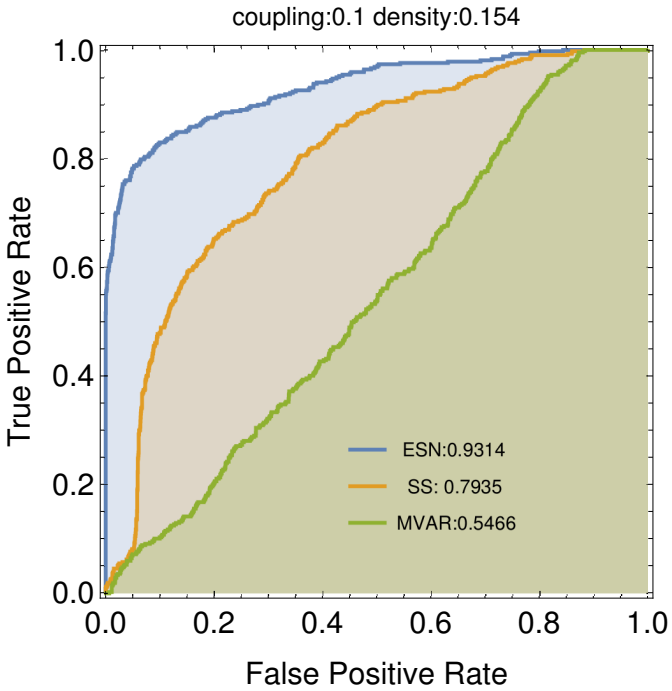

Figure 3. ROC curves ESN- SS- and MVAR-based GC. ROC curves built over the prediction of $(10^2 - 10) \times 30$ links (i.e 30 random networks), density $d_{\mathrm{n}} = 0.154$, coupling strength $w = 0.1$. All other parameters are kept constant (see Fig. 2).

9 coupling strengths = 2430 networks with 10 nodes each; $(10^2 - 10) \times 30 = 2700$ edges per network), the three estimation methods (ES-GC, SS-GC and MVAR-GC) were employed for ground-truth network reconstruction. For each $\{w, d_{\mathrm{n}}\}$ pair, the performance of each estimator was quantified through a receiver operating characteristic (ROC) curve built by varying the threshold in causality strength used for edge acceptance across the 2700 edges. In-house developed code for ES-GC estimation is available at [33].

For both SS-GC and MVAR-GC, the optimal autoregressive order p was chosen according to the Akaike information criterion (AIC) within the range 1-25 [34] ($p = 20$ for both estimators). The optimal p was not significantly sensitive to network density (data not shown). ES-GC hyperparameters were chosen by performing a 3-1 train-test split on univariate data generated from one network node and minimizing prediction error on the test set as a function of reservoir size M (interval: 1-500), leak-rate α (interval: 0.1-0.9) and spectral radius ρ (interval: 0.01-1) [35]. This procedure was repeated for varying data length (interval 2048-65536) This resulted in $M = 2500$ (corresponding to a reservoir of 250 neural units for each of the 10 nodes), $\alpha = 0.3$ and ρ=0.9 ;

The area under the ROC curve (AUC) was employed as a performance metric. Additionally, since SS-GC is an explicitly multi-scale method [28], SS-GC estimations were repeated at 18 different scales (1-18). In this paper, all AUC values presented for SS-GC are the highest value achieved amongst all scales.

E. Estimation of the human between-network connectome from fMRI data

As an example application to biological data, we use in-vivo fMRI data from 1003 subjects made available by the Human Connectome Project [36] as part of the S1200 PTN release. The subjects included underwent 4 sessions of 15-minute multi-band (repetition time (TR) = 0.72s) resting-state fMRI scans on a 3 Tesla scanner with isotropic spatial resolution of 2 mm, for a total of 4800 volumes per subjects. Preprocessing details can be found in [37]. After pre-processing, a group-principal component analysis [38] output was generated and fed into group-wise spatial independent component analysis (ICA) using FSL MELODIC tool [39] to obtain 15 distinct spatiotemporal components. Subject- and components- specific timeseries were then extracted, and a directed connectome was built for each subject through our ES-GC method. ES-GC hyperparameters were chosen as described above (using a train-test split of ICA-timeseries data), resulting in $M = 60$ (corresponding to a reservoir of 4 neurons for each of the 15 components) leak-rate $\alpha = 0.6$, and spectral radius ρ=0.9. Interestingly, we obtained a smaller optimal reservoir and a larger optimal leak-rate as compared to the synthetic data case, possibly indicating less rich 'dynamics' in fMRI data as compared to networks of nonlinear duffing oscillators.

III. Results

A. Synthetic validation results

Figure 3 shows the comparison between the ROC curves obtained when using ES-GC, SS-GC and MVAR-GC for exemplary density and coupling parameters $d_{\mathrm{n}} = 0.154$ and $w = 0.1$. ES-GC clearly outperforms both SS-GC (even at its optimal scales) and MVAR-GC (which only delivers chance-level performance). Additionally, the ROC curves show how for ES-GC true positive rates/false positive rates are larger/smaller (respectively) than for both other estimators at every discrimination threshold (i.e. operating point of the ROC curve).

Since performance of any causality estimator is expected to increase with coupling strength and to fluctuate with network density, we inspected AUC as a function of w with fixed d_{n} and vice versa (Fig. 4). For all estimators, the AUC increases with coupling strength (a higher coupling corresponds to a larger multivariate transfer entropy [25] up to the onset of generalized synchronization (data not shown)). ES-GC performs notably better than both other estimators at all network densities and all coupling strengths.

B. In-vivo human connectome results

ES-GC estimation in the full HCP sample resulted in $4 \times 1003 = 4012$ asymmetric adjacency matrices. For the purpose of visualization (see below), we calculated the element-wise median matrix across subjects and scans, which was then thresholded at the 90th percentile. The resulting directed, within-component connectome derived from 1003 healthy subjects is shown in Fig. 5 (see Figure caption for the physiological significance of each of the 15 components). These results

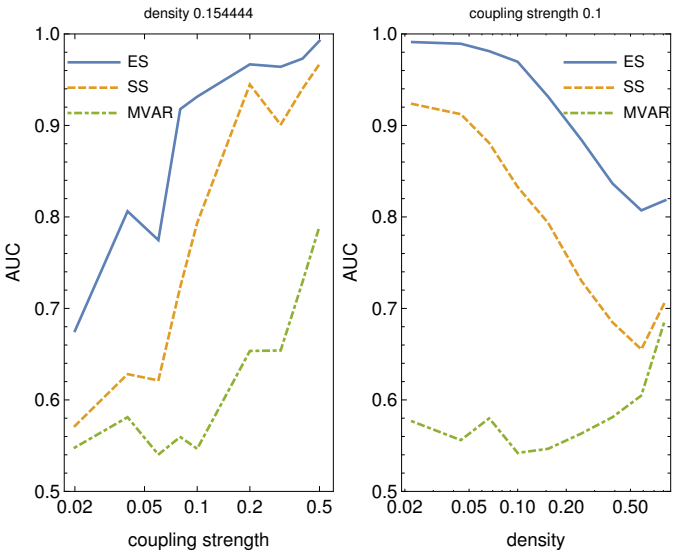

Figure 4. AUC comparison between ESN-GC, SS-GC and MVAR-GC with respect to coupling strength (left) and network density (right). Each AUC was built over the prediction of 2700 edges relative to 30 random networks.

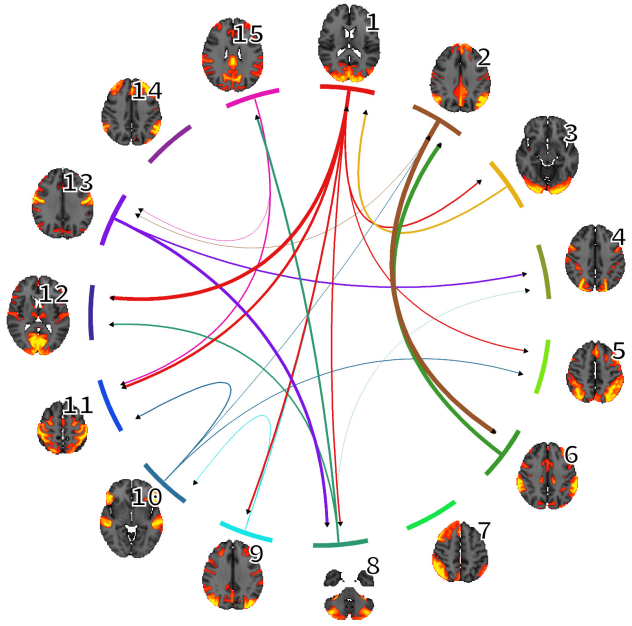

Figure 5. Graphical summary of in-vivo results. Directed, between-component brain connectivity network derived in 1003 HCP subjects (top 10% in median across subjects) is displayed. Color indicates the 'influencing' component. The significance of physiological networks can be summarized as follows [40]: 1: Visuo-Prefrontal Network, 2: Default Mode Network, 3: Extra-striate Visual Network , 4: Visuo-Parieto-Premotor Network, 5: Left Fronto-Parieto-Cerebellar Network, 6: "Salience" Network, 7: Right Fronto-Parieto-Cerebellar Network, 8: Hippocampal-Cerebellar Network, 9: Hippocampal-Posterior Cingulate Network, 10: Fronto-Temporal- Network, 11: Sensory-Motor Network, 12: Striate Visual Network, 13: Sensory/Motor-Limbic Network, 14: Fronto-Polar Network, 15: Cingulate Cortex Network.

suggest a strong bidirectional interaction between the Default Mode Network and the Salience network, a direct modulation of the Striate Visual Network by the Visuo-Prefrontal Network (but not vice versa) and a direct modulation of the Hippocampal-Cerebellar Network by the Sensory/Motor-Limbic-Network, which was only recently defined [40].

IV. Discussion and Conclusions

In this paper we transition away from classical causality quantifiers and define a novel approach to estimating nonlinear, directed network interactions through a specific class of RNNs (namely echo-state networks, or ESN) which do not suffer from training difficulties common in RNNs. We modify the current ESN formulation to represent nonlinear, multivariate signal coupling while decoupling internal model representations and using separate, heuristically motivated orthonormal bases for network weights. We then reformulate the classical GC framework in terms of ESN-based models for multivariate signals generated by complex networks. Our method is validated through extensive synthetic data simulation, where we find that ES-GC largely outperforms state-of-the art linear methods. Interestingly, in our model system, the coupling value $w = 0.1$ marks a transition across chance-level performance for classical MVAR-GC. This indicates that both ESN- and SS-based GC are capable of a non-random resolution of true links in a system where coupling is so weak that standard MVAR-GC analysis fails. Also, for all methods, performance mostly deteriorates with increasing network density. This is possibly due to the complex interplay between the dynamics of a high number of 'causal' nodes. Still, across the whole parameter space and model system parameters tested in this paper, ES-GC performs notably better than both other methods. While these results have been derived using forced, weakly coupled stochastic oscillators, the generality of the ES-GC framework allows to hypothesize that it could deliver superior performance in a larger class of dynamical situations like e.g. non-synchronized, weakly interacting, possibly chaotic and forced dynamical systems, in which causality detection is extremely challenging. Importantly, these circumstances are ubiquitous in biophysics and biomedical signals, where the detection of causal links in weakly interacting systems is often the stepping stone for physiological interpretation. In this context, our method was able to uncover direct functional links between sub-networks of the brain which have not been previously described. Relatedly, while the fMRI results in this paper are intended to demonstrate a possible application of our novel methods to real-world brain data, it is interesting to note the application of GC in neuroscience in general have been the object of constructive discussions [29], [41]–[44] which ultimately confirmed its applicability provided possible methodological pitfalls are avoided. For example, it is well known that fMRI signals are a surrogate of neuronal activity, and that the convolution with a locally varying HRF can confound causality results. Blind deconvolution methods have been proposed in this respect [45]. Still, the application of this type of pipeline is not yet widespread in fMRI studies using causality methods, and the importance and applicability of such methods in cases where fMRI time-series data is averaged over relatively large brain regions stemming from low-dimensional independent component analysis (line in this

paper) remains to be investigated. Also, accurate synthetic simulations of neuronal spiking and neurovascular coupling have shown that the top percentiles in the median causal adjacency matrix computed across subjects can be interpreted with extremely good positive predictive value [8]. Relatedly, it has been shown that GC in general is applicable to fMRI data and that HRF convolution retains monotonicity between fMRI causality and neural causality at realistic fMRI temporal resolution and noise level [46], which corroborates the idea of reliably interpreting the top percentiles of causal connections found in a large number of subjects. Also, the importance of accounting for nonlinearity in GC estimates in fMRI (which is included in our model) as well as the problem of non-overlapping regions of interest across subjects (which is circumvented by employing group-ICA) and of employing non-equilibrium timeseries (which, however, does not apply to independent components derived from resting spate data like in this paper) have been previously underlined [47].

In summary, ES-GC performs significantly better than commonly used and recently developed GC detection tools, even in complex networks with nonlinear signals and weak coupling, making it a valid tool for the analysis of e.g. multivariate biological networks. Future work will address the incorporation of structural priors from e.g. diffusion MRI [48] as well as the reformulation through recent, more sophisticated NN architectures (e.g. combinations of combination of ESN and LSTM models [49]) which have been built to facilitate training.

Author contributions statement

Conceptualization, AD and NT; Methodology, AD and NT; Formal analysis, AD; Funding acquisition, NT and MG; Investigation, AD; Supervision, NT; Writing – original draft, AD and NT; Writing – review and editing, AD, MG, NT. All authors critically reviewed the initial manuscript and approved the final manuscript as submitted.

Additional information

The authors declare no competing interests, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

[1] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[2] A. B. Barrett, L. Barnett, and A. K. Seth, "Multivariate granger causality and generalized variance," *Physical Review E*, vol. 81, no. 4, p. 041907, 2010.

[3] S. Guo, C. Ladroue, and J. Feng, "Granger causality: theory and applications," in *Frontiers in Computational and Systems Biology*. Springer, 2010, p. 83.

[4] S. Basu, A. Shojaie, and G. Michailidis, "Network granger causality with inherent grouping structure," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 417–453, 2015.

[5] L. Barnett and A. K. Seth, "Granger causality for state-space models," *Physical Review E*, vol. 91, no. 4, p. 040101, 2015.

[6] B. Lusch, P. D. Maia, and J. N. Kutz, "Inferring connectivity in networked dynamical systems: Challenges using granger causality," *Physical Review E*, vol. 94, no. 3, p. 032220, 2016.

[7] G. Deshpande, S. LaConte, G. A. James, S. Peltier, and X. Hu, "Multivariate granger causality analysis of fmri data," *Human brain mapping*, vol. 30, no. 4, pp. 1361–1373, 2009.

[8] A. Duggento, L. Passamonti, G. Valenza, R. Barbieri, M. Guerrisi, and N. Toschi, "Multivariate granger causality unveils directed parietal to prefrontal cortex connectivity during task-free mri," *Scientific reports*, vol. 8, no. 1, p. 5571, 2018.

[9] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear granger causality," *Physical review letters*, vol. 100, no. 14, p. 144103, 2008.

[10] B. Cheng and D. M. Titterington, "Neural networks: A review from a statistical perspective," *Statistical science*, pp. 2–30, 1994.

[11] M. Tshilidzi, *Neural Networks for Modeling Granger Causality*, 2015, ch. Chapter 5, pp. 87–103.

[12] A. Montalto, S. Stramaglia, L. Faes, G. Tessitore, R. Prevete, and D. Marinazzo, "Neural networks with non-uniform embedding and explicit validation phase to assess granger causality," *Neural Networks*, vol. 71, pp. 159–171, 2015.

[13] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing multiple nonlinear time series with extended granger causality," *Physics Letters A*, vol. 324, no. 1, pp. 26–35, 2004.

[14] F. Benhmad, "Modeling nonlinear granger causality between the oil price and us dollar: A wavelet based approach," *Economic Modelling*, vol. 29, no. 4, pp. 1505–1514, 2012.

[15] A. S. Chivukula, J. Li, and W. Liu, "Discovering granger-causal features from deep learning networks," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2018, pp. 692–705.

[16] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 312–340, 2019.

[17] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox, "Neural granger causality for nonlinear time series," *arXiv preprint arXiv:1802.05842*, 2018.

[18] A. Tank, I. Cover, N. J. Foti, A. Shojaie, and E. B. Fox, "An interpretable and sparse neural network model for nonlinear granger causality discovery," *arXiv preprint arXiv:1711.08160*, 2017.

[19] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm." IET, 1999.

[20] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[21] Y. Wang, K. Lin, Y. Qi, Q. Lian, S. Feng, Z. Wu, and G. Pan, "Estimating brain connectivity with varying-length time lags using a recurrent neural network," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1953–1963, 2018.

[22] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[23] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[24] ——, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[25] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.

[26] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, "Synchronization as adjustment of information rates: detection from bivariate time series," *Physical Review E*, vol. 63, no. 4, p. 046211, 2001.

[27] L. Barnett and T. Bossomaier, "Transfer entropy as a log-likelihood ratio," *Physical review letters*, vol. 109, no. 13, p. 138105, 2012.

[28] L. Faes, G. Nollo, S. Stramaglia, and D. Marinazzo, "Multiscale granger causality," *Physical Review E*, vol. 96, no. 4, p. 042150, 2017.

[29] L. Faes, S. Stramaglia, and D. Marinazzo, "On the interpretability and computational reliability of frequency-domain granger causality," *F1000Research*, vol. 6, 2017.

[30] H. Jaeger, *Short term memory in echo state networks*. GMD-Forschungszentrum Informationstechnik, 2001, vol. 5.

[31] P. Erdös and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[32] B. Bollobás and B. Béla, *Random graphs*. Cambridge university press, 2001, no. 73.

[33] A. Duggento, "Echostate-grangercausality," GitHub, 2019. [Online]. Available: https://github.com/andreaduggento/EchoState-GrangerCausality

[34] R. Shibata, "Selection of the order of an autoregressive model by akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[35] R. Kohavi and F. Provost, "Glossary of terms journal of machine learning," 1998.

[36] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[37] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni *et al.*, "The minimal preprocessing pipelines for the human connectome project," *Neuroimage*, vol. 80, pp. 105–124, 2013.

[38] S. M. Smith, A. Hyvärinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann, "Group-pca for very large fmri datasets," *Neuroimage*, vol. 101, pp. 738–749, 2014.

[39] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[40] N. Toschi, A. Duggento, and L. Passamonti, "Functional connectivity in amygdalar-sensory/(pre) motor networks at rest: new evidence from the human connectome project," *European Journal of Neuroscience*, vol. 45, no. 9, pp. 1224–1229, 2017.

[41] P. A. Stokes and P. L. Purdon, "A study of problems encountered in granger causality analysis from a neuroscience perspective," *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. E7063–E7072, 2017.

[42] L. Barnett, A. B. Barrett, and A. K. Seth, "Solved problems for granger causality in neuroscience: A response to stokes and purdon," *NeuroImage*, vol. 178, pp. 744–748, 2018.

[43] P. A. Stokes and P. L. Purdon, "Reply to barnett et al.: Regarding interpretation of granger causality analyses," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6678–E6679, 2018.

[44] L. Barnett, A. B. Barrett, and A. K. Seth, "Misunderstandings regarding the application of granger causality in neuroscience," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6676–E6677, 2018.

[45] G.-R. Wu, W. Liao, S. Stramaglia, J.-R. Ding, H. Chen, and D. Marinazzo, "A blind deconvolution approach to recover effective connectivity brain networks from resting state fmri data," *Medical image analysis*, vol. 17, no. 3, pp. 365–374, 2013.

[46] X. Wen, G. Rangarajan, and M. Ding, "Is granger causality a viable technique for analyzing fmri data?" *PloS one*, vol. 8, no. 7, p. e67428, 2013.

[47] J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour, "Six problems for causal inference from fmri," *neuroimage*, vol. 49, no. 2, pp. 1545–1558, 2010.

[48] P. Poulin, D. Jörgens, P.-M. Jodoin, and M. Descoteaux, "Tractography and machine learning: Current state and open challenges," *arXiv preprint arXiv:1902.05568*, 2019.

[49] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5900–5904.