Estimation of Granger causality through Artificial Neural Networks: applications to physiological systems and chaotic electronic oscillators

Yuri Antonacci [Corresp., 1, 2, 3], Ludovico Minati [4, 5], Luca Faes [6], Riccardo Pernice [6], Giandomenico Nollo [7], Jlenia Toppi [2, 3], Antonio Pietrabissa [3], Laura Astolfi [2, 3]

[1] Department of Physics and Chemistry "Emilio Segrè", University of Palermo, Palermo, Italy

[2] Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Fondazione Santa Lucia, Rome, Italy

[3] Department of Computer, Control and Management Engineering "Antonio Ruberti", University of Rome "La Sapienza", Rome, Italy

[4] Center for Mind/Brain Sciences (CIMeC), University of Trento, Trento, Italy

[5] Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan

[6] Department of Engineering, University of Palermo, Palermo, Italy

[7] Department of Industrial Engineering, University of Trento, Trento, Italy

Corresponding Author: Yuri Antonacci
Email address: yuri.antonacci@uniroma1.it

One of the most challenging problems in the study of complex dynamical systems is to find the statistical interdependencies among the system components. Granger causality (GC) represents one of the most employed approaches, based on modeling the system dynamics with a linear vector autoregressive (VAR) model and on evaluating the information flow between two processes in terms of prediction error variances. In its most advanced setting, GC analysis is performed through a state-space (SS) representation of the VAR model that allows to compute both conditional and unconditional forms of GC by solving only one regression problem. While this problem is typically solved through Ordinary Least Square (OLS) estimation, a viable alternative is to use Artificial Neural Networks (ANNs) implemented in a simple structure with one input and one output layer and trained in a way such that the weights matrix corresponds to the matrix of VAR parameters. In this work, we introduce an ANN combined with SS models for the computation of GC. The ANN is trained through the Stochastic Gradient Descent L1 (SGD-L1) algorithm, and a cumulative penalty inspired from penalized regression is applied to the network weights to encourage sparsity. Simulating networks of coupled Gaussian systems, we show how the combination of ANNs and SGD-L1 allows to mitigate the strong reduction in accuracy of OLS identification in settings of low ratio between number of time series points and of VAR parameters. We also report how the performances in GC estimation are influenced by the number of iterations of gradient descent and by the learning rate used for training the ANN. We recommend using some specific combinations

for these parameters to optimize the performance of GC estimation. Then, the performances of ANN and OLS are compared in terms of GC magnitude and statistical significance to highlight the potential of the new approach to reconstruct causal coupling strength and network topology even in challenging conditions of data paucity. The results highlight the importance of of a proper selection of regularization parameter which determines the degree of sparsity in the estimated network. Furthermore, we apply the two approaches to real data scenarios, to study the physiological network of brain and peripheral interactions in humans under different conditions of rest and mental stress, and the effects of the newly emerged concept of remote synchronization on the information exchanged in a ring of electronic oscillators. The results highlight how ANNs provide a mesoscopic description of the information exchanged in networks of multiple interacting physiological systems, preserving the most active causal interactions between cardiovascular, respiratory and brain systems. Moreover, ANNs can reconstruct the flow of directed information in a ring of oscillators whose statistical properties can be related to those of physiological networks.

1 # Estimation of Granger causality through
2 # Artificial Neural Networks: applications to
3 # physiological systems and chaotic
4 # electronic oscillators

5 **Yuri Antonacci**[1,2,3]**, Ludovico Minati**[4,5]**, Luca Faes**[6]**, Riccardo Pernice**[6]**,**
6 **Giandomenico Nollo**[7]**, Jlenia Toppi**[2,3]**, Antonio Pietrabissa**[2]**, and Laura**
7 **Astolfi**[2,3]

8 [1]**Department of Physics and Chemistry "Emilio Segrè", University of Palermo, 90128**
9 **Palermo, Italy**
10 [2]**Department of Computer, Control and Management Engineering "Antonio Ruberti",**
11 **University of Rome "La Sapienza", 00185 Rome, Italy**
12 [3]**Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Fondazione Santa Lucia,**
13 **00179 Rome, Italy**
14 [4]**Center for Mind/Brain Sciences (CIMeC), University of Trento, 38123 Trento, Italy**
15 [5]**Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503,**
16 **Japan**
17 [6]**Department of Engineering, University of Palermo, 90128 Palermo, Italy**
18 [7]**Department of Industrial Engineering, University of Trento, 38123 Trento, Italy**

19 Corresponding author:

20 Yuri Antonacci[1,2,3]

21 Email address: antonacci@diag.uniroma1.it; Bioengineering and Bioinformatics Laboratory,
22 BiBiLab, University of Rome "La Sapienza", Viale Ariosto 25, 00185 Rome, Italy

23 ## ABSTRACT

One of the most challenging problems in the study of complex dynamical systems is to find the statistical interdependencies among the system components. Granger causality (GC) represents one of the most employed approaches, based on modeling the system dynamics with a linear vector autoregressive (VAR) model and on evaluating the information flow between two processes in terms of prediction error variances. In its most advanced setting, GC analysis is performed through a state-space (SS) representation of the VAR model that allows to compute both conditional and unconditional forms of GC by solving only one regression problem. While this problem is typically solved through Ordinary least squares (OLS) estimation, a viable alternative is to use Artificial Neural Networks (ANNs) implemented in a simple structure with one input and one output layer and trained in a way such that the weights matrix corresponds to the matrix of VAR parameters. In this work, we introduce an ANN combined with SS models for the computation of GC. The ANN is trained through the Stochastic Gradient Descent l_1 (SGD-l_1) algorithm, and a cumulative penalty inspired from penalized regression is applied to the network weights to encourage sparsity. Simulating networks of coupled Gaussian systems, we show how the combination of ANNs and SGD-l_1 allows to mitigate the strong reduction in accuracy of OLS identification in settings of low ratio between number of time series points and of VAR parameters. We also report how the performances in GC estimation are influenced by the number of iterations of gradient descent and by the learning rate used for training the ANN. We recommend using some specific combinations for these parameters to optimize the performance of GC estimation. Then, the performances of ANN and OLS are compared in terms of GC magnitude and statistical significance to highlight the potential of the new approach to reconstruct causal coupling strength and network topology even in challenging conditions of data paucity. The results highlight the importance of a proper selection of regularization parameter which determines the degree of sparsity in the estimated network. Furthermore, we apply the two approaches to real data scenarios, to study the physiological network of brain and peripheral interactions in humans under different conditions of rest and mental stress, and the effects of the newly emerged concept of remote synchronization on the information exchanged in a ring of electronic oscillators. The results highlight how ANNs provide a mesoscopic description of the information exchanged in networks of multiple interacting physiological systems, preserving the most active causal interactions between cardiovascular, respiratory and brain systems. Moreover, ANNs can reconstruct the flow of directed information in a ring of oscillators whose statistical properties can be related to those of physiological networks.

INTRODUCTION

A fundamental problem in the study of dynamical systems in many domains of science and engineering is to investigate the interactions among the individual system components whose activity is represented by different recorded time series. The evaluation of the direction and strength of these interactions is often carried out employing the statistical concept of causality introduced by Wiener (Wiener, 1956) and formalized in terms of linear regression analysis by Granger (Granger, 1969). Wiener-Granger causality (GC) was firstly introduced in the framework of linear bivariate autoregressive modeling in its unconditional form for which a generic time series X is said to Granger-cause another series Y if the past of X contains information that helps to predict the future of Y above and beyond the information already contained in the past of Y (Granger, 1969). In the presence of more than two interacting system components, to take into account the presence of other time series which can potentially affect the two time series under analysis the bivariate formulation has been extended to the multivariate case through the use of vector autoregressive (VAR) models, leading to the computation of a conditional form of GC (Geweke, 1984). Due to its linear formulation, GC is very easy to implement, with very few parameters to be estimated if compared with model-free approaches and with a reduced computational cost (Porta and Faes, 2015).

GC from a driver to a target time series is typically quantified by comparing the prediction error variance obtained from two different linear regression models: (i) the "*full model*", in which the present sample of the target series is regressed on the past samples of all the time series in the dataset; (ii) the "*restricted model*", in which the present of the target is regressed on the past of all the time series excluding the driver (Barnett and Seth, 2014). However, this formulation does not take into account that, from a theoretical point of view, the order of the restricted model is infinite, leading to a strong bias or a very large variability associated with the estimation of GC, depending on the model order selected (Stokes and Purdon, 2017; Faes et al., 2017d; Barnett et al., 2018). To overcome the latter problem, an approach based on state-space (SS) modeling of the observed VAR process has been introduced (Barnett and Seth, 2015);

SS models provide a closed-form SS representation of the restricted VAR model and thus, starting from the identification of the full model only, GC in its conditional and unconditional form can be retrieved with high computational reliability directly from the SS parameters (Solo, 2016; Barnett and Seth, 2015; Faes et al., 2017d).

The literature provides different methodologies for VAR model identification, such as the solution of the Yule-Walker equations through Levison's recursion or the Burg algorithm (Kay, 1988) by using the closed-form solution of Ordinary least squares (OLS) estimator, or more sophisticated such as those based on Artificial Neural Networks (ANNs). ANNs have become very popular in recent years, and they have been extensively used as a modeling tool because they are data-driven self-adaptive methods and can work as universal functional approximators (Hornik et al., 1989; Hornik, 1991). The ANN structure used for linear regression comprises one input layer and one output layer which are linked by a matrix of weights obtained after training the network. During the training process, the inputs are presented to the network and the weights are adjusted to minimize the distance between the real and predicted output using error backpropagation techniques (Bishop, 1995).

However, regardless of the methodology used to approach the regression problem, the estimation may be problematic in the setting of many observed processes and short time series available (Antonacci et al., 2019a, 2020c). The literature reports that the stability and the existence of the solution for a linear regression problem are ensured when the number of data points is an order of magnitude greater than the number of VAR coefficients to be estimated (Schlögl and Supp, 2006; Lütkepohl, 2013). To cope with the issues arising in GC estimation when the ratio between data size and number of unknown parameters is low, different approaches have been proposed such as the use of time-ordered restricted VAR models (Siggiridou and Kugiumtzis, 2015), or the so-called partial conditioning (Marinazzo et al., 2012), and of penalized regression techniques based on the l_1-norm (LASSO regression) (Antonacci et al., 2020c; Tibshirani, 1996; Pagnotta et al., 2019). In the latter case, the solution of the linear regression problem is found adding a constraint to the cost function to be minimized, usually the Mean Squared Error (MSE), that induces variable selection of the VAR parameters with a consequent reduction of the MSE associated with the estimation process. Based on l_1-constrained problems, in recent years, different l_1-regularized algorithms have been developed to avoiding overfitting during the training of ANNs. Moreover, the l_1-norm can be applied directly on the weights of the network during the training phase in an efficient way through stochastic gradient descent l_1 (SGD-l_1) (Tsuruoka et al., 2009). While the use of ANNs as a VAR model for GC estimation has been proposed in both linear (Talebi et al., 2018) and non-linear frameworks (Montalto et al., 2015; Attanasio and Triacca, 2011; Duggento et al., 2019), the implementation of SGD-l_1 has never been tested for the purpose of reducing the effects of data paucity on the estimation of GC.

In the present work, an ANN used as a VAR model is embedded in the SS framework for the computation of GC (conditional and unconditional) and compared with the traditional OLS regression both in benchmark networks of simulated multivariate processes and in real-data scenarios. In simulations, we show how training parameters that are typically chosen in a heuristic way (i.e., learning rate and the number of iterations of gradient descent) can affect the estimation of GC in conditions of data paucity; after optimizing these parameters, we test the performance in the quantification of GC magnitude and statistical significance, reflecting respectively coupling strength and structure of the investigated directed functional network, comparatively with standard OLS identification. In real data analysis, we compare the two approaches first in physiological time series, reporting the evaluation of information flow and topology of the network of interactions between brain and peripheral systems probed in healthy subjects in different conditions of mental stress elicited by mental arithmetic and sustained attention tasks (Antonacci et al., 2020c; Zanetti et al., 2019), and then in signals produced by electronic circuits, showing how GC measures can describe the effect of remote synchronization previously observed in a ring of coupled chaotic oscillators (Gambuzza et al., 2013; Minati, 2015a; Minati et al., 2018).

The algorithms for the training of ANNs based on SGD-l_1 algorithm with the subsequent computation of GC by exploiting the SS framework are collected in the ANN-GC MATLAB toolbox, which can be downloaded from https://github.com/YuriAntonacci/ANN-GC-Toolbox (as Supplementary material)

METHODS

Vector Autoregressive Model Identification

Let us consider a dynamical system $\mathscr{Y}$ whose activity is mapped by a discrete-time stationary vector stochastic process composed of M real-valued zero-mean scalar processes, $\mathbf{Y} = [Y_1 \cdots Y_M]$. Considering

the time step n as the current time, the present and the past of the vector stochastic process are denoted as $\mathbf{Y}_n = [Y_{1,n} \cdots Y_{M,n}]$ and $\mathbf{Y}_n^- = [\mathbf{Y}_{n-1} \mathbf{Y}_{n-2} \cdots]$, respectively. Moreover, assuming that $\mathbf{Y}$ is a Markov process of order p, its whole past history can be truncated using p time steps, i.e., using the Mp-dimensional vector $\mathbf{Y}_n^p$ such that $\mathbf{Y}_n^- \approx \mathbf{Y}_n^p = [\mathbf{Y}_{n-1} \cdots \mathbf{Y}_{n-p}]$. Then, in the linear signal processing framework, the dynamics of Y can be described by the vector autoregressive (VAR) model:

$$\mathbf{Y}_n = \sum_{k=1}^{p} \mathbf{Y}_{n-k} \mathbf{A}_k + \mathbf{U}_n, \tag{1}$$

131   where $\mathbf{A}_k$ is an $M \times M$ matrix containing the VAR coefficients, and $\mathbf{U} = [U_1 \cdots U_M]$ is a vector of M
132   zero-mean white processes, denoted as innovations, with $M \times M$ covariance matrix $\Sigma \equiv \mathbb{E}[\mathbf{U}_n^T \mathbf{U}_n]$ ($\mathbb{E}$ is
133   the expected value).

Let us now consider a realization of the process $\mathbf{Y}$ involving N consecutive time steps, collected in the $N \times M$ data matrix $[\mathbf{y}_1; \cdots; \mathbf{y}_N]$, where the delimiter ";" stands for row separation, so that the ith row is a realization of $\mathbf{Y}_i$, i.e., $\mathbf{y}_i = [y_{1,i} ... y_{M,i}], i = 1, ..., N$, and the jth column is the time series collecting all realizations of Y_j, i.e., $[y_{j,1} ... y_{j,N}]^T, j = 1, ..., M,$. The Ordinary least squares (OLS) identification finds an optimal solution for the problem (1) by solving the following linear quadratic problem:

$$\widehat{\mathbf{A}} = \text{argmin}_{\mathbf{A}} ||\mathbf{y} - \mathbf{y}^p \mathbf{A}||_2^2, \tag{2}$$

134   where $\mathbf{y} = [\mathbf{y}_{p+1}; \cdots; \mathbf{y}_N]$ is the $(N-p) \times M$ matrix of the predicted values, $\mathbf{y}^p = [\mathbf{y}_{p+1}^p; \cdots; \mathbf{y}_N^p]$ is the
135   $(N-p) \times Mp$ matrix of the regressors and $\mathbf{A} = [\mathbf{A}_1; \cdots; \mathbf{A}_p]$ is the $Mp \times M$ coefficient matrix. The
136   problem has a solution in a closed form $\widehat{\mathbf{A}} = ([\mathbf{y}^p]^T \mathbf{y}^p)^{-1} [\mathbf{y}^p]^T \mathbf{y}$ for which the residual sum of squares
137   (RSS) is minimized (Lütkepohl, 2013).

138 **Artificial Neural Networks as a Vector Autoregressive Model**

139   Let consider a generic ANN described by the function $y = f(\mathbf{w}; \mathbf{x})$ which takes as input a vector $\mathbf{x} \in \mathfrak{R}^d$
140   and outputs a scalar value $y \in \mathfrak{R}$. In the following, we consider networks with a single output for the sake
141   of simplicity, but all the treatments can be extended to the case of multiple outputs. The output of the
142   network depends on a set of Q adaptable parameters (i.e., the weights connecting the layers), that are
143   collected in a single vector $\mathbf{w} \in \mathfrak{R}^Q$ to be optimized during the training process.

Given a training data set of N input/output pairs $S = \{\mathbf{x}_i, y_i\}$, the learning task aims at solving the following regularized optimization problem:

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, f(\mathbf{w}; \mathbf{x}_i)) + \lambda r(\mathbf{w}), \tag{3}$$

where $l(\cdot, \cdot)$ is a convex function $\in C^1$, i.e, continuously differentiable with respect to $\mathbf{w}$, while $r(\cdot)$ is a convex regularization term with a regularization parameter $\lambda \in \mathfrak{R}^+$. A typical loss function used for the linear regression problem is the squared error of the regression analysis. Inspired by the LASSO algorithm, a way to enforce sparsity in the vector of weights is to penalize the cumulative absolute magnitude of the weights by using the l_1 norm as regularization term:

$$r(\mathbf{w}) = ||\mathbf{w}||_1 = \sum_{k=1}^{Q} |\mathbf{w}_k|. \tag{4}$$

Then, a possible way to solve the problem (3) is to use Stochastic Gradient Descent (SGD) that exploits a small randomly-selected subset of the training samples to approximate the gradient of the objective function. The number of training samples used for this approximation is the batch size. In the present work, we adopt a full batch approach in which all samples are considered, so that SGD simply translates into gradient descent. For each training sample i, the network weights are updated as follows:

$$\mathbf{w}^{j+1} = \mathbf{w}^j + \eta_j \frac{\partial}{\partial \mathbf{w}} \left(l(y_i, f(\mathbf{w}; \mathbf{x}_i)) - \frac{\lambda}{N} \sum_{k=1}^{Q} |\mathbf{w}_k| \right), \tag{5}$$

144   where j is the iteration counter and η_j is the learning rate at each iteration. The difficulty with l_1
145   regularization is that the last term on the right-hand side in (5) is not differentiable when the weight is

146 zero. To solve this issue, following the procedure introduced in Tsuruoka et al. (2009) l_1 regularization
147 with cumulative penalty is applied directly on the weights of the network during the training process.

Let u_j be the absolute value of the total l_1 penalty received by each weight. Since the absolute value of the l_1 penalty does not depend on the weight and on the regularization parameter λ, it is the same for all the weights and is simply accumulated as:

$$u_j = \frac{\lambda}{N} \sum_{t=1}^{j} \eta_t. \tag{6}$$

At each training sample i, the weights of the network are updated as follows:

$$w_k^{j+\frac{1}{2}} = w_k^j + \eta_j \frac{\partial l(y_i, f(\mathbf{w}; \mathbf{x}_i))}{\partial \mathbf{w}} \Bigg|_{\mathbf{w}=\mathbf{w}^j}, \tag{7}$$

$$\text{if } w_k^{j+\frac{1}{2}} > 0 \text{ then } w_k^{j+1} = max(0, w_k^{j+\frac{1}{2}} - (u_k + q_k^{j-1})), \tag{8}$$

$$\text{else if } w_k^{j+\frac{1}{2}} < 0 \text{ then } w_k^{j+1} = min(0, w_k^{j+\frac{1}{2}} - (u_k - q_k^{j-1})), \tag{9}$$

where q_k^j is the total l_1-penalty that w_k has actually received:

$$q_k^j = \sum_{t=1}^{j} (w_k^{t+1} - w_k^{t+\frac{1}{2}}). \tag{10}$$

148 This method for updating the weights penalizes the weight according to the difference between u_j and
149 q_k^{j-1} and is called SGD-l_1.

Generalizing the whole procedure to a network with multiple outputs, in the linear signal processing framework the optimization problem (3) can be solved by using a linear function $f(\cdot; \cdot)$ linking the input layer with the output layer. In particular, the structure of the neural network necessary for solving the regularized problem (3) in the linear framework is reported in Figure (1) for the nth training sample. The input layer shows Mp neurons representing the past history of the considered stochastic process, truncated at p lags ($\mathbf{Y}_n^p$). The output layer is composed of M neurons representing the present state of the whole system ($\mathbf{Y}_n$). The $Mp \times M$ matrix $\mathbf{W}$ contains the weights of the networks that describe the relationships existent between the output and the input layer. Considering all the $(N-p)$ training samples, the loss function $l(\cdot, \cdot)$ becomes:

$$l(\mathbf{y}, \mathbf{y}^p \mathbf{W}) = ||\mathbf{y} - \mathbf{y}^p \mathbf{W}||_2^2, \tag{11}$$

150 which highlights that the weight $\mathbf{W}$ corresponds to the matrix $\mathbf{A}$ containing the parameters of the VAR
151 model (1). Thus, the described ANN is completely equivalent to a VAR model, except for the fact that the
152 training process induces sparsity into the weight matrix $\mathbf{W}$. A feed-forward neural network with no hidden
153 layers, like the one described above, is a generalized linear model that can be identified with an equivalent
154 least squares optimization problem with l_1 regularization applied to the estimated coefficients. If this
155 regularization is not applied, and by using the loss function (11), the problem stated in (3) is completely
156 equivalent to an OLS regression (Sun, 2000).

Determination of the regularization parameter

158 The determination of the regularization parameter λ is a key element of the estimation process, as its
159 selection strongly influences the performance of resulting regression. For a high value of λ, the SGD-l_1
160 algorithm provides a matrix of weights $\mathbf{W}$ in which all entries are zero. On the other hand, when $\lambda \to 0$,
161 the weights stored in $\mathbf{W}$ are all different from zero and the solution corresponds to the OLS solution
162 (Tibshirani, 1996). In this work, the optimal value for λ has been tested in the range $[\lambda_l, \lambda_u]$, where λ_l
163 and λ_u are the values leading to maximum density (no zero elements) and maximum sparseness (all zero
164 elements) of the weight matrix. Subsequently, following the procedure described in Sun et al. (2016),
165 with a hold out approach, we independently draw 90% of the samples available (rows of $\mathbf{y}$ and $\mathbf{y}^p$) as the

$Y_{1,n-1}$
$Y_{2,n-1}$
$Y_{M,n-1}$
$Y_{1,n-2}$
$Y_{2,n-2}$
$Y_{M,n-2}$
$Y_{1,n-p}$
$Y_{2,n-p}$
$Y_{M,n-p}$

$Y_{1,n}$
$Y_{2,n}$
$Y_{M,n}$

Input Layer   Output Layer
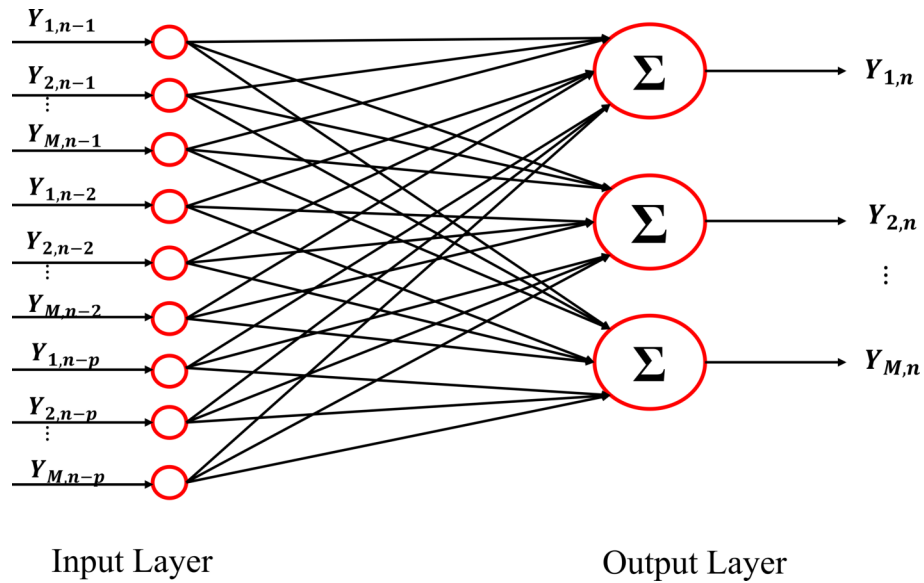
Figure 1. Schematic representation of the architecture of the Neural Network used as VAR model. The input and the output of the network are represented by the lagged variables and by the present states of all processes included in the analysis

166 training set and kept the remaining 10% for testing. Training and test sets were then normalized and, for
167 each assigned λ, the number of non-zero weights was counted in the matrix $\widehat{\mathbf{W}}$ estimated on the training
168 set, and the RSS was computed on the test set as well. This procedure was iterated for each λ, and the
169 optimal λ was taken as the value minimizing the ratio between RSS and the number of non-zero weights
170 (Sun et al., 2016; Antonacci et al., 2020c; Tibshirani and Taylor, 2012). The weight matrix $\mathbf{W}$ obtained
171 with the selected optimal λ was then used for the subsequent GC analysis.

172 **Measuring Granger Causality**

Given the vector process $\mathbf{Y} = [Y_1 \cdots Y_M]$, let us assume Y_j as the *target* process and Y_i as the *source* process, with the remaining $M-2$ processes collected in the vector $\mathbf{Y}_s$ where $s = \{1,...,M\} \setminus \{i,j\}$. Considering the past of the source process $Y_{i,n}^p$ and the past of the target process $Y_{j,n}^p$ we state that the ith process G-causes the jth process (conditional on the other s processes), if $Y_{i,n}^p$ conveys information about $Y_{j,n}$ above and beyond the information contained in $Y_{j,n}^p$ and in all other processes $\mathbf{Y}_{s,n}^p$. This definition is implemented regressing the present of the target on the past of all processes (full regression) and on the past of all processes except the driver (restricted regression), to yield respectively the prediction errors $E_{j|ijs,n} = Y_{j,n} - \mathbb{E}\left[Y_{j,n}|\mathbf{Y}_n^p\right]$ and $E_{j|js,n} = Y_{j,n} - \mathbb{E}\left[Y_{j,n}|Y_{j,n}^p, \mathbf{Y}_s^p\right]$. The resulting prediction error variances, $\lambda_{j|ijs} = \mathbb{E}[E_{j|ijs,n}^2]$ and $\lambda_{j|js} = \mathbb{E}[E_{j|js,n}^2]$ are then combined to obtain the definition of GC (in its conditional form) from Y_i to Y_j (Geweke, 1982):

$$F_{i \to j|s} = \ln \frac{\lambda_{j|js}}{\lambda_{j|ijs}}. \tag{12}$$

Following a similar reasoning, the GC in its original form (unconditional) from Y_i to Y_j is defined as (Granger, 1969):

$$F_{i \to j} = \ln \frac{\lambda_{j|j}}{\lambda_{j|ij}}, \tag{13}$$

173 where $\lambda_{j|j} = \mathbb{E}[E_{j|j,n}^2]$ and $\lambda_{j|ij} = \mathbb{E}[E_{j|ij,n}^2]$ are the prediction error variances of the linear regression
174 of $Y_{j,n}$ on $Y_{j,n}^p$ and on $[Y_{j,n}^p Y_{i,n}^p]$, respectively obtained from the errors $E_{j|j,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y_{j,n}^p]$ and
175 $E_{j|ij,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y_{j,n}^p, Y_{i,n}^p]$.

The prediction error variances needed for the determination of the GC measures can be computed from the identification of the model (1) or by the training of the presented neural network, i.e., from the parameters $(\mathbf{A}_1, ..., \mathbf{A}_p, \Sigma)$ estimated using OLS or from the weights $(\mathbf{W}, \Sigma)$ estimated through the SGD-l_1 training algorithm. Given that $E_{j|ijs,n} = U_{j,n}$, the error variance of the full regression can be obtained as the jth diagonal element of the error covariance matrix $\lambda_{j|ijs} = \Sigma(j, j)$. The other partial variances in (12) and (13) can be retrieved, starting from the identification of the full model, by exploiting the theory of State-Space (SS) models (Barnett and Seth, 2015; Faes et al., 2017a), according to which the VAR model (1) can be represented as an SS model relating the observed process $\mathbf{Y}$ to an unobserved process $\mathbf{Z}$ through the equations (Barnett and Seth, 2015; Solo, 2016):

$$\mathbf{Z}_{n+1} = \mathbf{Z}_n\mathbf{A} + \mathbf{E}_n\mathbf{K}, \tag{14}$$

$$\mathbf{Y}_n = \mathbf{Z}_n\mathbf{C} + \mathbf{E}_n, \tag{15}$$

where the innovations $\mathbf{E}_n = \mathbf{Y}_n - \mathbb{E}[\mathbf{Y}_n|\mathbf{Y}_n^p]$ are equivalent to the innovations $\mathbf{U}_n$ in (1) and thus have covariance matrix $\Phi = \mathbb{E}[\mathbf{E}_n^T\mathbf{E}_n] = \Sigma$. This representation, typically denoted as "innovation form" SS model (ISS) (Barnett and Seth, 2015), also evidences the Kalman Gain matrix $\mathbf{K}$, the state matrix $\mathbf{A}$ and the observation matrix $\mathbf{C}$, which can all be computed from the original VAR parameters in (1) as reported in (Faes et al., 2017a). The advantage of this representation is that it allows to form "submodels" which exclude one or more scalar processes from the observation equation (15) leaving the state equation (14) unaltered. In particular, the submodels excluding the driver process Y_i, the group of s processes $\mathbf{Y}_s$, or the the driver process Y_i and the group of s processes $\mathbf{Y}_s$, have the following observation equations:

$$\mathbf{Y}_{js,n} = \mathbf{Z}_n\mathbf{C}^{(js)} + \mathbf{E}_{js,n}, \tag{16}$$

$$\mathbf{Y}_{ji,n} = \mathbf{Z}_n\mathbf{C}^{(ji)} + \mathbf{E}_{ji,n}, \tag{17}$$

$$\mathbf{Y}_{j,n} = \mathbf{Z}_n\mathbf{C}^{(j)} + \mathbf{E}_{j,n}, \tag{18}$$

where the superscripts $(js), (ji)$ and (j) denote the selection of the columns with indices $(js), (ji)$ and (j) in a matrix. As shown by (Barnett and Seth, 2015), the submodels (14,16), (14,17) and (14,18) are not in ISS form, but can be converted into ISS by solving a Discrete Algebraic Riccati equation (DARE). Then, the covariance matrices of the innovations $\mathbf{E}_{js,n}, \mathbf{E}_{ji,n}$ and $\mathbf{E}_{j,n}$ include the desired error variances $\lambda_{j|js}$, $\lambda_{j|ji}$ and $\lambda_{j|j}$ as the first diagonal element.

In order to establish the existence of a direct link from the ith node to the jth node of the network represented by the observed vector process, the statistical significance of the conditional GC computed after OLS identification of the VAR model was tested using surrogate data. Specifically, one hundred sets of surrogate times series were first generated using the Iterative Amplitude Adjusted Fourier Transform (IAAFT) procedure (Schreiber and Schmitz, 1996); then, for each directed link $(i, j$ pair), the conditional GC $F_{i \to j|s}$ was estimated for each surrogate set, a threshold equal to the 95th percentile of its distribution on the surrogates was determined, and the link was considered as statistically significant when the estimated $F_{i \to j|s}$ was above the threshold. In the case of ANN identification, the statistical significance of the estimated conditional GC values was determined in a straightforward way exploiting the sparseness of the weights matrix $\mathbf{W}$ resulting from the training through SGD-l_1.

Simulation Experiments

This section reports three simulations designed to evaluate the performances of the proposed estimator of the GC based on ANNs trained with SGD-l_1 in comparison with the traditional VAR identification based on OLS. The first simulation evaluates the conditional GC computed by the ANN estimator in known structures of networks assessed with different amount of data samples, for different values of learning rate (η) and for different values of iterations of the SGD-l_1 algorithm. blueIn the second and in

197 the third simulation studies, after having extracted the best combination of learning rate and the number of
198 iterations of the gradient descent to be used in ANN-based estimation, we compare it with OLS estimation
199 as regards the ability to retrieve the true values of the conditional GC and to reconstruct the assigned
200 network topology. The effects of different values of signal-to-noise ratio (SNR) and of simulating a denser
201 network structure are evaluated respectively in the second and in the third study. In all simulations, the
202 topology is representative of the interaction of a ten-variate VAR process exhibiting a random interaction
203 structure with two different values of density of connected nodes (Toppi et al., 2016a; Antonacci et al.,
204 2020c; Pascucci et al., 2020).

Simulation Studies I-II

206 Simulated multivariate time series (M=10) were generated as a realization of a VAR(16) model fed
207 by zero-mean independent Gaussian noise with variance equal to 0.1. The simulated networks have a
208 ground-truth structure with a density of connected nodes equal to 15%, where non-zero AR parameters of
209 values chosen randomly in the interval [-0.8, 0.8] were set at lags assigned randomly in the range (1-16)
210 (Anzolin and Astolfi, 2018). The knowledge of the true AR parameters allows computing the theoretical
211 values of the conditional GC and the true network topology, as illustrated for an exemplary case in Figure
212 2. Simulations were generated for different values of: 1) the parameter K defined as the ratio between the
213 number of data samples available ($N \times M$) and the number of AR coefficients to be estimated ($M^2 \times p$);
214 2) the signal-to-noise ratio (SNR) defined as the ratio between the squared amplitude of the signal and the
215 square amplitude of additive white noise. One hundred networks were generated for each value of K in
216 the range (1,3,10,20); the length of the simulated time series was $N = 160$ when $K = 1$ and $N = 3200$
217 when $K = 20$. When additive noise was considered in the simulation study, SNR varies in the range (0.1,
218 1, 5, 10, 10^3).

219   First, considering ANN estimation performed for each value assigned to K and for each realization,
220 the learning rate η and the number of iterations for the SGD-l_1 during the training process were varied
221 respectively in the range ($10^{-3}, 10^{-4}, 10^{-5}$) and in the range (100, 1000, 2000). Importantly, for each
222 network structure a different neural network was trained initializing the weights according to the method
223 described in Glorot and Bengio (2010) that guarantees a faster convergence of the gradient descent
224 algorithm. After training, the conditional GC between each pair of processes was estimated from the
225 matrix of the weights $\mathbf{W}$ using the SS approach. Then, in order to assess which combination of learning
226 rate - number of iterations of the gradient descent is the best for a regression problem different measures
227 of performances were computed as explained in the following subsection. Second, by using the best
228 combination of learning rate\number of iterations of the gradient descent, the effects of K ratio and SNR
229 were assessed by comparing the performances of ANN and OLS in estimating conditional GC. In the
230 latter case, the same multivariate time series generated for the purposes of the first simulation study were
231 used, by simply adding white noise with amplitude tuned to get the desired SNR value.
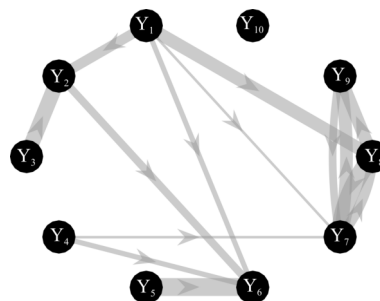
Figure 2. Graphical representation of one of the ground-truth networks of the simulation study. Arrows
represent the causal links randomly assigned between two network nodes via nonzero VAR coefficients.
The thickness of each arrow is proportional to the strength of the causal connection assessed by the
conditional GC, with minimum and maximum values equal to 0.0069 and 0.4. The number of
connections for each network is set to 14 out of 90.

Simulation Study III

Simulated multivariate time series (M=10) were generated as a realization of a reduced VAR(6) process in which coefficients of a VAR(1) model were placed in the first lag for the diagonal elements, while coefficients of a VAR(2) model were placed randomly with a variable delay (up to 6) for the off-diagonal elements (Rodrigues and Andrade, 2015). One-hundred surrogate networks were created assuming links in 80% of all possible connections and directed interactions were placed in a subset of existing links (50 %), with a final value of density of connected nodes $\sim$ 40%. Interactions were generated by randomly assigning both positive and negative values to the VAR(2) coefficients outside the diagonal. The magnitude of AR coefficients was randomly determined (range: 0.15-0.5 in steps of 0.01) (Pascucci et al., 2020). For each simulated dataset, the stochastic generation of a VAR model was reiterated until the system reached the asymptotic stability for which the real eigenvalues are lower than zero (Barnett and Seth, 2014). The knowledge of the true AR parameters allows computing the theoretical values of the conditional GC and the true network topology as illustrated in Figure 3. Simulations were generated for different values of the K ratio, as defined in the previous section, in the range (1, 3, 10, 20) with a resulting time series length $N = 60$ when $K = 1$ and $N = 1200$ when $K = 20$. In order to evaluate the differences between ANN and OLS estimation approaches, different measures of performance were computed as explained in the following subsection. For the ANN case we used the best combination of learning rate\number of iterations of the gradient descent obtained from simulation study I.
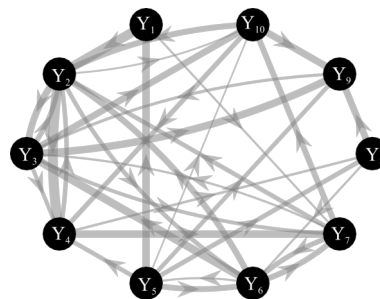
Figure 3. Graphical representation of one of the ground-truth networks of the simulation study III. Arrows represent the causal links randomly assigned between two network nodes via nonzero VAR coefficients. The thickness of each arrow is proportional to the strength of the causal connection assessed by the conditional GC, with minimum and maximum values equal to 0.03 and 0.31. The number of connections for each network is set to $\sim$ 38 out of 90.

Performance Evaluation

Performances were assessed both in terms of the accuracy in estimating the strength of the network links through the absolute values of the conditional GC measure, and in terms of the ability to reconstruct the network structure through the assessment of the statistical significance of the GC.

The bias of GC was computed comparing the estimated and theoretical GC values. For each pair of network nodes represented by the processes Y_i and Y_j, the theoretical GC obtained from the true VAR parameters, $F_{i \to j|s}$, was compared with the corresponding estimated GC value, $\widehat{F}_{i \to j|s}$ through the absolute bias measure (Kim and Kim, 2016):

$$bias = |F_{i \to j|s} - \widehat{F}_{i \to j|s}|. \tag{19}$$

The bias was assessed separately for null links and non-null-links, corresponding respectively to zero and non-zero values of the conditional GC, yielding the measures $bias_0$ and $bias_1$. For each network, these two measures were averaged across the non-null links (15 for the simulations I-II and 38 for the simulation III) and across the null links (75 for the simulations I-II and 52 for the simulation III) to get individual measures, denoted as $BIAS_1$ and $BIAS_0$. Finally, the distributions of the two parameters were obtained across the 100 simulated network structures.

The ability of ANN and OLS to detect the absence or presence of a network link based on the statistical significance of the GC was tested comparing two adjacency matrices representative of the estimated and theoretical network structures. This can be seen as a binary classification task where the existence

263 (class 1) or absence (class 0) of a causal connection is estimated using surrogate data for OLS and
264 looking at the presence/absence of non-zero weights for ANN, and is then compared with the underlying
265 ground-truth structure. Performances were assessed through the computation of false-negative rate (FNR,
266 measuring the fraction of non-null links with non-significant estimated GC), false-positive rate (FPR,
267 measuring the fraction of null links with significant estimated GC) and Area Under Curve (AUC) that
268 summarizes the information provided by FNR and FPR (Toppi et al., 2016b; Antonacci et al., 2019a).
269 In particular, the AUC parameter is obtained by applying a trapezoidal interpolation between a point on
270 the Receiver Operating Characteristic (ROC) space, extracted knowing false positives and true positives,
271 and the two extremes of the ROC space (0,0) and (1,1). These performance measures were computed
272 across the network links for each assigned network, and the corresponding distribution across the 100
273 simulated network structures was then obtained separately for OLS and ANN. In the case of ANNs,
274 the computation time (in seconds) required for the training of the ANN for different values of learning
275 rate, number of iterations of the gradient descent and data samples available was also considered as a
276 performance parameter. The average computation times over the 100 realizations were calculated using
277 an implementation of the algorithms in MATLAB$^{\circledR}$ environment on a PC with a six cores Intel Xeon
278 (CPU clock speed 3.7 GHz), 128· GB DDR4 RAM.

279   To establish which combination of learning rate and number of iterations of the gradient descent
280 guarantees the most accurate results for each value of the K-ratio, an indicator of the overall performance
281 (parameter S) was defined as the average of the two following performance parameters: i) the bias as
282 defined in (19) for non-null links, normalized with respect to the theoretical GC value; ii) the complement
283 to 1 of the AUC parameter, $1 - AUC$. These two parameters are both null in the case of perfect estimation,
284 and increase when the estimated GC values deviate from the theoretical (non-zero) values or when the
285 estimated network topology differs from the true topology. Both parameters were averaged across values
286 of the K-ratio, and then the S parameter was computed as their average.The distribution of S across the
287 100 realizations was investigated as a function of learning rate and number of iterations of SGD-l_1.

Statistical Analysis

289 For the first simulation, a three-way repeated-measures ANOVA was carried out for each performance
290 parameter ($BIAS_0$,$BIAS_1$,FNR,FPR,AUC), in order to evaluate the effects on the computed performance
291 parameters of different values of K (in the range [20, 10, 3, 1]), different values of the learning rate LR
292 (in the range [10^{-3},10^{-4},10^{-5}]) and different values of the number of iterations of SGD-l_1 (N_{train} in the
293 range [100, 1000, 2000]). Furthermore, with the aim of defining the best combination of learning rate and
294 number of SGD-l_1 iterations independently of the data size, a two-way repeated-measures ANOVA was
295 carried out for the parameter S using LR and N_{train} as factors and grouping data from all values of K, so as
296 to evaluate the effects of these two parameters on the overall performance.

297   For the second simulation, five different three-way repeated-measures ANOVA tests, one for each
298 performance parameter ($BIAS_0$,$BIAS_1$,FNR,FPR,AUC), were performed to evaluate the effects on the
299 performance of different values of K (in the range [20, 10, 3]), of different values of SNR (in the range
300 [0.1, 1, 5, 10, 10^3]) and of the two estimation methods ([OLS, ANN]).

301   For the last simulation, five different repeated measures two-way ANOVA tests, one for each perfor-
302 mance parameter ($BIAS_0$,$BIAS_1$,FNR,FPR,AUC), were performed to evaluate the effects on the perfor-
303 mance different values of K (in the range [20, 10, 3]) and different estimation methods ([OLS, ANN]).

304   The Greenhouse-Geisser correction for the violation of the spherical hypothesis was used in all
305 analyses. The Tukey's posthoc test was used for testing the differences between the sub-levels of the
306 ANOVA factors. The Bonferroni-Holm correction was applied for multiple ANOVAs computed on
307 different performance parameters.

Results of the Simulation Study I

309 The results of the three-way repeated-measures ANOVAs, expressed in terms of F-values and computed
310 separately on all the performance parameters considering K, LR and N_{train} as main factors, are reported in
311 Table 1.

312   The three-way ANOVAs revealed a strong statistical influence of the main factors N_{train}, *LR* and *K*
313 and of their interaction on all the performance parameters analyzed. The only non-significant effect was
314 that of the interaction between N_{train} and K on the AUC parameter.

315   Figure 4 reports the distribution of the parameters $BIAS_0$ and $BIAS_1$ according to the interaction
316 $N_{\text{train}} \times LR \times K$. In the analysis of the error associated with the estimation of the conditional GC along

Factors	DoF	$BIAS_0$	$BIAS_1$	FNR	FPR	AUC
N_{train}	(2, 198)	7.8***	711***	467***	68***	609***
LR	(2, 198)	69.6***	461***	325***	171***	656***
K	(3, 297)	16***	181***	309***	88***	344***
$N_{train} \times$ LR	(4, 396)	110.4***	101***	279***	156***	97.2***
$N_{train} \times$ K	(6, 594)	139.7***	2.6*	44***	98***	0.5
LR $\times$ K	(6, 594)	200.9***	13***	47***	132***	2.5*
$N_{train} \times$ LR $\times$ K	(12, 1188)	28.2***	71.6***	20***	15***	3.6***

Table 1. F-values and corresponding degrees of freedom (DoF) of the three-way repeated measures ANOVA.***, $p < 10^{-5}$;**, $10^{-5} < p < 0.01$; *, $0.01 < p < 0.05$.
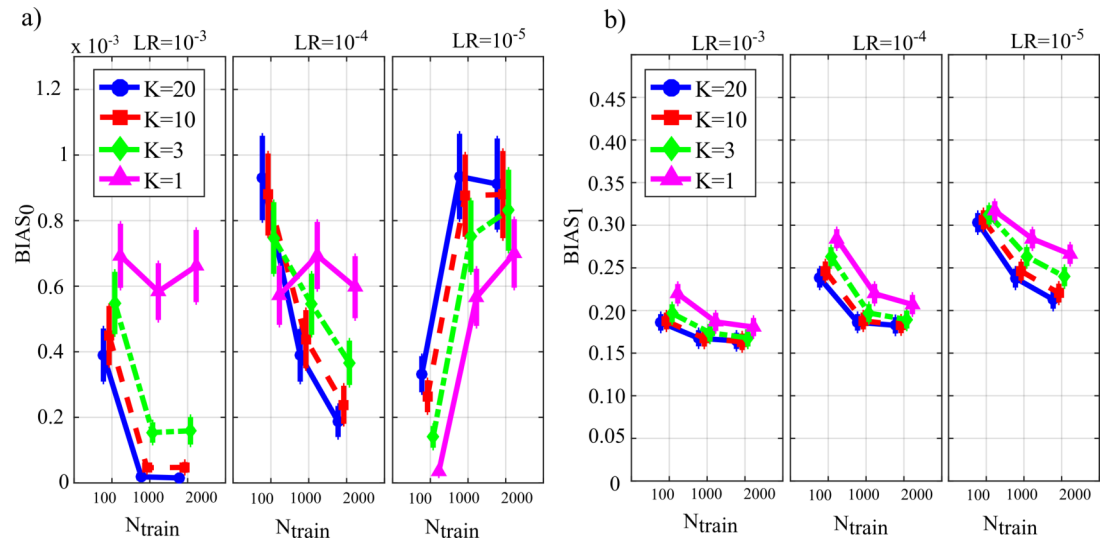
Figure 4. Distributions of the bias of conditional GC (value and 95% confidence interval across 100 simulated networks) estimated using ANNs for the first simulation study. Bias parameters computed for the null links ($BIAS_0$, panel a) and for the non-null links ($BIAS_1$, panel b) are plotted as a function of the number of iterations of the gradient descent (N_{train}) for different values of the ratio between data samples and model coefficients to be estimated (K) and of the learning rate (LR) of ANN training.

the null links ($BIAS_0$, Figure 4a)), an increase of the bias was observed at decreasing the number of data samples available (factor K), regardless of the learning rate (factor LR) and of the number of iterations of gradient descent (N_{train}).

Except for the case $LR = 10^{-5}$, increasing the number of iterations N_{train} reduced the bias for $LR = 10^{-3}$ and for $LR = 10^{-4}$, but not for $LR = 10^{-5}$ when the opposite behavior was observed. The bias analysis of the GC values computed along the non-null links (Figure 4b)) showed more clear patterns of the error, evidencing a decrease of $BIAS_1$ at increasing N_{train}, at increasing K, and at decreasing LR. The lowest mean values of $BIAS_1$ were obtained setting $LR = 10^{-3}$ and N_{train} equal to 1000 or 2000.

Figure 5 reports the distributions of the parameters FNR, FPR and AUC according to the interaction $N_{train} \times LR \times K$. The portion of non-null directed links incorrectly classified as null (FNR, Figure 5a)) was lower than 20% in all cases except for $N_{train} = 100$ and $K \leq 3$. The rate of false negative detections decreased at increasing K regardless of LR and N_{train}. A strong effect of the number of iterations on the FNR was observed in the most challenging condition of $K = 1$ (purple lines), especially when $LR = 10^{-5}$. The portion of null links incorrectly classified as non-null (FPR, Figure 5b)) was always lower than 20 %. The rate of false positive detections showed a tendency to increase at decreasing K, while it was almost stable at varying LR and N_{train}. The best scenario appears $LR = 10^{-3}$, showing a mean FPR under 0.1 for each value of $K > 1$. The overall accuracy measured by AUC (Figure 5c)) reached the highest values for $LR = 10^{-3}$ and $N_{train} \in \{1000, 2000\}$. In these conditions, a very accurate reconstruction of the network structure was obtained, as the accuracy was equal to 95 % for $K = 20$ and above 85% even when $K = 1$.

336 The performance showed a tendency to degrade at decreasing K, increasing LR and decreasing N_{train}.
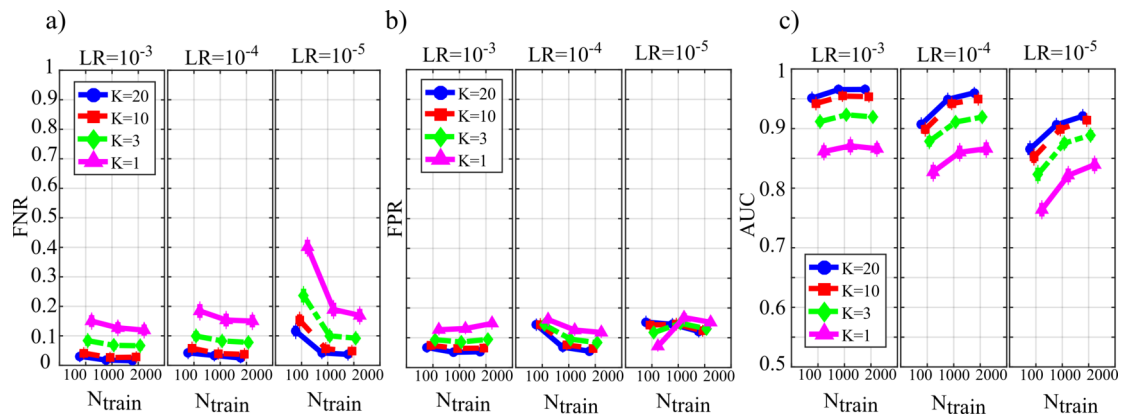
Figure 5. Distributions of the parameters assessing the quality of network reconstruction performed using ANNs for the first simulation study. Plots depict the distributions of FNR (a), FPR (b) and AUC (c) expressed as mean value and 95% confidence interval across 100 simulated networks as a function of the number of iterations of the gradient descent (N_{train}) for different values of the ratio between data samples and model coefficients to be estimated (K) and of the learning rate (LR) of ANN training.

337 Table 2 reports the computation time required for the training of the neural network in different
338 conditions of K ratio, learning rate and number of SGD-l_1 iterations averaged across the 100 realizations.
339 As expected, the computation time increases with the number of iterations of the gradient descent and
340 with the number of data samples available (K ratio). The least and most time-consuming settings were
341 $N_{\text{train}} = 100, K = 1$ and $N_{\text{train}} = 2000, K = 20$, respectively taking ~ 2 secs and ~ 210 secs.

	$LR = 10^{-3}$			$LR = 10^{-4}$			$LR = 10^{-5}$		
N_{train}	100	1000	2000	100	1000	2000	100	1000	2000
K=20	12.08	107.7	213.66	12	107.7	214.36	11.91	107.8	213.72
K=10	7.6	72.8	145.1	7.68	72.8	145.1	7.61	72.88	145.28
K=3	3.4	33.12	65.9	3.44	33.25	66.1	3.4	33.18	66.22
K=1	2.6	25.9	51.7	2.64	25.98	51.69	2.6	26	51.82

Table 2. Average computation time (in seconds, measured for 100 simulated networks) required to train the ANN for different values of K ratio, learning rate and number of iteration of gradient descent.

342 Figure 6 reports the distribution of the overall performance parameter S computed as a function of
343 the learning rate for different number of iterations of SDG-l_1 (interaction $N_{\text{train}} \times LR$). The results show
344 how the performance is affected significantly by both factors, with values of S that tend to decrease while
345 increasing the learning rate and the number of iterations of the gradient descent. The lower values of S,
346 indicating lowest bias of the estimated GC values and/or highest AUC in the classification of the network
347 structure, were observed for $LR = 10^{-3}$ and $N_{\text{train}} = 1000$ or $N_{\text{train}} = 2000$. As the improvement from
348 $N_{\text{train}} = 1000$ to $N_{\text{train}} = 2000$ was not statistically significant, we infer that the best setting is the least
349 computationally onerous combination, i.e., $LR = 10^{-3}, N_{\text{train}} = 1000$.

Results of the Simulation Study II

351 After the extraction of the best combination of the training parameters of the ANN, in the second simulation
352 study we compare the performance of OLS and ANN at varying the proportion between number of data
353 samples available and parameters to be estimated (K-ratio) as well as at varying the amplitude of white
354 noise added to the original time series (SNR). The results of the three-way repeated-measures ANOVAs,
355 expressed in terms of F-values and computed separately on all the performance parameters considering K,
356 SNR and $TYPE$ (i.e., the method used: OLS or ANN) as main factors, are reported in Table 3.
357 The three-way ANOVA highlights a strong statistical influence of the main factors K, SNR and $TYPE$
358 and of their interactions on all the performance parameters analyzed in this study. In this case the level
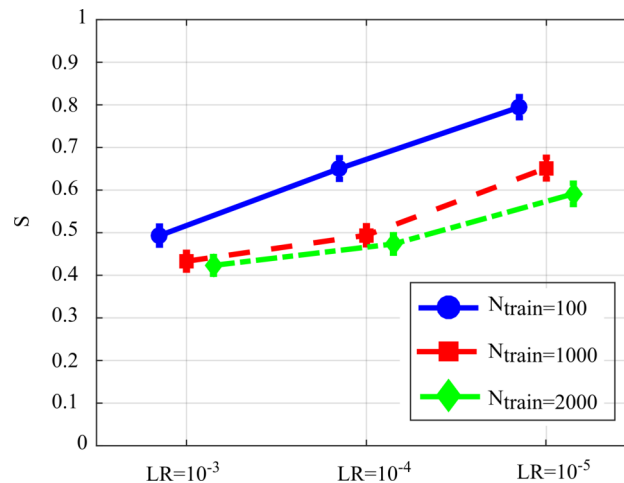
Figure 6. Distributions of S parameter considering the interaction factor $N_{\text{train}} \times LR$, expressed as mean value and 95% confidence interval of the parameter computed across 100 realizations of the first simulation study ($F(4, 396) = 128.09$, $p < 10^{-5}$).

Factors	DoF	BIAS$_0$	BIAS$_1$	FNR	FPR	AUC
TYPE	(1, 99)	5901***	77,8***	68,8***	27,4***	36,9***
SNR	(4, 396)	328,1***	1621,4***	645,1***	173,3***	761,2***
K	(2, 198)	9785,3***	0,2	2118,7***	10,2***	1881,1***
TYPE $\times$ SNR	(4, 396)	85,6***	199,7***	2,6**	46,2***	10,5***
TYPE $\times$ K	(2, 198)	8578***	99,9***	1093***	280,8***	570,1***
SNR $\times$ K	(8, 792)	33,3***	167,4***	50,4***	19,4***	30,3***
TYPE $\times$ K $\times$ SNR	(8, 792)	26,1***	128,8***	65,4***	13,3***	45,4***

Table 3. F-values and corresponding degrees of freedom (DoF) of the three-way repeated measures ANOVA investigating the effects of the factors K (ratio between data samples and number of model parameters), SNR (ratio between the squared amplitude of the signal and the square amplitude of the noise) and $TYPE$ (estimator used, i.e. OLS or ANN) on the performance parameters of GC estimation (BIAS$_0$, BIAS$_1$) and of network reconstruction (FNR, FPR, AUC).***, $p < 10^{-5}$.

$K = 1$ was not considered in the statistical comparison due to the non-convergence of the DARE equation for the OLS case.

Figure 7 reports the distribution of the parameters $BIAS_0$ and $BIAS_1$ according to the interaction factor $TYPE \times K \times SNR$. The comparison of OLS and ANN shows that the two estimation approaches have very different performance: in the computation of GC over the null links, the error of ANN is very close to zero even in the most challenging condition of K=1, while OLS shows an increasing bias with the decrease of the number of data samples available for the estimation of GC values (Figure 7a); in the computation of GC over the non-null links, the estimation bias is low but shows a tendency to increase for OLS, while it is remarkable but stable for the ANN. Concerning the additive noise, its impact is much more noticeable for the OLS case which shows a large increase of the bias measures with the decrease of SNR values; on the other hand, the trends of the two measures of bias for the ANN case seem to be rather constant. Only the bias in the computation of the GC on non-null links shows a slight reduction with increasing SNR. However, in a condition of sufficient data samples available (K=20) and a high value of signal-to-noise ratio ($SNR = 10^3$), OLS shows a bias associated with the non-null links which is very close to zero and considerably lower than that associated with ANN.

Figure 8 reports the distributions of the parameters FNR, FPR and AUC according to the interaction $TYPE \times K \times SNR$. When the value of SNR is equal to 10^3 the analysis of false negative detections of directed links (panel a) shows that the error committed increased with decreasing the number of data samples available. The error was comparable for OLS and ANN when $K = [20, 10]$, and then increased more markedly for OLS, while it remained lower than 10% even when $K = 1$ for ANN. On the other

379 hand, the analysis of false positive detections (panel b) showed an error quite low and stable with *K*
380 in the case of OLS, and an error slightly growing with *K* up to 15% in the case of ANN. The overall
381 performance evaluated through AUC showed high classification accuracy and absence of statistically
382 significant differences between the two estimation methods for $K = [20, 10]$, and a better performance
383 of ANN compared with OLS for lower values of *K*; a high AUC value ($\sim 85\%$) was reported for ANN
384 even when K=1. The situation becomes very different when the value of SNR decreases. Both false
385 negatives and false positives increase with the amplitude of the additive noise; the increase of FNR is
386 remarkable for the OLS method. The analysis of AUC trends for OLS case (panel c) highlights that
387 when SNR is very low and the number of data samples available is very scarce (K=3, green line) AUC
388 is less than 70%. This is not the case for ANN which shows an average value of AUC greater than
389 70% even when K=1 (purple line) and SNR=0.1 which represents the worst simulated scenario. Using a
390 quantile-based thresholding criteria approach for the AUC computation, as introduced in (Pascucci et al.,
391 2020), yields substantially overlapping trends of the performance measures (results reported in Figure S1
392 as Supplementary Material).
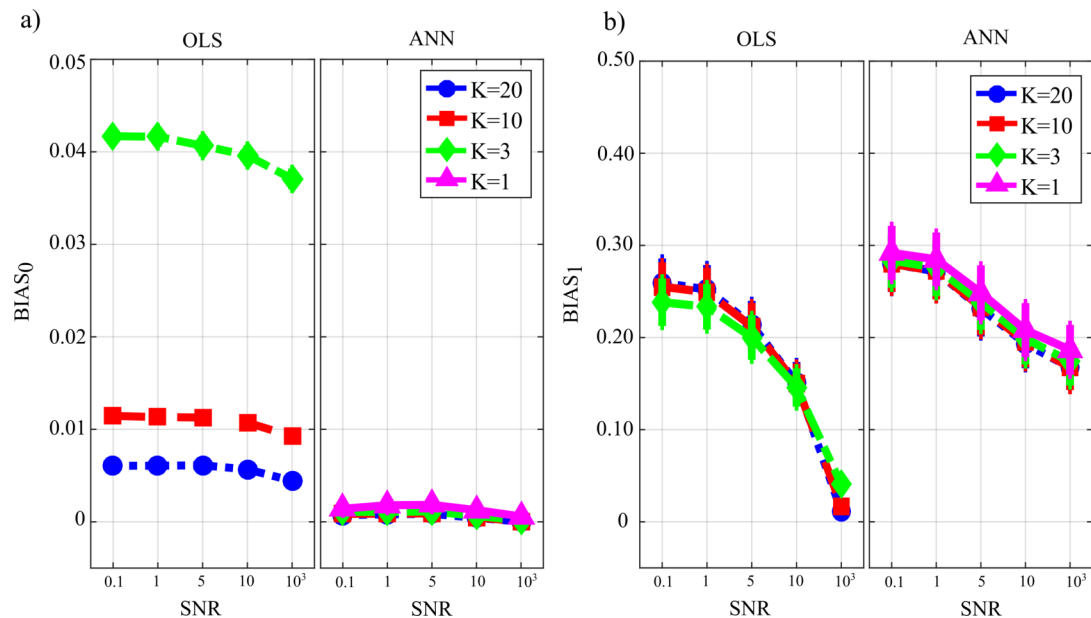
Figure 7. Distributions of the bias relevant to the estimation of GC on the null links ($BIAS_0$, panel a)
and on the non-null links ($BIAS_1$, panel b) plotted as a function of the ratio between data samples
available and number of parameters to be estimated (*K*) and of the ratio between signal amplitude and
noise amplitude (SNR), for OLS estimation and ANN estimation.

393 Table 4 reports the computation time required for the entire process of GC computation using the two
394 estimation approaches for different values of the *K* ratio when $SNR = 10^3$. OLS analysis includes SS
395 model identification and the subsequent evaluation of the null-case distribution for each couple of nodes
396 as described in the Methods section. ANN analysis includes SS model identification plus the training
397 process at $N_{\text{train}} = 1000, LR = 10^{-3}$. The analysis highlights the expected decrease of the computation
398 times with decreasing the *K* ratio and, more importantly, a strong reduction of the time requested for the
399 entire process when ANN is used in place of OLS. The computation time of OLS identification is not
400 reported for $K = 1$ due to the non-convergence of the solution to the DARE equation necessary for SS
401 model identification.

402 ***Results of the Simulation Study III***
403 In the last simulation study, we compare the performance of OLS and ANN at varying the proportion
404 between the number of data samples available and parameters to be estimated (*K*-ratio). The results of the
405 two-way repeated-measures ANOVAs, expressed in terms of F-values and computed separately on all the
406 performance parameters considering *K* and *TYPE* (i.e., the method used: OLS or ANN) as main factors,
407 are reported in Table 5. The two-way ANOVA analysis highlights a strong statistical influence of the main
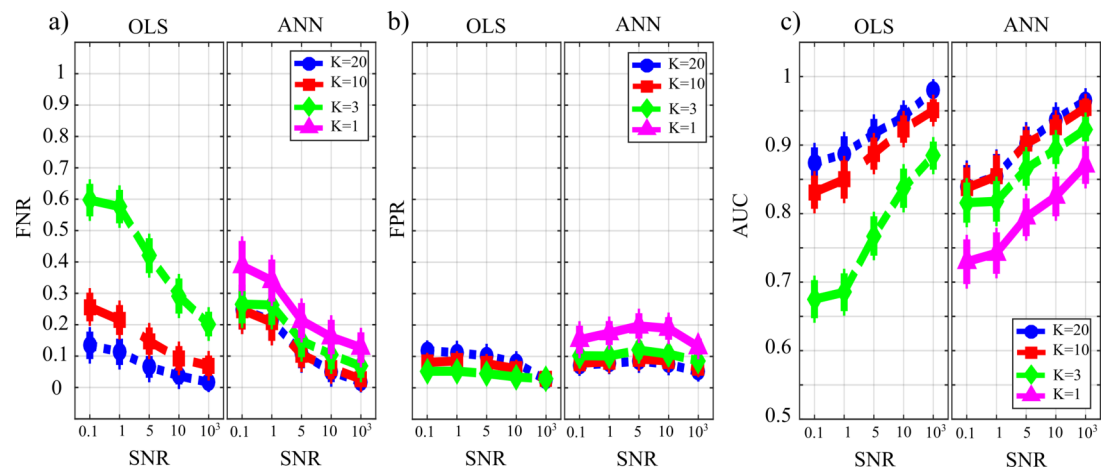
Figure 8. Distributions of the parameters assessing the performance of network reconstruction, i.e. the rate of false negatives (FNR, panel a) and of false positives (FPR, panel b) and of the area under the curve (AUC), plotted as a function of the ratio between data samples available and number of parameters to be estimated (K) and of the ratio between signal amplitude and noise amplitude (SNR), for OLS estimation and ANN estimation.

Method	OLS	ANN
$K = 20$	$9.1 \cdot 10^3$	142.18
$K = 10$	$4.5 \cdot 10^3$	107.28
$K = 3$	$1.3 \cdot 10^3$	67.6
$K = 1$	$-$	60.38

Table 4. Average computation time (in seconds, measured for 100 simulated networks) required by the OLS and ANN methods for the estimation of GC at different values of K ratio ($SNR = 10^3$)

408 factor K and TYPE and of their interaction ($TYPE \times K$) on all the performance parameters analyzed.
409 Also in this case the level $K = 1$ was not considered in the statistical analysis due to the non-convergence
410 of the DARE equation for the OLS case.
411   Figure 9 reports the distribution of the parameters $BIAS_0$ and $BIAS_1$ according to the interaction factor
412 $K \times TYPE$. The comparison of OLS (blue line) and ANN (red line) confirms the trends obtained for the
413 case $SNR = 10^3$ in the Simulation study II. In fact, the bias associated with ANN in the computation
414 over null links is very close to zero even in the most challenging condition of $K = 1$ with OLS showing
415 a very different trend with a strong increase associated with decreasing K-ratio values (panel a); in the
416 computation over the non-null links for ANN, the estimation bias displays a tendency to be stable but
417 remarkable if compared with OLS case.
418   Figure 10 reports the distribution of FNR, FPR, and AUC according to the interaction $K \times TYPE$.
419 The analysis of both false negatives (panel a) and false positives (panel b) shows a decrease with the
420 increase of the number of data samples available. The false negative rate is comparable for OLS and ANN
421 when $K = [20, 10]$, and then increases particularly for OLS while for ANN it assumes an average value
422 around 30% in the most challenging situation ($K = 1$) . The analysis of false positives (panel b) shows
423 a quite low and stable trend for the OLS case for all values of K, and an increasing trend for ANN up
424 to 20% ($K = 1$). Even in the best scenario of $K = 20$ the false positive rate assumes an average value of
425 $\sim$ 10%. The overall performance evaluated through AUC indicates high classification accuracy ($\geq 95\%$)
426 with a statistically significant difference between the two methods when $K = [20, 10]$ and a comparable
427 performance when $K = 3$ with no statistically significant differences highlighted by the post-hoc test.
428 However, in the most challenging situation of $K = 1$ the ANN method leads to an AUC value greater than
429 75%. As a general remark, there is a worsening of the performance in reconstructing the GC network if
430 compared with a sparser simulated network (Study II) that is more evident in the ANN case.

Factors	DoF	$BIAS_0$	$BIAS_1$	FNR	FPR	AUC
TYPE	(1, 99)	1295***	3518***	105,7***	491***	174,4***
K	(2, 198)	7454***	1196,2***	968,2***	111,1***	1468***
TYPE $\times$ K	(2, 198)	6770,5***	15,8**	268,5***	69,5***	102,8***

Table 5. F-values and corresponding degrees of freedom (DoF) of the two-way repeated measures ANOVA investigating the effects of the factors K (ratio between data samples and number of model parameters) and $TYPE$ (estimator used, i.e. OLS or ANN) on the performance parameters of GC estimation ($BIAS_0$, $BIAS_1$) and of network reconstruction (FNR, FPR, AUC).***, $p < 10^{-5}$.
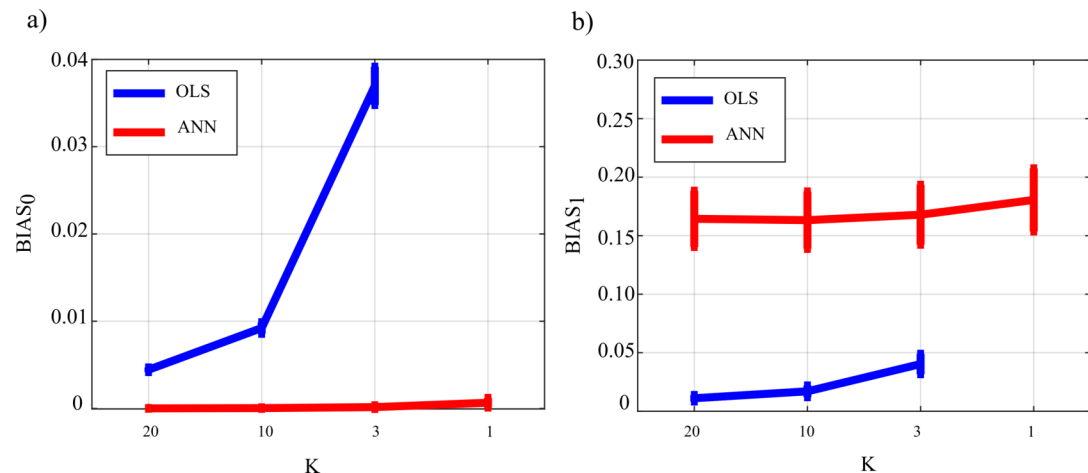
Figure 9. Distributions of the bias relevant to the estimation of GC on the null links ($BIAS_0$, panel a) and on the non-null links ($BIAS_1$, panel b) plotted as a function of the ratio between data samples available and number of parameters to be estimated (K), for OLS estimation (blue) and ANN estimation (red).

APPLICATION TO PHYSIOLOGICAL TIME SERIES

This section reports the application of the conditional GC, defined as in equation (12) and computed using OLS and ANN estimators, to the analysis of physiological networks formed by several time series reflecting the variability of heart rate, respiration, blood pulse propagation time, and of the amplitudes of different brain waves detected from EEG signals. The dataset used for the analysis was collected in a previous study on the interactions between various organ systems during different levels of mental stress (Zanetti et al., 2019).

Data acquisition and pre-processing

The experimental protocol involved eighteen healthy participants with age between 20 and 30 years, from whom different physiological signals were recorded during three tasks inducing different levels of mental stress: a resting condition lasting 12 minutes and consisting in watching a relaxing video (R); a mental arithmetic test during which the volunteer had to carry out the maximum number of 3-digit sums and subtractions (M); a sustained attention task that consisted in following a cursor on the screen while trying to avoid some obstacles (G). The experiment was approved by the Ethics Committee of the University of Trento, and all participants provided written informed consent. The study was in accordance with the Declaration of Helsinki.

The acquired physiological signals were the Electrocardiogram (ECG) signal, the respiratory signal (RESP) monitoring abdomen compartment movements, the blood volume pulse (BVP) signal measured through a photoplethysmographic technique, and Electroencephalogram (EEG) signals acquired using 14 channels Emotiv EPOC PLUS (international 10-20 locations). More details on the instrumentation and acquisition steps can be found in (Zanetti et al., 2019). The acquired physiological signals, representing the dynamical activity of different integrated physiological systems, were processed to extract synchronous time-series representing the time-course of different stochastic processes. Specifically, a template matching algorithm was employed to extract R peaks from the ECG and then measure R-R interval time series
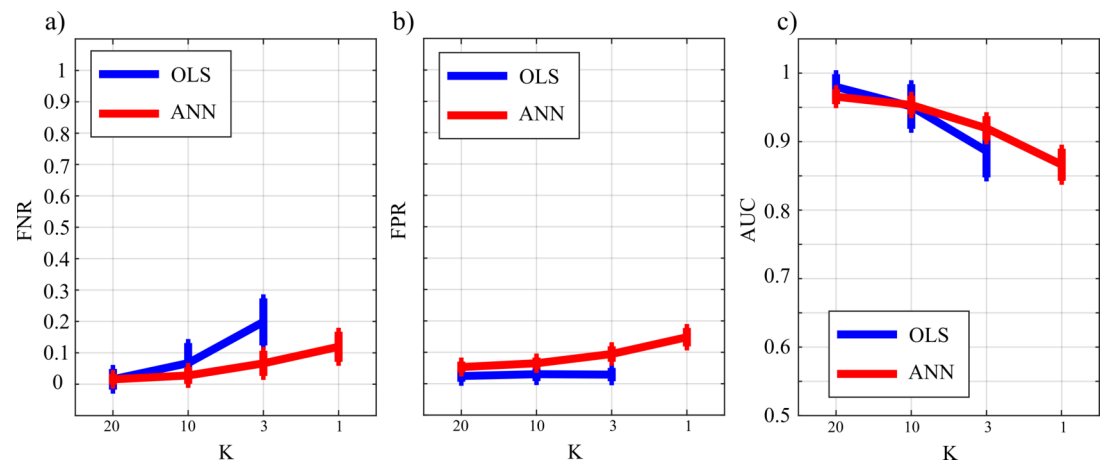
Figure 10. Distributions of the parameters assessing the performance of network reconstruction, i.e. the rate of false negatives (FNR, panel a) and of false positives (FPR, panel b) and of the area under the curve (AUC), plotted as a function of the ratio between data samples available and number of parameters to be estimated (K) for OLS estimation (blue) and ANN estimation (red).

(process η). The breath signal was sampled in correspondence of the R peaks to attain respiratory time series (process ρ). Moreover, the pulse arrival time was extracted as the time interval between the ECG R peak and the maximum derivative of the BVP signal (process π) for each cardiac cycle. With regard to brain activity, the power spectral density (PSD) of the EEG signals measured at the electrode F_z was calculated using a 2-s long sliding window with 50% overlap. Then, for each window, the PSD was integrated within four different frequency bands to obtain time series representative of the δ (0.5-3 Hz), θ (3-8 Hz), α (8-12 Hz) and β (12-25 Hz) brain wave amplitudes. The use of these frequency bands was motivated by studies which relate increasing levels of fatigue or alertness with higher PSD of the δ , θ and α processes and lower PSD of the β process (Sciaraffa et al., 2020; Tran et al., 2007; Trejo et al., 2007). The brain time series extracted in this way was synchronous with those obtained resampling at 1 Hz the three cardiovascular time series using spline interpolation (Zanetti et al., 2019). The rate of 1 Hz, which sets a time scale for the analysis which is compatible with the spectrum of heart rhythms, has already been used in previous studies in the field of network physiology for analyzing the time series from different body locations (Bashan et al., 2012; Bartsch et al., 2015). The uniformity of the final sampling rate and the synchronization of the signals acquired from different devices permitted to obtain seven synchronous time series for all the physiological districts.

Following the procedure described above, synchronous segments of the seven time series were selected inside each experimental condition (R, M, or G); each time series consisted of 300 samples, corresponding to five minutes of signal recording. All time series were checked for a restricted form of weak sense stationarity using the algorithm proposed in (Magagnin et al., 2011), which randomly extracts a given number of sub-windows from each time series and assesses the steadiness of mean and variance across the sub-windows. The seven time series extracted from each subject and experimental condition were considered to be a realization of a VAR process descriptive of the behavior of a dynamical system that describing the observed network of physiological interactions. For each subject and condition, the parameters of the VAR model fitting the seven observed time series, $A_1, ..., A_p, \Sigma$, were estimated with the two procedures described (i.e., OLS and ANN). The model order p was estimated for each experimental condition and subject through the Bayesian Information Criterion (BIC) (Schwarz, 1978).

Granger Causality Analysis

To assess the topological structure of the physiological network, the conditional Granger causality between each pair of nodes, $F_{i \to j|s}$, was computed through SS analysis applied to the VAR parameters estimated with the two presented methods (i.e, OLS and ANNs), and its statistical significance was assessed with the associated approach (i.e., using surrogate data for OLS and exploiting the intrinsic sparseness after the training process for ANN). The analysis was performed between each pair of processes as driver and target ($i, j = [\eta, \rho, \pi, \delta, \theta, \alpha, \beta], i \neq j$) and collecting the remaining five processes in the conditioning

489   vector with index s. Moreover, to confirm the results obtained in (Antonacci et al., 2020c) on the same
490   data, the in-strength - defined as the sum of all weighted inward links (Rubinov and Sporns, 2010) - was
491   computed for a specific network node (pulse arrival time π). The effect of the different experimental
492   conditions on the in-strength evaluated for the π node was assessed through the Kruskal-Wallis test
493   followed by the Wilcoxon rank-sum test between pairs of conditions. All analyses were performed with a
494   model of dimension Mp, where $M = 7$ and $p \sim 4$ (depending on the BIC) on time series of 300 points,
495   corresponding to $K \sim 10$ relating the amount of data sample available to the model dimension. The
496   performed analysis can be replicated by running the MATLAB script Test_Application in the released
497   toolbox for a single subject taken from the entire dataset (TimeSeriesStress).

498   ### *Results of Granger Causality Analysis*

499   Figure 11 depicts the network of physiological interactions reconstructed through the detection of the
500   statistically significant values of the conditional Granger causality ($F_{i \rightarrow j|s}$) computed for all pairs of
501   processes belonging to the analyzed network. The weighted arrows represent the most active connections
502   among the systems (arrows are present when at least three subjects show a statistically significant value of
503    $F_{i \rightarrow j|s}$). To ease interpretation and comparison between OLS and ANN estimates, the three sub-networks
504   representative of brain, body and brain-body interactions are depicted with arrows of different colors.
505   The networks estimated using OLS in the three experimental conditions (Figure 11.a-c) exhibit similar
506   structures to those estimated using ANN (Figure 11. (d-f)); the main difference is that networks estimated
507   with ANN show greater sparsity than those estimated with OLS.

508   A qualitative analysis of the networks illustrates the existence of a highly connected body sub-
509   network (red arrows), a weakly connected brain sub-network (purple arrows), and a pattern of brain-body
510   interactions (green arrows) that changes with the experimental condition. The body interactions are
511   characterized, consistently across the three conditions, by cardiovascular links (interactions from η to π)
512   and cardio-respiratory links (interactions between η and ρ), with a weaker coupling between ρ and π.
513   The use of ANN reveals a preferential direction from ρ to π that is not present in the condition M and is
514   bidirectional in the condition G. The topology of the brain sub-network assessed by the ANN method is
515   less stable across conditions, and looses consistency moving from R to G. On the contrary, in the OLS
516   case, the topology seems to be more consistent exhibiting weaker connections moving from R to M and
517   from M to G. The analysis of brain-body interactions reveals that such interactions are mostly directed
518   from the brain to the body sub-networks; in this case, the use of ANN clearly shows an increasing of
519   brain-body interactions during the condition G.

520   Figure 12 reports the distribution of the values of the in-strength index evaluated for the π node in each
521   experimental condition. For both OLS and ANN, the median value of the in-strength index is significantly
522   higher in the condition R with respect to the condition G. The use of ANN highlights lower values for the
523   in-strength parameter even if the trend is the same moving across the three experimental conditions. These
524   results show that both approaches detect a decrease of the information flow directed to the cardiovascular
525   node of the body subnetwork, documented by the reduction of the in-strength index in the G condition for
526   the process π.

527   ## APPLICATION TO A RING OF NON-LINEAR ELECTRONIC OSCILLATORS

528   In this section we investigate the application of GC, in its unconditional version, computed through OLS
529   and ANN by exploiting the SS approach, to a dataset of electronic non-linear chaotic oscillators, recorded
530   from a unidirectionally-coupled ring of 32 dynamic units, previously realized with the aim of studying
531   remote synchronization (Minati, 2015a; Minati et al., 2018). In the literature, it has been pointed out that a
532   single transistor oscillator can exhibit very complex activity and a ring of coupled oscillators can create a
533   community structure with statisical properties resembling physiological systems (Takahashi, 2013; Stam,
534   2005; Minati et al., 2015). The previous analysis has shown how it is possible to provide a mesoscopic
535   description of the information exchanged between different nodes of a network which represents the
536   activity of several physiological systems. On the other hand, the employment of an electronic circuit
537   comprising a ring of oscillators, provides a system of reduced scale and complexity, with respect to a
538   physiological one, yielding full access to the activity of each individual node. The resulting time series,
539   measured as voltage output by each oscillator, were considered as input for a VAR model and for an ANN,
540   descriptive of the behavior of the entire network ring.
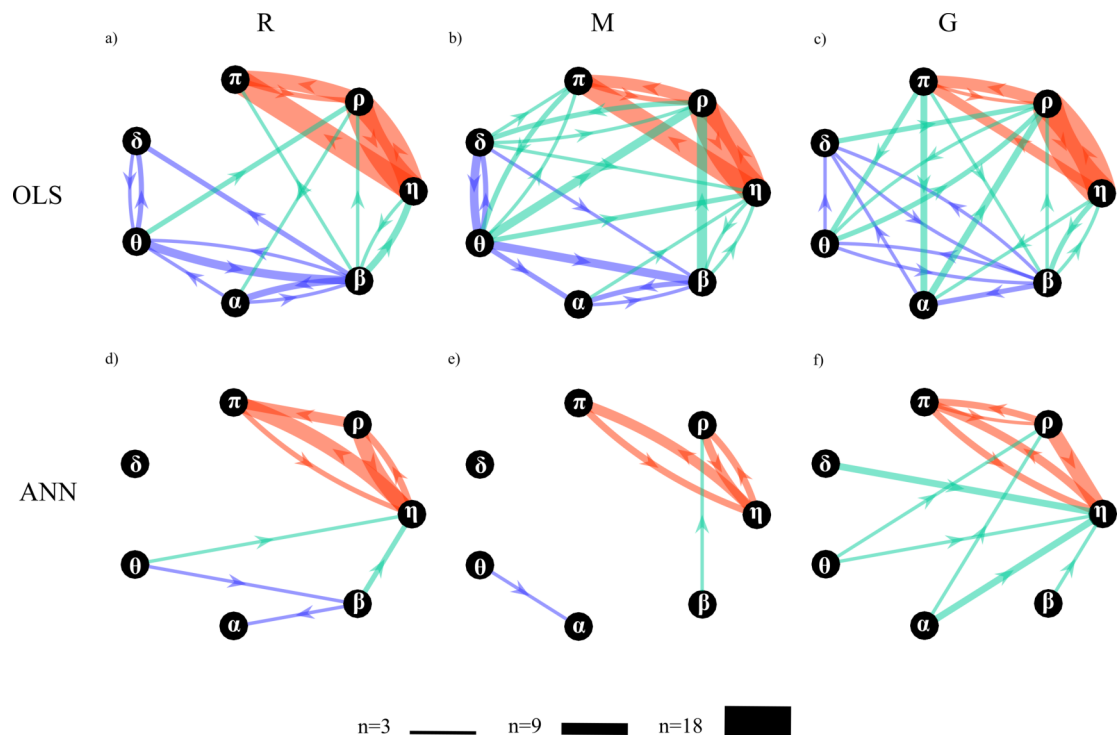
Figure 11. Topological structure of the network of physiological interactions reconstructed during the rest (R), mental arithmetic (M) and serious game (G) experimental conditions. Graphs depict significant directed interactions within the brain (purple arrows), body (red arrows) and brain-body (green arrows) sub-networks. Directed interactions were assessed counting the number of subjects for which the conditional Granger causality ($F_{i \rightarrow j|s}$) was detected as statistically significant using OLS (a-c) or ANN (d-f) in the estimation process. The arrow thickness is proportional to the number of subjects (n) for which the link is detected as statistically significant.

System description and synchronization analysis

The structural diagram of the oscillator circuit corresponding to each node in the network is reported in Figure 13.a and comprises four summing stages associated with low-pass filters. Three such stages with negative gains $G_1 = -3.6$, $G_2 = -3.12$, $G_4 = -3.08$ and filter frequency $F_1 = F_2 = F_3 = 2$ kHz are arranged as a ring oscillator. Two Integrator stages with integration constants $K_1 = 3.67 K_2 = 0.11 \mu s^{-1}$ with mixing gains $G_3 = -0.5$ and $G_5 = -0.71$ are overlapped to this structure. The ring is completed through fourth summing stages having $F_4 = 100$ kHz $\gg F_1$ with one input (gain $G_6 = 0.132$) which is necessary to close the internal ring itself and another (gain $G_i = -1.44$) connected to the previous oscillator in the ring network (Figure 13.b). To limit the voltage swing for the off-chip signal a gain inverter $G_0 = -0.4$ is installed. The recorded time series have a length $l = 65536$ points and are sampled with a sampling frequency $f_s = 100$ kHz and are freely available (Minati, 2015b).

The frequency spectrum of each node is represented by three peaks: the most prominent (central one) at $f_c \approx 2.8$ kHz and two weaker ones (sidebands) at $f_l = f_c/2 \approx 1.4$ kHz and $f_h = f_l + f_c \approx 4.2$ kHz. The higher sideband represents the mirror frequency of the lower one. As explained in (Minati et al., 2018), demodulation via envelope detection and subsequent interference occurs, and these phenomena lead to spatial fluctuations of the lower sideband amplitude that are closely related to the remote synchronization effect. In this system, remote synchronization is manifest as a non-monotonic decay of synchronization along the ring, wherein, with increasing distance from a given node, on average synchronization drops, then increases transitorily, and finally vanishes.

As in previous works (Minati, 2015a; Minati et al., 2018), we determined the instantaneous phase $\phi_m(t)$ and the envelope $A_m(t)$ of the output signal $v_m(t)$ of each oscillator m with the following relationship:

$$v_m(t) + i\hat{v}_m(t) = A_m(t)e^{i\phi(t)}, \qquad (20)$$
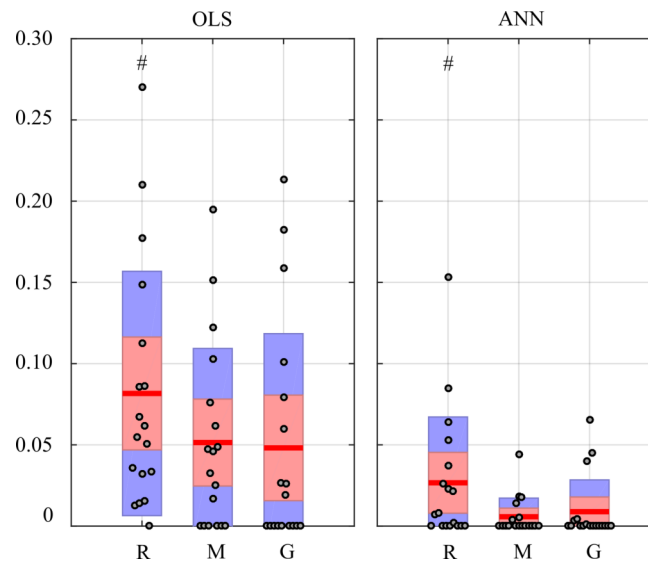
Figure 12. In-strength index computed for π node of the physiological network. Box plots report the distribution across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue bars) and the individual values (circles) of the in-strength computed at rest (R), during mental stress (M) and during serious game (G). Statistically significant differences between pairs of distributions are marked with # (R vs G).

560   where $\hat{v}_m(t)$ is the Hilbert transform of the recorded signal $v_m(t)$.

Given two generic time series Y_i and Y_j, amplitude synchronization for the envelope $A_m(t)$ was considered in terms of the maximum normalized cross-correlation coefficient for non-negative lags (that is, lags that take into account a possible propagation time along the direction of coupling, clock-wise in this system) $\max[C_{ij(\tau)}]_{\tau \geq 0}$ which is defined as:

$$C_{ij}(\tau) = \frac{k_{ij}(\tau)}{\sqrt{\sigma_i^2 \sigma_j^2}}, \tag{21}$$

561   where $k_{ij}(\tau) = \mathbb{E}[(Y_{i,n+\tau} - \mu_i)(Y_{j,n+\tau} - \mu_j)]$ is the time cross-covariance, $\mu_i = \mathbb{E}[Y_{i,n}]$ and $\mu_j = \mathbb{E}[Y_{j,n}]$ that
562   represent the mean of values of Y_i and Y_j; $\sigma_i^2 = \mathbb{E}[(Y_{i,n} - \mu_i)^2]$ and $\sigma_j^2 = \mathbb{E}[(Y_{j,n} - \mu_j)^2]$ which correspond
563   to the variances of Y_i and Y_j respectively.

564   In Figure 14 the analysis of cross-correlation coefficient performed for each pair of oscillators (i, j)
565   in the entire ring (panel a) is reported, alongside with the corresponding synchronization analysis for
566   three representative oscillator pairs (panel b) which exemplify the decay and transient recovery of
567   amplitude synchronization for three different distances from the node 1. The analysis of the cross-
568   correlation coefficient reveals that moving away from a node, synchronization initial decayed, then
569   gradually increased, rising till a distance $d \approx 8$, and eventually vanished as shown in Figure 14.a.
570   The structural coupling on the ring is only between first neighbors, as indicated by the master-slave
571   configuration, and the highlighted non-monotonic trend in the cross-correlation coefficient indicates a
572   situation of remote synchronization. The visual inspection of signal envelope for three different couples
573   of oscillators (panel b) confirms the analysis of cross-correlation with complete synchronization of the
574   couple $i = 1$, $j = 2$ (distance 1, $\max[C_{ij(\tau)}]_{\tau \geq 0} = 0.91$) that becomes a desynchronization for the couple
575   $i = 1$, $j = 6$ (distance 5, $\max[C_{ij(\tau)}]_{\tau \geq 0} = 0.19$); finally, the synchronization appears to be strong even for
576   the couple $i = 1$, $j = 9$ that means a physical distance of eight ($\max[C_{ij(\tau)}]_{\tau \geq 0} = 0.59$). The performed
577   analysis can be replicated by running the Matlab script Test_Oscillators in the released toolbox.

Granger Causality Analysis

579   From a theoretical point of view cross-correlation coefficient is a symmetric measure and thus, its value for
580   each time step is the same independently of the selected direction ($i \rightarrow j$, $j \rightarrow i$). For this reason, it is not
581   possible to assess if there is an information exchange between different oscillators. In order to test if there
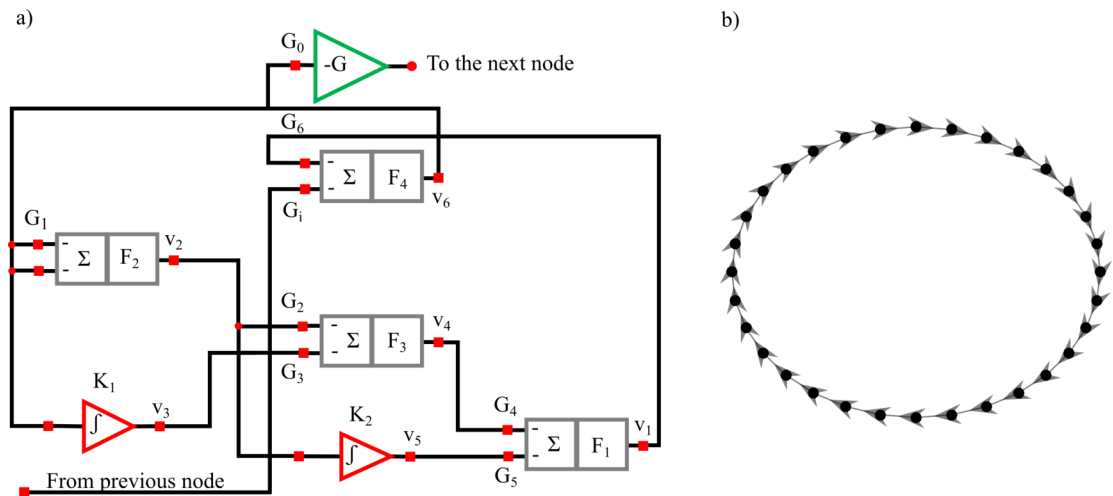
Figure 13. Diagram of the oscillator circuit corresponding to each node in the network (a). Master-Slave (unidirectional, clock-wise) structure of the ring comprising thirty two oscillators (b).

is information exchange between different oscillators, and if both methodologies can adequately capture the effects of "remote synchronization" restoring the results obtained in (Minati et al., 2018), Granger causality in its unconditional form was evaluated ($F_{i \to j}$) for each couple driver (i) target (j) belonging to the ring. Here, the past history of the target node j was approximated as $Y_{j,n}^p = [Y_{j,n-1}, \cdots, Y_{j,n-p}]$, i.e. with lagged components equally spaced in time. The past history of the driver node i was approximated as $Y_{i,n}^p = [Y_{i,n-1}, \cdots, Y_{i,n-p}]$. In the present analyses, the model order p was set to 16 with time series that were decimated firstly by a factor of 4 and subsequently by a factor 10. This process was needed in order to reduce the computational load and take into account the elimination of information storage and the propagation delays (Minati et al., 2018). In this condition, the ratio between the number of data samples and the number of VAR coefficients to be estimated is more or less equal to 3 ($K \approx 3$) and the partial variances needed for the evaluation of Granger causality were obtained through OLS and ANN by exploiting the theory of state-space models as described in the Methods section.

Figure 15 shows the results of the evaluation of unconditional GC ($F_{i \to j}$) performed for each couple (i, j) through OLS (Figure 15.a) and ANN (Figure 15.b). The estimated patterns are quite similar independently of the methodology used for estimation. The highest values of coupling estimated are linked to the previously described synchronization phenomenon: by considering a target (j) the coupling strength from the driver (i) to the considered target is very high nearby the position of the target; then decreases with the distance from the target with another peak at a distance approximately equal to 8 and finally vanishes. Another important feature is that this phenomenon is not bidirectional, but it is observable only in the direction $i \to j$ and not vice versa, as expected from the physical realization of the ring. Furthermore, the analysis of the pattern estimated through ANNs reveals more clearly the preferential synchronization clusters along the main diagonal. More in general, it is possible to observe a more sparse network when the analysis is performed through ANNs with the maximum value of observed coupling that is an order of magnitude smaller respect to the classical approach based on OLS (0.18 for OLS and 0.09 for ANNs).

The analysis of the computation time required for the estimation process, reveals a total temporal request of 28 hours (OLS $= 5.0605 \cdot 10^4$ s; ANNs $= 5.108 \cdot 10^4$ s) with the difference between the two methods ascribable to the training process of the ANN.

In order to test the degree of similarity between the two matrices, we computed the Spearman rank correlation coefficient that is a measure of the relationship between two variables when the data is in the form of rank orders. The Spearman rank correlation coefficient is in the range $[-1, 1]$ where 1 indicates complete agreement and -1 indicates complete disagreement. A value of 0 would indicate that the rankings were unrelated. Let R_i be the rank of the unconditional GC evaluated through OLS and S_i be the rank of the same analysis performed with ANN. Then, the rank-order correlation coefficient is defined to be the
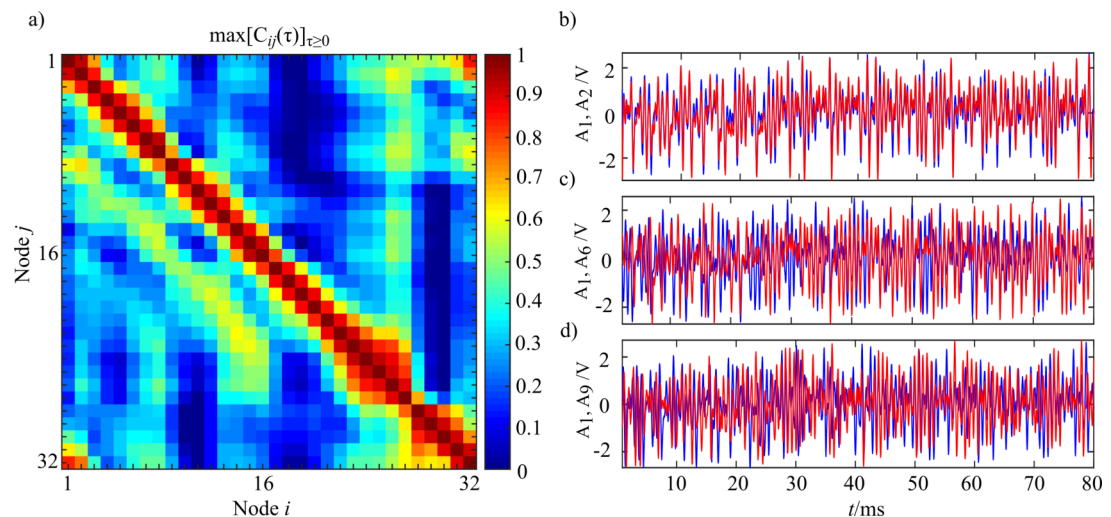
Figure 14. Instance of remote synchronization. The panel (a) reports the synchronization matrix for the entire ring intended as the maximum positive cross-correlation coefficient for the signal envelope $A_m(t)$. The panel (b) shows the signal envelope A_m for three different coupled of nodes demonstrating remote synchronization effects. The blue line represents A_1 with the red line that shows A_2 (panel b), A_6 (panel c) and A_9 (panel d). Time series were realigned to the lag for which the maximum value of cross correlation was observed.

linear correlation coefficient of the ranks, namely,

$$r_s = \frac{\sum_i (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2}\sqrt{\sum_i (S_i - \overline{S})^2}} \tag{22}$$

The significance of a nonzero value of r_s is tested by computing

$$t = r_s \sqrt{\frac{N-2}{1-r_s^2}}, \tag{23}$$

610 which is distributed approximately as Student's distribution with N-2 degrees of freedom (Hollander et al.,
611 2013). The result of this analysis reveals a value of $r_s = 0.84$ with a p-value $p < 10^{-5}$ indicating a strong
612 correspondence between the networks obtained through the two methodologies.

613 # DISCUSSION

614 ## Simulation study I

615 The first simulation study was designed to evaluate the effects of ANN training parameters on the GC
616 estimation process. We pointed out how the learning rate (*LR*) and the number of iterations (N_{train}) of the
617 gradient descent have an impact on the training process as regards both the regression problem and the
618 classification of significant network links (Zhang, 2006). The accuracy in the estimation of the regression
619 parameters, which reflects the accuracy in the magnitude of the estimated GC, was investigated while
620 varying the amount of data samples available for the estimation (Figure 3). As expected, the bias of GC
621 estimated over both null and non-null links increased in conditions of data paucity, while it was reduced
622 increasing the number of iterations of the gradient descent. An opposite trend was observed assessing the
623 bias along the null links for small learning rate ($LR = 10^{-5}$). This result was previously observed in the
624 context of classification analysis (Hoffer et al., 2017; Li et al., 2019) and is likely due to the fact that too
625 small learning rates can trap the ANN training process into local minima, resulting in our case in larger
626 differences between estimated and theoretical values of the conditional GC.
627   On the other hand, the analysis of the accuracy in reconstructing the network structure was tested
628 in terms of different classification parameters previously used to assess the structure of connectivity
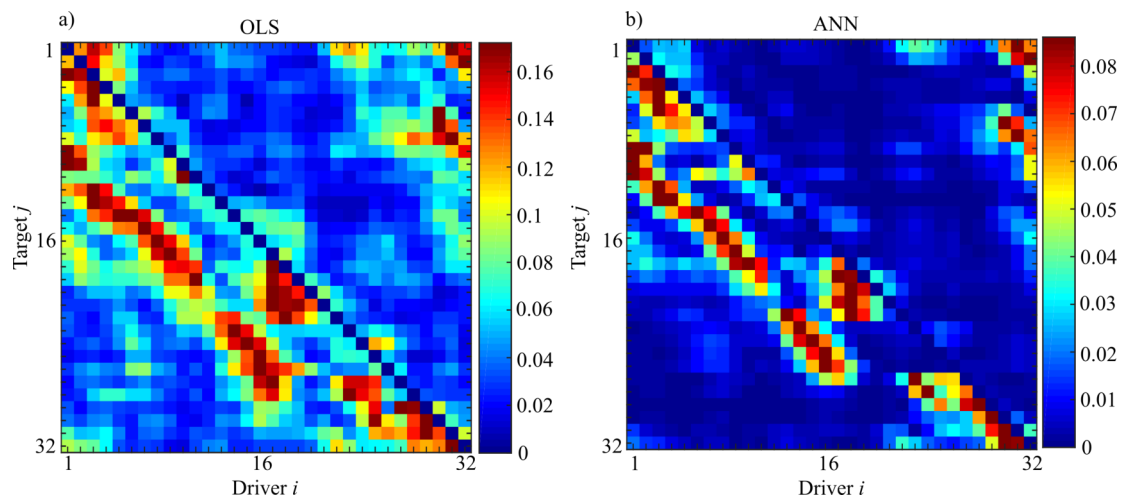
Figure 15. Unconditional Granger Causality Analysis performed on the network of 32 chaotic oscillators ($F_{i \rightarrow j}$). The matrices represent the analysis performed using OLS (panel a) and using ANNs (panel b) where each entry of the matrices corresponds to the strength of the causal influence from the driver i towards the target j. The value of Sperman rank correlation coefficient ($r_s = 0.84$) reveals a strong correlation between the two different patterns ($p < 10^{-5}$).

629 networks (Toppi et al., 2016b; Antonacci et al., 2019b, 2020c). The analysis (Figure 4) showed a general
630 improvement of the classification performance when increasing the number of data samples available and
631 the number of iterations of the SGD-l_1 algorithm, and when decreasing the learning rate. These results are
632 in line with previous studies analyzing the performance of estimators related with the concept of Granger
633 causality (Toppi et al., 2016a; Antonacci et al., 2020a, 2017), and help to optimize the parameter selection
634 for GC analysis based on ANN.

635  Such an optimization was performed in an objective way selecting the best combination of learning
636 rate and number of SGD-l_1 iterations that minimized the overall performance parameter S (Figure 5;
637 note that lower values of S indicate better performance). Varying the parameters N_{train} and LR within
638 ranges compatible with those suggested in a review of ANNs employed in classification analysis (Zhang,
639 2000), we identified the combination $LR = 10^{-3}$ and $N_{\text{train}} = 1000$ as the most suitable for optimizing the
640 performance of ANNs in the computation of magnitude and statistical significance of the conditional GC.
641 Overall, our simulation results lead to the following recommendations for GC estimation based on ANNs:

642  • the selection of the regularization parameter λ is crucial, and needs to be performed through
643   objective approaches such as the use of cross-validation employed in this study. In addition, a
644   careful selection of both the range and the number of λ values to be tested through cross-validation
645   is relevant; according to previous works and to the results obtained here, a range of three hundred
646   values seems to be sufficient.

647  • the factors which mostly affect the computation time are the number of data samples and the
648   number of iterations of the gradient descent (N_{train}). Although with a sufficient number of data
649   samples the impact of the number of iterations does not seem to be significant, we recommend to
650   set $N_{\text{train}} \geq 1000$.

651  • very small values of the learning rate should be avoided as they force the experimenter to increase
652   the number of iterations of the gradient descent to escape from local minima. We suggest the com-
653   bination $N_{\text{train}} = 1000$ and $LR = 10^{-3}$ as a good compromise between accuracy and computation
654   time.

Simulation studies II-III

656 The second and the third simulation studies were designed to analyze the performance of the proposed
657 ANN approach for GC estimation in comparison with the state-space analysis based on standard OLS
658 estimation of the VAR model (Barnett and Seth, 2015) in different experimental conditions. Simulation

study II has evaluated the effect of the number of data samples available (K-ratio) and the effect of the amplitude of white noise added to the original time series (SNR). Simulation study III was designed to compare the performance of the two methodologies on simulated networks with a smaller degree of sparsity with respect to the simulation study II (Pascucci et al., 2020). As in the first simulation, performances were assessed separately regarding the estimation bias and the statistical significance of the conditional GC. The bias analysis revealed the expected tendency to observe a larger difference between true and estimated GC values for decreasing the K ratio between amount of data samples and number of model parameters independently from the considered SNR value (Figures 7,9). This trend was marked for OLS-based GC estimates, confirming previous comparative studies (Schlögl and Supp, 2006; Antonacci et al., 2020b), and was much less evident for ANN-based estimates, which were more stable with respect to varying K. Considering the worst scenario in which the number of data samples available is equal to the number of VAR coefficients to be estimated ($K = 1$), the ANN estimation still yielded acceptable results, while OLS estimation was even not possible due to the non-convergence of the DARE equation contained in the SS estimation of GC (Antonacci et al., 2020c). The increasing bias observed for the OLS method while approaching the condition $K = 1$ is likely related to the fact that the matrix $[\mathbf{y}^p]^T \mathbf{y}^p$ (see methods) becomes progressively closer to singularity. On the other hand, a drawback of the ANN estimator is the substantial bias exhibited by the the conditional GC computed over the non-null links even in presence of sufficient amounts of data. This could be explained in part with the penalization directly applied on the matrix of coefficients that shrinks the values towards zero, and in part to the way by which the weights of the ANN are initialized (Scardapane and Wang, 2017). Figure 7 highlighted a smaller effect of SNR on the bias measures in the ANN case with respect to the OLS case, which reaches values of bias very close to zero when SNR is very high (SNR=10^3). This result could be explained by recalling the biased nature of regularization approaches and the tendency to counteract the effect of collinearity between regressors which may be induced by the additive noise (James et al., 2013).

Also the ability in reconstructing the network structure showed a tendency to decrease with the ratio K between the number of data samples and model parameters (Figures 8,10). In terms of overall accuracy, the ANN approach outperformed the OLS one for $K \leq 3$ and $SNR = 10^3$ resulting well-applicable ($AUC \approx 0.85$) even in the challenging condition $K = 1$. We ascribe this better performance to the use of the l_1 regularization introduced in the training of the ANN, which helps counteracting the collinearity between regressors induced by the decrease of the number of data samples available (Tibshirani, 1996; Silvey, 1969).

As expected, the AUC parameter reported in Figure 8 showed a tendency to decrease as a result of SNR reduction for both OLS and ANN. According to the results obtained in (Toppi et al., 2016a), for the OLS case the decreasing quality of the data leads to a strong increase of the estimated statistical thresholds, such that only a few connections survive to the assessment procedure. Otherwise, in the ANN case, such a trend can be explained by a wrong selection of the λ parameter during the training procedure triggered by white noise added and is in line with a previous study in which the effect of white noise in sparse regression methods was explored (Haufe et al., 2010). When a more dense network is analyzed, the results of Figure 10 showed a statistically significant difference between AUC values resulting from ANN and OLS estimation for K=[20,10] which become not statistically significant when K=3. If compared with the results obtained in Figure 8 (panel c and SNR=10^3) a deterioration of $\sim 10\%$ can be noticed in the reconstruction of the network structure, with an AUC value of $\sim 85\%$ when K=1 which becomes $\sim 75\%$ with a more dense network (Figure 10, panel c). The results here obtained are in line with those obtained in (Antonacci et al., 2020c) in which similar AUC trends are obtained for LASSO regression in simulated networks with a density of connected nodes equal to 50%.

When particularized to the rate of correct detection of null and non-null links, the performance under conditions of data paucity for both simulation studies differs for the two approaches, with ANN and OLS showing respectively better capability to correctly detect existing links (lower FNR) and better capability to correctly detect the absent links (lower FPR). The high rate of false negative detections exhibited by OLS when $K < 10$ is likely due to an inaccurate representation of the distribution of the GC under the null hypothesis of uncoupling, estimated empirically using surrogate time series (Antonacci et al., 2019a). On the other hand, the slightly higher rate of false positive detections exhibited by ANN is in line with previous findings in the context of information transfer estimation, in which the use of variable selection techniques showed few extra links, observed for different degrees of sparsity of the simulated network structure and values of K (Antonacci et al., 2020c; Haufe et al., 2010; Antonacci et al., 2020d). The latter

result is also confirmed by the value of false positives obtained for ANN case in Figure 10 (panel b) in which ANN proved to be more susceptible to false positives when a denser network structure is analyzed.

Concerning the effects of SNR (Figure 8), the decrease in the signal-to-noise ratio, regardless of the value of K-ratio considered, leads to a worsening of $\sim 20\%$ in the value of false negatives for both methods causing a strong reduction in the AUC value. However, ANN proved to be less affected by the reduction of signal-to-noise ratio and, even in the worst scenario of K=1 and SNR=0.1 the average value of AUC is above $\sim 75\%$ (Figure 8.c - purple line). These results are in line with different studies exploring the effects of SNR in Granger-based estimators in the time and frequency domain or in non-stationary regimes (Toppi et al., 2016a; Pascucci et al., 2020; Astolfi et al., 2007).

In sum, we provide the following remarks about the comparison between the two methods:

- if one is interested in the reconstruction of the network topology, ANNs can be used as a valid alternative to standard OLS approaches with a considerable computational cost reduction (Table 4).

- the capabilities in reconstructing the network topology of both methodologies are strongly influenced by the signal-to-noise ratio and the network density, with ANN performing better if sparse networks are considered and OLS which is more vulnerable to low SNR values.

- if one is interested in the assessment of coupling strength as measured by the GC values, ANNs are much more accurate than OLS in detecting small or zero GC values but are more biased in the detection of non-zero GC values.

- the use of ANNs with the parameter combination $N_{\text{train}} = 1000, LR = 10^{-3}$ guarantees a good level of accuracy in the estimation of GC even for conditions of strong data paucity.

Two final issues should be discussed concerning the structures of simulated networks and a technical aspect related to the methodology proposed. In functional brain networks analysis, the topology of network interactions is often represented by a full AR process for which AR coefficients are non-zero at each considered lag but fade out exponentially as the lags increase. Even if the structure of the AR process here simulated was not completely full, the network structure of simulation study III was used in a previous study that approximates properties of realistic brain networks, extending beyond classical approaches with a restricted number of nodes and fixed connectivity patterns (Pascucci et al., 2020). In fact, with a completely full AR process it should be better to use regression analysis with a structural constraint such as group LASSO which outperforms l_1 regularization techniques without structural constraint (i.e. LASSO regression and the methodology here proposed) as discussed in previous works (Mullen et al., 2015; Haufe et al., 2010). Nevertheless, it is worth stressing that the formulation here introduced can be easily extended to a similar form of group sparse development inspired by the group LASSO regression, by forcing all outgoing connections from a single neuron (corresponding to a group) to be either simultaneously zero or not as reported in (Scardapane et al., 2017).

As a final remark, we want to emphasize that, even if the l_1-regularized (SGD-l_1) and l_1-constrained (LASSO) algorithms target different objective functions, their behavior could be related since the idea at the basis of their functioning is the same (Tsuruoka et al., 2009). Nevertheless, the advantage of this type of formulation lies in the fact that it can be used indifferently with several types of loss functions (e.g. cross-entropy loss), or with different structures of the neural network designed to model non-linear relationships between input and output layers (i.e. the past states of the whole system and the present state of the target process) (Tsuruoka et al., 2009; Scardapane and Wang, 2017).

Application to Physiological Networks

Within the emerging field of network physiology, it is possible to analyze physiological interactions in a multivariate fashion, building complex networks whose nodes and edges represent different organ systems and their communication mechanisms (Bashan et al., 2012). However, identifying networks on the basis of the information exchanged between physiological signals is not a trivial task and requires the development of novel approaches (Faes et al., 2017c). As a main challenge is to interpret dense networks in terms of the underlying physiological mechanisms (Faes et al., 2015a; Porta and Faes, 2015), the study performed here was aimed to show the usefulness of GC measures based on ANNs for the description of brain, peripheral, and brain-heart interactions in a previously studied dataset (Zanetti et al., 2019). The usability of the proposed approach can be inferred linking the present results to those that we obtained in

recent studies where the possibility to describe the topology of physiological networks through penalized regressions was explored (Antonacci et al., 2020c,a). In particular, the very similar network topologies observed here and in (Antonacci et al., 2020c) using very different identification methods support the usefulness of sparse model identification approaches for the study of physiological interactions.

The analysis of the statistically significant values of the conditional GC led us to detecting specific topology structures (Figure 11). In the study of the peripheral sub-network of cardiovascular and respiratory interactions, we confirm the results of previous works highlighting the presence of significant interaction patterns which are observed consistently across physiological states (Zanetti et al., 2019; Porta et al., 2017; Antonacci et al., 2020c). These patterns comprise a strong information flow between η and ρ reflecting the mechanisms of respiratory sinus arrhythmia (Berntson et al., 1993) and cardio-respiratory synchronization (Schäfer et al., 1998), the causal interaction $\eta \rightarrow \pi$ reflecting the physiological effect of the heart rate on stroke volume and arterial pressure which modulates the arterial pulse wave velocity (Javorka et al., 2017), and the causal interaction $\rho \rightarrow \pi$ reflecting the influences of breathing on the intra-thoracic pressure, blood pressure and blood flow velocity (Drinnan et al., 2001). The main effect observed when changing the physiological state was the statistically significant decrease of the in-strength index of the vascular node π occurring with the transition from R to G (Figure 12); physiologically, this variation can be related to a reduced efferent nervous system activity from the cardiac and respiratory centers towards the vascular system during mental stress conditions (Antonacci et al., 2020c,a; Pernice et al., 2020). While the majority of these patterns were observed identically by OLS and ANN identification approaches, the interaction between ρ and η was detected as bidirectional using OLS and as unidirectional using ANN; the presence of unidirectional interactions $\rho \rightarrow \eta$ is physiologically more plausible with the mechanism of respiratory sinus arrhythmia (Berntson et al., 1993; Faes et al., 2015b).

As regards the analysis of the brain sub-network, we detected interaction patterns which are weaker and less consistent across physiological states. Using OLS, the total number of connections shows a tendency to decrease moving from R to M and to G. Using ANN, the brain sub-network is very sparse during R and M, and disconnected during G. The latter result is in line with our recent work in which the same dataset was analyzed through different measures of information dynamics computed through LASSO regression (Antonacci et al., 2020c). In such work, a different degree of disconnection was observed for the brain sub-network; given the general weakness of the connections, it is reasonable to assume that the results are influenced by the selection the regularization parameter λ that controls the amounts of shrinkage applied to the ANN weights, as in the optimization of λ the weaker connections have a higher probability to be discarded (Tibshirani, 1996; Tibshirani and Taylor, 2012). This confirms the importance of employing automatic strategies, such as that used in this work, for the selection the regularization parameter, in order to provide an objective quantification of the network topology. Here, the adoption of an automatic strategy led to detect a much more sparsely connected brain subnetwork using ANN than OLS, confirming results previously reported for this type of data (Zanetti et al., 2019).

The regularization approach implicitly present in ANN training allowed highlighting better than standard OLS analysis the modification of the structure of brain-body interactions across the considered physiological states. Indeed, while both OLS and ANN suggest an increase of the connections between brain and body during sustained attention (condition G), the results achieved with ANN highlight the emergence of causal interactions from brain to body moving from R and M to G. The rise of these connections, directed mostly to the ρ and η nodes of the peripheral sub-network, confirms the results of previous studies about the importance of the brain oscillations for attention tasks that can be correlated with the cardiac and respiratory activity (Tort et al., 2018; Kubota et al., 2001).

Application to chaotic electronic oscillators

The recorded time series and the master-slave unidirectional structure guarantee a higher level of stationarity and more elementary dynamics with a well known a-priori topological effect compared to physiological systems. For these reasons, it is reasonable to assume that electronic oscillators could represent a useful benchmark for testing in real settings new methods developed for the study of the interactions between dynamical systems.

The second application was therefore devised to demonstrate the validity of the proposed method, based on the combination of ANN and SS modeling, to compute GC from the output signals of a network of electronic oscillators. The analysis of the cross-correlation coefficient presented in Figure 14 revealed the existence of a preferential synchronization effect between groups of nodes that are not

819 directly connected via a physical link and, in particular, we found a maximum of the cross-correlation
820 coefficient at a distance $d \approx 8$. This result is in agreement with previous analyses performed in the same
821 ring of oscillators (Minati, 2015a; Minati et al., 2018) and with the recently introduced concept of remote
822 synchronization which reveals mutual synchronization between pairs of locally coupled groups of nodes
823 in a network. Thus, each group of nodes remotely synchronized is physically connected through a group
824 of intermediary nodes more weakly synchronized with them (Gambuzza et al., 2013).

825   In order to investigate if the observed remote synchronization corresponds to "remote" information
826 transfer, we performed unconditional GC analysis with both OLS and ANN. An inspection of Figure 15
827 clearly shows the good overlap between the networks estimated with the two methodologies; this result is
828 supported quantitatively by the analysis of the Spearman rank correlation coefficient ($r_s = 0.84$, $p < 10^{-5}$).
829 A similar analysis was performed on the same dataset by (Minati et al., 2018), who used uniform embed-
830 ding to approximate the history of target and driver time series as $Y_{j,n}^- = [Y_{j,n-\delta}, Y_{j,n-\tau-\delta}, \cdots, Y_{j,n-p\tau-\delta}]$,
831 $Y_{i,n}^- = [Y_{i,n-\delta-d}, Y_{i,n-\tau-\delta-d}, \cdots, Y_{i,n-p\tau-\delta-d}]$, where the additional time lag $\delta = 0.01$ ms was added to
832 ensure the full elimination of information storage (Wibral et al., 2013) and the lag d was introduced to
833 account for propagation delays and was set searching for the minimal prediction error over the range
834 $d \in [0, 2]$.

835   Here, we confirm the results obtained in Minati et al. (2018) with a different analysis that exploits the
836 SS representation of the VAR model and the ANN training. In particular, both methodologies can capture
837 the dynamical activity in a ring of electronic oscillators with a well-defined complexity and stability of the
838 network topology, since it is possible to obtain structures overlapped with those extracted performing the
839 analysis with different methodologies already reported in the literature. From a methodological point of
840 view, the strong overlap between the two networks can be motivated by the results of the simulation study
841 II for which at $K = 3$ the AUC parameter, indicating the capability in the reconstruction of the network
842 topology, showed a very small difference between the two methods. Furthermore, it is also important to
843 note that, as an effect of the l_1-norm applied to the weights of the network during the training process, the
844 maximum value of GC estimated with ANN is one order of magnitude less for ANN than OLS (Sun et al.,
845 2016).

CONCLUSIONS AND LIMITATIONS

847 This work documented that neural networks can be used in combination with state-space models for the
848 identification of linear parametric models, allowing computationally reliable and accurate estimation of
849 GC in its conditional and unconditional forms. In particular, we showed how this combined approach
850 leads to overcoming both the decrease in accuracy reported for traditional least-squares identification
851 when it needs to be performed in unfavorable conditions of data availability (Schlögl and Supp, 2006),
852 and the problems arising in the computation of GC estimated through different regression problems (Faes
853 et al., 2017d). ANNs are useful in particular to assess the statistical significance of GC estimates, favoring
854 the reconstruction of the network topology underlying the observed dataset without the need to employ
855 time-consuming asymptotic or empirical procedures for significance assessment.

856   The implementation of the proposed approach for the study of physiological networks and coupled
857 electronic oscillators documented its usefulness in practical applications, supported by the observation
858 of interaction patterns similar to those found in previous studies where the datasets were first studied in
859 terms of GC (Zanetti et al., 2019; Minati et al., 2018). All the findings in this work suggest that ANNs
860 are able to detect the strongest interactions providing output patterns of information dynamics which are
861 more straightforward and easy to interpret than those obtained with OLS.

862   An aspect not directly investigated in this work, that will be addressed with further studies, concerns the
863 effect of sparsity operated by l_1-constrained (e.g. LASSO regression) and l_1-regularized (e.g. ANN here
864 proposed) on GC measures that explicitly re-elaborate the VAR parameters. The induced sparsity in the
865 time domain might introduce uneven shrinking of the VAR coefficients over lags which eventually causes
866 undesired alterations in the frequency domain and this could impact the accuracy of several Granger-based
867 estimators in the frequency domain such as Partial Directed Coherence (Baccalá and Sameshima, 2001),
868 Directed Transfer function (Kamiński et al., 2001) or Granger causality in the frequency domain (Barnett
869 and Seth, 2014).

870   Future developments will aim at exploring the possibility of evaluating GC with non-linear ANNs
871 trained with SGD-l_1 to guarantee sparseness in the estimated patterns of causality. Although l_1-regularized
872 and l_1-constrained learning algorithms are not directly comparable due to their different objective functions,

873 a comparison of the two approaches in term of practicality is of interest in the field of stochastic
874 optimization(Tsuruoka et al., 2009).Furthermore, an extensive comparison between the well-known
875 LASSO regression and the ANN based approach here proposed, in different conditions of density of
876 connected nodes and signal-to-noise ratio, may provide useful insights in the use of either approach
877 (Pagnotta et al., 2019; Pascucci et al., 2020; Antonacci et al., 2020c).

878   Given the tight relation between information dynamics and the VAR representation of Gaussian
879 stochastic processes, future works can be envisaged to introduce ANNs for the estimation of measures of
880 information dynamics different than the GC (Faes et al., 2017c; Finn and Lizier, 2020), computed even
881 across multiple time scales (Faes et al., 2017a; Martins et al., 2020). Moreover, this new method will easily
882 find application even in different contexts, such as the study of dynamic information flow between stock
883 market indices (Scagliarini et al., 2020), between different brain regions with Granger-based estimators
884 (Astolfi et al., 2007), for time series analysis in climatology (Faes et al., 2017b), or for the study of gene
885 regulatory networks (Davidson and Levin, 2005).

REFERENCES

887 Antonacci, Y., Astolfi, L., Busacca, A., Pernice, R., Nollo, G., and Faes, L. (2020a). Model-based
888   transfer entropy analysis of brain-body interactions with penalized regression techniques. In *2020 11th*
889   *Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*, pages 1–2. IEEE.
890 Antonacci, Y., Astolfi, L., and Faes, L. (2020b). Testing different methodologies for granger causality
891   estimation: A simulation study. In *2020 28th European Signal Processing Conference (EUSIPCO)*,
892   pages 940–944. IEEE.
893 Antonacci, Y., Astolfi, L., Nollo, G., and Faes, L. (2020c). Information transfer in linear multivariate
894   processes assessed through penalized regression techniques: Validation and application to physiological
895   networks. *Entropy*, 22(7):732.
896 Antonacci, Y., Faes, L., and Astolfi, L. (2020d). Information dynamics analysis: A new approach based
897   on sparse identification of linear parametric models. In *2020 42nd Annual International Conference of*
898   *the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 26–29. IEEE.
899 Antonacci, Y., Toppi, J., Caschera, S., Anzolin, A., Mattia, D., and Astolfi, L. (2017). Estimating brain
900   connectivity when few data points are available: Perspectives and limitations. In *2017 39th Annual*
901   *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages
902   4351–4354. IEEE.
903 Antonacci, Y., Toppi, J., Mattia, D., Pietrabissa, A., and Astolfi, L. (2019a). Estimation of brain
904   connectivity through artificial neural networks. In *2019 41st Annual International Conference of the*
905   *IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 636–639. IEEE.
906 Antonacci, Y., Toppi, J., Mattia, D., Pietrabissa, A., and Astolfi, L. (2019b). Single-trial connectivity
907   estimation through the least absolute shrinkage and selection operator. In *2019 41st Annual International*
908   *Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6422–6425.
909   IEEE.
910 Anzolin, A. and Astolfi, L. (2018). Statistical causality in the eeg for the study of cognitive functions in
911   healthy and pathological brains. *Sapienza University of Rome*.
912 Astolfi, L., Cincotti, F., Mattia, D., Marciani, M. G., Baccala, L. A., de Vico Fallani, F., Salinari, S., Ursino,
913   M., Zavaglia, M., Ding, L., Edgar, C. J., Miller, G. A., He, B., and Babiloni, F. (2007). Comparison of
914   different cortical connectivity estimators for high-resolution eeg recordings. *Human brain mapping*,
915   28(2):143–157.
916 Attanasio, A. and Triacca, U. (2011). Detecting human influence on climate using neural networks based
917   granger causality. *Theoretical and Applied Climatology*, 103(1-2):103–107.
918 Baccalá, L. A. and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure
919   determination. *Biological cybernetics*, 84(6):463–474.
920 Barnett, L., Barrett, A. B., and Seth, A. K. (2018). Misunderstandings regarding the application of granger
921   causality in neuroscience. *Proceedings of the National Academy of Sciences*, page 201714497.
922 Barnett, L. and Seth, A. K. (2014). The mvgc multivariate granger causality toolbox: a new approach to
923   granger-causal inference. *Journal of neuroscience methods*, 223:50–68.
924 Barnett, L. and Seth, A. K. (2015). Granger causality for state-space models. *Physical Review E*,
925   91(4):040101.

926 Bartsch, R. P., Liu, K. K., Bashan, A., and Ivanov, P. C. (2015). Network physiology: how organ systems
927   dynamically interact. *PloS one*, 10(11):e0142143.

928 Bashan, A., Bartsch, R. P., Kantelhardt, J. W., Havlin, S., and Ivanov, P. C. (2012). Network physiology
929   reveals relations between network topology and physiological function. *Nature communications*,
930   3(1):1–9.

931 Berntson, G. G., Cacioppo, J. T., and Quigley, K. S. (1993). Respiratory sinus arrhythmia: auto-
932   nomic origins, physiological mechanisms, and psychophysiological implications. *Psychophysiology*,
933   30(2):183–196.

934 Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

935 Davidson, E. and Levin, M. (2005). Gene regulatory networks. *Proceedings of the National Academy of
936   Sciences*, 102(14):4935–4935.

937 Drinnan, M. J., Allen, J., and Murray, A. (2001). Relation between heart rate and pulse transit time during
938   paced respiration. *Physiological measurement*, 22(3):425.

939 Duggento, A., Guerrisi, M., and Toschi, N. (2019). Echo state network models for nonlinear granger
940   causality. *bioRxiv*, page 651679.

941 Faes, L., Marinazzo, D., Jurysta, F., and Nollo, G. (2015a). Linear and non-linear brain–heart and
942   brain–brain interactions during sleep. *Physiological measurement*, 36(4):683.

943 Faes, L., Marinazzo, D., and Stramaglia, S. (2017a). Multiscale information decomposition: Exact
944   computation for multivariate gaussian processes. *Entropy*, 19(8):408.

945 Faes, L., Nollo, G., Stramaglia, S., and Marinazzo, D. (2017b). Multiscale granger causality. *Physical
946   Review E*, 96(4):042150.

947 Faes, L., Porta, A., and Nollo, G. (2015b). Information decomposition in bivariate systems: theory and
948   application to cardiorespiratory dynamics. *Entropy*, 17(1):277–303.

949 Faes, L., Porta, A., Nollo, G., and Javorka, M. (2017c). Information decomposition in multivariate
950   systems: definitions, implementation and application to cardiovascular networks. *Entropy*, 19(1):5.

951 Faes, L., Stramaglia, S., and Marinazzo, D. (2017d). On the interpretability and computational reliability
952   of frequency-domain granger causality. *F1000Research*, 6.

953 Finn, C. and Lizier, J. T. (2020). Generalised measures of multivariate information content. *Entropy*,
954   22(2):216.

955 Gambuzza, L. V., Cardillo, A., Fiasconaro, A., Fortuna, L., Gómez-Gardenes, J., and Frasca, M. (2013).
956   Analysis of remote synchronization in complex networks. *Chaos: An Interdisciplinary Journal of
957   Nonlinear Science*, 23(4):043103.

958 Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal
959   of the American statistical association*, 77(378):304–313.

960 Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series.
961   *Journal of the American Statistical Association*, 79(388):907–915.

962 Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural
963   networks. In *Proceedings of the thirteenth international conference on artificial intelligence and
964   statistics*, pages 249–256.

965 Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods.
966   *Econometrica: journal of the Econometric Society*, pages 424–438.

967 Haufe, S., Müller, K.-R., Nolte, G., and Krämer, N. (2010). Sparse causal discovery in multivariate time
968   series. In *Causality: Objectives and Assessment*, pages 97–106.

969 Hoffer, E., Hubara, I., and Soudry, D. (2017). Train longer, generalize better: closing the generalization
970   gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*,
971   pages 1731–1741.

972 Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751.
973   John Wiley & Sons.

974 Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*,
975   4(2):251–257.

976 Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal
977   approximators. *Neural networks*, 2(5):359–366.

978 James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume
979   112. Springer.

980 Javorka, M., Krohova, J., Czippelova, B., Turianikova, Z., Lazarova, Z., Javorka, K., and Faes, L. (2017).

Basic cardiovascular variability signals: mutual directed interactions explored in the information domain. *Physiological Measurement*, 38(5):877.

Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics*, 85(2):145–157.

Kay, S. M. (1988). *Modern spectral estimation: theory and application*. Pearson Education India.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.

Kubota, Y., Sato, W., Toichi, M., Murai, T., Okada, T., Hayashi, A., and Sengoku, A. (2001). Frontal midline theta rhythm is correlated with cardiac autonomic activities during the performance of an attention demanding meditation procedure. *Cognitive brain research*, 11(2):281–287.

Li, Y., Wei, C., and Ma, T. (2019). Towards explaining the regularization effect of initial large learning rate in training neural networks. In *Advances in Neural Information Processing Systems*, pages 11674–11685.

Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.

Magagnin, V., Bassani, T., Bari, V., Turiel, M., Maestri, R., Pinna, G. D., and Porta, A. (2011). Non-stationarities significantly distort short-term spectral, symbolic and entropy heart rate variability indices. *Physiological measurement*, 32(11):1775.

Marinazzo, D., Pellicoro, M., and Stramaglia, S. (2012). Causal information approach to partial conditioning in multivariate data sets. *Computational and mathematical methods in medicine*, 2012.

Martins, A., Pernice, R., Amado, C., Rocha, A. P., Silva, M. E., Javorka, M., and Faes, L. (2020). Multivariate and multiscale complexity of long-range correlated cardiovascular and respiratory variability series. *Entropy*, 22(3):315.

Minati, L. (2015a). Remote synchronization of amplitudes across an experimental ring of non-linear oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(12):123107.

Minati, L. (2015b). Time series from ring oscillators. `http://www.lminati.it/listing/2015/a/time_series/figure_8/`. top_row.mat.

Minati, L., Chiesa, P., Tabarelli, D., D'Incerti, L., and Jovicich, J. (2015). Synchronization, non-linear dynamics and low-frequency fluctuations: analogy between spontaneous brain activity and networked single-transistor chaotic oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033107.

Minati, L., Faes, L., Frasca, M., Oświecimka, P., and Drożdż, S. (2018). Apparent remote synchronization of amplitudes: A demodulation and interference effect. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):063124.

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

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., Jung, T.-P., and Cauwenberghs, G. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry eeg. *IEEE Transactions on Biomedical Engineering*, 62(11):2553–2567.

Pagnotta, M. F., Plomp, G., and Pascucci, D. (2019). A regularized and smoothed general linear kalman filter for more accurate estimation of time-varying directed connectivity. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 611–615. IEEE.

Pascucci, D., Rubega, M., and Plomp, G. (2020). Modeling time-varying brain networks with a self-tuning optimized kalman filter. *PLoS computational biology*, 16(8):e1007566.

Pernice, R., Antonacci, Y., Zanetti, M., Busacca, A., Marinazzo, D., Faes, L., and Nollo, G. (2020). Multivariate correlation measures reveal structure and strength of brain-body physiological networks at rest and during mental stress. *Frontiers in Neuroscience*, 14:1427.

Porta, A., Bari, V., De Maria, B., and Baumert, M. (2017). A network physiology approach to the assessment of the link between sinoatrial and ventricular cardiac controls. *Physiological Measurement*, 38(7):1472.

Porta, A. and Faes, L. (2015). Wiener–granger causality in network physiology with applications to cardiovascular control and neuroscience. *Proceedings of the IEEE*, 104(2):282–309.

Rodrigues, J. and Andrade, A. (2015). Synthetic neuronal datasets for benchmarking directed functional

1036   connectivity metrics. *PeerJ*, 3:e923.

1037   Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: uses and
1038   interpretations. *Neuroimage*, 52(3):1059–1069.

1039   Scagliarini, T., Faes, L., Marinazzo, D., Stramaglia, S., and Mantegna, R. N. (2020). Synergistic
1040   information transfer in the global system of financial markets. *Entropy*, 22(9):1000.

1041   Scardapane, S., Comminiello, D., Hussain, A., and Uncini, A. (2017). Group sparse regularization for
1042   deep neural networks. *Neurocomputing*, 241:81–89.

1043   Scardapane, S. and Wang, D. (2017). Randomness in neural networks: an overview. *Wiley Interdisciplinary*
1044   *Reviews: Data Mining and Knowledge Discovery*, 7(2):e1200.

1045   Schäfer, C., Rosenblum, M. G., Kurths, J., and Abel, H.-H. (1998). Heartbeat synchronized with
1046   ventilation. *nature*, 392(6673):239–240.

1047   Schlögl, A. and Supp, G. (2006). Analyzing event-related eeg data with multivariate autoregressive
1048   parameters. *Progress in brain research*, 159:135–147.

1049   Schreiber, T. and Schmitz, A. (1996). Improved surrogate data for nonlinearity tests. *Physical review*
1050   *letters*, 77(4):635.

1051   Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

1052   Sciaraffa, N., Liu, J., Aricò, P., Flumeri, G. D., Inguscio, B., Borghini, G., and Babiloni, F. (2020).
1053   Multivariate model for cooperation: bridging social physiological compliance and hyperscanning.
1054   *Social Cognitive and Affective Neuroscience*.

1055   Siggiridou, E. and Kugiumtzis, D. (2015). Granger causality in multivariate time series using a time-
1056   ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 64(7):1759–
1057   1773.

1058   Silvey, S. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society:*
1059   *Series B (Methodological)*, 31(3):539–552.

1060   Solo, V. (2016). State-space analysis of granger-geweke causality measures with application to fmri.
1061   *Neural computation*, 28(5):914–949.

1062   Stam, C. J. (2005). Nonlinear dynamical analysis of eeg and meg: review of an emerging field. *Clinical*
1063   *neurophysiology*, 116(10):2266–2301.

1064   Stokes, P. A. and Purdon, P. L. (2017). A study of problems encountered in granger causality analysis from
1065   a neuroscience perspective. *Proceedings of the national academy of sciences*, 114(34):E7063–E7072.

1066   Sun, K., Huang, S.-H., Wong, D. S.-H., and Jang, S.-S. (2016). Design and application of a variable
1067   selection method for multilayer perceptron neural network with lasso. *IEEE transactions on neural*
1068   *networks and learning systems*, 28(6):1386–1396.

1069   Sun, X. (2000). *The Lasso and its implementation for neural networks*. PhD thesis, National Library of
1070   Canada= Bibliothèque nationale du Canada.

1071   Takahashi, T. (2013). Complexity of spontaneous brain activity in mental disorders. *Progress in Neuro-*
1072   *Psychopharmacology and Biological Psychiatry*, 45:258–266.

1073   Talebi, N., Nasrabadi, A. M., and Mohammad-Rezazadeh, I. (2018). Estimation of effective connectivity
1074   using multi-layer perceptron artificial neural network. *Cognitive Neurodynamics*, 12(1):21–42.

1075   Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical*
1076   *Society: Series B (Methodological)*, 58(1):267–288.

1077   Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*,
1078   pages 1198–1232.

1079   Toppi, J., Mattia, D., Risetti, M., Formisano, R., Babiloni, F., and Astolfi, L. (2016a). Testing the signifi-
1080   cance of connectivity networks: Comparison of different assessing procedures. *IEEE Transactions on*
1081   *Biomedical Engineering*, 63(12):2461–2473.

1082   Toppi, J., Sciaraffa, N., Antonacci, Y., Anzolin, A., Caschera, S., Petti, M., Mattia, D., and Astolfi,
1083   L. (2016b). Measuring the agreement between brain connectivity networks. In *2016 38th Annual*
1084   *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages
1085   68–71. IEEE.

1086   Tort, A. B., Ponsel, S., Jessberger, J., Yanovsky, Y., Brankačk, J., and Draguhn, A. (2018). Parallel
1087   detection of theta and respiration-coupled oscillations throughout the mouse brain. *Scientific reports*,
1088   8(1):1–14.

1089   Tran, Y., Thuraisingham, R., Wijesuriya, N., Nguyen, H., and Craig, A. (2007). Detecting neural changes
1090   during stress and fatigue effectively: a comparison of spectral analysis and sample entropy. In *2007 3rd*

1091   _International IEEE/EMBS Conference on Neural Engineering_, pages 350–353. IEEE.

1092 Trejo, L. J., Knuth, K., Prado, R., Rosipal, R., Kubitz, K., Kochavi, R., Matthews, B., and Zhang, Y.
1093   (2007). Eeg-based estimation of mental fatigue: convergent evidence for a three-state model. In
1094   _International Conference on Foundations of Augmented Cognition_, pages 201–211. Springer.

1095 Tsuruoka, Y., Tsujii, J., and Ananiadou, S. (2009). Stochastic gradient descent training for l1-regularized
1096   log-linear models with cumulative penalty. In _Proceedings of the Joint Conference of the 47th Annual
1097   Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the
1098   AFNLP_, pages 477–485.

1099 Wibral, M., Pampu, N., Priesemann, V., Siebenhühner, F., Seiwert, H., Lindner, M., Lizier, J. T., and
1100   Vicente, R. (2013). Measuring information-transfer delays. _PloS one_, 8(2):e55809.

1101 Wiener, N. (1956). The theory of prediction. _Modern mathematics for engineers_.

1102 Zanetti, M., Faes, L., Nollo, G., De Cecco, M., Pernice, R., Maule, L., Pertile, M., and Fornaser, A. (2019).
1103   Information dynamics of the brain, cardiovascular and respiratory network during different levels of
1104   mental stress. _Entropy_, 21(3):275.

1105 Zhang, G. P. (2000). Neural networks for classification: a survey. _IEEE Transactions on Systems, Man,
1106   and Cybernetics, Part C (Applications and Reviews)_, 30(4):451–462.

1107 Zhang, G. P. (2006). Avoiding pitfalls in neural network research. _IEEE Transactions on Systems, Man,
1108   and Cybernetics, Part C (Applications and Reviews)_, 37(1):3–16.