SAPIENZA
UNIVERSITÀ DI ROMA

Department of Computer, Control and Management Engineering
University of Rome La Sapienza

Ph.D. Program in Automatic control, Bioengineering and Operations Research

Curriculum: Bioengineering - XXXIII Cycle

Estimating Functional Networks in Dynamical Biological Systems through Convex Optimization Theory

Yuri Antonacci

Thesis Advisor
Laura Astolfi, Ph.D.

Ph.D. program coordinator
Giuseppe Oriolo, Ph.D.

Academic Year 2019/2020

Contents

List of Figures

List of Tables

Acronyms

ACC Accuracy

ANNs Artificial Neural Networks

AR Autoregressive

AUC Area Under Curve

BCI Brain Computer Interface

BIC Bayesian Information Criterion

cTE conditional Transfer Entropy

DARE discrete algebraic Riccati equation

DC Directed Coherence

DTF Directed Transfer Function

E-NET Elastic Net

EEG Electroencephalography

F-LASSO Fused LASSO

FFNN Feed-Forward Neural Network

FNR false negative rate

FPR false positive rate

GC Granger causality

GCV Generalized Cross Validation

IAAFT Iterative Amplitude Adjusted Fourier Transform

ISS innovation form SS

jTE joint Transfer Entropy

LASSO Least Absolute Shrinkage and Selecting Operator

MAPE Mean Absolute Percentage Error

mBTS modified Backward in Time Selection

MI Motor Imagery

MSE Mean Squared Error

MuI Mutual Information

MVAR Multivariate Autoregressive model

OLS Ordinary Least Square

PC Partial Conditioning

PDC Partial Directed Coherence

PID Partial Information Decomposition

RMSE Root Mean Squared Error

RR Ridge regression

RSS Residuals Sum of Squares

SG-LASSO Sparse Group LASSO

SGD Stochastic Gradient Descent

SGD-l_1 Stochastic Gradient Descent l_1

SS State-Space

TE Transfer Entropy

VAR Vector Autoregressive model

Introduction

Complex systems abound in the natural and social world, e.g across systems as apparently diverse such as the human brain, the immune system, economics and the world wide web. Yet despite three decades of intense research activity in studying complexity, many big issues remain only partially resolved, including a good quantitative definition for a complex system. Quantitatively, complex systems are often described as collections of entities, where the global behaviour is a non-trivial result of the local interactions of the individual elements [1]. Many approaches, from time series analysis and stochastic modelling, have been proposed to model the behaviour of complex systems based on observed time series by separating the systems's behaviour between observed macroscopic and hidden microscopic scales [2]. A dynamical system is a stochastic process of the form $(X_n, Y_n,)_n$, where X_{n+1} depends only on X and possibly some noise, and Y_n depends only on X_n and possibly some noise. We think of X_n as the true state of the system at time n and Y_n as our observation of the system at time n [3].

There are a variety of tools we consider to study the "information processing" in a dynamical system such as model identification, state estimation, prediction or forecasting, estimation or detection. One of the most flexible and easy to implement, in the context of linear signal processing, is the Vector Autoregressive Model (VAR) whose identification process (i.e. the estimation of autoregressive parameters matrix) is at the basis of the most used estimators for analysing the statistical dependencies between different time series representing the activity of the entire dynamical system [4–7]. However, the identification procedure for specific combinations of number of processes-number of observations in the time series, could lead to severe correlation between the regressors resulting in high bias and variance in the used estimator, which can be counteracted with the use of penalized regression techniques [8, 9].

Chapter 1: The first part of this thesis work has been focused in introducing and testing multivariate convex regression methodologies, as a tool for estimating the statistical dependencies among different dynamical systems. Since that there are no extensive studies available that assess the performance of different penalized regression techniques in different experimental conditions, in Chapter 1 I report a comparative analysis among different penalized regression techniques in the context of convex optimization which guarantees the existence of a solution to the VAR identification problem [10]. Here, we tested their ability to estimate autoregressive parameters in different condition of data samples available. Then, to prove the validity of the simulation study results, all the penalized regression techniques have been applied to electroencephalographic signals recorded from the scalp of an healthy volunteer during the execution of a motor imagery task [11].

Chapter 2: Another important tool for investigating and quantifying the information processing represented by the Information theory that has already been proved to be a useful framework for the

design and analysis of complex self-organized systems[12, 13]. Through the framework of information dynamics it is possible to dissect the information processing into basic elements that reflect the new information produced at each time instant about a target system, the information stored in the target systems and the modification of the information flowing towards the target systems [14, 15]. In this context, it has been recently introduced a tool able to compute any measure of information dynamics from the parameters of a VAR model used to characterizes an observed multivariate Gaussian process even in combination with state-space modelling [16, 17]. The tools of information dynamics have contributed substantially to the development of the new field of Network physiology, for which the human body can be modelled as an ensemble of complex physiological systems [18, 19]. Motivated by the fact that penalized regression techniques were not yet introduced and tested for the decomposition of information processing, in Chapter 2 it is investigated the possibility to integrate the so-called LASSO regression, in a framework for the computation of these measures. LASSO regression is selected on the basis of the trade-off between accuracy in the identification procedure and computational time required for its use from the study performed in the previous chapter. In this way, it has been possible to analyse information dynamics with a very high number of processes even if in the context of network physiology where typically only short realizations are available due to the process of synchronization of time series representative of different organ activities. The new tool has been tested firstly on simulation settings and explored in eighteen healthy subjects during different tasks inducing different levels of mental stress.

Chapter 3: LASSO regression is a method based on a specific mathematical model and requires a process of training. The results of Chapters 1 and 2 clearly demonstrated that could be computationally very onerous, especially if combined with state space modelling and in conditions of very long time series and dynamical systems with a very high number of processes. For this reason, in the Chapter 3 we tried to overcome this computational limitations by introducing an Artificial Neural Network equivalent to a VAR model [20]. In particular, thanks to a new training algorithm based on Stochastic gradient descent it has been possible to induce sparsity in the weights matrix of the network during the training phase, but with a less computational cost if compared with traditional LASSO implementation [21]. This new tool was then combined with a state-space model and using for Granger causality (GC) estimation. Firstly, it was tested in estimating GC on simulated dataset with a second step of validation on different real world datasets. The first analysis was performed, on the dataset used in the Chapter 2, in the context of Network physiology to have a comparative endpoint between LASSO regression and ANNs. This is based on the recently results which explicitly link Transfer Entropy and Granger causality for Gaussian processes [6].

The new tools has been tested in the GC estimation between different signals produced by electronic circuits studied as a link with the human brain's complex system. In fact, in the last years it has been demonstrated that a single transistor oscillator can exhibit a very complex activity depending on a easily tunable control parameter oscillating periodically, chaotically, or very close to criticality [22]. An experimental investigation of a ring of thirty diffusely coupled oscillators (each consisting of a bipolar junction transistor, three reactive components and a resistor) has demonstrated the spontaneous formation of multi-scale community structure as a function of coupling strength with elements of similarity to the organization observed in brain networks analysed through BOLD signals from functional Magnetic Resonance [23]. Furthermore, coupled single-transistor oscillators represent an undemanding experimental platform with which one may

attempt to recapture associations between brain connectivity and the dynamics observed through BOLD time series either empirically or in numerical simulation.

Chapter 4: Given the results of Chapters 1, 2 and 3 in the Chapter 4 an extensive analysis of the performance of different methods in estimating GC, was performed. In particular, due to the high dimension of the observed data, the "curse of dimensionality" may arise leading to unreliable estimation of direct causality. With the aim of carrying out an extensive comparative study, the performance of different methodologies, available in the current literature and explored in this thesis work, for the estimation of GC have been compared. Furthermore, we provided an implementation in combination with space state models for the methods that were not previously tested with this strategy. [24]. The performance of all the methods for GC estimation combined or not with state-space models have been tested in two different simulation studies, the first one representing a large network, with few links with the presence of redundancy between processes, and the second one with a small network highly interconnected.

A conclusion summarizing the main contributions of this Ph.D. project, together with their impact and limitations, closes this dissertation. Finally, two chapters are dedicated to a list of publications and CV originated from this Ph.D. course.

Chapters 1 and 4 has been carried out in collaboration with Neuroelectrical Imaging and BCI Lab (NEILab, PI: Donatella Mattia, MD, Ph.D.) at Fondazione Santa Lucia, IRCCS, Rome, Italy. Chapters 2 and 3 were performed in collaboration with the department of Engineering, University of Palermo (Prof. Luca Faes), Department of Industrial engineering and Center for Mind/Brain Sciences, University of Trento (Prof. Giandomenico Nollo and Ludovico Minati).

Chapter 1

Measuring Connectivity in Linear Multivariate Processes with Penalized Regression Techniques

1.1 Introduction

In the framework of linear signal processing, Vector Autoregressive model (VAR) has been proved to be a robust and reliable tool for analysing temporal dependencies among time series in several research fields, ranging from economics to biomedical sciences. In neuroscience, these models are extensively used for evaluating brain connectivity in order to understand how different brain areas communicate [25]. The general term of brain connectivity includes the concept of "functional connectivity", that refers to the evaluation of statistical dependencies between different neuronal units, and "effective connectivity", that is used for giving a description of a network with directional effects. By using the VAR model in a multivariate fashion, connectivity can be exploited in terms of coupling, i.e the presence of interactions, and causality, i.e the presence of temporal dependencies between time series [26]. The popularity of the VAR models lies in the fact that they are closely related to the frequency-domain representation of physiological time series, that are rich of oscillatory content, such as Electroencephalography (EEG) rhythms [27]. Connectivity is very often formalised with a Multivariate Autoregressive model (MVAR) which allows to derive time and frequency domain representations by the model coefficients and their spectral representation, respectively.

A very popular estimator defined in time and frequency domain is the Granger causality (GC), which represents a versatile tool for analysing the cause-effect relationships between different time series. GC was firstly formulated in the framework of bivariate autoregressive modelling, stating that a time series X G-causes another time 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 [28]. To account for the influence of other time series, the bivariate formulation has been extended to the multivariate case by using MVAR model, leading to the evaluation of the well-known conditional form of GC [29]. The most used measures, able to quantify causality in the frequency domain, are the Directed Transfer Function (DTF) [30], the Directed Coherence (DC) [31], the Partial Directed Coherence (PDC) and all its versions [32, 33]. All these measures are widely used and accepted

for the analysis of interactions among EEG time series and more in general between physiological time series. Another issue of great importance is the assessment of the statistical significance of the selected connectivity measure. The statistical procedures provide a significance threshold for the connectivity values estimated: if the value of the element of the matrix is below the threshold it is considered null. The threshold value can be obtained either empirically, based on surrogate time series, or derived theoretically [34].

The computation of all the aforementioned connectivity estimators and its assessment procedure lie on the identification procedure of the MVAR model used for the investigation of time lagged dependencies. This includes solving a linear regression problem in order to estimate Autoregressive (AR) coefficients or the variance of residual error term, depending on which type of estimator was selected. One possible approach is based on the Ordinary Least Square (OLS) method that finds the optimal solution starting from the VAR model hypothesis of independence between regressors and residuals of the regression [25] . The OLS accuracy, i.e, the quality of the AR parameters estimation, is strongly influenced by the number of data samples available for the estimation process and, in particular, the mean squared error of the MVAR parameters estimation decreases with the increase of available data samples [35]. As a rule of thumb, it is necessary that the ratio between the number of data samples available and the number of parameters to be estimated, known as k-ratio, is at least 10 in order to ensure the accuracy of OLS estimator. Otherwise, particularly when the k-ratio is close to one, the estimation problem becomes ill-posed and under-determined [36]. In such situation, the OLS does not guarantee the uniqueness of the solution, leading to an ambiguity in the VAR representation of the data. Moreover, the number of MVAR parameters to be estimated and the number of data samples available for the estimation process have a strong impact also on the accuracy of statistical assessment procedures needed for discarding spurious links [37]. To overcome these problems, it is possible to use other classes of regression techniques, such as penalized regression techniques, as an alternative to the OLS estimator. The idea is to add a constraint (based on the l_1-norm and/or the l_2-norm) to regularize the OLS estimator. The effect of the l_2-norm is to shrink of the MVAR parameters towards zero, whereas the effect of the l_1-norm is to select only specific coefficients, setting the others to zero. Both l_2- and l_1-norm methods can improve the OLS accuracy when the k-ratio is lower than 10, with the side-effect of also reducing the mean square error [38].

Penalized regression techniques have been used for GC estimation [39, 40], for neuro-imaging data analysis [41] and for brain connectivity analysis between EEG time series in a single trial or in real-time settings [11, 42]. The aforementioned works have some limitations: the different penalized regression methods were compared only for a single k-ratio value; convex and non-convex methods were directly compared even if it is well known that non-convex methods do not achieve a stable solution; in real-time application, only the group-LASSO regression was applied. Even if penalized regression techniques have been proposed in different studies, they are not commonly used in EEG studies. Motivated by the fact that there are no available extensive studies that assess the performance of the different regression techniques in different conditions, this work presents a comparison among the regression techniques under different experimental conditions (i.e., different values of the k-ratio, including those with which the OLS fails), both based on surrogate and real data, with the following objectives:

1. evaluate the accuracy of methods based on the l_2 and l_1- norms in estimating the values of

the MVAR parameters.

2. evaluate the accuracy of the penalized regression methods based on the l_1-norm in the selection of the estimated MVAR parameters;

3. provide guidelines for the use of penalized regression algorithms.

In particular, the performance of different penalized regression techniques will be compared on simulated EEG time series. The selected techniques are those that solve a linear convex optimization problem, in which the existence of a solution is guaranteed and can be found by computationally efficient algorithms, also in iterative versions (with further advantages in terms of computational and storage requirements): Ridge regression (RR) [43], Least Absolute Shrinkage and Selecting Operator (LASSO) regression [44], Elastic Net (E-NET) regression [45], Fused LASSO (F-LASSO) regression [46] and Sparse Group LASSO (SG-LASSO) regression [47]. The latter methods, due to the effect of l_1-norm, are also called variable selection techniques. A preliminary version of this work has been reported in [48], where the bias in the estimation of PDC through OLS and LASSO was compared. In the current work, we generalize the study by including other penalized regression techniques and extending the simulation studies. Furthermore, this work is not focused on a specific connectivity estimators but the performance will be evaluated during the MVAR model identification.

As described in the second part of the manuscript, penalized regression techniques were also applied to real EEG data, recorded from an healthy subject, performing a Motor Imagery (MI) task [11]. MI plays an important role in clinical and neuroscience studies and for this reason the neuronal representation of MI and motor execution have been studied intensively for years using brain imaging techniques such as functional magnetic resonance [49], EEG [50] and positron emission tomography [51]. Given the cooperation of different brain districts during the MI task, it is reasonable to assume that brain connectivity provides useful information for understating the phenomenon. For example, in the context for brain computer interface, it has been explored the feasibility of using brain connectivity as additional feature for the discrimination between different mental tasks [11, 52]. However, the authors pointed out how the accuracy during the entire process of connectivity estimation, i.e, the estimation of MVAR parameters and the assessment procedure, drops dramatically when few data samples are available. The idea here is to show how it is possible to estimate brain connectivity through penalized regression techniques, even when few data samples are available, and to discriminate between two different MI tasks with features extracted from a brain connectivity analysis. The algorithms for the VAR model identification based on penalized regression techniques are collected in the PID-LASSO MATLAB toolbox, which can be downloaded from https://github.com/YuriAntonacci/S-MVAR.

1.2 Methods

1.2.1 MVAR model Identification

Let us consider a dynamical system $\mathbf{Y}$ composed of M real-valued zero-mean stationary vector stochastic processes, $\mathbf{Y} = [Y_1 \cdots Y_M]$. Considering the time step n as the current time, the present and the past of the 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 completely described by the VAR model:

$$\mathbf{Y}_n = \sum_{k=1}^{p} \mathbf{Y}_{n-k}\mathbf{A}_k + \mathbf{U}_n, \tag{1.1}$$

where $\mathbf{A}_k \in \Re^{M\times M}$ is the matrix containing the autoregressive (AR) coefficients, and $\mathbf{U} = [U_1 \cdots U_M] \in \Re^{M\times 1}$ is a zero-mean white processes, denoted as innovations, with covariance matrix $\mathbf{\Sigma} \equiv \mathbb{E}[\mathbf{U}_n^T\mathbf{U}_n] \in \Re^{M\times M}$ ($\mathbb{E}$ is the expectation value).

Let us now consider a realization of the process $\mathbf{Y}$ involving N_s consecutive time steps, collected in the data matrix $[\mathbf{y}_1; \cdots ; \mathbf{y}_{N_s}] \in \Re^{N_s\times M}$, where the operator ";" stands for row separation, so that the i^{th} row is a realization of $\mathbf{Y}_i$, i.e. $\mathbf{y}_i = [y_{1,i}...y_{M,i}], i = 1, ..., N_s$, and the j^{th} column is the time series collecting all realizations of Y_j, i.e. $[y_{j,1}...y_{j,N_s}]^T, j = 1, ..., M$. OLS finds an optimal solution for the problem (1.1) by solving the following linear quadratic problem [25]:

$$\widehat{\mathbf{A}} = argmin_{\mathbf{A}}||\mathbf{y} - \mathbf{y}^p\mathbf{A}||_2^2, \tag{1.2}$$

where $\mathbf{y} = [\mathbf{y}_{p+1}; \cdots ; \mathbf{y}_{N_s}] \in \Re^{(N_s-p)\times\ M}$ is the matrix of the responses values, $\mathbf{y}^p = [\mathbf{y}_{p+1}^p; \cdots ; \mathbf{y}_{N_s}^p] \in \Re^{(N_s-p)\times\ Mp}$ is the matrix of the regressors and $\mathbf{A} = [\mathbf{A}_1; \cdots ; \mathbf{A}_p] \in \Re^{Mp\times M}$ is the matrix of coefficients. The 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 is minimized. When $N_s - p \leq Mp$ the OLS does not guarantee the uniqueness of the solution since the matrix $([\mathbf{y}^p]^T\mathbf{y}^p)$ becomes singular [53].

1.2.2 Penalized linear regression techniques

From a mathematical point of view, regularizing the OLS problem means adding an additional term (constraint) to the problem (1.2). In the Lagrangian form, the constraints are written in the cost function as additional weighted costs. The multi-objective optimization problems reads as:

$$\widehat{\mathbf{A}} = argmin_{\mathbf{A}}(||\mathbf{y} - \mathbf{y}^p\mathbf{A}||_2^2 + \mathcal{F}_{\mathbf{p}}(\mathbf{A}), \tag{1.3}$$

where $\mathcal{F}_{\mathbf{p}}$ is a penalty function applied to each column of regression coefficients $\mathbf{A}$. As anticipated in Section 1.1, in this work we analyze the linear penalty function described in Table I.

The first analyzed regression method is RR, characterized by l_2-norm based term that shrinks the estimated AR parameters towards zero, favouring the Mean Squared Error (MSE) [43]. The regularization parameters λ controls the amount of penalization to be applied, if $\lambda = 0$, RR reduces to OLS. The minimization of the RR functional leads to the closed form solution $\hat{\mathbf{A}} = [(\mathbf{y}^p)^T\mathbf{y}^p + \lambda\mathbf{I}]^{-1}(\mathbf{y}^p)^T\mathbf{y}$, where $\mathbf{I} \in \Re^{Mp\times Mp}$ is the identity matrix. Besides the reduction of the MSE, it would be of interest the reduction of the number of selected parameters (i.e., parameters where non-null values are imposed) especially when $N_s < Mp$, to identify only those AR parameters that have an effect on the response matrix $\mathbf{y}$. This result could be achieved by using a penalization cost based on the l_0-norm, which is equal to the number of non-zero elements in a vector [38]. Unfortunately, such a penalization function would render the optimization problem non-convex [54]. The effect of

Table 1.1: Penalty term associated with the problem 1.3

Method	Norm	Penalty term
OLS	-	-
RR	l_2	$\lambda\|\|\mathbf{A}\|\|_2^2$
LASSO	l_1	$\lambda\|\|\mathbf{A}\|\|_1$
E-NET	l_1,l_2	$\lambda_1\|\|\mathbf{A}\|\|_1 + \lambda_2\|\|\mathbf{A}\|\|_2^2$
F-LASSO	l_1,l_2	$\lambda_1\|\|\mathbf{A}\|\|_1 + \lambda_2 \sum_{i=1}^{Mp-1} \|\mathbf{A}_{i+1} - \mathbf{A}_i\|$
SG-LASSO	l_1,l_2	$\lambda_1\|\|\mathbf{A}\|\|_1 + \lambda_2 \sum_{i=1}^{g} w_i^g \|\|\mathbf{A}_{G_i}\|\|_2$

the l_0-norm can be approximated by using penalization terms based on the l_1-norm, such as the LASSO: the penalty term acts so as to shrink some coefficients and set others to 0 [44]. In LASSO, the parameter λ sets the trade-off between the number of non-null coefficients selected in the matrix $\mathbf{A}$ and Residuals Sum of Squares (RSS).

Similar to the LASSO, the E-NET simultaneously performs automatic variable selection, regulated by the parameter λ_1, weighting the term based on the l_1-norm, and coefficient shrinkage, regulated by the parameter λ_2 weighting the term based on the l_2-norm [45]. E-NET coincides with RR if $\lambda_1 = 0$ and with LASSO if $\lambda_2 = 0$. F-LASSO includes in its penalty function a further term respect to LASSO, which computes the l_1-norm of the vector of the differences between the coefficients of successive predictors. This term is used to enforce smoothness along the predictors, i.e. along the columns of the matrix $\mathbf{A}$ of the AR parameters. When λ_2=0 F-LASSO coincides with LASSO [46].

Finally, SG-LASSO is a convex combination of the LASSO and the group LASSO penalties [47]. This procedure imposes a structural constraint on the AR coefficients matrix, besides the basic sparsity one (l_1 term), to model the assumption that the predictors can be associated in groups of a given size. In the penalty term of SG-LASSO, g is the number of groups, decided a prior, and w_i^g denotes the weight for the $i-th$ group (in this study, as specified in [10], the weight of each group was set to 1). For multivariate regression analysis, as pointed out in [39], M groups of p elements for each column of $\mathbf{A}$ can be used . The p elements of each group, $[a_{ij}(1), \cdots, a_{ij}(p)] \in \Re^{p\times 1}$ with $(i,j) \in (1, \cdots, M)$, can be represented by one AR parameter from the time lag 1 to time lag p. In practice, the second term of the SG-LASSO penalization cost selects a subset of groups by setting to zero all the coefficients in some groups (sparsity between groups), whereas the first term encourage sparsity within each group.

Since the different rows of the matrices $\mathbf{y}^p$ and $\mathbf{y}$ can be considered as independent from each other, by diving them into train and test sets (in this work, 50% of the rows were used for training and 50% for testing), it is possible to estimate the optimal value of lambda (λ_{opt}). Training and test sets are standardized and the optimal values for λ are selected by using a Generalized Cross Validation (GCV) criterion [41, 55].

1.3 Comparing performances of different penalized linear regression methods

1.3.1 Simulation experiment

Problem (1.3) was solved for the different penalty terms listed in Table 1.1. For OLS ans RR cases, the SCoT package (Python) [11] was used; for the other cases, the SLEP package (MatLab®) [10] was used and the multivariate linear regression solved with accelerated gradient methods [56, 57] and parallel computing toolbox of MatLab.

The simulation study included the following steps:

1. Generation of simulated data-sets, fitting predefined ground-truth network under different values of the k-ratio $(0.5, 0.8, 1, 1.5, 2, 3)$. The k-ratio is computed as the ratio between the number of available data samples N_s, and the number of parameters to be estimated Mp.

2. Selection of the regularization parameters for each regression method by means of GCV criterion. The process was iterated ten times for each λ by randomly changing the training and testing sets.

3. Estimation of AR parameters by using the six regression methods, i.e., OLS, LASSO, E-NET, F-LASSO and SG-LASSO.

4. Evaluation of the performances by comparing the estimated AR parameters with those imposed in the corresponding ground-truth network.

To increase the robustness of the statistical analysis, the entire procedure was repeated 50 times with randomly generated data-sets.

1.3.2 Signal generation

The simulated data-sets were generated according to different ground-truth networks by means of an MVAR model used as generator filter [34]. The simulated multivariate time series ($M = 10$) were generated as realizations of a VAR(10) process fed by Gaussian noises with variance equal to 1. Furthermore, to generate signals replicating the spectral properties of EEG signals, an autoregressive component extracted from a real EEG signal was imposed in the model. The EEG signals were acquired with a sampling frequency of 200 Hz at C_z location during a resting state condition for a healthy subject [58]. The simulated networks were randomly generated with a density of the connected nodes of 15% (14 out of 90 possible connections). AR parameters were set by assigning randomly the lag in the range (1-10) and the coefficient value in the interval [-0.9, 0.9]. Under these constraint, 50 realizations of the VAR(10) processes were generated with different values of the k-ratio parameter in the range $(0.5, 0.8, 1, 1.5, 2, 3)$, so that the length of the simulated time series was $N_s = 50$ when $k = 0.5$ and $N_s = 300$ when $k = 3$.

1.3.3 Performances Evaluation

The performances of all the regression methods were assessed in terms of accuracy in estimating the strength of the network link (values of MVAR coefficients) and in terms of ability to reconstruct

the network structure. The first analysis was performed separately for non-null and null links, computing different measures of bias through the comparison between the estimated and theoretical values of AR coefficients values. Specifically, the values of the theoretical AR matrix $\mathbf{A}$ was compared with the estimated values stored in $\hat{\mathbf{A}}$ using the the Mean Absolute Percentage Error (MAPE) [59], if the theoretical value is different from zero, and the Root Mean Squared Error (RMSE) , if the theoretical value is zero:

$$MAPE = \frac{100}{M} \sum_{m \in M} \left| \frac{\mathbf{A}_m - \hat{\mathbf{A}}_m}{\mathbf{A}_m} \right|, \tag{1.4}$$

$$RMSE = \sqrt{\frac{\sum_{n \in N} (\mathbf{A}_n - \hat{\mathbf{A}}_n)^2}{N}}, \tag{1.5}$$

where M and N represent the set of non-null and null elements. Finally, the distribution of MAPE and RMSE were assessed across the 50 simulated network structures and presented separately for each method.

Second, the ability in assessing the variable selection procedure was tested comparing the two AR matrices representative of the estimated and theoretical network structure. This can be seen as a binary classification task where the existence (class 1) or absence (class 0) of each estimated connection is assessed and compared to the underlying ground-truth structure. Performances were assessed through computation of the false positive rate (FPR) (measuring the fraction of null links for which an AR coefficient different from zero was detected), false negative rate (FNR) (measuring the fraction of non-null links for which an AR coefficient different from zero was detected) and Area Under Curve (AUC) that summarized the information provided by FNR and FPR [20, 60]. Each of these performance parameters was obtained across the network links for each individual network, and its distribution across the 50 simulated network structures was then presented separately for each regression method.

The last considered performance parameter is the computational time (in seconds), required for the computation of the AR matrix $\hat{\mathbf{A}}$. For penalized regression methods, the estimation process is dived in two different steps: 1) selection of the regularization parameters, taking T_{sel} seconds; 2) computation of $\hat{\mathbf{A}}$ taking T_{comp} seconds. The process for the selection of regularization parameters was performed by means of GCV criterion within a range of 350 values for λ_1 and 100 values for λ_2 for each method and for each value of k-ratio (where applicable, see Table 1.1). To speed-up the entire process, the parallel computing toolbox implemented in MatLab®2016a was used.

1.3.4 Statistical Analysis

In the present work five different repeated measures two-way ANOVA tests were performed, one for each performance parameter (MAPE, RMSE, FNR, FPR and AUC), to evaluate the effects of different values of k-ratio (varied in the range $(0.5, 0.8, 1, 1.5, 2, 3)$ - factor K) and different regression methods (OLS, RR, LASSO, E-NET, F-LASSO, SG-LASSO - factor TYPE). For FNR, FPR and AUC, OLS and RR were not considered in the ANOVA test because they do not produce sparse AR matrices. The Greenhouse-Geisser correction for the violation of the spherical hypothesis was used in all the analyses. Tukey's post-hoc test was used for testing the differences between sub-levels of ANOVA factors. Bonferroni-Holm correction was applied for multiple ANOVAs computed on

Table 1.2: F-values of the two-way repeated measures ANOVA

Factor	MAPE	RMSE	FNR	FPR	AUC
K	367.2^{**}	324^{**}	140.6^{**}	23.7^{**}	143.7^{**}
TYPE	149.8^{**}	1863^{**}	85.3^{**}	277.5^{**}	129.1^{**}
$K \times TYPE$	95.6^{**}	321^{**}	13.5^{**}	10.1^{**}	11.5^{**}

different performance parameters.

1.3.5 Results of the Simulation Study

The results of the two-way repeated measures ANOVAs computed separately on all the performance parameters, expressed in terms of F-values considering K and TYPE as within main factors are reported in Table 1.2 (** is associated with $p < 10^{-5}$). The two-way ANOVA performed on MAPE and RMSE reveals a strong statistical influence of the main factors K and TYPE and of their interaction on the two performance parameters. Fig. 1.1 reports the distribution of the parameters MAPE and RMSE according to the interaction $K \times TYPE$. The comparison of the six different VAR identification procedures shows that the trends for penalized regression techniques (RR, LASSO, E-NET, F-LASSO and SG-LASSO) and OLS are very different. The analysis of error committed in the estimation of the non-null links (MAPE, 1.1.a) highlights that, independently from the used methods the error decreases with the increasing of the number of data samples available for the estimation. The six identification methods exhibit different performances as function of the value of k-ratio: when such number is the highest (K=3) there are no statistically significant differences between all the methods, with an error of $\sim 20\%$; when $K \leq 2$, LASSO (black line) and E-NET (purple line) assume smaller values of MAPE, with no significant differences between them as highlighted by Tukey's post-hoc test. On the other hand, OLS shows the highest values of MAPE, that becomes very high for the challenging conditions $K \leq 1$, in which these values suffers a sharp rise from 100% up to 280%. In the same conditions, the remaining methods show a slight increase with MAPE values remaining below 100%. Regarding the analysis of the error in the estimation of null-links (RMSE, 1.1.b), the error off all the methods that perform a variable selection (LASSO, E-NET, F-LASSO and SG-LASSO) is almost zero for all the level of K (even for the most challenging conditions of $K \leq 1$). In the set of penalized regression, RR (red-line) performs better than OLS but shows a mean value of RMSE higher respect the other penalized regressions. OLS (blu line) shows a similar trend of RMSE if compared with MAPE, with a sharp increase of bias when the amount of data samples become smaller. Also in this case, there is a rough discontinuity in the trend describing the RMSE, between $K = 1.5$ and $K = 1$, as highlighted also for MAPE.

The two-way ANOVA performed on FNR, FPR and AUC reveals a strong statistical influence of the main factor K and TYPE and of their interaction on FNR, FPR and AUC. Fig.1.2 reports the distributions of the parameters FPR, FNR and AUC according to the interaction $K \times TYPE$ for the VAR identification methods performing variable selection. The analysis of the rate of false negatives (Fig. 1.2.a) shows that the number of links incorrectly classified as null shows different trends depending on the regression methods used. All the methods show a decreasing number of false negatives with the increasing of the number of data samples available. However, in the most

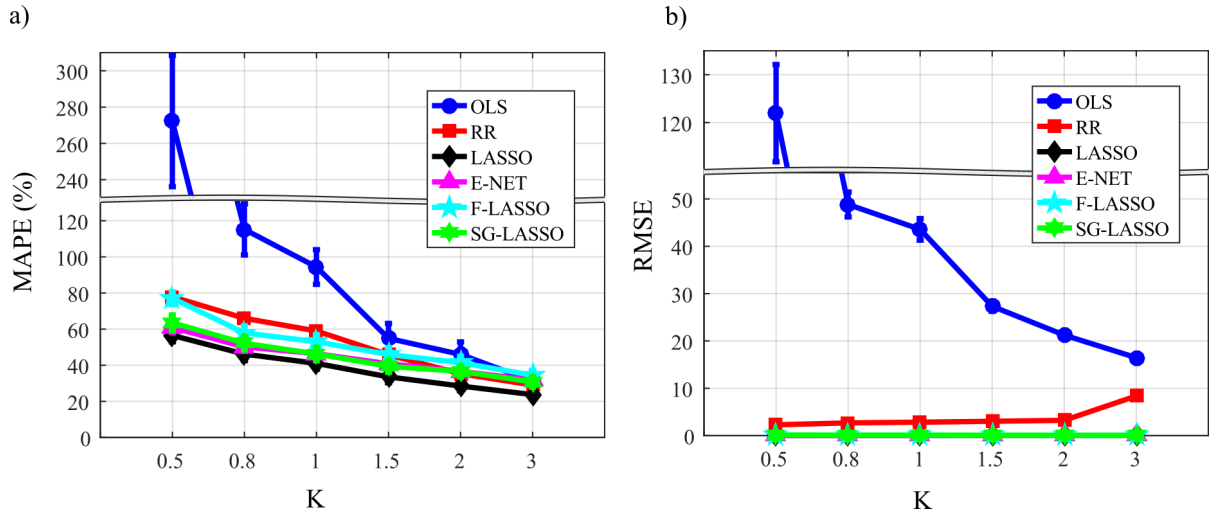

Figure 1.1: Distribution of the bias parameters computed for the non-null links (MAPE, a) and for the null (RMSE, b) considering the interaction $K \times TYPE$, expressed as mean value and 95% confidence interval of the parameter computed across 50 realizations for the six VAR identification methods and for different values of K.

challenging condition of $K = 0.5$ (the unknown AR parameters are twice the data samples) the highest values of FNR is showed by F-LASSO (azure line) that classifies as null links the half of those imposed as different from zero. SG-LASSO (green line) shows a comparable trend, even if, in the most challenging condition of $K = 0.5$, it performs better than F-LASSO. LASSO and E-NET (black and purple line, respectively) show the best performances, with overlapped trends and no significant differences for all the values of k-ratio as highlighted from Tukey's post-hoc tests; the values of false negatives for the latter methods range from almost zero for $K = 3$ to below 30% for $K = 0.5$.

On the other hand, the analysis of the rate of false positives (Fig. 1.2.b) shows that the number of links incorrectly classified as non-null is stable and almost negligible for all the methods except for F-LASSO (azure line) that shows higher average values of FPR even if below 0.1 (corresponding to 10%). The overall performance assessed through the AUC parameters is far better using LASSO and E-NET (Fig.1.2.c): the rate of correctly detected links in all the analyzed conditions is larger for LASSO and E-NET if compared with F-LASSO and SG-LASSO that perform slightly worse with F-LASSO that shows the worst performance. For $K = 3$ LASSO and E-NET show an average value of AUC close to 1 (that means perfect reconstruction of the network structure). In any case, such methods show the possibility to reconstruct the network structure with a very good accuracy ($\sim$ 0.75 in the worst scenario) even when the conditions are very challenging. Table 1.3reports the computational time (in seconds) required for the process of selection of the optimal value of lambda/s (T_{sel}) and for the computation of the MVAR parameters after the selection of lambda (T_{comp}). The two values are computed for each regression method and for each value of k-ratio. In the case of OLS, it was not possible to evaluate T_{sel} parameter since it is not included in the set of penalized regression methods. Times were recorded on a PC with an IntelCore i7-6700 processor, clock speed 3.40 GHz, 8 Gb RAM DDR4 (1.33 MHz), Intel (R) HD Graphics 530, 1024 Mb dedicated VRAM.

Table 1.3 shows the differences between the computational times of the six regression methods. Both, T_{sel} and T_{comp} increase with K. For all the values of k-ratio analyzed, the methods

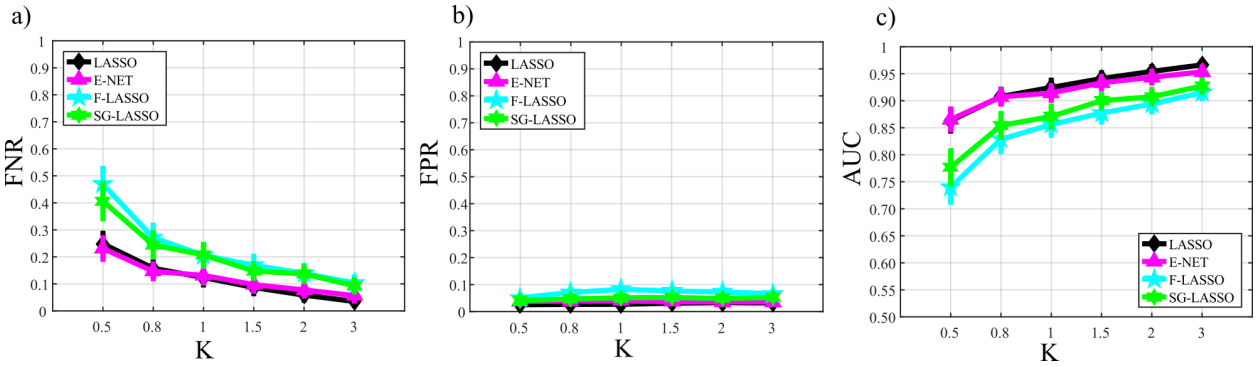

Figure 1.2: Distribution of FNR (a), FPR (b) and AUC (c) parameters considering the interaction $K \times TYPE$, expressed as mean value and 95% confidence interval of the parameter computed across 50 realizations for the four VAR identification methods, performing variable selection, and for different values of K.

Table 1.3: Computational time required for each regression method

Regression Method		K=0.5	K=0.8	K=1	K=1.5	K=2	K=3
OLS	T_{sel}	-	-	-	-	-	-
	T_{comp}	0.03	0.02	0.03	0.05	0.05	0.06
RR	T_{sel}	7.7	9.4	10.3	13.1	15.3	20.4
	T_{comp}	0.01	0.02	0.02	0.03	0.04	0.06
LASSO	T_{sel}	6.7	10.8	12.9	19.2	23.8	35.7
	T_{comp}	0.02	0.03	0.03	0.04	0.05	0.06
E-NET	T_{sel}	315	495	572	872	927	1400
	T_{comp}	0.03	0.03	0.04	0.04	0.05	0.09
F-LASSO	T_{sel}	588	860	882	1200	1300	1700
	T_{comp}	0.05	0.06	0.06	0.07	0.08	0.09
SG-LASSO	T_{sel}	870	1200	1500	1800	2000	2300
	T_{comp}	0.03	0.04	0.05	0.07	0.08	0.19

based on only one regularization parameter (i.e., LASSO and RR) need a shorter time for the selection step if compared with those based on two regularization parameters (i.e., F-LASSO, E-NET and SG-LASSO). Furthermore, among the methods based on l_1-norm, LASSO shows the lowest computational time required for the estimation process. The most time consuming method is the SG-LASSO.

To summarize all the trends analyzed so far, all the performance parameters are directly influenced by the amount of data available with a direct proportionality (the accuracy in the estimation process improves with the increase of the K value). Despite the cases of scarce amount of data available, all the penalized regression methods analyzed are able to reach excellent performance for both the estimation of AR parameters and assessment of the connectivity structure. Among those based only on l_1-norm i.e, LASSO and EL-NET regression methods, showed the best performance (with no differences in a statistical sense) and LASSO regression requires a computational time far less than that required from EL-NET.

1.4 Testing of linear regression techniques on real EEG data

In this section, we show that, relying on features extracted from a brain connectivity analysis, it is possible i) to estimate brain connectivity through penalized regression techniques even when few data samples are available and ii) to discriminate between two different motor imagery (MI) tasks.

1.4.1 Data description and pre-processing

The data were recorded by the authors of [11]. The data set includes 45 EEG channels with sintered Ag/AgCl ring electrodes. The locations of EEG channels corresponded to the international 10-20 system. The signals were recorder at a sampling rate of 300 Hz with three synchronized g.USBamp amplifiers (g.tec, Guger Technologies OEG, Graz, Austria). The amplifiers filtered the raw data with a 0.5-100 Hz band-pass and a 50 Hz notch filter. By the recorded electrooculogram, Ocular components in the EEG were reduced with a regression based approach [61]. The EEG data were further resampled at 100 Hz. To test the methods on real EEG signals, an healthy subject (male, right hand), with no prior experience in Brain Computer Interface (BCI) control, was selected to record one session, consisting of 90 trials of right hand motor imagery (HAND) and 90 trials and foot motor imagery (FOOT). Further details about the experimental paradigm and the pre-processing process are available in [11].

1.4.2 Single-trial Connectivity Estimation

For this analysis, in order to reproduce the condition of simulation study, 11 out of the 45 available channels were selected, specifically, C_5, C_3, C_1, C_2, C_4, C_6, CP_3, CP_4, C_z, CP_z, FP_z, and, as suggested in [11], 100 samples between the third and the fourth second were selected. The eleven time series obtained from each trial and for each condition (HAND - FOOT) were interpreted as a realization of a VAR process whose matrix of parameters $\hat{\mathbf{A}}$ was estimated with the six different identification methods under analysis (i.e. OLS, RR, LASSO, EL-NET, F-LASSO, SG-LASSO). The model order p was estimated for each experimental condition, using the Final Prediction Error (FPE) criterion [25]. All the analysis were performed by identifying VAR models of dimension Mp, where M=11 and $p \sim 15$ on time series of 100 points, which brought to work with values of k-ratio close to 1.
The estimated AR parameters were subsequently calculated in the frequency domain by means of the Fourier transform obtaining information about the existing connectivity relationships between time series [62].

1.4.3 Classification Task

To verify 1) that penalized regressions represent reliable methods for estimating connectivity between EEG time series and 2) that the features derived from connectivity analysis provide useful information, a classification task was performed.

The classification HAND vs. FOOT was repeated by using as features the frequency version of estimated AR parameters (real part of the Fourier transform of the AR parameters) in two different frequency bands, typically related with MI tasks [63]: α (8-12 Hz) and β (13-30 Hz). By considering f_i as the frequency interval under investigation and $M = 11$, it was possible to extract $M^2 f_i = 605$

features in α band ($f_i = [8 - 12]$) and 2178 in β band ($f_i = [13 - 30]$).
In order to ensure a ratio of 5 between the number of training cases and the number of classifier parameters to be estimated, i.e., to avoid overfitting during the training phase, a sub-selection of the features was performed [64]. In particular, we selected the 25^{th} highest -ranked ones according to the associated t-value (independent samples t-test, HAND vs FOOT) for each frequency bands [65]. Note that all the diagonal elements of the AR matrices were removed to maintain only information related to brain connectivity analysis.

The data were split in 70% for the training process (126 cases out of 180), 15% for validation (27 cases out of 180) and 15% for testing (27 cases out of 180). Each of the three data sets contained the same number of observations per class (HAND-FOOT). By using an holdout approach, at each iteration a different Feed-Forward Neural Network (FFNN) was trained, validated and tested [64]. The structure chosen for the FFNN is one of the most widely used for classification purposes [66, 67] that includes one hidden layer with one neuron (sigmoid activation function) and an output layer with a softmax activation function [68]. The initial weights of the network were randomly generated and the training was performed with gradient descent algorithm (with learning rate set to 10^{-3}), with cross-entropy used as cost function [69]. The training process was stopped by means of the early stopping criterion [70]. The process was repeated 100 times for each considered frequency band. As a performance parameter, we computed the classification Accuracy (ACC) on the test set [71].

For this study, a two-way repeated measures ANOVA test on the classification accuracy (ACC) was performed in order to evaluate the performance of classification accuracy depending on the VAR identification method (factor TYPE: OLS, RR, E-NET, LASSO, F-LASSO, SG-LASSO) and the frequency band (factor BAND: α,β). Moreover, to assess statistically differences between pair of distributions independent samples t-test were performed.

1.4.4 Results on real EEG data

Figs. 1.3 and 1.4 report the topological structures of the 25^{th} highest ranked features extracted for α and β frequency band, with each VAR identification method. The 25 features reported in each network represent the connections maximizing the statistical difference between HAND and FOOT according with t-value.

Fig. 1.3 shows that OLS and RR (Fig. 1.3a,b) share the selection of some connections between the right and the left hemisphere. However, only RR shows a pattern mostly involving the left hemisphere. LASSO, E-NET, F-LASSO and SG-LASSO (Fig. 1.3c,d,e,f) show two sub-networks, one over the left motor areas (around the C_3 electrode) and the other one over the ipsilateral motor areas, in which there is an involvement of fewer channels. The observation of these patterns highlights a deep involvement of the left centro-parietal areas located around the C_3 electrode and a small participation of the right hemisphere (C_4) in discriminating between the two MI tasks.This behaviour is more evident in the case of LASSO and E-NET which show a smaller number of weak connections (blue) and a larger number of strong connections (purple and green) if compared with other methods . F-LASSO and SG-LASSO show also strong connections in the right hemisphere (e.g $C_6 \rightarrow CP_4$ and $FP_z \rightarrow C_6$) resulting in a less readable pattern.

As reported in Fig.1.4, OLS shows a strong involvement of the right hemisphere with CP_4 as main hub of different connections (Fig.1.4a). On the other hand, RR shows an high involvement

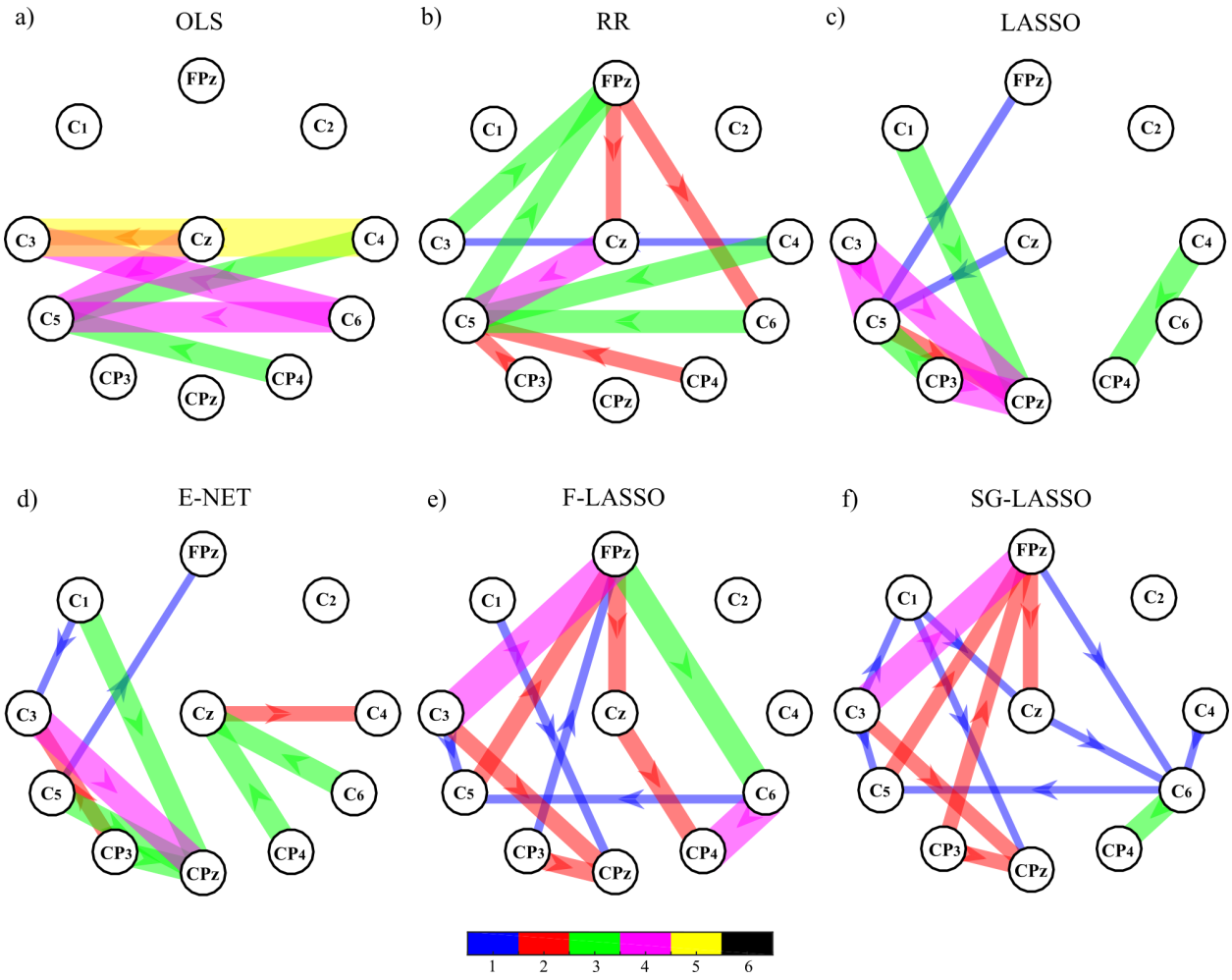

Figure 1.3: Topological structure of the 25 highest ranked features extracted in α band for: OLS (a), RR (b), LASSO (c), E-NET (d), F-LASSO (e) and SG-LASSO (f)VAR identification methods. The arrows represents the AR parameters, in its frequency version, that maximize the statistical difference between HAND and FOOT. The colour of the arrows encodes the number of times that the connections was selected as statistically significant in the frequency bins $8 - 12$ Hz (blue-1 times, black 6 times).

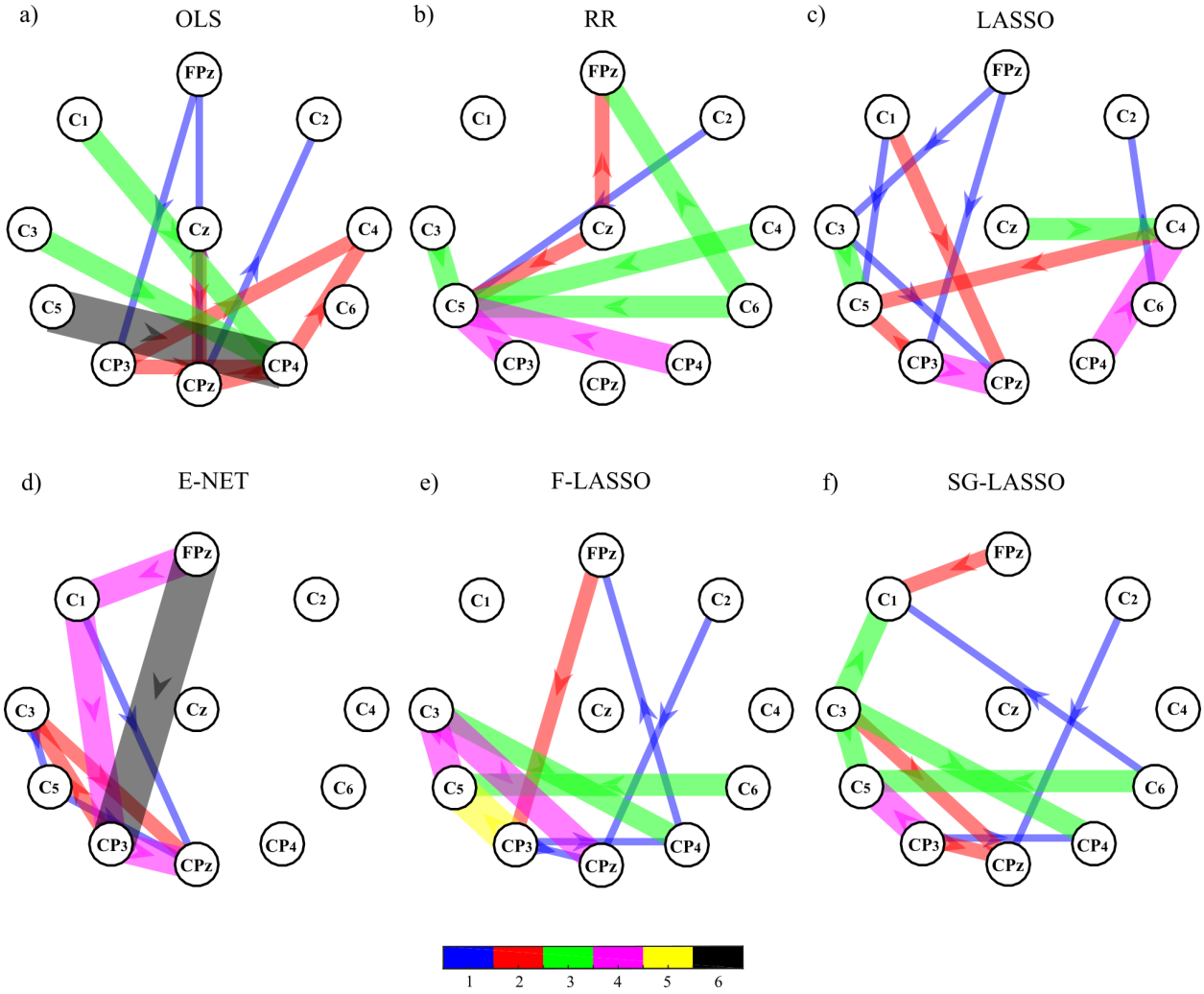

Figure 1.4: Topological structure of the 25 highest ranked features extracted in β band for: OLS (a), RR (b), LASSO (c), E-NET (d), F-LASSO (e) and SG-LASSO (f)VAR identification methods. The arrows represents the AR parameters, in its frequency version, that maximize the statistical difference between HAND and FOOT. The colour of the arrows encodes the number of times that the connections was selected as statistically significant in the frequency bins $13 - 30$ Hz (blue-1 times, black 6 times).

of the left hemisphere with C_5 receiving from different electrodes (Fig.1.4b). LASSO, E-NET, F-LASSO and SG-LASSO reveal the presence of a sub-network involving C_3,C_5, CP_3 and CP_z electrodes in the left controlateral hemisphere (Fig.1.4c,d,e,f). Also in this case there is a small involvement of the right hemisphere with E-NET that shows most active connections in the left hemisphere with no connections selected in the ipsilateral hemisphere (Fig.1.4d). More in general, with respect to the α band, the strength of some connections is increased when penalized regressions are used, as in the case of F-LASSO and SG-LASSO for which the connection $C_3 \rightarrow CP_z$ becomes stronger (Figs.1.3-1.4e,f). Furthermore, in such case the structure of the networks become sparser that those obtained in α band even if the number of features selected is the same.

Fig. 1.5 reports the distributions of the classification accuracy evaluated on the test set according with the interaction $BAND \times TYPE$ for each regression methods and for α and β frequency bands.

For all the regression methods, except for RR and E-NET, ACC is significantly greater in β band (blue boxes) with respect to α band (orange boxes). RR and E-NET do not show significantly differences between ACC evaluated in α and β frequency bands. Moving towards the analysis of

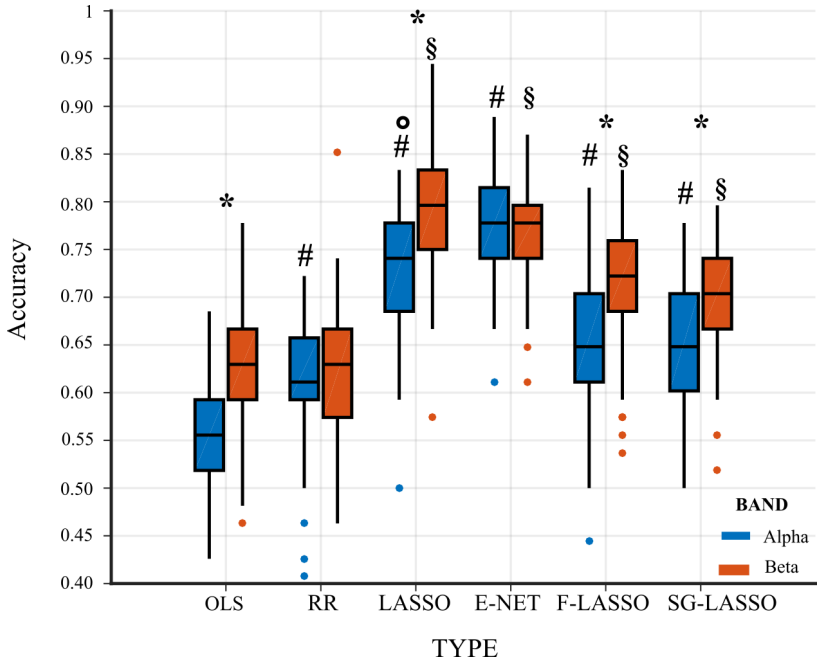

Figure 1.5: Classification accuracy evaluated on the testing set for each regression methods obtained in each frequency bands (α and β). Box plots report the distribution of classification accuracy across the 100 iterations by considering the interaction $TYPE \times BAND$ (F=18.57, $p < 10^{-5}$). On each box, the central line indicated the median, edge of the box indicate 25^{th} and 75^{th} percentiles and outliers are marked with a circle.Statistically significant differences between pairs of distributions are marked with $*$ (α vs β), with $\#$ ($\cdot$ vs OLS for α band), with § ($\cdot$ vs OLS for β band) and with $\circ$ (LASSO vs E-NET in α band)

α band it is possible to see that all penalized regression techniques show a significant increment of ACC if compared with OLS (distributions marked with #); this is not the case for the β band in which RR does not statistically differ from OLS. All others regression methods (i.e. LASSO, E-NET, F-LASSO and SG-LASSO) show a significant increment of ACC with respect to that obtained with OLS (distribution marked with §).

Among all, LASSO and E-NET show the highest values of ACC with an average value between 0.75 (LASSO α band) and 0.8 (LASSO β band) and a significant increment of E-NET with respect to LASSO only in the α band (distribution marked with $\circ$). On the other hand, OLS shows the worst performance in discriminating between HAND and FOOT with average value of ACC ~ 0.55 in α band and ~ 0.65 for β band.

Table 1.4 reports the time required for estimating the AR coefficients matrix $\hat{\mathbf{A}}$ for each VAR identification methods. Also in this case the process of cross validation described in Section III.C was followed for the selection of different optimal values of lambda/s with the same PC used for the simulation experiment. As a first observation, the computational times required are comparable to those obtained in the simulation study (see Table 1.3) performed on $M = 10$ surrogate EEG data for the case $K = 1$. Table 1.4 shows that, OLS (which does not require the selection of a regularization parameter) is the fastest method. Among penalized regression methods, RR and LASSO show a difference of few seconds for the selection of λ_{opt} ($\sim 7s$, $\sim 8s$ respectively). E-NET, F-LASSO and SG-LASSO represent the slowest regression methods with several minutes needed for the selection of λ_1 and λ_2 ($\sim 12min$ for E-NET and $\sim 32min$ SG-LASSO).

Table 1.4: Computational time required for each regression method

Regression Method		$K \sim 1$
OLS	T_{sel}	-
	T_{comp}	0.04
RR	T_{sel}	6.9
	T_{comp}	0.05
LASSO	T_{sel}	8.1
	T_{comp}	0.11
E-NET	T_{sel}	719
	T_{comp}	0.15
F-LASSO	T_{sel}	997
	T_{comp}	0.16
SG-LASSO	T_{sel}	1978
	T_{comp}	0.18

1.5 Discussion

1.5.1 Simulation experiment

The simulation study was designed to compare the performances of OLS and five different penalized regression techniques (RR, LASSO, E-NET, F-LASSO and SG-LASSO) in the VAR identification procedure. The performances were evaluated, firstly, in the estimation of the strength of the network links (values of the MVAR coefficients). Then, the performances of penalized regressions based on l_1-norm were explored in detecting the network structure. Lastly, the computational time required for solving all the linear problems needed were explored.

The accuracy in the estimation of the strength of the network links was investigated across different k-ratio values by means of MAPE and RMSE used as performance parameters (Fig. 1.1). As expected, both performance parameters for all the regression methods show a tendency to increase as the k-ratio decreases. This tendency is more evident for OLS, as already documented testing different VAR parameters identification approaches (e.g., the Levinson recursion for the solution of Yule Walker equations) in the context of signal processing [36]. The situation becomes worse when the problem approach the case $K = 1$. OLS does not guarantee the uniqueness of the solution because the matrix $([\mathbf{y}^p]^T\mathbf{y}^p)$, necessary for the solution of the problem (1.2), approaches singularity. As a consequence, OLS exhibits a strong bias, as reported in Fig.1.1, and it is necessary the use of penalized regressions [44, 48].

Here, it was documented that all the penalized regressions lead to trends of bias (MAPE, RMSE) which are consistently very low for any value of k-ratio in the estimation of null links (Fig.1.1b), and rise with k-ratio values but without any brisk discontinuities even for $K \leq 1$ (Fig. 1.1a). These performances confirm the high tolerance of penalized regressions to collinearity between regressors caused by the reduction of data samples available [20, 72]. Furthermore, the trends obtained for MAPE highlighted the best performances of LASSO and E-NET without any statistical differences between them.

The ability in the reconstruction of the network structure was investigated by analyzing the performance of the regression methods based on l_1-norm (which perform a variable selection) in terms of AUC and rates of false negative and false positive detection. In fact, AUC appeared to be

the best suited indicator in terms of its capability to synthesize the similarity of two networks also in the condition of class imbalance, a typical condition in sparse networks, as those simulated in this work [60]. Focusing on Fig.1.2, all the penalized regression approaches were characterized by a low percentage of false positives (FPR ~ 0.05) and values of false negatives and AUC that increase with the decreasing of K value. In particular, the maximum AUC (~ 0.95) value was obtained by LASSO and E-NET regression for each analysed level of K ratio. Even in the worst condition ($K = 0.5$), these two methods reached high AUC values (above 0.85). On the other hand, F-LASSO showed the worse performances for all the K values, reaching unacceptable AUC values for K equal to 0.5. SG-LASSO showed better performances than F-LASSO but worse than LASSO and EL-NET. AUC trends are in line with the results obtained in [41]. In particular, the authors showed how penalized regression techniques can efficaciously detect the existence of a causal relationships even when the number of nodes in the network is greater than the number of data samples available, with AUC values that varying between 0.6 and 0.9.

The better performances of LASSO and E-NET with respect to the other penalized regressions might be explained with a mathematical consideration. In [45], it has been highlighted how, in a situation in which regressors are grouped ($\mathbf{y}^p$ in this work), LASSO and EL-NET are not the ideal methods for a VAR model identification. In fact, F-LASSO and SG-LASSO introduce structural constraints in order to account for the grouped variable (this is a typical condition, for instance, when the predictors are genes). Hence, it is reasonable to assume that LASSO and EL-NET work better than SG-LASSO and F-LASSO in the case of brain connectivity estimation, in which no prior information on the matrix of estimated parameters exist, apart the constraint of time-lagged variables [39].

Even if Group-LASSO regression is included in the convex linear problems, it was excluded from this work. The connectivity structure were enforced only in a specific temporal lag but the Group LASSO regression is able to set to zero an entire group of predictors without introducing sparsity within groups. For this reason, it is reasonable to assume that if the structure of the problem becomes more favourable to the Group LASSO (e.g., if the connectivity structure is not simulated as an impulse response), the GCV criterion could select an optimal value of λ_1 close to zero for SG-LASSO in order to perform Group LASSO regression [10] we recall that Group-LASSO and SG-LASSO are the same when $\lambda_1 = 0$ (See Table 1.1).

Finally, the computational time required for applying the analysed methods was computed. In this work two different computational times are reported, one for the GCV procedure for lambda/s selection (depending on whether the methods presents one or two parameters for regularization) and the other one for the MVAR parameters evaluation (Table 1.2). The total time required for each tested algorithm increased with the K ratio [34]. As expected, the methods using two different regularization parameters (E-NET, F-LASSO and SG-LASSO) need more time for the GCV application, since all the combinations of the two parameters should be tested. Moreover, the lambda selection is the most consuming process, whereas the MVAR parameters computation is very fast even for high K values. The less time-consuming methods are LASSO (among those based on l_1 norm) and the E-NET regressions (among those based on a linear combination of l_1 and l_2-norm). Computational time is an increasing function of the number of lambda/s to be tested. Thus, by limiting the number of lambda/s to test it is possible to shorten the computational times.

1.5.2 Application of linear regression techniques on real EEG data

Results obtained for simulated EEG data showed that penalized regressions are reliable tools for estimating brain connectivity in conditions of data paucity where VAR identification based on OLS fails. Previous studies highlighted the possibility of discriminating different experimental conditions by using features related to brain connectivity, such as the estimated AR parameters and their frequency version [73, 74]. However, it was pointed out that, with the current methodology based on OLS estimator, it is impossible to reach an appropriate accuracy and it is necessary to move towards penalized regression techniques [11]. The combined use of features derived from brain connectivity with neural network classifiers showed an increase in classification accuracy [75] with most of the results found in α and β frequency bands related with MI tasks [50, 74]. Interestingly, the results here reported show comparable classification accuracy even for different tasks and for a different number of available data samples. In fact, as reported in Fig. 1.5 the highest values of ACC were reached by LASSO and EL-NET in α and β bands (~ 0.8). Furthermore, as proof of unsuitability of OLS in a condition of strong data paucity ($K \sim 1$), Fig. 1.5 showed the lowest values of ACC in both frequency bands analyzed.It is possible to speculate that the performances of ACC described in Fig. 1.5 can be related with the results of simulation study. Even if not directly comparable, ACC follows the trends of performance parameters of Figs.1.1,1.2 in the case of $K = 1$: highest values reached by LASSO and E-NET followed by F-LASSO, SG-LASSO, RR and, finally, OLS.

These results can be also partially explained by analyzing the selected features (Figs.1.3,1.4). In fact, classification performances are increased when the connections selected as features have a physiological meaning such as those selected in the controlateral hemisphere (left). From a physiological point of view, the subnetworks aroused from the use of penalized regressions could be related with the key role of the supplementary motor area and M_1 during the execution of a motor imagery tasks as showed in [76]. Furthermore, the lowest performance of OLS in the classification task could be also related with the selection of features that do not have physiological importance such as those showed in Figs.1.3,1.4a. The results here obtained document not only that penalized regressions could overcome the problem induced by the data paucity but that can also estimate different connectivity patterns with a plausible physiological meaning.

Even if LASSO and EL-NET showed the highest performances, there is a huge difference in the computational time required for the estimation of the optimal regularization parameter/s: $\sim 8s$ for LASSO and about $700s$ for EL-NET. To reduce the EL-NET computational time, a smaller number of values to be tested could be considered, even if this could result in an inaccurate estimate of the optimal value of the regularization parameters.

1.6 Conclusions

The aim of this work was to evaluate the usefulness of penalized regression methods for connectivity estimation. The simulation study showed how LASSO and E-NET can estimate with high accuracy not only the value of AR coefficients, but also the related connectivity structure even in conditions of data paucity in which OLS fails (e.g. when collinearity between regressors arises for the lack of data points).

The application of six different regression methods to the study of brain connectivity evidenced

consistent patterns strictly related with the physiology of a MI task. The results regard the possibility to discriminate between two tasks through a classifier trained with features extracted from a connectivity analysis. LASSO and E-NET showed the best performances in terms of accuracy of classification. It is worth of note that these results were obtained for $k = 1$, a condition in which OLS fails, E-NET and LASSO showed comparable performances in the simulation study. This findings suggests that, when OLS cannot be used LASSO might represents the most suitable method, even when not so much computational power is available.

The overall results pave the way to the use of sparse identification procedures for connectivity estimation in all those conditions in which few data samples are available, as in the estimation of brain networks, at the level of single trial or even in real-time applications. Since all the linear connectivity estimators are based on the identification of a VAR model, it is reasonable to assume that penalized regressions could be used for the computation of all the connectivity estimators in both time and frequency domains.

Chapter 2

Information Transfer in Linear Multivariate Processes Assessed through Penalized Regression Techniques: Validation and Application to Physiological Networks

2.1 Introduction

Physiological systems such as the cerebral, cardiac, vascular and respiratory system exhibit a dynamic activity which results from the continuous modulation of multiple control mechanisms and changes transiently across different physiological states. Accordingly, the human body can be modeled as an ensemble of complex physiological systems, each with its own regulatory mechanisms, that dynamically interact to preserve the physiological functions [18]. These interactions are commonly studied in a non-invasive way by recording physiological signals that are subsequently elaborated to extract time series of interest which reflect the dynamic state of the system under analysis [77, 78]. Many studies in the literature have provided strong evidence about the existence of a relationship between the properties of time series extracted and the physiological functions, even if most of these evidences come from the analysis of the dynamics within a single system (i.e., variability of heart rate, activity or connectivity within brain networks [79, 80]) or at most between two systems (cardiovascular, cardio-respiratory and brain–heart interactions [81, 82]). Only recently, with the introduction of the concept of network physiology grounded on a system-wide integration approach, it has been possible to analyze the physiological interactions in a fully multivariate fashion. With this approach, the various physiological systems that compose the human organism are considered to be the nodes of a complex network [83]. Nevertheless, identifying a network comprised of different dynamic physiological systems is a non-trivial task that requires the development of methodological approaches able to take into account the intrinsically multivariate nature of the network, and to describe the different aspects of network activity and connectivity dealing with complex dynamics and intricate topological structures.

Recent studies in the context of information theory have shown how the information processing in a network of multiple interacting dynamical systems, described by multivariate stochastic processes, can be dissected into basic elements of computation defined with the so-called framework of information dynamics [84]. These elements essentially reflect the new information produced at each moment in time about a target system in the network, the information stored in the target system, the information transferred to it from other connected systems and the modification of the information flowing from multiple sources to the target [85]. In particular, the information transfer defines the information that a group of systems designed as "sources" provide about the present state of the target [86]; information modification is strongly related to the concept of redundancy and synergy between two source systems sharing information about a target system, which refers to the existence of common information about the target that can be recovered when the sources are used separately (redundancy) or when they are used jointly (synergy) [87]. Thus, positive values of information modification indicate net synergy, which reflects the concept of information independence of the sources. On the other hand, negative values of information modification indicate redundancy, which reflects the fact that no additional information is conveyed about the target system when the two sources are considered together rather than in isolation [88]. Operational definitions of these concepts have been recently proposed, also showing how—for Gaussian processes modeled within a linear multivariate framework—the information transferred between two network nodes conditioning to the remaining nodes corresponds to the well-known measure of Granger Causality (GC) formulated in a multivariate context [6], and the measures of redundancy and synergy can be obtained as separate measures through a so-called Partial Information Decomposition (PID) [89].

The tools of information dynamics have contributed substantially to the development of the field of Network Physiology, with particular regard to the description of complex organ system interactions in various physiological states and conditions. In fact, measures information transfer and information modification have proven useful to the understanding of the dynamic interactions that are essential to produce different physiological states, e.g., wake and sleep [82, 83, 90, 91], rest and physiological stress [92, 93], relaxed conditions and mental workload [94, 95], neutral states and emotion elicitation [96, 97]. However, despite its growing appeal and widespread use in physiology and in diverse branches of science [98–101], the field of information dynamics is still under development and different aspects have to be further explored to fully exploit its potential. Recent developments have led to the formulation of a computational framework for the analysis of information dynamics which makes use of the State-Space (SS) formulation of vector autoregressive models (VAR) and of the formation of reduced linear regression models [17, 102] whose prediction error variance is related to the entropies needed for the computation of GC and PID measures [103]. The framework exhibits high computational reliability when compared with classical regression approaches for the estimation of Granger-causal measures [103], and is being increasingly used to assess information dynamics in the context of Network Physiology [78, 93].

Nevertheless, being based entirely on linear parametric modeling, it suffers from the known vulnerability to the lack of data of the standard VAR identification techniques such as the Ordinary Least Square (OLS) or the Levison's recursive algorithm for the solution of Yule-Walker equations. This issue exposes the identification process to increased bias and variance of the estimated parameters [9], and may result in ill-posed regression problems when the regressor's matrix approaches singularity [43]. As pointed out in the literature, the ratio between the number of data samples

available and the number of regression coefficients to be estimated should be at least equal to 10 to guarantee the accuracy of the estimation procedure [9, 104, 105]. This implies that the length of the time series used for VAR identification needs to increase proportionally with the number of processes jointly analyzed, which imposes a limitation to the size of the network that can be investigated if short datasets are available for the analysis. This is the case of common Network Physiology applications, where typically only short realizations of stationary multivariate physiological processes are available due to the different temporal scales and dynamics of the physiological signals involved.

To cope with the reduction of accuracy in the estimation process when dealing with a large number of time series and/or a small amount of data samples available, different strategies have been proposed in the literature such as the so-called partial conditioning [106] or the use of time-ordered restricted VAR models that are specifically built only for the computation of GC [107]. A former, more general solution is the use of penalized regression techniques that regularize a linear regression problem using one or more constraints [108]. Among them, the Least Absolute Shrinkage and Selection Operator (LASSO) uses a constraint based on the l_1 norm that if applied directly on the regression problem, yields to a sparse coefficients matrix which leads to a reduction of the mean square error in conditions of data paucity [44]. Penalized regression techniques implemented for GC analysis have been successfully applied in many different contexts, ranging from simulation studies [39] to the analysis of electroencephalographic signals [11, 105, 109], neuroimaging data [41] and Macroeconomic data [110]. In the present work, the LASSO regression is embedded in the VAR-SS framework for the computation of information dynamics, and is compared with the traditional OLS regression as regards its capability to estimate conditional information transfer and PID measures both in benchmark networks of simulated multivariate processes and in real networks of multiple physiological time series.

We show that it is possible, also in conditions of data paucity, to accurately reconstruct both the topology and the patterns of information transfer in networks of several coupled Gaussian systems exhibiting complex interactions, and to extract physiologically plausible patterns of interaction between the cardiovascular, respiratory and brain systems explored in healthy subjects during different conditions of mental stress elicited by sustained attention or mental arithmetic tasks [78, 95, 111].

The algorithms for the VAR-SS model identification based on the LASSO regression, with the subsequent computation of conditional information transfer and PID measures, are collected in the PID-LASSO MATLAB toolbox, which can be downloaded from http://github.com/YuriAntonacci/PID-LASSO-toolbox and http://lucafaes.net/PIDlasso.html

2.2 Materials and Methods

2.2.1 Vector Autoregressive Model Identification

Let us consider a dynamical system $\mathcal{Y}$, 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 completely described by the Vector autoregressive (VAR) model:

$$\mathbf{Y}_n = \sum_{k=1}^{p} \mathbf{Y}_{n-k}\mathbf{A}_k + \mathbf{U}_n, \tag{2.1}$$

where $\mathbf{A}_k$ is an $M \times M$ matrix containing the autoregressive (AR) coefficients, and $\mathbf{U} = [U_1 \cdots U_M]$ is a vector of M zero-mean white processes, denoted as innovations, with $M \times M$ covariance matrix $\mathbf{\Sigma} \equiv \mathbb{E}[\mathbf{U}_n^T \mathbf{U}_n]$ ($\mathbb{E}$ is the expectation 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 operator ";" stands for row separation, so that the i^{th} row is a realization of $\mathbf{Y}_i$, i.e., $\mathbf{y}_i = [y_{1,i} ... y_{M,i}], i = 1, ..., N$, and the j^{th} 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 Square (OLS) identification finds an optimal solution for the problem (2.1) by solving the following linear quadratic problem:

$$\widehat{\mathbf{A}} = argmin_{\mathbf{A}} ||\mathbf{y} - \mathbf{y}^p \mathbf{A}||_2^2, \tag{2.2}$$

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 $(N-p) \times Mp$ matrix of the regressors and $\mathbf{A} = [\mathbf{A}_1; \cdots ; \mathbf{A}_p]$ is the $Mp \times M$ coefficient matrix. The 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 is minimized (RSS) [104, 112]. When $N - p \leq Mp$ the OLS does not guarantee the uniqueness of the solution since the matrix $([\mathbf{y}^p]^T \mathbf{y}^p)$ becomes singular [105, 112]. Even in this situation, it is possible to solve the problem stated in Equation (2.1) through the Least Absolute Shrinkage and Selection Operator (LASSO) which introduces a constraint in the linear quadratic problem (2.2) [108]:

$$\widehat{\mathbf{A}} = argmin_{\mathbf{A}} (||\mathbf{y} - \mathbf{y}^p \mathbf{A}||_2^2 + \lambda ||\mathbf{A}||_1). \tag{2.3}$$

In Equation (2.3), the additional term based on the l_1 norm forces a sparse a solution such that some of the VAR coefficients are shrunk to zero, with the shrinkage parameter λ controlling the trade-off between the number of non-zero coefficients selected in the matrix $\widehat{\mathbf{A}}$ and the residual sum of squares (RSS). Even if the problem (2.3) admits a solution, it will not be in a closed form since the l_1 norm is not differentiable at zero [44]. The optimal value of λ for the solution of the problem (2.3) requires a cross-validation approach for its determination. Typically, a predefined interval of values for λ is defined such that the biggest value provides an estimated AR matrix of zeroes and the lowest provides a dense AR matrix [113] (in this work, 300 values of λ were selected). Subsequently, using an hold-out approach, as described in [114], it is possible to independently draw 90% of the observations of the predicted values and of the regressors (rows of $\mathbf{y}$ and $\mathbf{y}^p$) as training set and keeping the remaining 10% for the testing set. Training and test sets are then reduced to zero mean and unit variance and, for each assigned λ, the number of non-zero coefficients is evaluated for the matrix $\widehat{\mathbf{A}}$ estimated from the training set, and the corresponding RSS is computed on the test set. After repeating this operation several times (10 in this work) by randomly changing the training and testing sets, the optimal value of λ is chosen as the one that minimizes the ratio between RSS and the number of non-zero VAR coefficients [55]. The matrix of AR coefficients $\widehat{\mathbf{A}}$ is then estimated by using the estimated optimal value of λ.

2.2.2 Measures of Information Transfer

Considering the overall observed 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 $s = \{1, ..., M\} \backslash \{i, j\}$. Then, the Transfer Entropy (TE) from Y_i to Y_j quantifies the amount of information that the past of the source, $Y^p_{i,n}$, provides about the present of the target, $Y_{j,n}$, over and above the information already provided by the past of the target itself , $Y^p_{j,n}$, and is defined as follows [77, 115]:

$$T_{i \to j} = I(Y_{j,n}; Y^p_{i,n} | Y^p_{j,n}) = H(Y_{j,n} | Y^p_{j,n}) - H(Y_{j,n} | Y^p_{j,n}, Y^p_{i,n}) \tag{2.4}$$

where $I(\cdot;\cdot|\cdot)$ represents the conditional mutual information and $H(\cdot|\cdot)$ represents the conditional entropy [1]. In the presence of two sources Y_i and Y_k, the information transferred towards the target Y_j from the two sources taken together is quantified by the joint Transfer Entropy (jTE):

$$T_{ik \to j} = I(Y_{j,n}; Y^p_{i,n}, Y^p_{k,n} | Y^p_{j,n}) = H(Y_{j,n} | Y^p_{j,n}) - H(Y_{j,n} | Y^p_{j,n}, Y^p_{i,n}, Y^p_{k,n}) \tag{2.5}$$

where $Y^p_{k,n}$ represents the past of the source k. Then, a possible way to decompose the jTE is that provided by the so-called partial information decomposition (PID). The PID expands the information transferred jointly from two sources to a target in four different quantities, reflecting the unique information transferred from each individual source to the target, measured by the unique TEs $U_{i \to j}$ and $U_{k \to j}$, and the redundant and synergistic information transferred from the two sources to the target, measured by the redundant TE $R_{ik \to j}$ and the synergistic TE $S_{ik \to j}$ [116]. These four measures are related to each other and to the joint and individual TEs from each source to the target by the following equations:

$$T_{ik \to j} = U_{i \to j} + U_{k \to j} + R_{ik \to j} + S_{ik \to j}, \tag{2.6}$$

$$T_{i \to j} = U_{i \to j} + R_{ik \to j} \tag{2.7}$$

$$T_{k \to j} = U_{k \to j} + R_{ik \to j} \tag{2.8}$$

In the PID defined above, the terms $U_{i \to j}$ and $U_{k \to j}$ quantify the parts of the information transferred to the target process Y_j which are unique to the source processes Y_i and Y_k, respectively, mirroring the contributions to the predictability of the target that can be obtained from one of the sources but not from the other. Each of these unique contributions sums up with the redundant TE to retrieve the information transfer defined by the classical measure of the bivariate TE, thus indicating that $R_{ik \to j}$ pertains to the part of the information transferred individually, yet redundantly from a source to the target. The term $S_{ik \to j}$ refers to the synergy between the two sources while they transfer information to the target, intended as the information that is uniquely obtained taking the two sources Y_i and Y_k together, but not considering them alone. While several implementations of the PID exists depending on how a fourth equation is formulated to complete the definitions (2.6-2.8), in the case of joint Gaussian processes it has been shown that an unifying formulation is that defining the redundant transfer as the minimum information transferred individually by each source to the

target, i.e., $R_{ik\to j} = min(T_{i\to j}, T_{k\to j})$ [89].

In addition to the measures defining the PID, another important information measure used to detect the topological structure of direct interactions in a network of M interacting processes is the conditional Transfer Entropy (cTE). With the notation introduced above for the overall vector process $\mathbf{Y}$, the cTE from a driver process Y_i to a target process Y_j computed considering the other processes in the network collected in $\mathbf{Y}_s$, is defined as:

$$T_{i\to j|s} = I(Y_{j,n}; Y^p_{i,n}|Y^p_{j,n}, \mathbf{Y}^p_{s,n}) = H(Y_{j,n}|Y^p_{j,n}, \mathbf{Y}^p_{s,n}) - H(Y_{j,n}|\mathbf{Y}^p_n) \tag{2.9}$$

The cTE quantifies the amount of information contained in the present state of the target process that can be predicted by the past states of the source process, above and beyond the information that is predicted already by the past states of the target and of the all other processes [6]. An implication of this definition is that non-zero values of the cTE $T_{i\to j|s}$ correspond to the presence of a direct causal interaction from Y_i to Y_j, which is typically depicted, in a network representation where nodes are associated with processes and edges with significant causal interactions, with an arrow connecting the i^{th} and j^{th} nodes.

2.2.3 Computation of the Measures of Information Transfer for Multivariate Gaussian Processes

When the observed multivariate process $\mathbf{Y}$ has a joint Gaussian distribution, the information-theoretic measures described in Section 2.2 can be formulated in an exact way based on the linear VAR representation provided in Section 2.1. Indeed, it has been shown that the covariance matrices of the observed vector process and of the residuals of the formulation (1.1) contain, in the case of jointly distributed Gaussian processes, all of the entropy differences which are needed to compute the information transfer [117]. In turn, these entropy differences are expressed by the concept of partial covariance formulated in the context of linear regression analysis. Specifically, defining $E_{j|j,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y^p_{j,n}]$ and $E_{j|ij,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y^p_{i,n}, Y^p_{j,n}]$ as the prediction errors of a linear regression of $Y_{j,n}$ performed respectively on $Y^p_{j,n}$ and $[Y^p_{i,n}Y^p_{j,n}]$, the conditional entropies $H(Y_{j,n}|Y^p_{j,n})$ and $H(Y_{j,n}|Y^p_{j,n}, Y^p_{i,n})$ can be expressed as functions of the prediction error variances $\lambda_{j|j} = \mathbb{E}[E^2_{j|j,n}]$ and $\lambda_{j|ij} = \mathbb{E}[E^2_{j|ij,n}]$ as follows [6, 118]:

$$H(Y_{j,n}|Y^p_{j,n}) = \frac{1}{2} ln 2\pi e \lambda_{j|j}, \tag{2.10a}$$

$$H(Y_{j,n}|Y^p_{j,n}, Y^p_{i,n}) = \frac{1}{2} ln 2\pi e \lambda_{j|ij}, \tag{2.10b}$$

from which the TE from Y_i to Y_j can be retrieved using (2.7):

$$T_{i\to j} = \frac{1}{2} ln \frac{\lambda_{j|j}}{\lambda_{j|ij}}. \tag{2.11}$$

Following similar reasoning, the jTE from (Y_i, Y_k) to Y_j can be defined as:

$$T_{ik\to j} = \frac{1}{2} ln \frac{\lambda_{j|j}}{\lambda_{j|ijk}}, \tag{2.12}$$

where $\lambda_{j|ijk} = \mathbb{E}[E^2_{j|ijk,n}]$ is the variance of the prediction error of a linear regression of $Y_{j,n}$ on $(Y^p_{i,n}, Y^p_{j,n}, Y^p_{k,n})$ with prediction error $E_{j|ijk,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y^p_{i,n}, Y^p_{j,n}, \mathbf{Y}^p_{s,n}]$, and the cTE from Y_i to Y_j given $\mathbf{Y}_s$ can be defined as:

$$T_{i\to j|s} = \frac{1}{2} ln \frac{\lambda_{j|js}}{\lambda_{j|ijs}}, \tag{2.13}$$

where $\lambda_{j|js} = \mathbb{E}[E^2_{j|js,n}]$ is the variance of the prediction error of a linear regression of $Y_{j,n}$ on $(Y^p_{j,n}, \mathbf{Y}^p_{s,n})$ with prediction error $E_{j|js,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y^p_{j,n}, Y^p_{s,n}]$ and $\lambda_{j|ijs} = \mathbb{E}[E^2_{j|ijs,n}]$ is the variance of the prediction error of a linear regression of $Y_{j,n}$ on $\mathbf{Y}^p_n$ with prediction error $E_{j|ijs,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|\mathbf{Y}^p_n]$. Moreover, from the definitions in Section 2.2 it is then possible to obtain the redundant TE, the synergistic TE and the unique TEs in addition to the cTE. Therefore, the computation of all the information measures amounts to calculate the partial variances to be inserted in Equations (2.11)–(2.13). In the following subsection we report how to derive such partial variances exploiting the State–Space formulation of the VAR model (1.1) [103].

Formulation of State–Space Models

A discrete state–space (SS) model is a linear model in which a set of input, output and state variables are related by first order difference equations [17]. The VAR model (1.1) can be represented equivalently as an SS model ([119]) which relates the observed process $\mathbf{Y}$ to an unobserved state process $\mathbf{Z}$ through the observation equation

$$\mathbf{Y}_n = \mathbf{C}\mathbf{Z}_n + \mathbf{E}_n, \tag{2.14}$$

and describes the update of the state process through the state equation

$$\mathbf{Z}_{n+1} = \mathbf{A}Z_n + \mathbf{K}E_n. \tag{2.15}$$

The innovations $\mathbf{E}_n$ of Equations (2.14) and (2.15) are equivalent to the innovations $\mathbf{U}_n$ in (1.1) and thus have covariance matrix $\mathbf{\Phi} \equiv \mathbb{E}[\mathbf{E}^T_n \mathbf{E}_n] = \mathbf{\Sigma}$. This representation, typically denoted as innovation form SS (ISS) model , also demonstrates 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 (1.1) as reported in ([119]) . Starting from the parameters of an ISS model is possible to compute any partial variance $\lambda_{j|a}$, where the subscript a denotes any combination of indexes $\in (1, ..., M)$, by evaluating the innovation of a "submodel" obtained removing from the observation Equation (2.14) the variables not included in a. Furthermore, in this formulation the state Equation (2.15) remains unaltered and the observation equation of relevant submodel becomes:

$$\mathbf{Y}^{(a)}_n = \mathbf{C}^{(a)}\mathbf{Z}_n + \mathbf{E}^{(a)}_n, \tag{2.16}$$

where the subscript ^{a} denotes the selection of the rows with indices a of a vector or a matrix. As demonstrated in [102, 103], the submodel (2.15) and (2.16) is not in ISS form, but can be converted into ISS by solving a Discrete Algebraic Riccati equation (DARE). Then, the covariance matrix of the innovations $\mathbf{\Phi}^{(a)} = \mathbb{E}[\mathbf{E}^{(a)T}_n \mathbf{E}^{(a)}_n]$ includes the desired error variance $\lambda_{j|a}$ as diagonal element corresponding to the position of the target Y_j. Thus, it is possible to compute all the partial variances needed for the evaluation of all the information measures introduced, starting from a set

of ISS parameters. In particular, these parameters can be directly extracted by the knowledge of the parameters of the original VAR model (i.e., $\boldsymbol{A}_1, ..., \boldsymbol{A}_p, \boldsymbol{\Sigma}$) , which in this study are estimated by identifying the VAR model (1.1) making use of either the OLS method or the LASSO regression.

2.2.4 Testing the Significance of the Conditional Transfer Entropy

Since the cTE $T_{i\to j|s}$ is a measure of the information transferred directly (i.e., without following indirect paths) from the source Y_i to the target Y_j, and for Gaussian processes is equivalent to conditional Granger causality [6], it is of interest to perform the assessment of its statistical significance with the aim to establish the existence of a direct link from the i^{th} node to the j^{th} node of the observed network of interacting processes. In this work, the significance of cTE, computed after OLS identification of the VAR model, was tested generating sets of surrogate time series which share the same power spectrum of the original time series but are otherwise uncorrelated. Specifically, 100 sets of surrogate time series were generated using the Iterative Amplitude Adjusted Fourier Transform (IAAFT) procedure [120]; then, the cTE was estimated for each surrogate set, a threshold equal to the 95^{th} percentile of its distribution on the surrogates was determined for each directed link, and the link was detected as statistically significant when the original cTE was above the threshold. In the case of LASSO, the statistical significance of the estimated cTE values was determined exploiting the sparseness of the identification procedure. Since LASSO model identification always produces a sparse matrix with several VAR coefficients equal to zero, the cTE values result exactly zero when the coefficients along the investigated direction are zero at each time lag; on the contrary, cTE is positive, and was considered to be statistically significant in this study, when at least one coefficient is non-zero along the considered direction.

2.3 Simulation Experiments

This section reports two simulation studies performing a systematic evaluation of the performances of the two VAR identification methodologies (OLS and LASSO) employed for the practical computation of the measures of information transfer in known networks assessed with different amount of data samples available. First, we study the behaviour of the measures of information transfer and information modification in a four-variate VAR process specifically configured to reproduce coexisting forms of redundant and synergistic interactions between source processes sending information towards a target [89, 103]. Second, with specific focus on the estimation of the cTE and of its statistical significance, we compared the ability of OLS and LASSO to reconstruct an assigned network topology in a ten-variate VAR process exhibiting a random interaction structure with fixed density of connected nodes [34, 105]

2.3.1 Simulation Study I

Simulation Design and Realization

Simulated multivariate time series (M=4) were generated as realizations of the following VAR(2) process depicted in Figure 2.1 [77, 103, 121]:

$$Y_{1,n} = 2\rho_1 \cos(2\pi f_1) Y_{1,n-1} - \rho_1^2 Y_{1,n-2} + U_{1,n}, \tag{2.17a}$$

$$Y_{2,n} = 2\rho_2 \cos(2\pi f_2) Y_{2,n-1} - \rho_2^2 Y_{2,n-2} + Y_{1,n-1} + U_{2,n}, \tag{2.17b}$$

$$Y_{3,n} = 2\rho_3 \cos(2\pi f_3) Y_{3,n-1} - \rho_3^2 Y_{3,n-2} + Y_{1,n-1} + U_{3,n}, \tag{2.17c}$$

$$Y_{4,n} = \frac{1}{2} Y_{2,n-1} + \frac{1}{2} Y_{3,n-1} + U_{4,n}, \tag{2.17d}$$

In (2.17), $\mathbf{U} = [U_1 \dots U_4]$ is a vector of zero-mean uncorrelated white noises with unit variance (i.e., with covariance $\mathbf{\Sigma} \equiv \boldsymbol{I}$). The VAR parameters are selected to allow autonomous oscillations for Y_1, Y_2, and Y_3 by placing, in the VAR representation in the $Z-$domain, complex-conjugate poles with modulus ρ_i and phase $2\pi f_i, i = 1, 2, 3$; here we set pole modulus $\rho_1 = \rho_2 = \rho_3 = 0.95$ and pole frequency $f_1 = 0.1$, $f_2 = f_3 = 0.25$. Moreover, interactions between different processes were set to allow a common driver effect $y_2 \leftarrow y_1 \rightarrow y_3$ and unidirectional couplings $y_2 \rightarrow y_4$ and $y_3 \rightarrow y_4$, with weights indicated in Fig. 1.1. With these settings, 100 realizations of the processes were generated under different values of the parameter K defined as the ratio between the number of data samples available (N) and the number of AR coefficients to be estimated (Mp); the parameter K was varied in the range $(1, 2, 5, 10, 30)$, so that the length of the simulated time series was $N = 8$ when $K = 1$ and $N = 240$ were when $K = 30$. For each realization and for each value of K, all the measures appearing in the PID of the information transfer were computed by exploiting the SS approach applied to the VAR parameters estimated through OLS or LASSO identification; PID analysis was performed considering either Y_4 or Y_1 as the target process, and both Y_2 and Y_3 as the source processes. Then, the bias and variance of each estimated PID measure were assessed, for each K and separately for OLS and LASSO, respectively as the absolute difference between the mean value of the measure over the 100 realizations and its theoretical value computed using the true values imposed for the VAR parameters, and as the sample variance estimated over the 100 realizations.

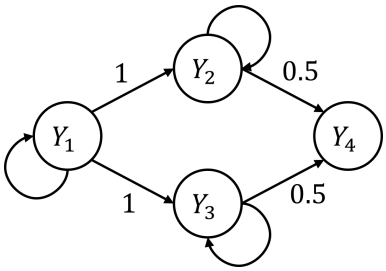

Figure 2.1: Graphical representation of the four-variate VAR (Vector Autoregressive) process realized in the first simulation according to Equation (2.17). Network nodes represent the four simulated processes, and arrows represent the imposed causal interactions (self-loops depict influences from the past to the present sample of a process).

Simulation Results

Figures 2.2 and 2.3 show the trends of bias and variance associated with the estimation of TE ($T_{2\rightarrow j}, T_{3\rightarrow j}$), redundant TE ($R_{23\rightarrow j}$), synergistic TE ($S_{23\rightarrow j}$) and unique TEs ($U_{2\rightarrow j}$, $U_{3\rightarrow j}$) respectively when $j = 4$ (target process Y_4) and $j = 1$ (target process Y_1), computed after VAR model identification using OLS (blue) and LASSO (red) and depicted as a function of the ratio K between time series length and number of model parameters.

As a general result, both figures show that the accuracy of all estimates of the PID measures is strongly influenced by the amount of data available, with a progressive increase of both the bias and the variance of the estimates with the decrease of the parameter K. The LASSO regression exhibits a substantially better performance in the estimation of the PID measures particularly when the amount of data samples is scarce ($K \leq 2$). In the most challenging condition of $K = 1$ (number of AR coefficients equal to the number of data points) the results are reported only for the LASSO regression since in this condition for OLS it was impossible to evaluate the PID measures due to the non-convergence of the DARE equation solution during the computation. In the other cases ($K \in (5, 10, 30)$) the two identification methods show comparable trends, with slightly better performance exhibited by OLS identification in the assessment of non-zero PID measures (Fig. 2.2), and by LASSO identification in the assessment of zero PID measures (Fig. 2.3).

In fact, when Y_4 is taken as target process, the sources Y_2 and Y_3 send the same amount information towards the target and this information is entirely redundant ($T_{2\to4} = T_{3\to4} = R_{23\to4} = 0.63, U_{2\to4} = U_{3\to4} = 0$); moreover, a non-negligible amount of synergistic information transfer is present ($S_{23\to4} = 0.56$) [103]. As reported in Fig. 2.2, the estimates of the non-zero quantities ($T_{2\to4}, T_{3\to4}, R_{23\to4}, S_{23\to4}$) assessed through LASSO-VAR identification exhibit higher variance than those assessed through the OLS, as well a slight negative bias which becomes relevant only in the case of the synergistic TE ; in such a case the underestimation of $S_{23\to4}$ is present also after OLS identification when $K = 2$ (Figure 2.2c).

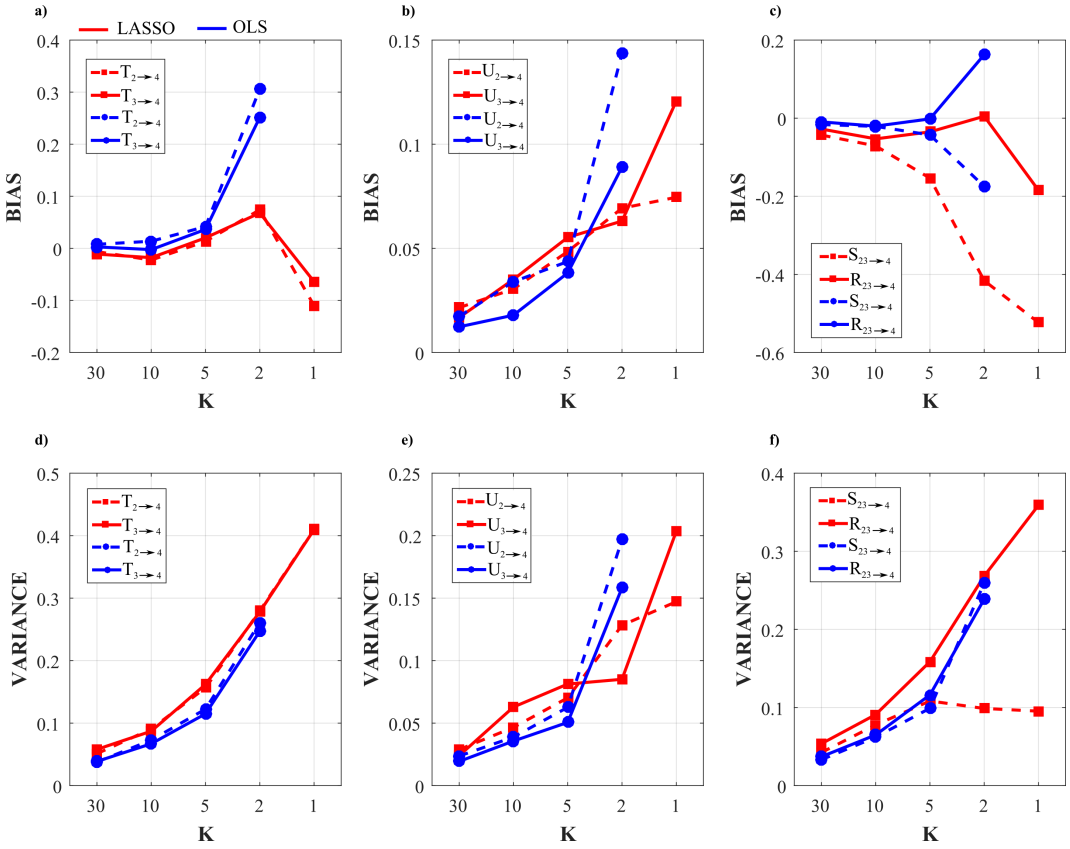

Figure 2.2: Accuracy of PID (Partial Information Decomposition) measures computed for the VAR processes of Simulation I when Y_4 is taken as the target process. Panels report the bias (**a**, **b**, **c**) and the variance (**d**, **e**, **f**) relevant the computation of the TE (Transfer Entropy) from Y_2 to Y_4 and from Y_3 to Y_4 (a,d), the unique TE from Y_2 to Y_4 and from Y_3 to Y_4 (b,e) and the redundant and synergistic TE from Y_2 and Y_3 to Y_4 (c,f).

When the process Y_1 is taken as the target, all the PID measures are null ($T_{2\to1} = T_{3\to1} = U_{2\to1} = U_{3\to1} = S_{23\to1} = R_{23\to1} = 0$) because no causal interactions are directed towards Y_1. As shown in Figure 2.3, in this case the LASSO identification outperforms the OLS method, showing lower bias and variance for all values of K with evident improvement in the performance when $K \leq 2$. Interestingly, for low values of K the LASSO regression detected the absence of synergy with more accuracy than that of redundancy (Figure 2.3c,f).

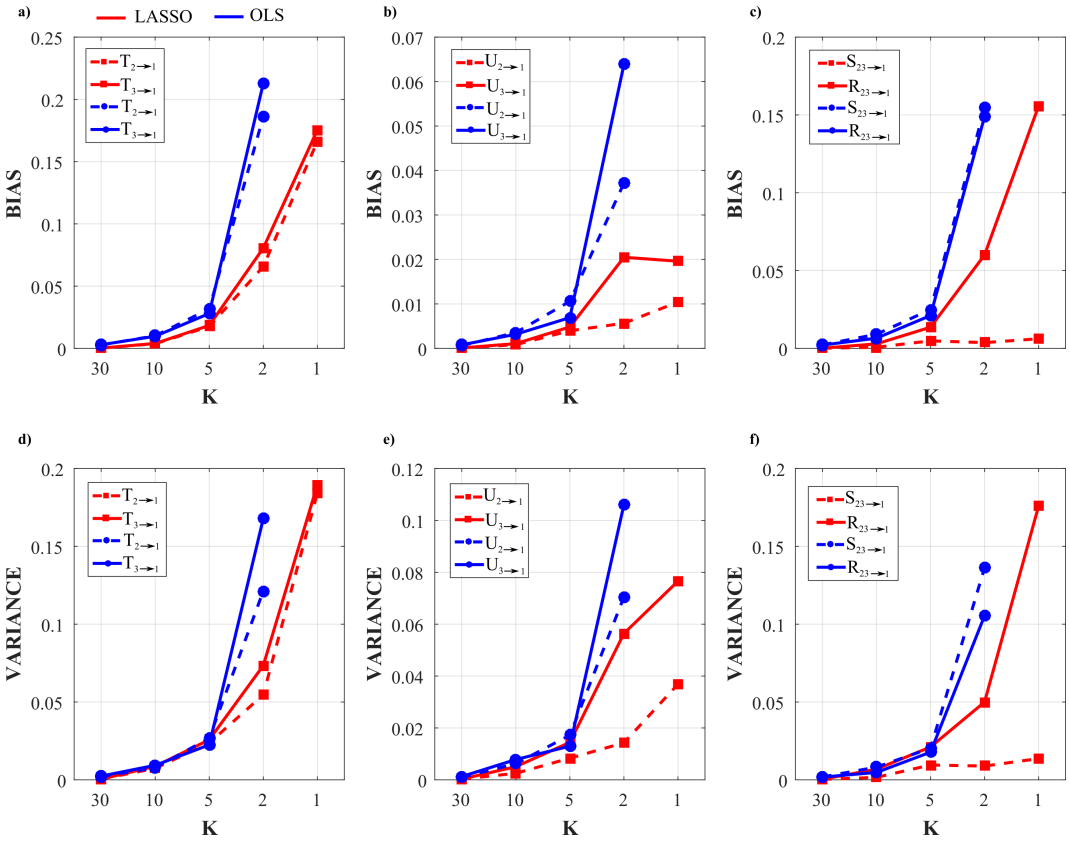

Figure 2.3: Accuracy of PID measures computed for the VAR processes of Simulation I when Y_1 is taken as the target process. Panels report the bias (a,b,c) and the variance (d,e,f) relevant the computation of the TE from Y_2 to Y_1 and from Y_3 to Y_1 (a,d), the unique TE from Y_2 to Y_1 and from Y_3 to Y_1 (b,e) and the redundant and synergistic TE from Y_2 and Y_3 to Y_1 (c,f).

2.3.2 Simulation Study II

Simulation Design and Realization

Simulated multivariate time series ($M = 10$) were generated as realizations of a VAR(10) model fed by white Gaussian noises with variance equal to 0.1. The simulated networks have a ground-truth structure with a density of connected nodes equal to 50% in which non-zero AR parameters were set assigning randomly the lag in the range (1-10) and the coefficient value in the interval $[-0.6, 0.6]$ [58]. A representative example of one possible generated network is shown in Figure 2.4, where the strength of the directed links is provided by the theoretical cTE computed between two processes starting from the true AR parameters. Under these constraints, 100 realizations (each with its specific network structure) of the VAR(10) process were generated with different values of the parameter K in the range $(1, 2, 5, 10, 30)$, so that the length of the simulated time series was $N = 100$ when $K = 1$ and $N = 3000$ were when $K = 30$. For each realization and for each value of

K, the cTE between each pair of processes was computed by exploiting the SS approach applied to the VAR parameters estimated through OLS or LASSO identification. Then, the bias and variance of the cTE estimates obtained through OLS and LASSO identification were assessed separately for the connections with zero and non-zero cTE as explained in the following subsection.

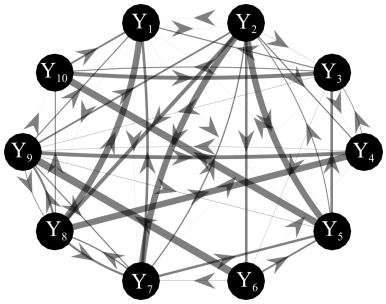

Figure 2.4: Graphical representation for one of the ground-truth networks of Simulation II. Arrows represent the existence of a link, randomly assigned, between two nodes in the network. The thickness of the arrows is proportional to the strength of the connection, with a maximum value for the cTE equal to 0.15. The number of connections for each network is set to 45 out of 90.

Performance Evaluation

The performances of LASSO and OLS were assessed both in terms of the accuracy in estimating the strength of the network links through the absolute values of the cTE measure, and in terms of the ability to reconstruct the network structure through the assessment of the statistical significance of cTE. The first analysis was performed separately for non-null and null links computing the bias of cTE through the comparison between the estimated and theoretical cTE values. Specifically, for each pair of network nodes represented by the processes Y_i and Y_j, the theoretical cTE obtained from the true VAR parameters, $T_{i\to j|s}$, was compared with the corresponding estimated cTE value, $\widehat{T}_{i\to j|s}$, using a measure of absolute bias ($bias$) if the theoretical link is null, and a normalized measure of bias ($bias_N$) if the theoretical link is non-null [59]:

$$bias = |T_{i\to j|s} - \widehat{T}_{i\to j|s}|, \tag{2.18a}$$

$$bias_N = \left| \frac{T_{i\to j|s} - \widehat{T}_{i\to j|s}}{T_{i\to j|s}} \right|. \tag{2.18b}$$

Then, for each network, the values of $bias$ and $bias_N$ were averaged respectively across the 45 non-null links and across the 45 null links to get individual measures, denoted as $BIAS$ and $BIAS_N$. Finally, the distributions of $BIAS$ and $BIAS_N$ were assessed across the 100 simulated network structures and presented separately for OLS and LASSO.

Second, the ability of OLS and LASSO to detect the absence or presence of network links based on the statistical significance of the cTE was tested comparing the two adjacency matrices representative of the estimated and theoretical network structures. This can be seen as a binary classification task where the existence (class 1) or absence (class 0) of each estimated connection is assessed (using surrogate data for OLS and looking for zero/non-zero estimated coefficients for LASSO) and compared with the underlying ground-truth structure. Performances were assessed through the computation of the false positive rate (FPR, measuring the fraction of null links for

which a statistically significant cTE was detected), false negative rate (FNR, measuring the fraction of non-null links for which the cTE was detected as non-significant) and accuracy (ACC, measuring the fraction of false detections) parameters [109, 122]. Each of these performance measures was obtained across the network links for each individual network, and its distribution across the 100 simulated network structures was then presented separately for OLS and LASSO.

Statistical Analysis

For this simulation study, five different repeated measures two-way ANOVA tests, one for each performance parameter ($BIAS$,$BIAS_N$,FNR,FPR,ACC) were performed, to evaluate the effects of different values of K (varied in the range $[30, 10, 5, 2]$) and different identification methodologies ($[OLS, LASSO]$) on performance parameters.

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

Results of the Simulation Study

The results of the two-way repeated measures ANOVAs, expressed in terms of F-values and computed separately on all the performance parameters considering K and TYPE (identification method used) as main factors, are reported in Table 2.1.

Table 2.1: F-values of the two-way repeated measures ANOVA. ** is associated with $p < 10^{-5}$

Factor	$BIAS$	$BIAS_N$	FNR	FPR	ACC
K	8582**	1694**	2204**	197.2**	2492**
TYPE	1640**	377**	3538**	223.4**	1575**
K x TYPE	8633**	848**	1055**	114.5**	339**

The two-way ANOVAs reveal a strong statistical influence of the main factors K and TYPE and of their interaction on all the performance parameters analyzed. It is worth of note that the level $K = 1$ was not considered in the statistical analysis due to the non-convergence of the DARE equation for the OLS case.

Figure 2.5 reports the distribution of the parameters $BIAS$ and $BIAS_N$ according to the interaction factor K x TYPE.

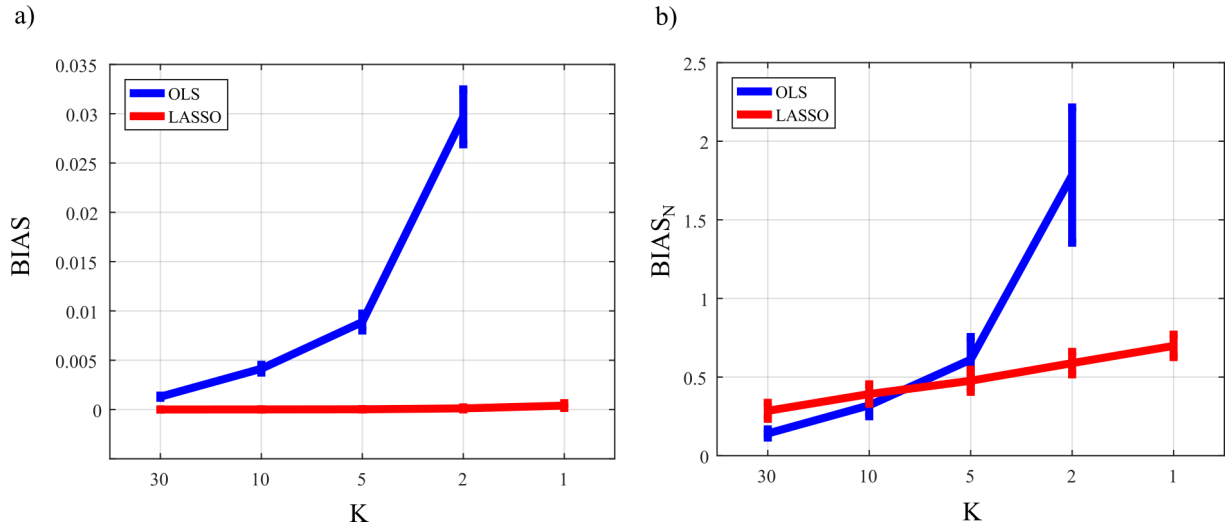

Figure 2.5: Distribution of the bias parameters computed for the null links ($BIAS$, **a**) and for the non-null links ($BIAS_N$, **b**) considering the interaction factor K x TYPE, expressed as mean value and 95% confidence interval of the parameter computed across 100 realizations of simulation II for OLS (blue line) and LASSO (red line) for different values of K.

The comparison of the two VAR identification procedures shows that the trends for LASSO (red line) and OLS (blue line) are very different. In the analysis of the error committed in the estimation of the null links (parameter $BIAS$) the error of LASSO estimates is almost zero for all levels of K (even for $K \leq 2$ that are the most challenging situations), while OLS estimates show a sharp increase of the error with the decrease of data samples available for the estimation of cTE (2.5 a). The analysis of the error committed in the estimation of the non-null links (parameter $BIAS_N$,2.5 b) highlights that for both methods the error increases with decreasing the value of K. The two identification methods exhibit different performance as a function of the number of data samples available for the estimation procedure: when such number is high ($K = 30$), the OLS assumes a significantly smaller bias than LASSO; when $10 \leq K \leq 5$ there are no significant differences between the two methods; in the most challenging conditions with $K < 5$ OLS exhibits a drastic rise of $BIAS_N$ towards 2 (which means an overestimation up to 200%), while LASSO identification allows limitation of the bias which remains below 1 even when $K = 1$.

Figure 2.6 reports the distributions of the parameters FPR, FNR and ACC according to the interaction K x TYPE. The analysis of the rate of false negatives (Figure 2.6a) shows that the number of links incorrectly classified as null increases while decreasing the amount of data available (K decreasing from 10 to 2), with values of FNR rising from about 0.1 to about 0.6 using the OLS, and remaining much lower (between 0 and 0.2) using LASSO identification. On the other hand, the analysis of the rate of false positives (Figure 2.6b) returns opposite trends, with several absent links incorrectly classified as non-null which is stable and almost negligible using OLS, and exhibits a slight growth that leads the FPR value from 0 with K=30 to about 0.25 for K=1. The overall performance assessed through the ACC parameter is better using LASSO identification (Figure 2.6c): the rate of correctly detected links is comparable in the favorable condition $K = 30$, while when $K \leq 10$ LASSO shows better performance (significantly higher values of ACC) than OLS and can reconstruct the network structure with a very good accuracy ($\sim 80\%$) even in the challenging condition of $K = 1$.

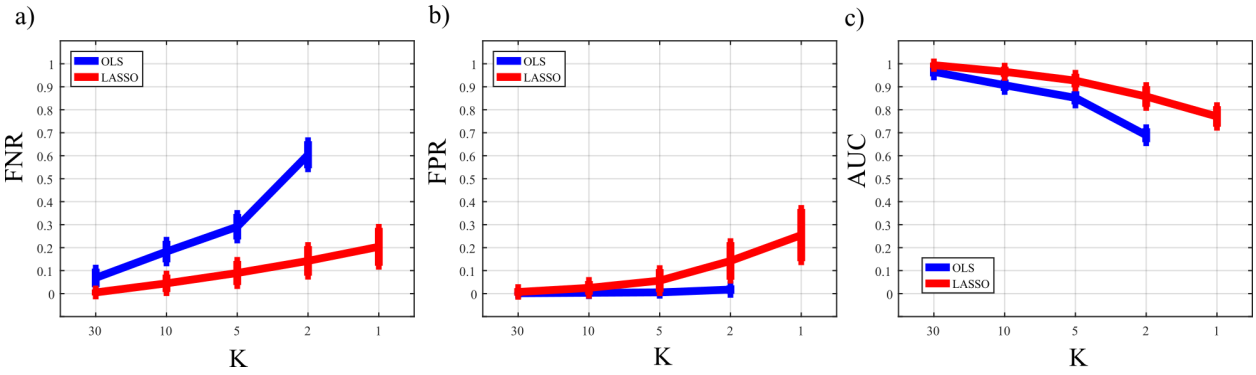

Figure 2.6: Distributions of FNR (**a**), FPR (**b**) and ACC (**c**) parameters considering the interaction factor K x TYPE, expressed as mean value and 95% confidence interval of the parameter computed across 100 realizations of simulation II for OLS (blue line) and LASSO (red line) for different values of K.

2.4 Application to Physiological Time Series

This section reports the application of the measures of information transfer, based on VAR models, estimated through OLS or LASSO identification, to a dataset of physiological time series previously collected with the aim of studying organ system interactions during different levels of mental stress [78]. The physiological time series measured for each subject were considered to be a realization of a vector stochastic process descriptive of the behavior of a composite dynamical system which forms a network of physiological interactions. Such network is composed of two distinct sub-networks, which are in turn formed by three nodes ("body" or peripheral sub-network) and four nodes (brain sub-network). The dynamic activity at each network node is quantified by a scalar process, as specifically defined in the next subsection.

2.4.1 Data Acquisition and Pre-Processing

Eighteen healthy participants with an age between 18 and 30 years were recorded during three different tasks inducing different levels of mental stress: a resting condition induced watching a relaxing video (R); a condition of mental stress induced by the execution of a mental arithmetic task (M) using an online tool in which the participants had to perform sums and subtractions of 3-digit numbers and write the solution in a text-box using the keyboard; a condition of sustained attention induced playing a serious game (G) which consisted of following a point moving on the screen using the mouse and trying to avoid different obstacles. All participants provided written informed consent. The experiment was approved by the Ethics Committees of the University of Trento. The study was in accordance with the Declaration of Helsinki.

The acquired physiological signals were the Electrocardiogram (ECG) signal, the respiratory signal (RESP) measured monitoring abdominal movements, the blood volume pulse (BVP) signal measured through a photoplethysmographic technique, and 14 Electroencephalogram (EEG) signals recorded at different locations in the scalp. After a pre-processing step performed in MatLab R2016b (Mathworks, Natick, MA, USA), seven physiological time series, each consisting of 300 data points and taken as a realization of the stochastic process representing the activity of specific physiological (sub)systems, were extracted from the recorded signals as follows: 1) the R-R tachogram, represented by the sequence of the time distances between consecutive R peaks of the ECG (process

η) ; 2) The series of respiratory amplitude values, sampled at the onset of each detected R-R interval (process ρ): 3) the pulse arrival time (process π) obtained computing the time elapsed between each R peak in the ECG and the corresponding point of maximum derivative in BVP signal; the sequences of the EEG power spectral density, measured in consecutive time windows (lasting 2 s with 1 s overlap) of the EEG signal acquired at the electrode F_z, integrated within the bands $0.5 - 3Hz$ (process δ), $3 - 8Hz$ (process θ), $8 - 12Hz$ (process α), and $12 - 25Hz$ (process β). Before VAR modeling, the time series were reduced to zero mean and unit variance and checked for a restricted form of weak sense stationarity using the algorithm proposed in [123], which divides each time series into a given number of randomly selected sub-windows, assessing for each of them the stationarity of mean and variance. A detailed description of signal recording, experimental protocol and time series extraction can be found in [78, 95].

2.4.2 Information Transfer Analysis

The seven time series obtained from each subject and from each condition were interpreted as a realization of a VAR process whose parameters $\boldsymbol{A}_1, ..., \boldsymbol{A}_p, \boldsymbol{\Sigma}$ were estimated with the two different identification methods under analysis (i.e., OLS and LASSO). The model order p was estimated, for each experimental condition and for each subject, using the Bayesian Information Criterion [124]. Then, two different analyses were performed through the application of the SS approach:

1. First, a PID analysis was performed for OLS and LASSO through the computation of the joint information transfer $T_{ik\to j}$ and the terms of its decomposition $U_{i\to j}$, $U_{k\to j}$, $R_{ik\to j}$, $S_{ik\to j}$. The analysis was performed collecting in the first source (index i) the processes $[\eta, \rho, \pi]$ forming the so-called "body" sub-network that accounts for cardiac, cardiovascular and respiratory dynamics, and in the second source (index k) the processes $[\delta, \theta, \alpha, \beta]$ forming the "brain" sub-network that accounts for the different brain wave amplitudes; the analysis was repeated considering each one of the seven processes as the target process ($j = [\eta, \rho, \pi, \delta, \theta, \alpha, \beta]$) and excluding it from the set of sources.

2. Second, the topological structure of the network of physiological interactions was detected computing the conditional transfer entropy $T_{i\to j|s}$ based on the two VAR identification methods combined with their method for assessing the statistical significance of cTE (i.e., using surrogate data for OLS and exploiting the intrinsic sparseness for LASSO). 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 vector with index s. As a quantitative descriptor of the network was used the in-strength, defined as the sum of all weighted inward links connected to one node [125]. Moreover, to describe the overall brain–body interactions the in-strength of the body sub-network due to brain sub-network (and vice-versa) was computed considering as link weights the percentage of subjects showing at least one statistically significant brain-to-body connection (and vice-versa). To study the involvement of each specific node in the network, the in-strength of each node was computed considering as link weights the cTE values of all network links pointing into the considered node.

2.4.3 Statistical Analysis

The effect of the different experimental conditions (R,M,G) on each PID measure computed for each target process ($j = [\eta, \rho, \pi, \delta, \theta, \alpha, \beta]$) and for each VAR identification method (OLS, LASSO) was assessed with a Kruskal-Wallis test followed by a Wilcoxon rank sum test to assess statistical differences between pairs of conditions. Moreover, the Wilcoxon rank sum test was performed also to assess statistical differences between the two unique TEs ($U_{i\to j}$,$U_{k\to j}$) or between the redundant and synergistic TEs ($R_{ik\to j}$,$S_{ik\to j}$) assessed for a given experimental condition and for a given target process and identification method. Finally, in order to assess the effect of the experimental condition on the in-strength evaluated for each node in the network, a Kruskal-Wallis test was performed, followed by the Wilcoxon rank sum test between pairs of conditions.

2.4.4 Results of Real Data Application

The results of PID analysis, describing how information is transferred within the observed network of brain–body interactions, are reported respectively in Figures 2.7 (OLS results) and 2.8 (LASSO results) for the targets belonging to the body sub-network (η,ρ,π), and in Figures 2.9 (OLS results) and 2.10 (LASSO results) for the targets belonging to the brain sub-network (δ,θ,α,β). The results of cTE analysis, illustrating the topology of the detected physiological networks, are reported in Figures 2.11(direct links), 2.7(brain–body interactions) and 2.13 (in-strength). All analyses are performed identifying VAR models of dimension Mp, where $M = 7$ and $p \sim 4$ (depending on the Bayesian Information Criterion) on time series of 300 points, which brought us to work with values $K \sim 10$ for the parameter relating the amount of data sample available to the model dimension.

Partial Information Decomposition

Figures 2.7 and 2.8 report, respectively for OLS and LASSO estimation, the distributions across subjects of the joint TE ($T_{ik\to j}$, left panels) directed to each target j belonging to the body sub-network from the two other body sources (index i) and from the four brain sources (index k), as well as of its decomposition into unique TEs ($U_{i\to j}$ and $U_{k\to j}$, middle panels) and redundant and synergistic TEs ($R_{ik\to j}$, $S_{ik\to j}$, right panels), evaluated at rest (R), during mental stress (M) and serious game (G).

Figure 2.7 shows that for each target in the body sub-network, the trends of the joint TE ($T_{ik\to j}$, Figure 2.7a,d,g) are mostly determined by the processes belonging to the same sub-network, as documented by the substantial values of the unique information transfer $U_{i\to j}$ and the negligible values of the unique transfer $U_{k\to j}$ (Figure 2.7b,e,h, with statistically significant difference between $U_{i\to j}$ and $U_{k\to j}$) and by the low values of the information transferred to η, ρ and π in a synergistic or redundant way from the brain and body sub-networks (Figure 2.7c,f,i). While for the targets η and ρ the PID measures did not vary significantly across conditions, the information transferred jointly from the brain and body sources towards the target π (Figure 2.7g) as well as the unique information transferred to π internally in the body sub-network (Figure 2.7h) decreased significantly moving from R to M and from R to G. This result documents a reduction of the causal interactions from RR interval and respiration towards the pulse arrival time during conditions of mental stress.

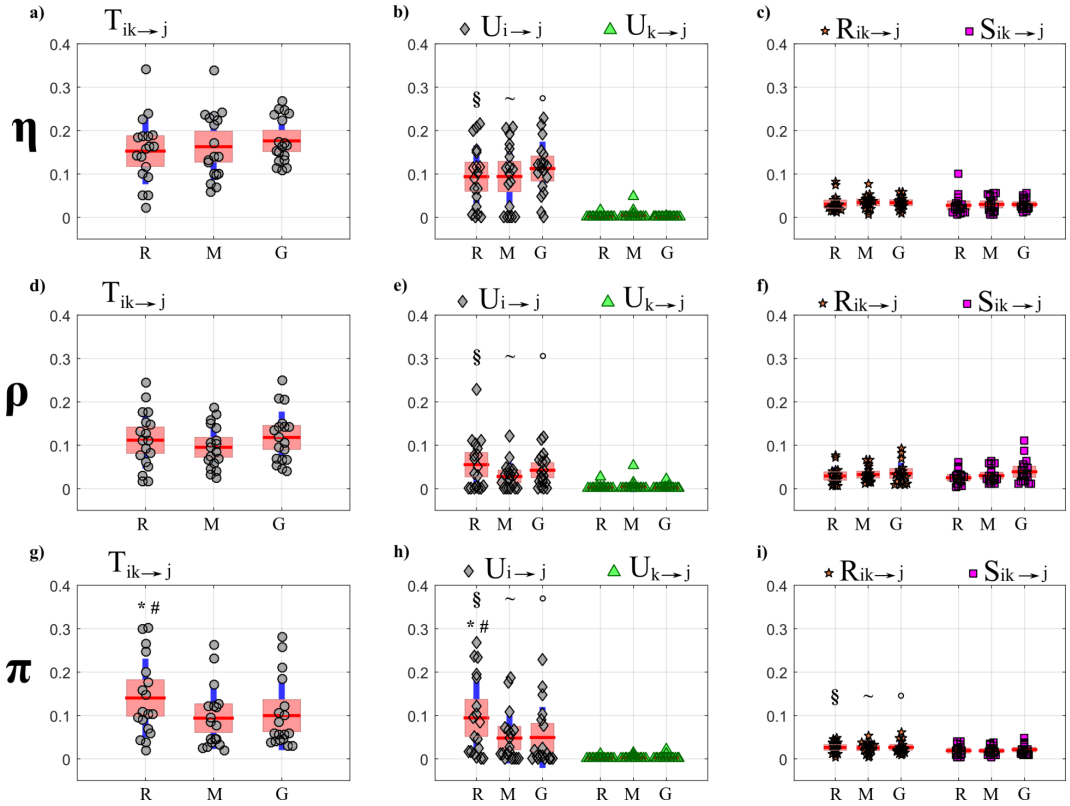

Figure 2.7: Partial Information Decomposition of brain–body interactions directed to the body nodes of the physiological network, assessed using OLS VAR identification. Box plots report the distributions across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue lines) as well as the individual values (circles or triangles) of the PID measures (**a**, **d**, **g**: joint information transfer; **b**, **e**, **h**: unique information transfer; **c**, **f**, **i**: synergistic and redundant transfer) computed at rest (R), during mental stress (M) and during serious game (G) considering the RR interval (η), the respiratory amplitude (ρ), or the pulse arrival time (π) as the target process j, and the body and brain sub-networks as source processes i and k. Statistically significant differences between pairs of distributions are marked with $*$ (R vs M), with # (R vs G), with § (R vs R), with $\sim$ (M vs M) and with $\circ$ (G vs G)

.

As reported in Figure 2.8, the trends of the joint TEs computed after LASSO identification when the processes η and π (a-g) are taken as target are comparable to those obtained with OLS identification and shown in Figure 2.7. In particular, also in this case a significant reduction of the joint TE directed to π is observed during the conditions M and G compared to R (Figure 2.3g), which is mostly due to a decrease of the unique information transferred to π from the body source ($U_{i\to j}$, Figure 2.8h). Moreover, also in this case the unique TE directed towards η and π from the brain sub-network ($U_{k\to j}$, Figure 2.8b,h) shows values very close to zero (b-h) and significantly lower than those of the unique TE $U_{i\to j}$. While the synergistic TE $S_{ik\to j}$ is almost zero for any target, the redundant TE $R_{ik\to j}$ is significantly higher than $S_{ik\to j}$ when the target is the vascular process π (Figure 2.8i). A result demonstrated specifically using the LASSO identification method is the absence of joint TE directed to the respiration process ρ (Figure 2.8d), documenting the absence of interactions directed toward respiration in all physiological conditions.

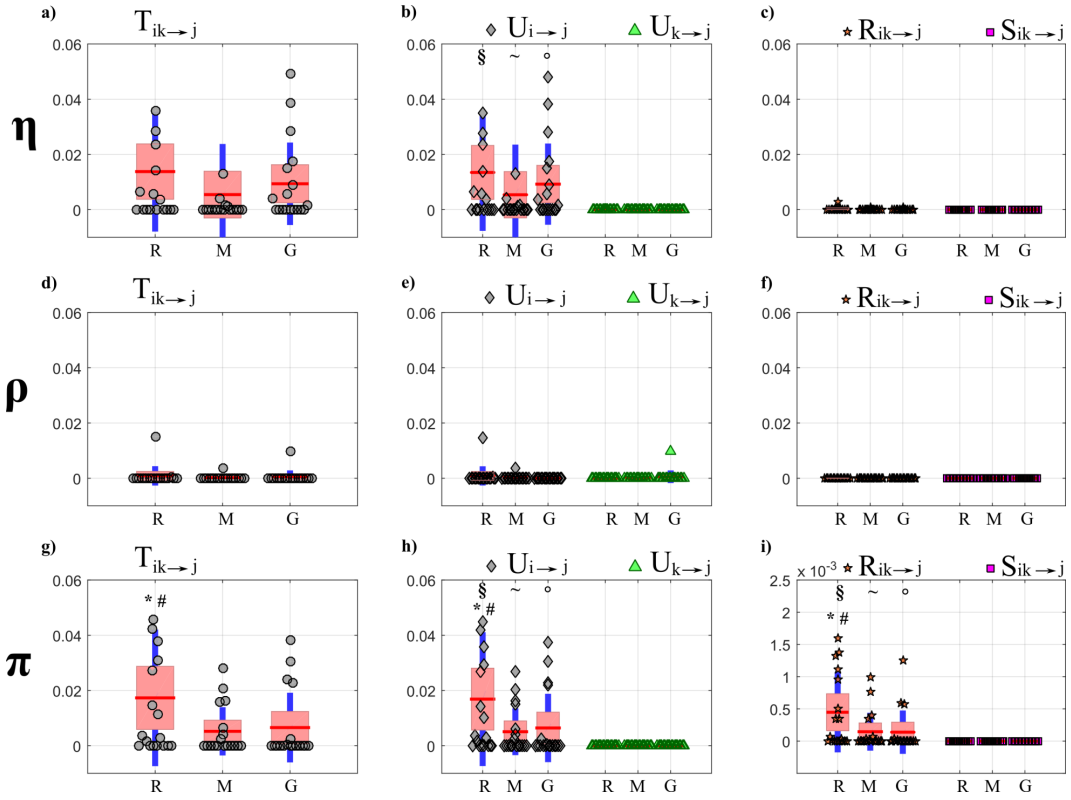

Figure 2.8: Partial Information Decomposition of brain–body interactions directed to the body nodes of the physiological network, assessed using LASSO-VAR identification. Box plots report the distributions across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue lines) as well as the individual values (circles or triangles) of the PID measures (**a**, **d**, **g**: joint information transfer; **b**, **e**, **h**: unique information transfer; **c**, **f**, **i**: synergistic and redundant transfer) computed at rest (R), during mental stress (M) and during serious game (G) considering the RR interval (η), the respiratory amplitude (ρ), or the pulse arrival time (π) as the target process j, and the body and brain sub-networks as source processes i and k. Statistically significant differences between pairs of distributions are marked with $*$ (R vs M), with # (R vs G), with § (R vs R), with $\sim$ (M vs M) and with $\circ$ (G vs G).

Figures 2.9 and 2.10 report, respectively for OLS and LASSO estimation, the distributions across subjects of the joint TE ($T_{ik\to j}$, left panels) directed to each target j belonging to the brain sub-network from the three other brain sources (index k) and from the three body sources (index i), as well as of its decomposition into unique TEs ($U_{i\to j}$ and $U_{k\to j}$, middle panels) and redundant and synergistic TEs ($R_{ik\to j}$, $S_{ik\to j}$, right panels), evaluated at rest (R) and during mental stress (M) and serious game (G).

Considering the joint TE exchanged toward the brain rhythms, in contrast to what observed for the body sub-network (2.7 a-e-g), the joint TE assessed through OLS identification shows a tendency to increase during M and especially during G compared to R (2.9 a,d,g,l); the increase is statistically significant for the δ (2.9 a), and is supported by a significant increase of the redundant and synergistic TEs $R_{ik\to j}$ and $S_{ik\to j}$ which suggests an increased contribution of brain–body interactions to the rhythmic variations of the δ brain wave amplitude. An increase of the redundant brain–body interactions during stress states is observed also for the θ brain wave amplitude (2.9 f). The analysis of the unique information transfer (2.9b,e,h,m) shows that the unique information provided by the brain sub-network ($U_{k\to j}$) is generally larger than that provided by the body sub-network ($U_{k\to j}$), with statistically significant differences during R and when the target of the unique transfer is given by the processes θ, α and β.

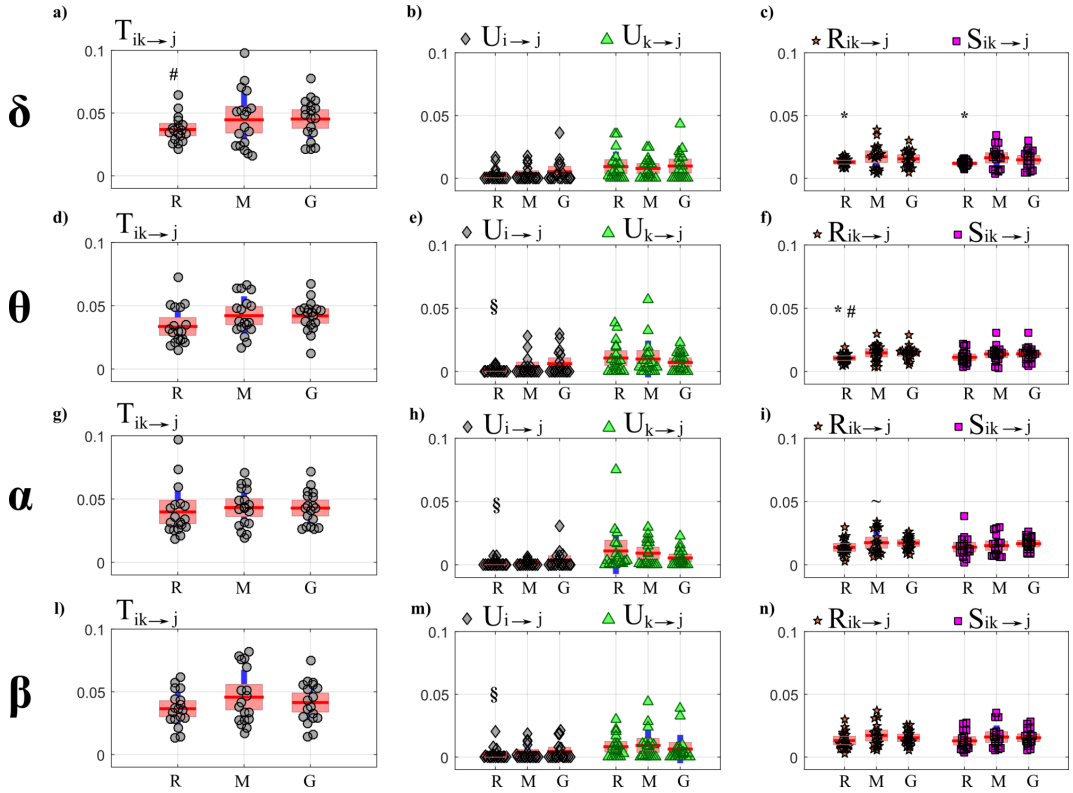

Figure 2.9: Partial Information Decomposition of brain–body interactions directed to the brain nodes of the physiological network, assessed using OLS VAR identification. Box plots report the distributions across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue lines) as well as the individual values (circles or triangles) of the PID measures (**a**, **d**, **g**, **l**: joint information transfer; **b**, **e**, **h**, **m**: unique information transfer; **c**, **f**, **i**, **n**: synergistic and redundant transfer) computed at rest (R), during mental stress (M) and during serious game (G) considering the δ, θ, α, or β brain wave amplitude as the target process j, and the body and brain sub-networks as source processes i and k. Statistically significant differences between pairs of distributions are marked with $*$ (R vs M), with # (R vs G), with § (R vs R), with $\sim$ (M vs M) and with $\circ$ (G vs G).

When PID directed towards the brain processes is computed using LASSO (Figure 2.10), a main result is that interactions are weak and do not vary significantly across physiological states. Notably, the joint TE and all PID terms relevant to the target δ are almost equal to zero in all conditions (Figure 2.5a,b,c). Similarly, also the values of the unique TE from the body sub-network to any brain process ($U_{i\to j}$, Figure 2.10b,e,h,m) and of both the redundant and synergistic TE ($R_{ik\to j}$, $S_{ik\to j}$, Figure 2.10c,f,i,n) are zero in almost all subjects and conditions, indicating that the LASSO approach does not detect interactions directed from body to brain in this dataset.

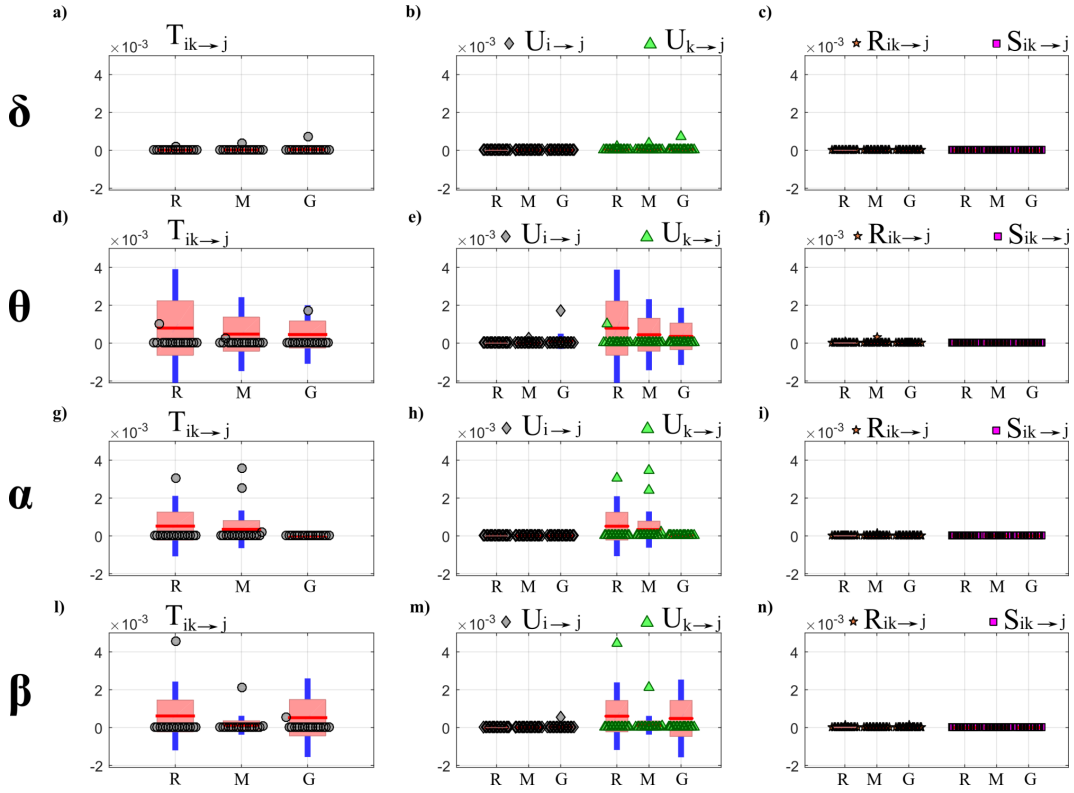

Figure 2.10: Partial Information Decomposition of brain–body interactions directed to the brain nodes of the physiological network, assessed using LASSO-VAR identification. Box plots report the distributions across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue lines) as well as the individual values (circles or triangles) of the PID measures (**a**, **d**, **g**, **l**: joint information transfer; **b**, **e**, **h**, **m**: unique information transfer; **c**, **f**, **i**, **n**: synergistic and redundant transfer) computed at rest (R), during mental stress (M) and during serious game (G) considering the δ, θ, α, or β brain wave amplitude as the target process j, and the body and brain sub-networks as source processes i and k. Statistically significant differences between pairs of distributions are marked with $*$ (R vs M), with $\#$ (R vs G), with $\S$ (R vs R), with $\sim$ (M vs M) and with $\circ$ (G vs G).

Conditional Information Transfer

Figure 2.11 reports the network of physiological interactions reconstructed through the detection of the statistically significant values of the conditional transfer entropy ($T_{i\to j|s}$) computed for any pair of processes belonging to the brain and body sub-networks. The weighted arrows, depicting the most active connections among systems (arrows are present when at least 3 subjects show significant values of $T_{i\to j|s}$) show a similar structure when estimated in the three analyzed conditions using OLS (2.11a,b,c) and LASSO (2.11d,e,f) . The main distinctive features are the existence of a densely connected sub-network of body interactions (red arrows), of a weakly connected sub-network of brain interactions (yellow arrows), and of changing patterns of brain–body interactions (blue arrows). In general, LASSO shows, for each condition analyzed, a greater sparsity in the estimated networks, preserving only the most active links detected by OLS.

Within body interactions are characterized mainly by cardiovascular links (interactions from η to π) and cardio-respiratory links (interactions between η and ρ), with a weaker coupling between ρ and π which exhibits a preferential direction from ρ to π; the use of LASSO elicits the unidirectional nature of cardio-respiratory interactions (from ρ to η). On the other hand, the topology of the brain sub-network is less stable in the three conditions and appears to lose consistency passing from REST

to GAME; also in this case the use of LASSO leads to a greater sparsity, with nodes almost fully disconnected. As to brain–body interactions, they occur almost exclusively along the direction from brain to body; in this case the use of LASSO demonstrates that interactions from brain to body increase during the GAME condition.

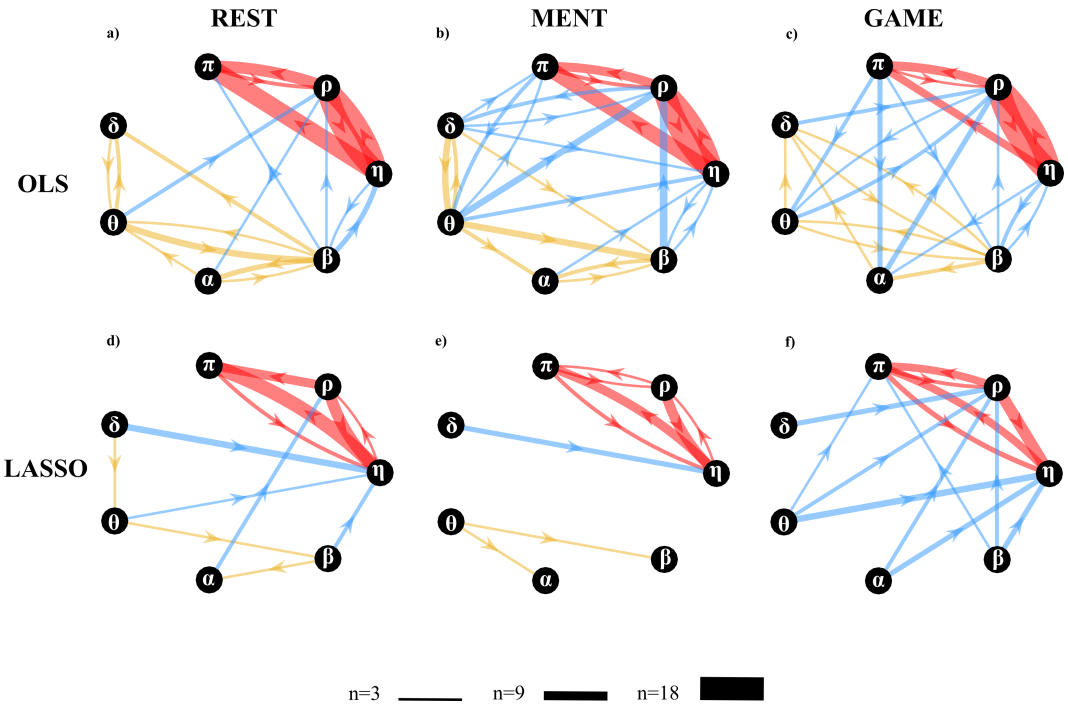

Figure 2.11: Topological structure for the networks of physiological interactions reconstructed during the three analyzes physiological states. Graphs depict significant directed interactions within the brain (yellow arrows) and body (red arrows) sub-networks as well as interactions between brain and body (blue arrows). Directed interactions were assessed counting the number of subjects for which the conditional transfer entropy ($T_{i\to j|s}$) was detected as statistically significant using OLS (**a**, **b**, **c**) or LASSO (**d**, **e**, **f**) to perform VAR model identification. The arrow thickness is proportional to the number of subjects (n) for which the link is detected as statistically significant.

To quantify the overall extent of the brain–body interactions from the above estimated cTE networks, was computed the percentage of subjects with statistically significant values of the cTE along the direction from brain to body and in the opposite direction from body to brain. This was obtained considering the brain sub-network and the body sub-network as single nodes, and computing the in-strength to one sub-network by considering only the connections coming from the other sub-network. The average values are shown in Figure 2.12.

The results reported in Figure 2.12 show that interactions are found more consistently along the direction from brain to body than along the opposite direction. In particular, LASSO does not show any link directed from body to brain in any of the three analyzed conditions. In the resting condition (R), the percentage of active links directed from brain to body is similar for the two VAR identification methods. Then, OLS identification results in a larger number of links moving from R to M, and a decrease during G. Conversely, LASSO shows a decrease of the percentage of significant links during M and a sharp increase during G.

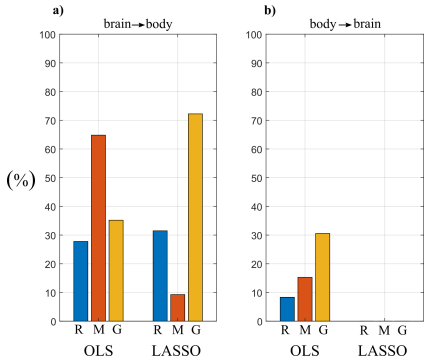

Figure 2.12: Bar plots reporting the in-strength index extracted from the cTE networks of Figure 2.11 by considering as link weights the percentage of subjects showing a brain-to-body connection (**a**) or a body-to-brain connection (**b**), computed at rest (R), during mental stress (M) and during serious game (G) for the two VAR identification methods. Please note that the in-strength computed along the direction from body to brain using LASSO is null in all conditions.

Figure 2.13 reports the distribution of the values of the in-strength index evaluated for each node of the network in each experimental condition. For both OLS and LASSO, the median value of the in-strength index (Figure 2.13a,b,c,h,i,l) is higher for the network nodes of the body sub-network than for those belonging to the brain sub-network (Figure 2.13 d,e,f,g,m,n,o,p). An exception to this difference is the in-strength of the links directed towards the node ρ, which is very close to zero when assessed using LASSO identification (Figure 2.13i). Moreover, the estimated in-strength values are, on average, lower when assessed through LASSO than through OLS. Considering the in-strength of individual nodes, a statistically significant reduction is observed moving from R to G for the weights of the connections directed towards π (Figure 2.13c,l), for both OLS and LASSO methods.

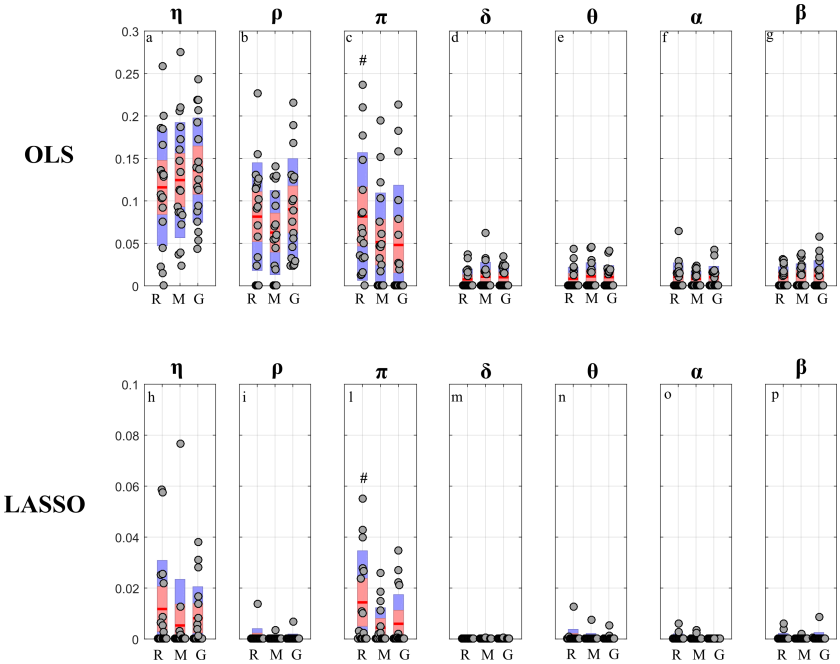

Figure 2.13: In-strength index computed for each node of the physiological network. Box plots report the distributions across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue bars) as well as the individual values (circles) of the in-strength index (a-g) OLS, h-p LASSO) computed at rest (R), during mental stress (M) and during serious game (G) for each node (η,ρ,π,δ,θ,α,β). Statistically significant differences between pairs of distributions are marked with # (R vs G).

2.5 Discussion

2.5.1 Simulation Study I

The first simulation study was designed to compare the performance of the traditional OLS approach and the LASSO regression, implemented for the identification of VAR models in their state–space formulation [102], in estimating the information measures related to PID. The decomposition of the information transferred jointly from two sources to a target process allows investigation of how information is modified in a non-trivial way through redundant and synergistic interactions between the sources [126] . In particular, the model structure adopted in our simulation highlights the coexistence of synergistic and redundant contributions to the target Y_4 from the two sources Y_2 and Y_3 even if they are not directly coupled [103]. In situations such as this, the adoption of PID is fundamental to elicit how the two sources contribute to the target with both redundant and synergistic information transfer: the redundant contribution refers to the common information that both sources convey to the target; the synergistic contribution is considered an extra information transferred towards the target and is ascribed to the weakest source in the system [89].

The analysis in Figures 2.2 and 2.3 shows an evident dependence of both the bias and the variance of all partial information decomposition measures on the factor K. This result is expected and reflects the well-known decrease of the prediction accuracy with the number of data samples available. In this context, our results document that the LASSO regression performs better in challenging conditions when the number of model parameters approaches the sample size ($K \leq 5$). In these conditions it has been pointed out how OLS is not suitable for the solution of a regression problem and that its solution could even not exist [109, 127]. On the other hand, LASSO shows high robustness to the lack of data points, which results in limited values of bias and variance [38]. We note that despite this better performance of LASSO, in the condition $K = 1$ all the PID measures that were different from zero ($T_{2\to4}$,$T_{3\to4}$,$S_{23\to4}$,$R_{23\to4}$) exhibit a consistent negative bias (Figure 2.2). This severe under estimation was previously highlighted in different scenarios, in which LASSO shrinkage produces biased estimation for the large coefficients and thus in some conditions could be sub-optimal in terms of estimation risk [128, 129].

When the amount of data sample is not scarce compared to the number of model parameters ($K > 5$) the performance of the two identification methods is comparable, with slight differences depending on the true value of the PID measures. In the case of non-zero PID measures (Figure 2.2) OLS showed better performance than LASSO in terms of bias and variance. This result is mainly due to the effect of the constraint based on the l_1 norm that performs a variable selection but with an increased bias and variance in the performed estimation. [44, 105].

On the other hand, in the scenario in which all the PID measures are equal to zero (Figure 2.3), LASSO performs better than OLS in all the conditions analyzed as regards both the bias and the variance of the estimates of information transfer. This can be explained with the continuous shrinkage and selection of the most relevant coefficients that set to zero most of the estimated AR coefficients [55].

2.5.2 Simulation Study II

The second simulation was designed to compare the performance of OLS and LASSO identification in estimating the cTE in a network of multiple interacting processes. The tested measure

is highly relevant, as it is equivalent to the multivariate (conditional) Granger causality measure estimated within the most accurate framework available, i.e., that of vector state–space models[102]. Within this framework, we assessed both the statistical significance and the accuracy of the estimated values of the cTE, thus comparing OLS and LASSO regarding their accuracy in detecting the network structure and the coupling strength.

The accuracy in the estimation of the cTE values was investigated across different K ratio levels by means of $BIAS$ and $BIAS_N$ used as performance parameters (Figure 2.5). As expected, both parameters show a tendency to increase as the K ratio decreases. This tendency is evident particularly for OLS estimation, as already documented testing different VAR parameter identification approaches (e.g., the Levinson recursion for the solution of Yule-Walker equations) in the context of signal processing [9]. The situation becomes worse when approaching the condition $K = 1$, in which the matrix $([\mathbf{y}^p]^T\mathbf{y}^p)^{-1}$ approaches singularity. Consequently, in this case the solution to the DARE equation necessary to convert the SS model into the ISS form did not converge, thus impeding OLS-based estimation of the cTE. In such conditions it is necessary to move to the use of penalized regression techniques [44, 105, 109]. Here we document that the LASSO regression leads to trends of the cTE bias which are consistently very low for any value of K in the estimation of the null links (Figure 2.5a), and rise with K but without exhibiting abrupt increases even for $K = 1$ in the estimation of the non-null links (Figure 2.5b). These good performances of LASSO identification confirm its higher tolerance to collinearity between regressors caused by the reduction of data samples available [72].

The reliability in the reconstruction of the network structure was investigated analyzing the performance of the two identification methods in terms of overall accuracy and rates of false negative and false positive detections. The ACC parameter appeared to be the best-suited indicator to synthesize the similarity between the estimated network and the ground-truth network [122]. Moreover, with the network structure simulated here, ACC is not affected by the class imbalance problem, a typical condition in sparse networks [130]. As expected, the ACC parameter decreased with the K ratio, with LASSO performing progressively better than OLS (Figure 2.6 c). These results are in line with previous studies reporting the performance of different methods for the assessment of the statistical significance of causal interactions in different methodological contexts [34, 104, 105].

When the test was particularized to the rate of correct detection of null and non-null links, the performance under conditions of data paucity differ for the two identification methods, with LASSO showing better capability to correctly detect existing links (lower FNR) and OLS showing slightly better capability to correctly detect the absent links (lower FPR). In particular, by analyzing the trends of FNR (Figure 2.6 a) LASSO showed better performance than OLS for $K \leq 10$, especially when the conditions for the estimation become very challenging ($K \leq 5$). This behavior is related to the shrinkage of the VAR parameters. In fact, the selected lambda tends to rise if the number of data samples decreases and this implies a greater sparsity of the estimated network with a high probability of producing false negatives [131]. In the same conditions, the value of FNR for OLS was around 60%. This poor performance is likely due to an inaccurate representation of the distribution of the cTE under the null hypothesis of uncoupling, estimated empirically using uncoupled surrogate time series, performed with very few data samples. On the contrary, while both methods display a low number of false positives for $K > 5$, LASSO tends to produce an over-selection of the estimated links when $K \leq 5$. This result is in line with previous findings in the context of GC estimation, in

which LASSO showed few extra links, observed for different combinations of degree of sparsity of the simulated network structure and K ratio [39, 41].

2.5.3 Real Data Application

Partial Information Decomposition Analysis

The main results of the partial decomposition of the information transfer within the network of brain and body interactions are that: (i) a significant information is transferred within the body sub-network, composed by the processes representative of the cardiac (η, heart period), vascular (π, pulse arrival time) and respiratory (ρ) dynamics, which is directed towards the η and π nodes as a result of respiration-related and cardiovascular effects; (ii) the information transferred to the nodes of the brain sub-network, representing the amplitude variations of the δ, θ, β, and α EEG waves, is lower and due almost exclusively to internal dynamics within this sub-network; (iii) a negligible amount of information is transferred between the two sub-networks as a result of their redundant or synergistic interaction. While these results are observed consistently using the two VAR identification methods (see Figures 2.7- 2.8 and Figs. 2.9- 2.10, respectively), the use of the LASSO regression allows the elicitation of them more clearly. From a methodological point of view, this behavior is a result of the inclination towards sparseness of the LASSO method, which shrinks towards zero most of the VAR parameters that have a small effect on the target dynamics [44]. Such inclination puts also in evidence other behaviors, such as the substantial absence of information directed to the ρ node of the body network and to the δ node of the brain network. While in the first case the result is physiologically plausible since cardio-respiratory interactions are known to be almost unidirectional in nature (i.e., previous studies have found that respiration significantly affects the cardiovascular variables without being affected by them [77, 121, 132]), in the second case it could be related to an underestimation of the information transfer with the LASSO technique, since the δ waves seem to play a role in the organization of brain dynamics [18, 82, 133].

As the results reported above were observed consistently independently on the analyzed physiological state, they could be interpreted as a hallmark of how the networks of brain and body interactions organize their dynamic communication evaluated in terms of information transfer. Nevertheless, the conditions of mental stress evoked by the mental arithmetic task and the sustained attention task were able to induce, when compared with the resting condition set as baseline, some significant modifications in the amount of information transferred toward some specific nodes. In particular, a significant reduction of the joint brain–body TE computed when π was taken as the target process was observed during the two stress conditions compared to rest. This joint information transfer was due almost exclusively to contributions of unique transfer from the η and ρ nodes of the body sub-network (Figure 2.7 h,2.8 h), with a small amount of redundant brain–body information transfer (Figure 2.7 h,2.8 i) and negligible amounts of synergistic transfer or unique transfer from the brain sub-network; the unique transfer reflects cardiac and respiratory effects on the variability of the pulse arrival time, while the redundant transfer is related to common mechanisms whereby such variability is influenced by the brain rhythms one side and the cardio-respiratory rhythms on the other side. In this context, the results here obtained are in line with those obtained in [78] where a significant reduction of total information transferred towards π was found while playing a serious game with respect to a resting condition. Analyzing the same dataset in terms of mutual

information, the authors of [111] found a significant reduction of the information shared between the pulse arrival time (π) and the cardio-respiratory system (η, ρ) during the conditions M and G compared with R. The significant decrease of the static mutual information computed in [111] and the dynamic measure of the joint and unique TE computed in the present study can be viewed as different aspects of the weakening of cardiovascular and cardio-respiratory interactions during mental stress. Physiologically, the underlying mechanisms could include an increased modulation of peripheral vascular resistance during stress which, as highlighted in [118, 134], could dampen the modulation of the pulse arrival time due to heart rate variability and respiration.

When the target process belongs to the brain sub-network, the information transfer estimated through the LASSO regression was almost null when directed towards δ and very small when directed towards θ, α or β (Figure 2.10 a-c). This result may reflect the lack or significant connectivity towards the brain sub-network, or the lower sensitivity of penalized regression methods to weak connectivity. In fact, using OLS a certain amount of information transfer to the nodes of the brain network was detected, with a significant increment of the joint transfer entropy from R to G when δ is the target process (Figure 2.9 a), that is mostly due to the significant increment of redundant and synergistic TEs (Figure 2.9 c). Furthermore, a significant increase of the redundant TE ($R_{ik\to j}$) was also observed during M and G with respect to R when θ is the target process (2.9 f). The involvement of the brain waves during mental stress tasks was also investigated using information measures in [111], finding a larger involvement of δ and θ activity compared to rest that agrees with the results obtained here in terms of redundant TE computed after OLS identification.

Conditional Information Transfer Analysis

The analysis of the statistically significant values of the conditional information transfer (cTE measure) led us to detect specific topology structures for the sub-networks that compose the overall physiological network of brain and body interactions (Figure 2.11). First, a quite consistent topology was found across different physiological states for the interactions between the cardiovascular and respiratory systems (Figure 2.11a,b,c, and Figure 2.11d,e,f, red arrows), which is in line with a recent similar work performed in the context of information dynamics [78, 92]. In particular, the strong link connection between η and ρ reflects a marked coupling between the heart rate variability and respiration, which is due to the well-known mechanisms such as respiratory sinus arrhythmia (RSA) [135] and cardio-respiratory synchronization [136]. This connection was detected as bidirectional using OLS, and as unidirectional from ρ to η using LASSO, confirming that the preferential direction of the cardio-respiratory interactions is that documenting the effect of respiration on the heart rate (RSA) [77, 115, 136]. Second, the information transferred from η to π reflects the well-known effect of the heart rate on stroke volume and arterial pressure which has a modulating effect on the arterial pulse wave velocity [137]. Moreover, the influence of respiration ρ on the pulse arrival time variability π reflects the breathing influences on the intra-thoracic pressure, blood pressure and blood flow velocity [137].

A further result relevant to the peripheral sub-network is the significant decrease of the in-strength relevant to the vascular node π observed for both OLS and LASSO moving from rest to the serious game condition independently (Figure 2.13 c,l). This weaker topology is likely related to the significantly lower amount of information transferred towards π during the condition G compared to R (Figure 2.7-g and Figure 2.8-g). From a physiological point of view, this lower

transfer mediated by weaker topology could suggest a reduction of the efferent nervous system activity from the cardiac and respiratory centers and directed towards the vascular system during conditions of mental attention.

Compared with the body sub-network, the links of the brain sub-network form a structure which seems less consistent across the different experimental conditions (Figure 2.11, yellow arrows). While OLS estimation shows an apparent decrease in the number of connections moving from R to M and especially to G, the LASSO regression yields an almost disconnected sub-network of brain-brain interactions. In contrast to that observed in this work, in [78] a more connected brain sub-network was found during the mental arithmetic task with respect to the resting condition. This difference can be partially methodological, as different model order selection criteria (Akaike vs. Bayesian) and methods to assess the statistical significance of cTE (F-test vs. surrogate data) were used in [78] and in the present work. These choices could indeed affect the estimation procedure and provide slightly different results especially in the presence of weak connections as in this case [34, 138, 139].

Finally, exploration of the network of dynamical interactions between the brain and the peripheral systems led us to investigate how the EEG dynamics, mostly determined by the central nervous system, interact with the cardiovascular and respiratory dynamics regulated by the autonomic nervous system (Figure 2.11, blue arrows, and Figure 2.12). Although quantitative statistical comparison cannot be performed for the results reported in Figures 2.11,2.12 they document that brain–heart interactions are mostly oriented in the direction from brain to heart. This suggests that efferent autonomic commands directed to the peripheral systems follow in time the neural modulation of the brain wave amplitudes. Moreover, we find that the two mental stress conditions induce an enhancement of brain–body interactions, with a substantial increase of the number of significant links directed from the brain to the body sub-network and assessed using OLS during the mental arithmetic condition, or using LASSO during the serious game condition. The results based on OLS resemble those obtained recently on the same dataset [78], and recall previous findings highlighting significant correlations between the amplitude of brain oscillations (especially in the β band) and the heart rate and respiration dynamics [82, 140]. The results based on LASSO highlight the emergence during sustained attention evoked by serious game playing of causal interactions from brain to the peripheral systems, mostly originating from the θ, α and β nodes and directed to the ρ and η nodes. These findings are supported by previous studies suggesting that the neural mechanisms responsible for the generation of α and θ brain oscillations are crucial for attention tasks and can be correlated with the cardiac autonomic activity and to its respiratory determinants [141–143].

2.6 Conclusions

The aim of this work was to test the usefulness of penalized regression techniques for the computation of different parametric measures of information transfer in networks of coupled stochastic processes. In particular, we considered the LASSO regression, a well-known technique that has been extensively used in different research fields, and implemented it for the first time within the most advanced framework for the linear parametric estimation of information dynamics, i.e., that based on the state–space computation of conditional Granger causality and partial information decomposition in vector stationary stochastic processes [89, 102, 103]. Our comparative validation with the traditional least squares identification of vector stochastic processes (OLS estimator) highlighted

that LASSO allows highly accurate estimation of not only the amount of information transferred between coupled processes, but also the topological structure of the underlying network, especially in conditions of data paucity which make OLS estimation unreliable or even not applicable. On the other hand, in favorable conditions of data size related to the dimension of the model to be identified the results of classical and penalized regression were fully overlapped, confirming the appropriateness of embedding LASSO into the framework for the linear parametric analysis of information dynamics.

The application of the two identification methods to the study of the network of physiological interactions within and between brain and peripheral dynamics has demonstrated consistent patterns of information transfer and similar network structures. Here, the main findings regard the detection of significant information transfer within the body sub-network sustained by cardiovascular and respiratory dynamics, with reduced cardio-respiratory effects on the vascular dynamics in the presence of mental stress, and the existence of weak but significant brain–body interactions directed from the brain rhythms to the peripheral dynamics, with enhanced link strength in conditions of mental stress. It is worth noting that these results were obtained for K=10, a condition in which the two identification procedures showed comparable performance in the simulation studies. This finding suggests that even in conditions that allow the use of OLS, LASSO is able to detect the strongest interactions among those determined by the combined activity of the central and autonomic nervous systems, providing as outcome estimated patterns of information dynamics which are more straightforward and easy to interpret than those obtained with OLS.

The directed links between different physiological systems observed in this study can reflect either well-defined physiological mechanisms, such as the respiratory and heart rate effects on the pulse arrival time [134, 144], or statistical associations with likely common determinants of physiological origin, like the brain–heart interactions which are thought to be mediated by dynamic alterations of the sympatho-vagal balance [82, 96, 145]. In either case, approaches like ours that allow the probing of the dynamic interaction among different organ systems can be very useful to show how an imbalanced interaction may have a negative impact on health [145]. Previous studies have indeed demonstrated pathological changes in brain–body interactions with clinical significance, for instance related to sleep stages and insomnia [146], to sleep apneas [147] or to schizophrenia [132]. However, the analysis of brain–body interactions in different experimental conditions such as those analyzed in this paper, is somehow still unexplored and further studies need to be performed in order to strengthen the validity of the results obtained in the present and in previous studies.

Chapter 3

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

3.1 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 [148] and formalized in terms of linear regression analysis by Granger [28]. 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 [28]. 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 [29]. 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 [149].

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 [7]. 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 [139, 150, 151]. To overcome the latter problem, an approach based on state-space (SS)

modeling of the observed VAR process has been introduced [102]; 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 [17, 102, 139].

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 [152] by using the closed-form solution of Ordinary Least Square (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 [153, 154]. 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 [155].

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 [8, 20]. 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 [36, 112]. 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 [107], or the so-called partial conditioning [106], and of penalized regression techniques based on the l_1-norm (LASSO regression) [8, 44, 156]. 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) [21]. While the use of ANNs as a VAR model for GC estimation has been proposed in both linear [157] and non-linear frameworks [158–160], 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 [8, 78], 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 [161–163].

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 NN-GC MATLAB toolbox, which can be downloaded from https://github.com/YuriAntonacci/ANN-GC-Toolbox.

3.2 Methods

3.2.1 Vector Autoregressive Model Identification

Let us consider a dynamical system $\mathcal{Y}$ 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:

$$\mathbf{Y}_n = \sum_{k=1}^{p} \mathbf{Y}_{n-k}\mathbf{A}_k + \mathbf{U}_n, \tag{3.1}$$

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 zero-mean white processes, denoted as innovations, with $M \times M$ covariance matrix $\mathbf{\Sigma} \equiv \mathbb{E}[\mathbf{U}_n^T\mathbf{U}_n]$ ($\mathbb{E}$ is 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 i^{th} row is a realization of $\mathbf{Y}_i$, i.e., $\mathbf{y}_i = [y_{1,i}...y_{M,i}], i = 1, ..., N$, and the j^{th} 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 Square (OLS) identification finds an optimal solution for the problem (3.1) by solving the following linear quadratic problem:

$$\widehat{\mathbf{A}} = \text{argmin}_{\mathbf{A}} ||\mathbf{y} - \mathbf{y}^p\mathbf{A}||_2^2, \tag{3.2}$$

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 $(N-p) \times Mp$ matrix of the regressors and $\mathbf{A} = [\mathbf{A}_1; \cdots ; \mathbf{A}_p]$ is the $Mp \times M$ coefficient matrix. The 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 Residuals Sum of Squares (RSS) is minimized [112].

3.2.2 Artificial Neural Networks as a Vector Autoregressive Model

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

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, f(\mathbf{w}; \mathbf{x}_i)) + \lambda r(\mathbf{w}), \tag{3.3}$$

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 $\lambda \in \Re^+$. 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:

$$r(\mathbf{w}) = ||\mathbf{w}||_1 = \sum_{k=1}^{Q} |\mathbf{w}_k|. \tag{3.4}$$

Then, a possible way to solve the problem (3.3) 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:

$$\mathbf{w}^{j+1} = \mathbf{w}^j + \eta_j \frac{\partial}{\partial \mathbf{w}} \left(l(y_i, f(\mathbf{w}; \mathbf{x}_i)) - \frac{\lambda}{N} \sum_{k=1}^{Q} |\mathbf{w}_k| \right), \tag{3.5}$$

where j is the iteration counter and η_j is the learning rate at each iteration. The difficulty with l_1 regularization is that the last term on the right-hand side in (3.5) is not differentiable when the weight is zero. To solve this issue, following the procedure introduced in [21] l_1 regularization 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:

$$u_j = \frac{\lambda}{N} \sum_{t=1}^{j} \eta_t. \tag{3.6}$$

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

$$w_k^{j+\frac{1}{2}} = w_k^j + \eta_j \frac{\partial l(y_i, f(\mathbf{w}; \mathbf{x}_i)}{\partial \mathbf{w}} \bigg|_{\mathbf{w}=\mathbf{w}^j}, \tag{3.7}$$

$$\textbf{if } \; w_k^{j+\frac{1}{2}} > 0 \; \textbf{ then } \; w_k^{j+1} = max(0, w_k^{j+\frac{1}{2}} - (u_k + q_k^{j-1})), \tag{3.8}$$

$$\textbf{else if } \; w_k^{j+\frac{1}{2}} < 0 \; \textbf{ then } \; w_k^{j+1} = min(0, w_k^{j+\frac{1}{2}} - (u_k - q_k^{j-1})), \tag{3.9}$$

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

$$q_k^j = \sum_{t=1}^{j} (w_k^{t+1} - w_k^{t+\frac{1}{2}}). \tag{3.10}$$

This method for updating the weights penalizes the weight according to the difference between u_j and 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 (3.3) 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 (3.3) in the linear framework is reported in Figure (3.1) for the n^{th} 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:

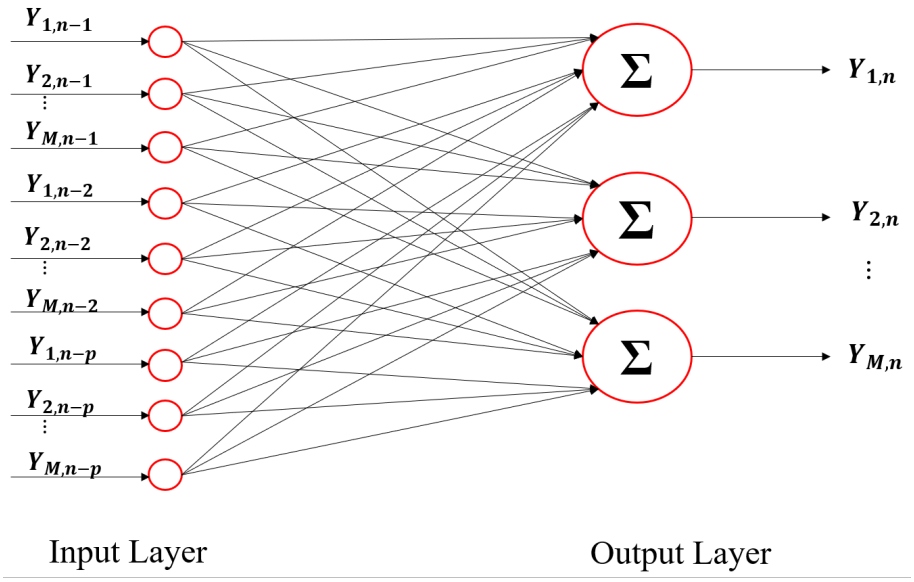

Figure 3.1: Schematic representation of the architecture of the Neural Network used as VAR model. The input and the output of the network are represented by the lagged variables and by the present states of all processes included in the analysis

$$l(\mathbf{y}, \mathbf{y}^p\mathbf{W}) = ||\mathbf{y} - \mathbf{y}^p\mathbf{W}||_2^2, \tag{3.11}$$

which highlights that the weight $\mathbf{W}$ corresponds to the matrix $\mathbf{A}$ containing the parameters of the VAR model (3.1). Thus, the described ANN is completely equivalent to a VAR model, except for the fact that the training process induces sparsity into the weight matrix $\mathbf{W}$.

Determination of the regularization parameter

The determination of the regularization parameter λ is a key element of the estimation process, as its selection strongly influences the performance of resulting regression. For a high value of λ, the

SGD-l_1 algorithm provides a matrix of weights $\mathbf{W}$ in which all entries are zero. On the other hand, when $\lambda \to 0$, the weights stored in $\mathbf{W}$ are all different from zero and the solution corresponds to the OLS solution [44]. In this work, the optimal value for λ has been tested in the range $[\lambda_l, \lambda_u]$, where λ_l and λ_u are the values leading to maximum density (no zero elements) and maximum sparseness (all zero elements) of the weight matrix. Subsequently, following the procedure described in [164], with a hold out approach, we independently draw 90% of the samples available (rows of $\mathbf{y}$ and $\mathbf{y}^p$) as the training set and kept the remaining 10% for testing. Training and test sets were then normalized and, for each assigned λ, the number of non-zero weights was counted in the matrix $\widehat{\mathbf{W}}$ estimated on the training set, and the RSS was computed on the test set as well. This procedure was iterated for each λ, and the optimal λ was taken as the value minimizing the ratio between RSS and the number of non-zero weights [8, 55, 164]. The weight matrix $\mathbf{W}$ obtained with the selected optimal λ was then used for the subsequent GC analysis.

3.2.3 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 $Y^p_{i,n}$ and the past of the target process $Y^p_{j,n}$ we state that the i^{th} process G-causes the j^{th} process (conditional on the other s processes), if $Y^p_{i,n}$ conveys information about $Y_{j,n}$ above and beyond the information contained in $Y^p_{j,n}$ and in all other processes $\mathbf{Y}^p_{s,n}$. 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 $E_{j|ijs,n} = Y_{j,n} - \mathbb{E}\left[Y_{j,n}|\mathbf{Y}^p_n\right]$ and $E_{j|js,n} = Y_{j,n} - \mathbb{E}\left[Y_{j,n}|Y^p_{j,n}, \mathbf{Y}^p_s\right]$. The resulting prediction error variances, $\lambda_{j|ijs} = \mathbb{E}[E^2_{j|ijs,n}]$ and $\lambda_{j|js} = \mathbb{E}[E^2_{j|js,n}]$ are then combined to obtain the definition of GC (in its conditional form) from Y_i to Y_j [165]:

$$F_{i\to j|s} = \ln \frac{\lambda_{j|js}}{\lambda_{j|ijs}}. \tag{3.12}$$

Following a similar reasoning, the GC in its original form (unconditional) from Y_i to Y_j is defined as [28]:

$$F_{i\to j} = \ln \frac{\lambda_{j|j}}{\lambda_{j|ij}}, \tag{3.13}$$

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

The prediction error variances needed for the determination of the GC measures can be computed from the identification of the model (3.1) 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 j^{th} diagonal element of the error covariance matrix $\lambda_{j|ijs} = \Sigma(j, j)$. The other partial variances in (3.12) and (3.13) can be retrieved, starting from the identification of the full model, by exploiting the theory of State-Space (SS) models [102, 166], according to which the VAR model (3.1) can be represented as an SS model relating the observed process $\mathbf{Y}$ to an

unobserved process $\mathbf{Z}$ through the equations [17, 102]:

$$\mathbf{Z}_{n+1} = \mathbf{Z}_n\mathbf{A} + \mathbf{E}_n\mathbf{K}, \tag{3.14}$$

$$\mathbf{Y}_n = \mathbf{Z}_n\mathbf{C} + \mathbf{E}_n, \tag{3.15}$$

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 (3.1) 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) [102], 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 (3.1) as reported in [166]. The advantage of this representation is that it allows to form "submodels" which exclude one or more scalar processes from the observation equation (3.15) leaving the state equation (3.14) 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:

$$\mathbf{Y}_{js,n} = \mathbf{Z}_n\mathbf{C}^{(js)} + \mathbf{E}_{js,n}, \tag{3.16}$$

$$\mathbf{Y}_{ji,n} = \mathbf{Z}_n\mathbf{C}^{(ji)} + \mathbf{E}_{ji,n}, \tag{3.17}$$

$$\mathbf{Y}_{j,n} = \mathbf{Z}_n\mathbf{C}^{(j)} + \mathbf{E}_{j,n}, \tag{3.18}$$

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 [102], the submodels (3.14,3.16), (3.14,3.17) and (3.14,3.18) 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 i^{th} node to the j^{th} 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 [120]; 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 95^{th} 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.

3.2.4 Simulation Study

This section reports two 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. In the second simulation, after having extracted the best combination of learning rate and number of iterations of the gradient descent to be used in ANN-based estimation, we compare it with OLS estimation as regards the ability to retrieve the true values of the conditional GC and to reconstruct the assigned network topology. In both simulations, the topology is representative of the interaction of a ten-variate VAR process exhibiting a random interaction structure with fixed density of connected nodes [8, 34].

Simulation Design

Simulated multivariate time series (M=10) were generated as a realization of a VAR(16) model fed by zero-mean independent Gaussian noise with variance equal to 0.1. The simulated networks have a ground-truth structure with a density of connected nodes equal to 15%, where non-zero AR parameters of values chosen randomly in the interval [-0.8, 0.8] were set at lags assigned randomly in the range (1-16) [58]. The knowledge of the true AR parameters allows computing the theoretical values of the conditional GC and the true network topology, as illustrated for an exemplary case in Fig. 3.2. Simulations were generated for different values of the parameter K defined as the ratio between the number of data samples available ($N \times M$) and the number of AR coefficients to be estimated ($M^2 \times p$). One hundred networks were generated for each value of K in the range (1,3,10,20); the length of the simulated time series was $N = 160$ when $K = 1$ and $N = 3200$ when $K = 20$.

First, considering ANN estimation performed for each value assigned to K and for each realization, the learning rate η and the number of iterations for the SGD-l_1 during the training process were varied respectively in the range $(10^{-3}, 10^{-4}, 10^{-5})$ and in the range $(100, 1000, 2000)$. Importantly, for each network structure a different neural network was trained initializing the weights according to the method described in [167] that guarantees a faster convergence of the gradient descent algorithm. After training, the conditional GC between each pair of processes was estimated from the matrix of the weights $\mathbf{W}$ using the SS approach. Then, in order to assess which combination of learning rate - number of iterations of the gradient descent is the best for a regression problem, and to evaluate the differences between the NN and OLS methodologies, different measures of performances were computed as explained in the following subsection.

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 [59]:

$$bias = |F_{i\to j|s} - \widehat{F}_{i\to j|s}|. \tag{3.19}$$

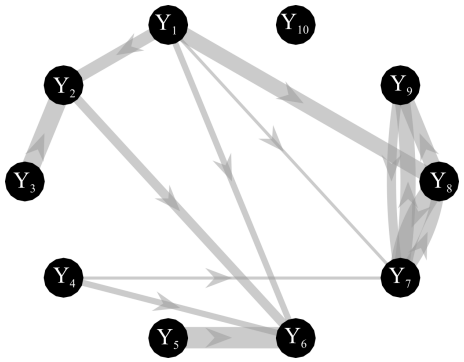

Figure 3.2: Graphical representation of one of the ground-truth networks of the simulation study. Arrows represent the causal links randomly assigned between two network nodes via nonzero VAR coefficients. The thickness of each arrow is proportional to the strength of the causal connection assessed by 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.

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 15 non-null links and across the 75 null links 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 (class 1) or absence (class 0) of a causal connection is estimated using surrogate data for OLS and looking at the presence/absence of non-zero weights for ANN, and is then compared with the underlying ground-truth structure. Performances were assessed through the computation of false-negative rate (FNR, measuring the fraction of non-null links with non-significant estimated GC), false-positive rate (FPR, measuring the fraction of null links with significant estimated GC) and Area Under Curve (AUC) that summarizes the information provided by FNR and FPR [20, 60]. These performance measures were computed across the network links for each assigned network, and the corresponding distribution across the 100 simulated network structures was then obtained separately for OLS and NN. In the case of ANNs, the computation time (in seconds) required for the training of the ANN for different values of learning rate, number of iterations of the gradient descent and data samples available was also considered as a performance parameter. The average computation times over the 100 realizations were calculated using an implementation of the algorithms in MATLAB® environment on a PC with a six cores Intel Xeon (CPU clock speed 3.7 GHz), 128· GB DDR4 RAM.

To establish which combination of learning rate and number of iterations of the gradient descent guarantees the most accurate results for each value of the K-ratio, an indicator of the overall performance (parameter S) was defined as the average of the two following performance parameters: i) the bias as defined in (3.19) for non-null links, normalized with respect to the theoretical GC value; ii) the complement to 1 of the AUC parameter, $1 - AUC$. These two parameters are both null in the case of perfect estimation, and increase when the estimated GC values deviate from the theoretical (non-zero) values or when the estimated network topology differs from the true

topology. Both parameters were averaged across values of the K-ratio, and then the S parameter was computed as their average.The distribution of S across the 100 realizations was investigated as a function of learning rate and number of iterations of SGD-l_1.

Statistical Analysis

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

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

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

Results of the Simulation Study I

The results of the three-way repeated-measures ANOVAs, expressed in terms of F-values and computed separately on all the performance parameters considering K, LR and N_{train} as main factors, are reported in Table 3.1.

Factors	**DoF**	**BIAS$_0$**	**BIAS$_1$**	**FNR**	**FPR**	**AUC**
$\mathbf{N}_{\text{train}}$	(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***
$\mathbf{N}_{\text{train}} \times \mathbf{LR}$	(4, 396)	110.4***	101***	279***	156***	97.2***
$\mathbf{N}_{\text{train}} \times \mathbf{K}$	(6, 594)	139.7***	2.6*	44***	98***	0.5
$\mathbf{LR} \times \mathbf{K}$	(6, 594)	200.9***	13***	47***	132***	2.5*
$\mathbf{N}_{\text{train}} \times \mathbf{LR} \times \mathbf{K}$	(12, 1188)	28.2***	71.6***	20***	15***	3.6***

Table 3.1: F-values and corresponding degrees of freedom (DoF) of the three-way repeated measures ANOVA.***, $p < 10^{-5}$;**, $10^{-5} < p < 0.01$; *, $0.01 < p < 0.05$.

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

Figure 3.3 reports the distribution of the parameters $BIAS_0$ and $BIAS_1$ according to the interaction $N_{\text{train}} \times LR \times K$. In the analysis of the error associated with the estimation of the

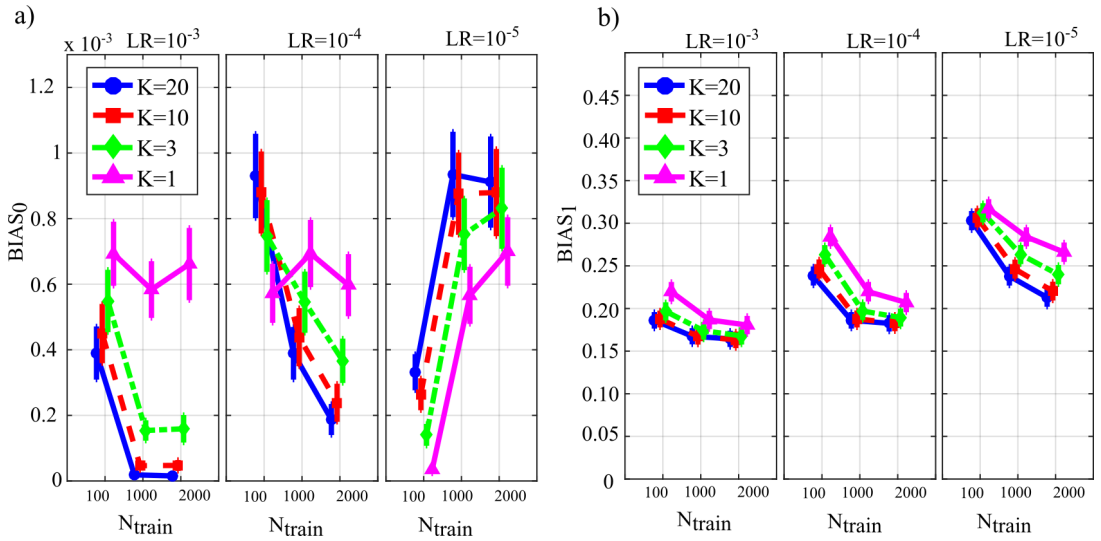

Figure 3.3: Distributions of the bias of conditional GC (value and 95% confidence interval across 100 simulated networks) estimated using ANNs for the first simulation study. Bias parameters computed for the null links ($BIAS_0$, panel a) and for the non-null links ($BIAS_1$, panel b) are plotted as 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.

conditional GC along the null links ($BIAS_0$, Fig. 3.3a)), 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 (Fig. 3.3b)) 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 3.4 reports the distributions of the parameters FNR, FPR and AUC according to the interaction $N_{\text{train}} \times LR \times K$. The portion of non-null directed links incorrectly classified as null (FNR, Fig. 3.4a)) was lower than 20% in all cases except for $N_{\text{train}} = 100$ 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 obseved 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, Fig. 3.4b)) 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 (Fig. 3.4c)) reached the highest values for $LR = 10^{-3}$ and $N_{\text{train}} \in \{1000, 2000\}$. 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 $K = 1$. The performance showed a tendency to degrade at decreasing K, increasing LR and decreasing N_{train}.

Table 3.2 reports the computation time required for the training of the neural network in different conditions of K ratio, learning rate and number of SGD-l_1 iterations averaged across the 100 realizations. As expected, the computation time increases with the number of iterations of the

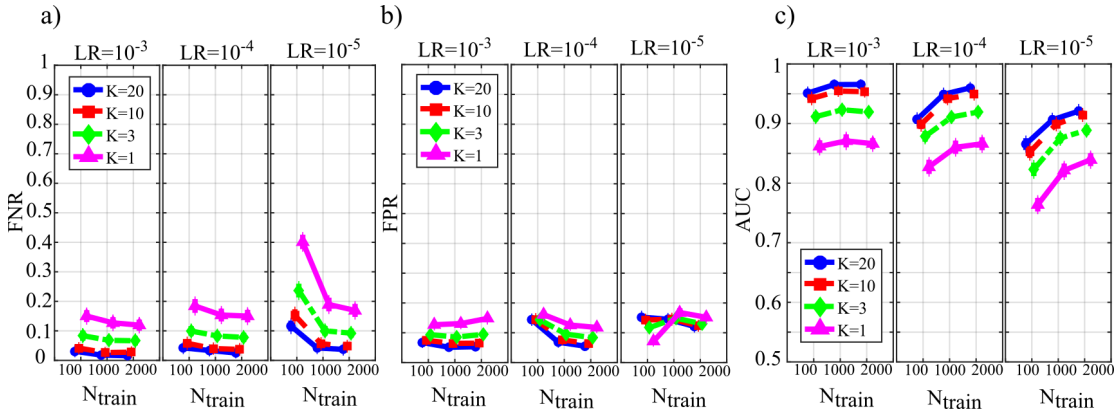

Figure 3.4: Distributions of the parameters assessing the quality of network reconstruction performed using ANNs for the first simulation study. Plots depict the distributions of FNR (a), FPR (b) and AUC (c) expressed as mean value and 95% confidence interval across 100 simulated networks as a function 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.

gradient descent and with the number of data samples available (K ratio). The least and most time-consuming settings were $N_{\text{train}} = 100, K = 1$ and $N_{\text{train}} = 2000, K = 20$, respectively taking ~ 2 secs and ~ 210 secs.

	$LR = 10^{-3}$			$LR = 10^{-4}$			$LR = 10^{-5}$		
$\mathbf{N}_{\text{train}}$	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 3.2: 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.

Figure 3.5 reports the distribution of the overall performance parameter S computed as a function of the learning rate for different number of iterations of SDG-l_1 (interaction $N_{\text{train}} \times LR$). The results show how the performance is affected significantly by both factors, with values of S that tend to decrease while increasing the learning rate and the number of iterations of the gradient descent. The lower values of S, indicating lowest bias of the estimated GC values and/or highest AUC in the classification of the network structure, were observed for $LR = 10^{-3}$ and $N_{\text{train}} = 1000$ or $N_{\text{train}} = 2000$. As the improvement from $N_{\text{train}} = 1000$ to $N_{\text{train}} = 2000$ was not statistically significant, we infer that the best setting is the least computationally onerous combination, i.e., $LR = 10^{-3}, N_{\text{train}} = 1000$.

Results of the Simulation Study II

After the extraction of the best combination of the training parameters of the ANN, in the second simulation study we compare the performance of OLS and ANN at varying the proportion between number of data samples available and parameters to be estimated (K-ratio). The results of the two-way repeated-measures ANOVAs, expressed in terms of F-values and computed separately on all the performance parameters considering K and $TYPE$ (i.e., the method used: OLS or ANN) as main factors, are reported in Table 3.3.

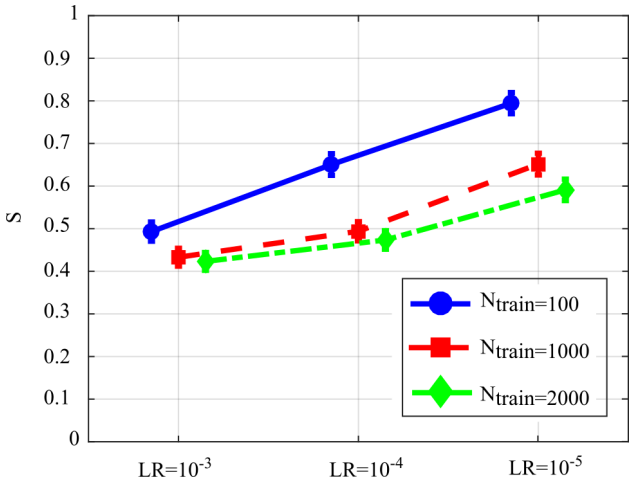

Figure 3.5: Distributions of S parameter considering the interaction factor $N_{\text{train}} \times LR$, expressed as mean value and 95% confidence interval of the parameter computed across 100 realizations of the first simulation study ($F(4, 396) = 128.09$, $p < 10^{-5}$).

Factors	**DoF**	**BIAS$_0$**	**BIAS$_1$**	**FNR**	**FPR**	**AUC**
TYPE	(1, 99)	3170***	882***	82.6***	153***	4.7***
K	(2, 198)	2310***	128***	362***	42.1***	472***
TYPE $\times$ K	(2,198)	2150***	75***	149***	31.6***	60.2***

Table 3.3: 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).***, $p < 10^{-5}$.

The two-way ANOVA analysis highlights a strong statistical influence of the main factor K and $TYPE$ and of their interaction ($TYPE \times K$) on all the performance parameters analyzed in this study. In this case the level $K = 1$ was not considered in the statistical comparison due to the non-convergence of the DARE equation for the OLS case.

Figure 3.6 reports the distribution of the parameters $BIAS_0$ and $BIAS_1$ according to the interaction factor $K \times TYPE$. The comparison of OLS (blue line) and ANN (red line) 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 (Fig. 3.6a)); 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.

Figure 3.7 reports the distributions of the parameters FNR, FPR and AUC according to the interaction $K \times TYPE$. 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 hand, the analysis of false positive detections (panel b) showed an error quite low and stable with K in the case of OLS, and an error slightly growing with K up to 15% in the case of ANN. The overall performance evaluated through AUC showed high classification accuracy and absence of statistically significant

differences between the two estimation methods for $K = [20, 10]$, and a better performance of ANN compared with OLS for lower values of K; a high AUC value ($\sim 85\%$) was reported for ANN even when K=1.

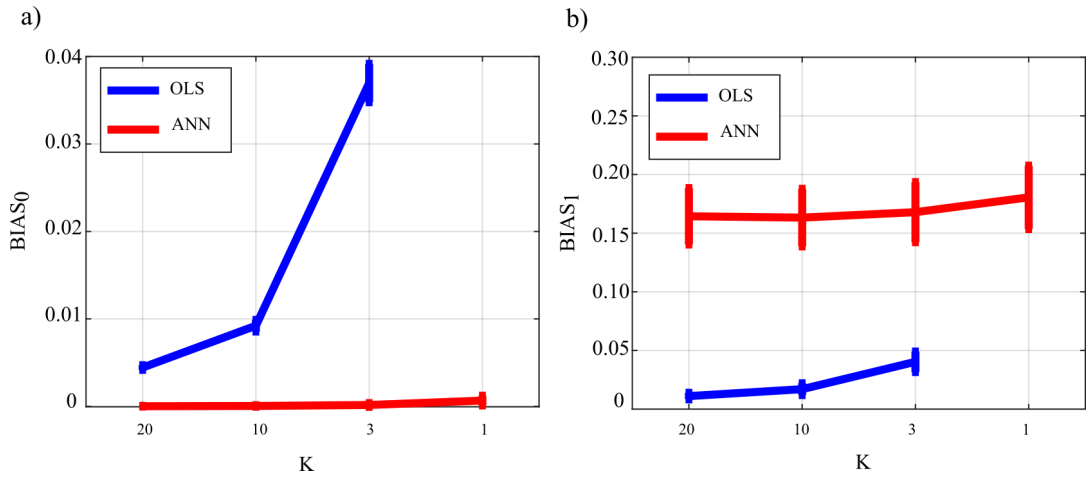

Figure 3.6: Distributions of the bias relevant to the estimation of GC on the null links ($BIAS_0$, panel a) and on the non-null links ($BIAS_1$, panel b) plotted as a function of the ratio between data samples available and number of parameters to be estimated (K), for OLS estimation (blue) and ANN estimation (red).

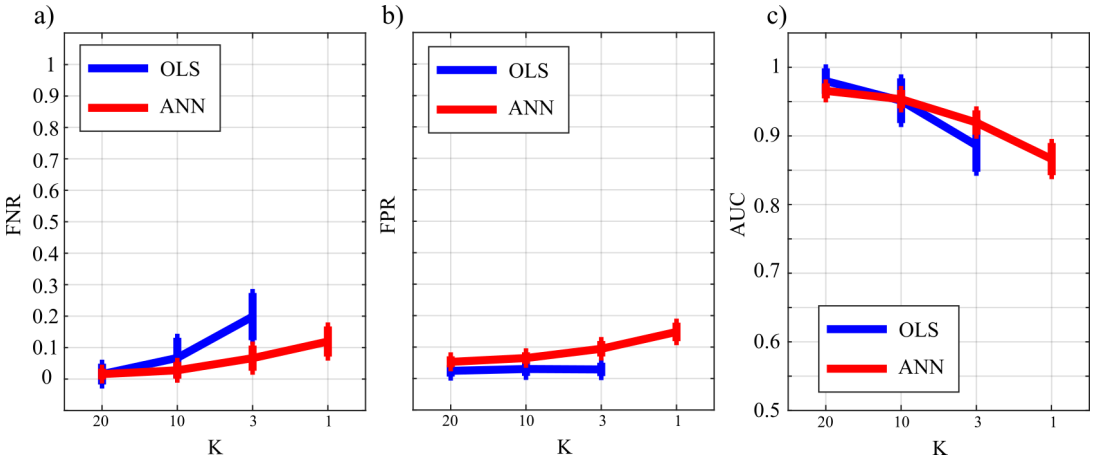

Figure 3.7: Distributions of the parameters assessing the performance of network reconstruction, i.e. the rate of false negatives (FNR, panel a) and of false positives (FPR, panel b) and of the area under the curve (AUC), plotted as a function of the ratio between data samples available and number of parameters to be estimated (K), for OLS estimation (blue) and ANN estimation (red).

Table 3.4 reports the computation time required for the entire process of GC computation using the two estimation approaches for different values of the K ratio. OLS analysis includes SS model identification and the subsequent evaluation of the null-case distribution for each couple of nodes as described in the Methods section. ANN analysis includes SS model identification plus the training process at $N_{\text{train}} = 1000, LR = 10^{-3}$. The analysis highlights the expected decrease of the computation times with decreasing the K ratio and, more importantly, a strong reduction of the time requested for the entire process when ANN is used in place of OLS. The computation time of OLS identification is not reported for $K = 1$ due to the non-convergence of the solution to the DARE equation necessary for SS model identification.

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 3.4: 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

3.3 Application to Physiological Time Series

This section reports the application of the conditional GC, defined as in equation (3.12) 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 [78].

3.3.1 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 [78]. 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 (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 [19, 168, 169].

To obtain synchronous timing of the seven measured time series, as all the brain time series resulted as sampled at 1 Hz, the cardiovascular time series were synchronously resampled at 1 Hz using spline interpolation. Stationary time series of 300 samples (5 min recordings), reduced to zero mean and unit variance, were considered for the analysis. The time series extracted from each subject 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, $\boldsymbol{A}_1, ..., \boldsymbol{A}_p, \boldsymbol{\Sigma}$, 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) [124].

3.3.2 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 vector with index s. Moreover, to confirm the results obtained in [8] on the same data, the in-strength - defined as the sum of all weighted inward links [125] - was computed for a specific network node (pulse arrival time π). The effect of the different experimental conditions on the in-strength evaluated for the π node was assessed through the Kruskal-Wallis test followed by the Wilcoxon rank-sum test between pairs of conditions. All analyses were performed with a model of dimension Mp, where $M = 7$ and $p \sim 4$ (depending on the BIC) on time series of 300 points, corresponding to $K \sim 10$ relating the amount of data sample available to the model dimension.

Results of Granger Causality Analysis

Figure 3.8 depicts the network of physiological interactions reconstructed through the detection of the statistically significant values of the conditional Granger causality ($F_{i\to j|s}$) computed for all pairs of processes belonging to the analyzed network. The weighted arrows represent the most active connections among the systems (arrows are present when at least three subjects show a statistically significant value of $F_{i\to j|s}$). To ease interpretation and comparison between OLS and ANN estimates, the three sub-networks representative of brain, body and brain-body interactions are depicted with arrows of different colors. The networks estimated using OLS in the three experimental conditions (Figure 3.8.a-c) exhibit similar structures to those estimated using ANN (Figure 3.8. (d-f)); the main difference is that networks estimated with ANN show greater sparsity than those estimated with OLS.

A qualitative analysis of the networks illustrates the existence of a highly connected body sub-network (red arrows), a weakly connected brain sub-network (purple arrows), and a pattern of brain-body interactions (green arrows) that changes with the experimental condition. The body

interactions are characterized, consistently across the three conditions, by cardiovascular links (interactions from η to π) and cardio-respiratory links (interactions between η and ρ), with a weaker coupling between ρ and π. The use of ANN reveals a preferential direction from ρ to π that is not present in the condition M and is bidirectional in the condition G. The topology of the brain subnetwork assessed by the ANN method is less stable across conditions, and looses consistency moving from R to G. On the contrary, in the OLS case, the topology seems to be more consistent exhibiting weaker connections moving from R to M and from M to G. The analysis of brain-body interactions reveals that such interactions are mostly directed from the brain to the body sub-networks; in this case, the use of ANN clearly shows an increasing of brain-body interactions during the condition G.

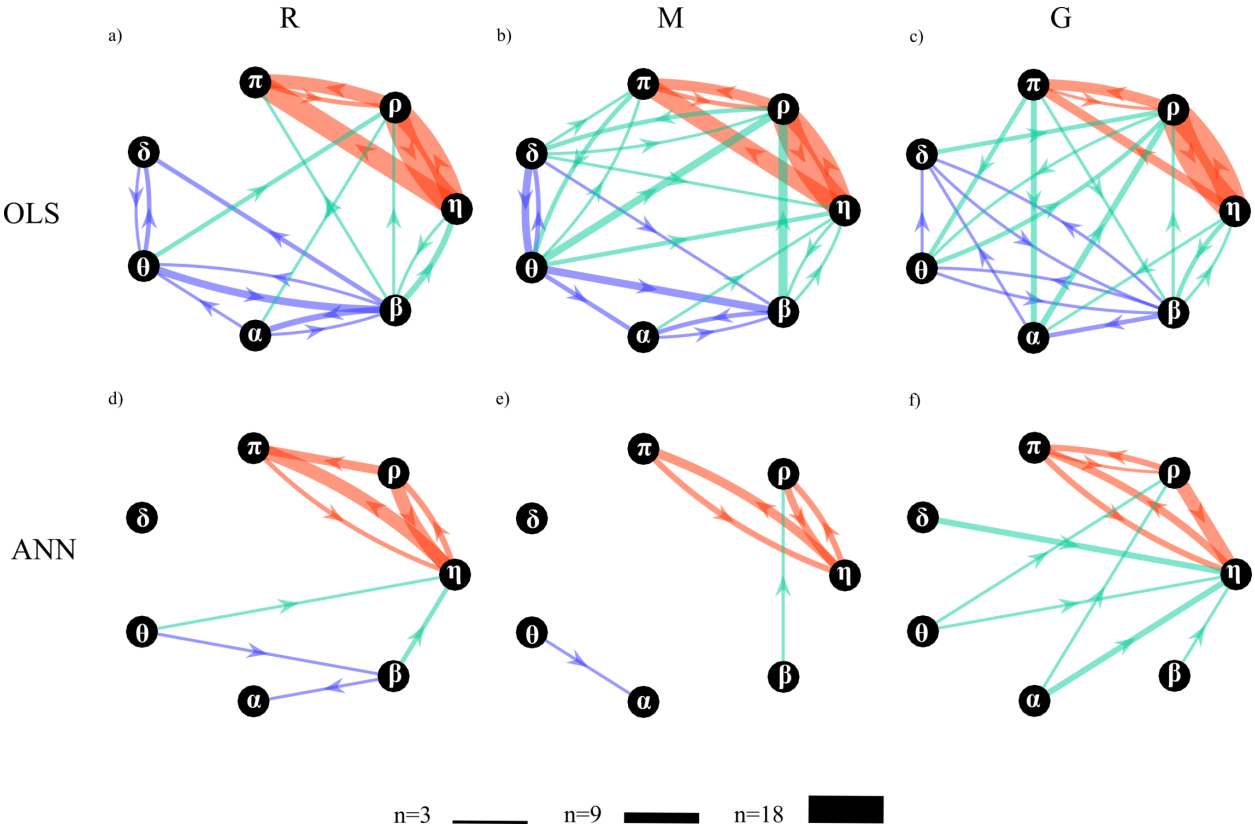

Figure 3.8: Topological structure of the network of physiological interactions reconstructed during the rest (R), mental arithmetic (M) and serious game (G) experimental conditions. Graphs depict significant directed interactions within the brain (purple arrows), body (red arrows) and brain-body (green arrows) sub-networks. Directed interactions were assessed counting the number of subjects for which the conditional Granger causality ($F_{i \to j|s}$) 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.

Figure 3.9 reports the distribution of the values of the in-strength index evaluated for the π node in each experimental condition. For both OLS and ANN, the median value of the in-strength index is significantly higher in the condition R with respect to the condition G. The use of ANN highlights lower values for the in-strength parameter even if the trend is the same moving across the three experimental conditions. These results show that both approaches detect a decrease of the information flow directed to the cardiovascular node of the body subnetwork, documented by the reduction of the in-strength index in the G condition for the process π.

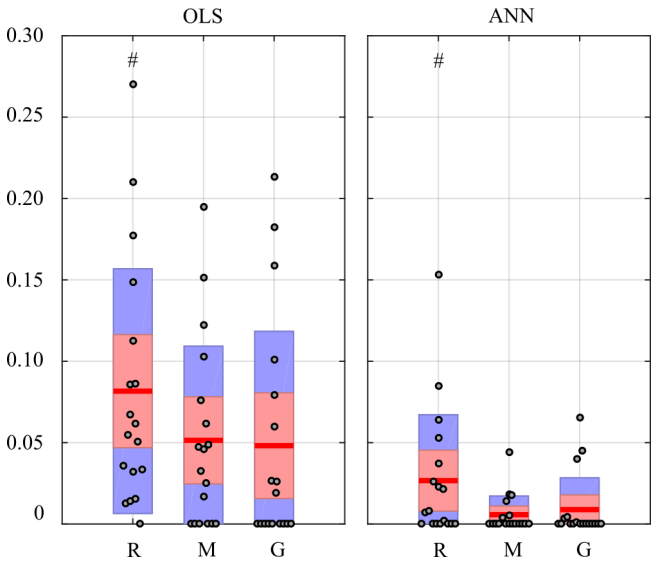

Figure 3.9: In-strength index computed for π node of the physiological network. Box plots report the distribution across subjects (median: red lines; interquartile range: box; $10^{th} - 90^{th}$ percentiles: blue bars) and the individual values (circles) of the in-strength computed at rest (R), during mental stress (M) and during serious game (G). Statistically significant differences between pairs of distributions are marked with # (R vs G).

3.4 Application to a ring of non-linear electronic oscillators

In this section we investigate the application of GC, in its unconditional version, computed through OLS and ANN by exploiting the SS approach, to a dataset of electronic non-linear chaotic oscillators, recorded from a unidirectionally-coupled ring of 32 dynamic units, previously realized with the aim of studying remote synchronization [162, 163]. In the literature, it has been pointed out that a single transistor oscillator can exhibit very complex activity and a ring of coupled oscillators can create a community structure with statisical properties resembling physiological systems [22, 170, 171]. The previous analysis has shown how it is possible to provide a mesoscopic description of the information exchanged between different nodes of a network which represents the activity of several physiological systems. On the other hand, the employment of an electronic circuit comprising a ring of oscillators, provides a system of reduced scale and complexity, with respect to a physiological one, yielding full access to the activity of each individual node. The resulting time series, measured as voltage output by each oscillator, were considered as input for a VAR model and for an ANN, descriptive of the behavior of the entire network ring.

3.4.1 System description and synchronization analysis

The structural diagram of the oscillator circuit corresponding to each node in the network is reported in Figure 3.10.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 $K_1 = 3.67K_2 = 0.11\mu s^{-1}$ 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 3.10.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 [172].

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 [163], 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 transitority, and finally vanishes.

As in previous works [162, 163], 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:

$$v_m(t) + i\hat{v}_m(t) = A_m(t)e^{i\phi(t)}, \tag{3.20}$$

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:

$$C_{ij}(\tau) = \frac{k_{ij}(\tau)}{\sqrt{\sigma_i^2 \sigma_j^2}}, \tag{3.21}$$

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 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 to the variances of Y_i and Y_j respectively.

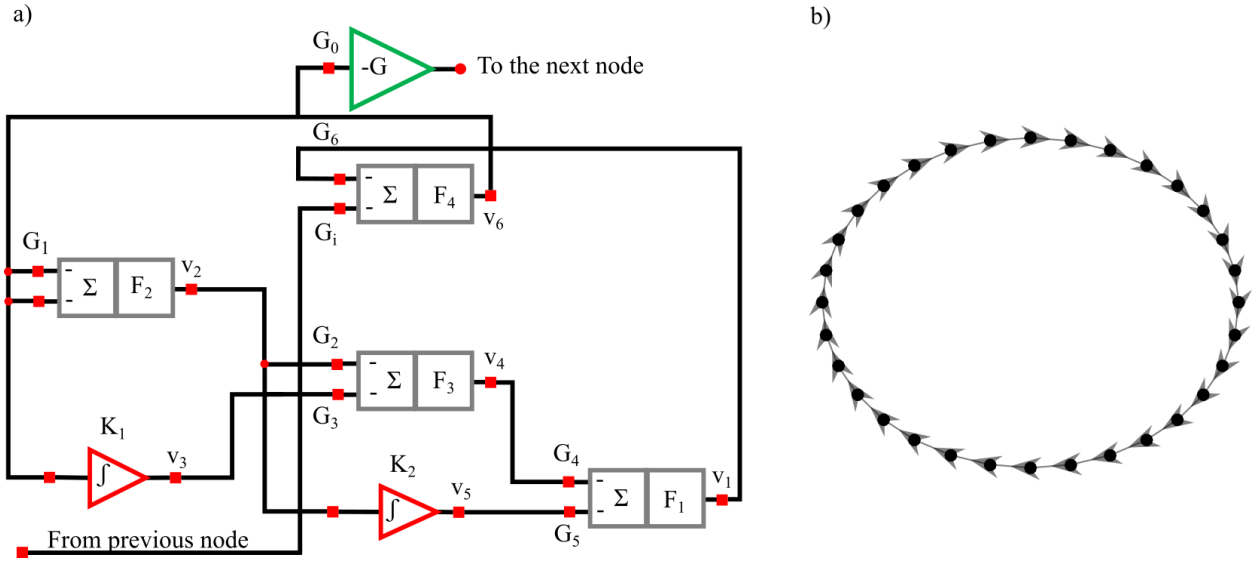

Figure 3.10: Diagram of the oscillator circuit corresponding to each node in the network (a). Master-Slave (unidirectional, clock-wise) structure of the ring comprising thirty two oscillators (b).

In Figure 3.11 the analysis of cross-correlation coefficient performed for each pair of oscillators (i, j) in the entire ring (panel a) is reported, alongside with the corresponding synchronization analysis for three representative oscillator pairs (panel b) which exemplify the decay and transient recovery of amplitude synchronization for three different distances from the node 1. The analysis of the cross-correlation coefficient reveals that moving away from a node, synchronization initial decayed, then gradually increased, rising till a distance $d \approx 8$, and eventually vanished as shown in Figure 3.11.a. The structural coupling on the ring is only between first neighbors, as indicated by the master-slave configuration, and the highlighted non-monotonic trend in the cross-correlation coefficient indicates a situation of remote synchronization. The visual inspection of signal envelope for three different couples of oscillators (panel b) confirms the analysis of cross-correlation with complete synchronization of the couple $i = 1$, $j = 2$ (distance 1, $\max[C_{ij(\tau)}]_{\tau \geq 0} = 0.91$) that becomes a desynchronization for the couple $i = 1$, $j = 6$ (distance 5, $\max[C_{ij(\tau)}]_{\tau \geq 0} = 0.19$); finally, the synchronization appears to be strong even for the couple $i = 1$, $j = 9$ that means a physical distance of eight ($\max[C_{ij(\tau)}]_{\tau \geq 0} = 0.59$). The performed analysis can be replicated by running the Matlab script Test_Oscillators in the released toolbox.

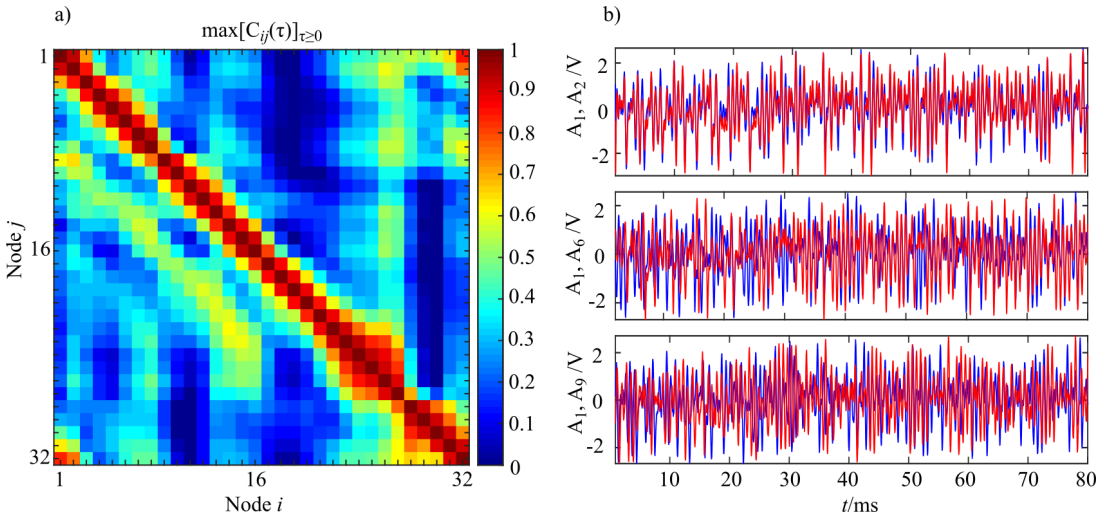

Figure 3.11: Instance of remote synchronization. The panel (a) reports the synchronization matrix for the entire ring intended as the maximum positive cross-correlation coefficient for the signal envelope $A_m(t)$. The panel (b) shows the signal envelope A_m for three different coupled of nodes demonstrating remote synchronization effects. The blue line represents A_1 with the red line that shows A_2 (on the top), A_6 (middle panel) and A_9 (on the bottom). Time series were realigned to the lag for which the maximum value of cross correlation was observed.

3.4.2 Granger Causality Analysis

From a theoretical point of view cross-correlation coefficient is a symmetric measure and thus, its value for each time step is the same independently of the selected direction ($i \to j$, $j \to i$). For this reason, it is not possible to assess if there is an information exchange between different oscillators. In order to test if there is information exchange between different oscillators, and if both methodologies can adequately capture the effects of "remote synchronization" restoring the results obtained in [163], 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 $Y^p_{j,n} = [Y_{j,n-1}, \cdots, Y_{j,n-p}]$, 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 [163]. 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 3.12 shows the results of the evaluation of unconditional GC ($F_{i \to j}$) performed for each couple (i, j) through OLS (Figure 3.12.a) and ANN (Figure 3.12.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 linear correlation coefficient of the ranks, namely,

$$r_s = \frac{\sum_i (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2} \sqrt{\sum_i (S_i - \overline{S})^2}} \tag{3.22}$$

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

$$t = r_s \sqrt{\frac{N-2}{1-r_s^2}}, \tag{3.23}$$

which is distributed approximately as Student's distribution with N-2 degrees of freedom [173]. The result of this analysis reveals a value of $r_s = 0.84$ with a p-value $p < 10^{-5}$ indicating a strong correspondence between the networks obtained through the two methodologies.

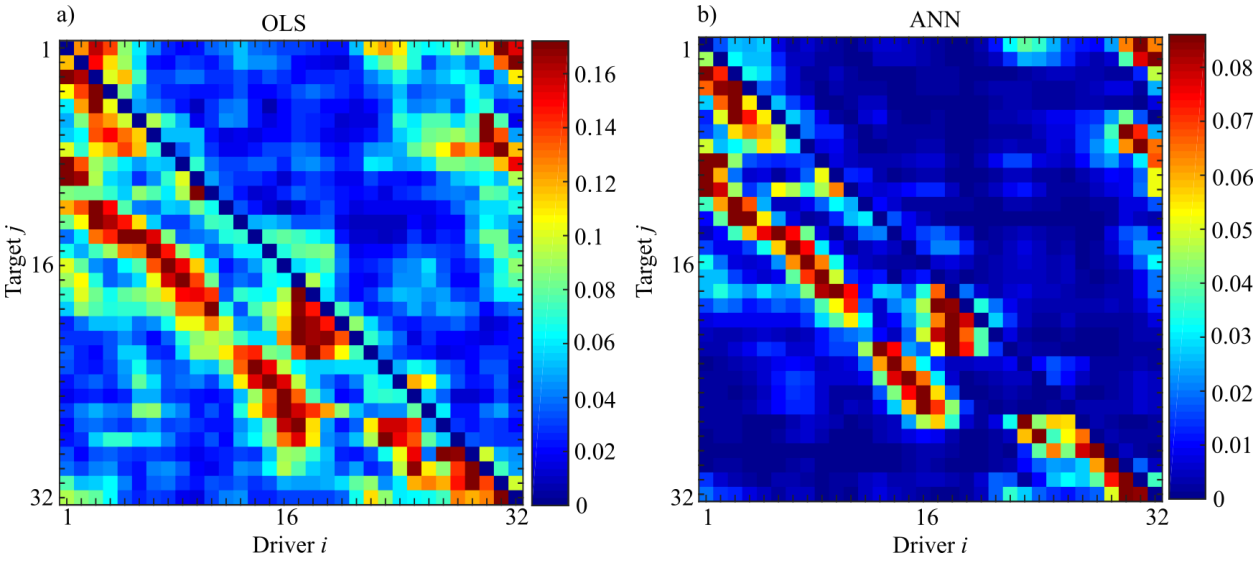

Figure 3.12: Unconditional Granger Causality Analysis performed on the network of 32 chaotic oscillators ($F_{i\to j}$). The matrices represent the analysis performed using OLS (panel a) and using ANNs (panel b) where each entry of the matrices corresponds to the strength of the causal influence from 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 ($p < 10^{-5}$).

3.5 Discussion

3.5.1 Simulation study I

The first simulation study was designed to evaluate the effects of ANN training parameters on the GC estimation process. We pointed out how the learning rate (LR) and the number of iterations (N_{train}) of the gradient descent have an impact on the training process as regards both the regression problem and the classification of significant network links [174]. The accuracy in the estimation of the regression parameters, which reflects the accuracy in the magnitude of the estimated GC, was investigated while varying the amount of data samples available for the estimation (Figure 3.3). As expected, the bias of GC estimated over both null and non-null links increased in conditions of data paucity, while it was reduced increasing the number of iterations of the gradient descent. An opposite trend was observed assessing the bias along the null links for small learning rate ($LR = 10^{-5}$). This result was previously observed in the context of classification analysis [175, 176] and is likely due to the fact that too small learning rates can trap the ANN training process into local minima, resulting in our case in larger differences between estimated and theoretical values of the conditional GC.

On the other hand, the analysis of the accuracy in reconstructing the network structure was tested in terms of different classification parameters previously used to assess the structure of connectivity networks [8, 48, 60]. The analysis (Figure 3.4) showed a general improvement of the classification performance when increasing the number of data samples available and the number of iterations of the SGD-l_1 algorithm, and when decreasing the learning rate. These results are in line with previous studies analyzing the performance of estimators related with the concept of Granger causality [34, 37, 177], and help to optimize the parameter selection for GC analysis based on ANN.

Such an optimization was performed in an objective way selecting the best combination of learning rate and number of SGD-l_1 iterations that minimized the overall performance parameter S (Figure 3.5; note that lower values of S indicate better performance). Varying the parameters N_{train}

and LR within ranges compatible with those suggested in a review of ANNs employed in classification analysis [69], we identified the combination $LR = 10^{-3}$ and $N_{\text{train}} = 1000$ as the most suitable for optimizing the performance of ANNs in the computation of magnitude and statistical significance of the conditional GC. Overall, our simulation results lead to the following recommendations for GC estimation based on ANNs:

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

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

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

3.5.2 Simulation study II

The second simulation study was designed to analyze the performance of the proposed ANN approach for GC estimation in comparison with the state-space analysis based on standard OLS estimation of the VAR model [102]. 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 (Figure 3.6). This trend was marked for OLS-based GC estimates, confirming previous comparative studies [36], 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 [8]. 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 [178].

Also the ability in reconstructing the network structure showed a tendency to decrease with the ratio K between the number of data samples and of model parameters (Figure 3.7). In terms of overall accuracy, the ANN approach outperformed the OLS one for $K \leq 3$ and resulted 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 [44, 127]. When particularized to the rate of correct detection of null and non-null links, the performance under conditions of data paucity 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 [20]. 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 [8, 39]. Note 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 [21]. 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 3.4).
- 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.

3.5.3 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 [18]. 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 [118]. As a main challenge is to interpret dense networks in terms of the underlying physiological mechanisms [149, 179], 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 [78]. 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 [8, 177]. In particular, the very similar network topologies observed here and in [8] 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 3.8). 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 [8, 78, 92]. These patterns comprise a strong information flow between η and ρ reflecting the mechanisms of respiratory sinus arrhythmia [135] and cardio-respiratory synchronization [136], 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 [180], and the causal interaction $\rho \rightarrow \pi$ reflecting the influences of breathing on the intra-thoracic pressure, blood pressure and blood flow velocity [137]. 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 3.9); 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 [8, 177]. 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 [77, 135].

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 [8]. 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 [44, 55]. 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 [78].

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 [141, 143].

3.5.4 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 3.11 revealed the existence of a preferential synchronization effect between groups of nodes that are not directly connected via a physical link and, in particular, we found a maximum of the cross-correlation coefficient at a distance $d \approx 8$. This result is in agreement with previous analyses performed in the same ring of oscillators [162, 163] and with the recently introduced concept of remote synchronization which reveals mutual synchronization between pairs of locally coupled groups of nodes in a network. Thus, each group of nodes remotely synchronized is physically connected through a group of intermediary nodes more weakly synchronized with them [161].

In order to investigate if the observed remote synchronization corresponds to "remote" information transfer, we performed unconditional GC analysis with both OLS and ANN. An inspection of Figure 3.12 clearly shows the good overlap between the networks estimated with the two methodologies; this result is supported quantitatively by the analysis of the Spearman rank correlation coefficient ($r_s = 0.84$, $p < 10^{-5}$). A similar analysis was performed on the same dataset by [163], who used uniform embedding to approximate the history of target and driver time series as $Y_{j,n}^{-} = [Y_{j,n-\delta}, Y_{j,n-\tau-\delta}, \cdots, Y_{j,n-p\tau-\delta}]$, $Y_{i,n}^{-} = [Y_{i,n-\delta-d}, Y_{i,n-\tau-\delta-d}, \cdots, Y_{i,n-p\tau-\delta-d}]$, where the additional time lag $\delta = 0.01$ ms was added to ensure the full elimination of information storage [181] and the lag d was introduced to account for propagation delays and was set searching for the minimal prediction error over the range $d \in [0, 2]$.

Here, we confirm the results obtained in [163] with a different analysis that exploits the SS representation of the VAR model and the ANN training. In particular, both methodologies can capture the dynamical activity in a ring of electronic oscillators with a well-defined complexity and stability of the network topology, since it is possible to obtain structures overlapped with those extracted performing the analysis with different methodologies already reported in the literature. From a methodological point of view, the strong overlap between the two networks can be motivated by the results of the simulation study II for which at $K = 3$ the AUC parameter, indicating the capability in the reconstruction of the network topology, showed a very small difference between the two methods. Furthermore, it is also important to note that, as an effect of the l_1-norm applied to the weights of the network during the training process, the maximum value of GC estimated with ANN is one order of magnitude less for ANN than OLS [164].

3.6 Conclusions

This work documented that neural networks can be used in combination with state-space models for the identification of linear parametric models, allowing computationally reliable and accurate estimation of GC in its conditional and unconditional forms. In particular, we showed how this combined approach leads to overcoming both the decrease in accuracy reported for traditional least-squares identification when it needs to be performed in unfavorable conditions of data availability [36], and the problems arising in the computation of GC estimated through different regression problems [139]. ANNs are useful in particular to assess the statistical significance of GC estimates, favoring the reconstruction of the network topology underlying the observed dataset without the

need to employ time-consuming asymptotic or empirical procedures for significance assessment.

The implementation of the proposed approach for the study of physiological networks and coupled electronic oscillators documented its usefulness in practical applications, supported by the observation of interaction patterns similar to those found in previous studies where the datasets were first studied in terms of GC [78, 163]. All the findings in this work suggest that ANNs are able to detect the strongest interactions providing output patterns of information dynamics which are more straightforward and easy to interpret than those obtained with OLS.

Chapter 4

Testing different methodologies for Granger Causality estimation in Multivariate Time series: a comparative study

4.1 Introduction

The evaluation of the direction and and strength of the interactions, among simultaneously observed dynamical systems, is an important topic currently under investigation in many field of science such as understanding the functionality of brain [26, 182] inferring about the causal relationship among different financial markets indices [183, 184] and the seismic activity [185].

Granger Causality (GC) is a very versatile tool for analysing the cause-effect relationships between different time series descriptive of the system dynamics [28]. In principle, GC is inferred utilizing linear bivariate regression stating that given a time series X that causes another time series Y, the causal information is stored in the past of X which helps the prediction of the present state of Y. In multivariate time series, bivariate causality measures may capture indirect effect between X and Y driving by another time series Z which is not included in the bivariate formulation, e.g. if there is a direct causal effect $X \rightarrow Z$ and $Z \rightarrow Y$, the indirect causal effect $X \rightarrow Y$ may arises [29, 165]. Thus, bivariate analysis cannot distinguish between direct and indirect causal effects and when applied to multivariate systems could results in erroneous estimated patterns. Therefore, to account for the influence of other time series the bivariate formulation has been extended to the multivariate case through the use of Vector Autoregressive models (VAR) which lead to the computation of a conditional form of GC [117]. In the latter case, the direct causality is explored between any given pair of time series of the analyzed system by conditioning on the information conveyed from the past of the remaining time series in the dynamical system.

Although multivariate analysis incorporates dimensional contributions of high orders, the "curse of dimensionality" may arise, meaning that the estimation of GC in a multivariate fashion may be problematic due to the high dimension of the observed data, leading to unreliable estimation problems [186]. Another important topic that would influence the estimation process is the presence of redundancy between time series: given A and B as two different set of processes, if we are

interested in studying the causal relationship $B \rightarrow A$, redundancy occurs if the same information being shared by the processes in B and the its presence can lead to under-estimation of their causality when a standard multivariate approach is applied [186, 187].When analyzing interactions in a multivariate time series, redundancy may arise if some processes are influenced by another one not included in the regression problem or the existence of a certain degree of synchronization between stochastic processes [101].

The bias in the estimation process due to over-fitting or redundancy has been addressed by incorporating dimension reduction techniques such as LASSO regression [8, 188, 189], Artificial Neural Networks (ANNs) trained with specific algorithm [158] or by developing techniques based on subset regression, model reduction and non-uniform embedding [190, 191]. Variable selection techniques have also been used in order to limit the dimensionality within the estimation procedure of multivariate causality measures such as Group LASSO or others regression methods based on convex optimization [20, 39]. Beside these methods, recently a partial conditioning scheme has been also introduced for the computation of GC to cope with the effects of redundant variables and data paucity that could be present as a result of the high number of processes in the dynamical system under analysis and the way in which interact each other [106].

GC from a driver to a target time series is typically quantified, in the multivariate setting, by comparing the prediction error variance obtained from two different linear regression models: a model where the present state of the target is regressed on the past samples of all time series in the dataset (full model), and a model where the present of the target is regresses on the past of all time series excluding the driver (restricted model)[7]. Unfortunately, since the restricted model is theoretically of infinite order, the finite sample estimation of GC based on performing two separate regressions leads to a strong bias or very large variability of the estimate, depending on whether small or high orders are used to identify the restricted model [139, 192]. To counteract this problem, approaches based on the identification of a single VAR model have been recently introduced to estimate the GC. Finite order linear state space (SS) models is closed under the operations of aggregation or sub-sampling and the conditional GC in both time and frequency domains, can be easily derived from SS parameters via the solution of a single discrete algebraic Riccati equation (DARE) [17, 102].

In the literature there are different comparative studies aiming at evaluating and comparing different causality measures. However, each study focuses on different type of measure or specific feature of the analyzed dataset. Many different causality measures are compared with the standard GC measures, or transfer entropy [86] (which represent the non-linear analogue of Granger causality)[30, 193]. Other measures have been developed and presented, especially in the context of brain study [4, 5], and there are many comparative studies on electroencephalograms [194–198]. Although different studies have been performed in this context, in the literature an extensive comparative studies in different simulation settings is still missing. Furthermore, not all the GC measures introduced so far, have been developed in the context of State Space modelling.

In this Chapter, I performed an extensive comparative study in which five different strategies for the computation of Granger causality are presented as originally defined in the related works. Furthermore, all of them will be then embedded in the state space representation in order to overcome the limitations of existing GC estimation methods. The performance in the quantification of GC magnitude and statistical significance are compared in two different simulation studies by

known coupled systems with different network size and density of connected nodes also in condition of strong data paucity.

4.2 Methods

4.2.1 Granger Causality

Let us consider a discrete-time, stationary vector stochastic process composed of M real-valued zero-mean scalar processes, $\mathbf{Y}_n = [Y_{1,n} \dots Y_{M,n}]^T$. Assuming that $\mathbf{Y}_n$ is a Markov process of order p, in the linear signal processing framework it can be completely described by the VAR model:

$$\mathbf{Y}_n = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{Y}_{n-k} + \mathbf{U}_n, \tag{4.1}$$

where $\mathbf{A}_k$ is an $M \times M$ matrix containing the autoregressive (AR) coefficients $a_{ji,k}$ that relate $Y_{j,n}$ to $Y_{i,n-k}$ $(i, j \in (1, \dots, M), k \in (1, \dots, p))$, and $\mathbf{U}_n = [U_{1,n} \dots U_{M,n}]^T$ is a vector of M zero-mean gaussian innovation processes with covariance matrix $\mathbf{\Sigma} \equiv \mathbb{E}[\mathbf{U}_n \mathbf{U}_n^T]$ (where $\mathbb{E}$ is the expectation value). The problem 4.1 can be solved by means of ordinary least squares (OLS), computing the matrix of coefficients that minimizes the residual error term [112].

Let us assume the process $Y_{j,n}$ as the *target* and the process $Y_{i,n}$ as the *driver* process, with the remaining $M-2$ processes collected in the vector $\mathbf{Y}_{k,n}$, where $k = \{1, .., M\} \setminus \{i, j\}$. Then, denoting $\mathbf{Y}_{m,n}^- = [Y_{m,n-1} Y_{m,n-2} \dots]^T$ as the past history of the generic process Y_m, we state that the i^{th} process G-causes the j^{th} process (conditional on the other k processes), if $Y_{i,n}^-$ conveys information about $Y_{j,n}$ above and beyond all information contained in $Y_{j,n}^-$ and $\mathbf{Y}_{k,n}^-$. This definition leads to perform a regression of the present of the target on the past of all processes, yielding to the prediction error $e_{j|ijk,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|\mathbf{Y}_n^-]$, and on the past of all processes except the driver, yielding to the prediction error $e_{j|jk,n} = Y_{j,n} - \mathbb{E}[Y_{j,n}|Y_{j,n}^-, \mathbf{Y}_{k,n}^-]$. The prediction error variances resulting from these "full" and "restricted" regressions, $\lambda_{j|ijk} = \mathbb{E}[e_{j|ijk,n}^2]$ and $\lambda_{j|jk} = \mathbb{E}[e_{j|jk,n}^2]$ are then combined to obtain the definition of GC from Y_i to Y_j [103]:

$$F_{i \to j} = ln \frac{\lambda_{j|jk}}{\lambda_{j|ijk}}. \tag{4.2}$$

4.2.2 Computation of Granger causality

In this section I describe five different approaches for the computation of $F_{i \to j}$ as defined in 4.2. For all of them given that $e_{j|ijk,n} = U_{j,n}$ the error variance of the full regression can be obtained as the j^{th} diagonal element of the error covariance matrix, $\lambda_{j|ijk} = \mathbf{\Sigma}(j, j)$, that can be retrieved by solving the full and the restricted regression problems with the five different procedures named: Ordinary Least Square (OLS) regression, modified Backward in Time Selection (mBTS) based on time ordered restricted VAR models, Partial Conditioning (PC), Artificial Neural Networks (ANNs) and LASSO regression. All the methods introduced can be used for the solution of both full and restricted regressions as described in the following or by using the theory of the state-space models. Note that the solution of the different regression problems require that the whole past history of $\mathbf{Y}$ can be truncated using p time steps, i.e., using the Mp-dimensional vector $\mathbf{Y}_n^p$ such that $\mathbf{Y}_n^- \approx \mathbf{Y}_n^p = [\mathbf{Y}_{n-1} \cdots \mathbf{Y}_{n-p}]$.

GC based on OLS (DR-GC) [29]

By considering N samples of the multivariate process $\mathbf{Y}$ the prediction error variance of the full linear regression problem can be computed from the identification of the model 4.1 from the parameters $(\mathbf{A}_1, \ldots, \mathbf{A}_p, \Sigma)$ obtained by the closed form solution of OLS [8]. By considering a single time step n the restricted model is instead formulated as follows:

$$Y_{j,n} = \sum_{m=1, m\neq i}^{M} \sum_{k=1}^{p} c_{mk} Y_{m,n-k} + e_{j|jk,n}, \tag{4.3}$$

where c_{mk} represents the $(M-1)p$ vector of coefficients with $e_{j|jk,n}$ residuals of the restricted regression problem. Then, the desired error variance $\lambda_{j|jk}$ is computed simply as the variance of the estimated residuals by solving the problem 4.3 through OLS regression.

GC based on mBTS (DR-mBTS) [107]

The modified backward-in-time-selection (mBTS) method is a bottom-up strategy designed for the analysis of multivariate time series to reduce the terms in the VAR model 4.1. The rationale is to evaluate progressively the inclusion of the lagged variables stored in $\mathbf{Y}_{n-k}$ starting with the most current and moving backward in time. First, for each of the M processes stored in $\mathbf{Y}$ is created the vector of all lagged variables that are candidate to be included in the model:

$$\mathbf{W} = [Y_{1,n-1}, \ldots, Y_{1,n-p}, Y_{2,n-1}, \ldots, Y_{M,n-p}], \tag{4.4}$$

and has Mp components. The algorithm aims at finding an explanatory vector $\mathbf{w}_j$ formed from the most significant lagged variables of $\mathbf{W}$ in predicting the current state of the target process $Y_{j,n}$. The explanatory vector $\mathbf{w}_j$ is built progressively adding one lagged variable at each cycle. The models formed by the candidate explanatory vectors at each cycle are assessed with the Bayesian Information Criterion (BIC) [124]. Upon termination after P_j cycles, the algorithm gives the final explanatory vector $\mathbf{w}_j$ of size P_j for Y_j. It is noted that $\mathbf{w}_j$ may not have lagged components of all M processes. For the computation of $F_{i\to j}$ we consider the following representation for the explanatory vector $\mathbf{w}_j$:

$$\mathbf{w}_j = [\mathbf{w}_{j,1}, \mathbf{w}_{j,2}, \ldots, \mathbf{w}_{j,M}], \tag{4.5}$$

meaning that $\mathbf{w}_j$ is decomposed to vectors of lags from each process Y_m. For each $m = 1, \ldots, M$, $\mathbf{w}_{j,m}$ is:

$$\mathbf{w}_{j,m} = \{Y_{m,n-\tau(1)}, \ldots, Y_{m,n-\tau(p_m)}\}, \tag{4.6}$$

and $\tau(l)(l = 1, \ldots, p_m)$, denote the p_m selected lags variable Y_m by mBTS. The length of $\mathbf{w}_j$ is $P_j = \sum_{m=1}^{M} p_m$. With this notation, the full model for the target Y_j becomes:

$$Y_{j,n} = \sum_{m=1}^{M} \mathbf{c}_{j,m} \mathbf{w}_{j,m}^T + e_{j|\mathbf{w}_{j,n}}, \tag{4.7}$$

where $\mathbf{c}_{j,m}$ is a row vector of p_m coefficients and the ^{T} sets $\mathbf{w}_{j,m}$ in column form. $e_{j|\mathbf{w}_j,n}$ represents the prediction error resulting from the regression of the present of the target $Y_{j,n}$ on the time lagged

variables selected by mBTS for each process. The entire procedure is then repeated by excluding from the M processes the driver process Y_i. As result, the restricted model can be written as:

$$Y_{j,n} = \sum_{m=1, m\neq i}^{M} \tilde{\mathbf{c}}_{j,m} \mathbf{w}_{j,m}^{\prime T} + e_{j|\mathbf{w}^{\prime}_{j,n}}, \tag{4.8}$$

where $\tilde{\mathbf{c}}_{j,m}$ is a row vector of coefficients as $\mathbf{c}_{j,m}$ and $e_{j|\mathbf{w}^{\prime}_{j,n}}$ is the prediction error resulting from the regression of the present of the target $Y_{j,n}$ on the time lagged variables selected by mBTS for each process except those belonging to the driver Y_i. The two models are then fitted with OLS and the two desired error variances $\lambda_{j|ijk}$, $\lambda_{j|jk}$ are then obtained as $\mathbb{E}[e^2_{j|\mathbf{w}_{j,n}}]$ and $\mathbb{E}[e^2_{j|\mathbf{w}^{\prime}_{j,n}}]$

GC based on Partial Conditioning (DR-PCGC) [106]

Partial conditioning is an approach whereby GC is computed, for any assigned pair of driver and target processes, including only a subset of the M observed processes in the VAR representation. Specifically, to compute GC from the i^{th} to the j^{th} process in $\mathbf{Y}$, a VAR model in the form of 4.1 is identified starting from a vector process which comprises Y_i, Y_j and the n_d processes which are deemed as most informative for the driver Y_i. Such other processes, which are collected in the vector Y_c, are identified maximizing the Mutual Information (MuI) between $Y^-_{i,n}$ and $Y^-_{c,n}$ [106]. Given two multivariate variables $\mathbf{V}$ and $\mathbf{W}$ the variance of a linear regression of $\mathbf{V}$ on $\mathbf{W}$ is given by [6, 118]:

$$\boldsymbol{\Sigma}(\mathbf{V}|\mathbf{W}) = \boldsymbol{\Sigma}(\mathbf{V}) - \boldsymbol{\Sigma}(\mathbf{V};\mathbf{W})\boldsymbol{\Sigma}(\mathbf{W})^{-1}\boldsymbol{\Sigma}(\mathbf{V};\mathbf{W})^T, \tag{4.9}$$

where $\boldsymbol{\Sigma}(\cdot)$,$\boldsymbol{\Sigma}(\cdot;\cdot)$ denote respectively covariance matrix and cross-covariance matrix. Mutual information between the two variables $\mathbf{V}$ and $\mathbf{W}$ is defined as the difference between two entropy terms as follows:

$$I(\mathbf{V};\mathbf{W}) = H(\mathbf{V}) - H(\mathbf{V}|\mathbf{W}), \tag{4.10}$$

where $H(\mathbf{V})$ represents the Shannon's entropy of $\mathbf{V}$ and the term $H(\mathbf{V}|\mathbf{W})$ represents the conditional entropy of $\mathbf{V}$ given $\mathbf{W}$. Under Gaussian assumption, for a multivariate random variable we have the well-known expression [199]:

$$H(\mathbf{V}) = \frac{1}{2}\ln\left(|\boldsymbol{\Sigma}(\mathbf{V})|\right) + \frac{1}{2}n\ln\left(2\pi e\right) \tag{4.11}$$

with n which represents the dimension of $\mathbf{V}$. In the same way, the conditional entropy $H(\mathbf{V}|\mathbf{W})$ for two jointly multivariate Gaussian variables can be expressed in terms of determinant of the corresponding partial covariance matrix:

$$H(\mathbf{V}|\mathbf{W}) = \frac{1}{2}\ln\left(\boldsymbol{\Sigma}(\mathbf{V}|\mathbf{W})\right) + \frac{1}{2}n\ln 2\pi e \tag{4.12}$$

To see this, we have:

$$H(\mathbf{V}|\mathbf{W}) = H(\mathbf{V},\mathbf{W}) - H(\mathbf{W}) = \frac{1}{2}\ln\left(|\boldsymbol{\Sigma}(\mathbf{V},\mathbf{W})|\right) - \frac{1}{2}\ln\left(|\boldsymbol{\Sigma}|\right) + \frac{1}{2}n\ln 2\pi e \tag{4.13}$$

where:

$$\mathbf{\Sigma}(\mathbf{V},\mathbf{W}) = \begin{pmatrix} \mathbf{\Sigma}(\mathbf{V}) & \mathbf{\Sigma}(\mathbf{V};\mathbf{W}) \\ \mathbf{\Sigma}(\mathbf{V};\mathbf{W})^T & \mathbf{\Sigma}(\mathbf{W}) \end{pmatrix} \tag{4.14}$$

and with the block determinant identity [200]

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D||A - BD^{-1}C| \tag{4.15}$$

it is possible to obtain:

$$|\mathbf{\Sigma}(\mathbf{V},\mathbf{W})| = |\mathbf{\Sigma}(\mathbf{W})| \cdot |\mathbf{\Sigma}(\mathbf{V}|\mathbf{W})|. \tag{4.16}$$

Substituting $\mathbf{V}$ and $\mathbf{W}$ as $Y^-_{i,n}$ and $Y^-_{c,n}$ mutual information becomes:

$$I(Y^-_{i,n}; Y^-_{c,n}) = \frac{1}{2} \ln \left(\frac{|\mathbf{\Sigma}(Y^-_{i,n})|}{|\mathbf{\Sigma}(Y^-_{i,n}|Y^-_{c,n})|} \right). \tag{4.17}$$

The number of conditioning processes is selected automatically finding the knee of the curve that measures the MI values as a function of the number of processes in Y_c (up to $M-1$). Then, a double regression limited to the subset of identified processes is performed defining the full and restricted regression models:

$$Y_{j,n} = \sum_{m=1}^{n_d+2} \sum_{k=1}^{p} c_{mk} Y_{m,n-k} + e_{j|ijc,n}, \tag{4.18}$$

$$Y_{j,n} = \sum_{m=1, m\neq i}^{n_d+2} \sum_{k=1}^{p} \tilde{c}_{mk} Y_{m,n-k} + e_{j|jc,n}. \tag{4.19}$$

After the identification of these models through the OLS estimator, the variance of the residuals $\lambda_{j|ijc} = \mathbb{E}[e^2_{j|ijc,n}]$ and $\lambda_{j|jc} = \mathbb{E}[e^2_{j|jc,n}]$ are estimated and used to compute the GC measure

$$F^C_{i\to j} = ln \frac{\lambda_{j|jc}}{\lambda_{j|ijc}}. \tag{4.20}$$

GC based on LASSO regression (DR-LASSO)[44]

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 i^{th} row is a realization of $\mathbf{Y}_i$, i.e., $\mathbf{y}_i = [y_{1,i} ... y_{M,i}], i = 1, ..., N$, and the j^{th} 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 Square (OLS) identification finds an optimal solution for the problem (4.1) by solving the following linear quadratic problem:

$$\widehat{\mathbf{A}} = argmin_{\mathbf{A}} ||\mathbf{y} - \mathbf{y}^p \mathbf{A}||_2^2, \tag{4.21}$$

where $\mathbf{y} = [\mathbf{y}_{p+1}; \cdots; \mathbf{y}_N]$ is the $(N-p) \times M$ matrix of the predicted values, $\mathbf{y}^p = [\mathbf{y}^p_{p+1}; \cdots; \mathbf{y}^p_N]$ is the $(N-p) \times Mp$ matrix of the regressors and $\mathbf{A} = [\mathbf{A}_1; \cdots; \mathbf{A}_p]$ is the $Mp \times M$ coefficient matrix. The 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 is minimized (RSS) [104, 112]. With sparse regression analysis, it is possible to solve the full regression problem stated in Equation (4.1) through the Least Absolute Shrinkage and

Selection Operator (LASSO) which introduces a constraint in the linear quadratic problem (4.21) [108]:

$$\widehat{\mathbf{A}} = argmin_{\mathbf{A}}(||\mathbf{y} - \mathbf{y}^p\mathbf{A}||_2^2 + \lambda||\mathbf{A}||_1). \tag{4.22}$$

In Equation (4.22), the additional term based on the l_1 norm forces a sparse a solution such that some of the VAR coefficients are shrunk to zero, with the shrinkage parameter λ controlling the trade-off between the number of non-zero coefficients selected in the matrix $\widehat{\mathbf{A}}$ and the residual sum of squares (RSS). Even if the problem (4.22) admits a solution, it will not be in a closed form since the l_1 norm is not differentiable at zero [44]. The optimal value of λ for the solution of the problem (4.22) requires a cross-validation approach for its determination (the procedure here used it is fully described in [8]). The full model 4.1 is identified with the aforementioned procedure and the variance of the residuals $\lambda_{j|ijk}$ is obtained as the j^{th} element of the error covariance matrix $\mathbf{\Sigma}(j,j)$. The restricted model formulated as in 4.3 can be identified by solving the following linear quadratic problem:

$$\widehat{\mathbf{A}_r} = argmin_{\mathbf{A}_r}(||\mathbf{y}_r - \mathbf{y}_r^p\mathbf{A}_r||_2^2 + \lambda||\mathbf{A}_r||_1). \tag{4.23}$$

in which $\mathbf{A}_r$ represents the $(N-p)\ \times\ (M-1)p$ coefficient matrix with $\mathbf{y}_r$ and $\mathbf{y}_r^p$ representing the matrices of predicted values and the matrix of regressors for which the driver Y_i is not considered. As discussed in [39] the optimal value for λ used for the identification of the full model is used also for the restricted model. After solving the problem 4.23 the variance of the residuals $\lambda_{j|jk} = \mathbb{E}[e_{j|jk,n}^2]$ is obtained from the residuals covariance matrix of the restricted model as i^{th} diagonal element [7].

GC based on Artificial Neural Networks (DR-ANN) [44]

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

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \frac{1}{N}\sum_{i=1}^{N} l(y_i, f(\mathbf{w};\mathbf{x}_i)) + \lambda r(\mathbf{w}), \tag{4.24}$$

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 $\lambda \in \Re^+$. 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:

$$r(\mathbf{w}) = ||\mathbf{w}||_1 = \sum_{k=1}^{Q} |\mathbf{w}_k|. \tag{4.25}$$

Then, a possible way to solve the problem (4.24) 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 following the procedure fully described in [20, 21].

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

$$\mathbf{w}^{j+1} = \mathbf{w}^{j} + \eta_j \frac{\partial}{\partial \mathbf{w}} \left(l(y_i, f(\mathbf{w}; \mathbf{x}_i)) - \frac{\lambda}{N} \sum_{k=1}^{Q} |\mathbf{w}_k| \right), \tag{4.26}$$

where j is the iteration counter and η_j is the learning rate at each iteration. The difficulty with l_1 regularization is that the last term on the right-hand side in (4.26) is not differentiable when the weight is zero. To solve this issue, following the procedure introduced in [21] l_1 regularization with cumulative penalty is applied directly on the weights of the network during the training process. Generalizing the whole procedure to a network with multiple outputs, in the linear signal processing framework the optimization problem (4.24) can be solved by using a linear function $f(\cdot;\cdot)$ linking the input layer with the output layer. The structure of the neural network comprises one input layer with 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:

$$l(\mathbf{y}, \mathbf{y}^p\mathbf{W}) = ||\mathbf{y} - \mathbf{y}^p\mathbf{W}||_2^2, \tag{4.27}$$

which highlights that the weight $\mathbf{W}$ corresponds to the matrix $\mathbf{A}$ containing the parameters of the VAR model (4.1). Thus, the described ANN is completely equivalent to a VAR model, except for the fact that the training process induces sparsity into the weight matrix $\mathbf{W}$.

The determination of the regularization parameter λ is a key element of the estimation process, as its selection strongly influences the performance of resulting regression. Following the procedure described in [164], with a hold out approach, we independently draw 90% of the samples available (rows of $\mathbf{y}$ and $\mathbf{y}^p$) as the training set and kept the remaining 10% for testing. Training and test sets were then normalized and, for each assigned λ, the number of non-zero weights was counted in the matrix $\widehat{\mathbf{W}}$ estimated on the training set, and the RSS was computed on the test set as well. This procedure was iterated for each λ, and the optimal λ was taken as the value minimizing the ratio between RSS and the number of non-zero weights [8, 55, 164].The full model (4.1) is identified with ANNs and the variance of the residuals $\lambda_{j|ijk}$ is obtained as the j^{th} element of the error covariance matrix $\mathbf{\Sigma}(j,j)$. For the restricted model, by following the procedure of LASSO regression, it is possible to define and train a new ANN which structure, without the driver process Y_i, identifies the problem 4.23. For each couple (i,j) an ANN is trained, with Learning rate equal to 10^{-3} and number of iterations for the gradient descent equal to 1000 that were also used for the full-model as discussed in the previous section. After the identification of the restricted model, the variance of the residuals $\lambda_{j|jk}\,\mathbb{E}[e^2_{j|jk,n}]$ is obtained from the residuals covariance matrix as i^{th} diagonal element.

4.2.3 GC based on State Space models

The VAR model (4.1) can be represented equivalently as an SS model which relates the observed process Y_n to an unobserved state process Z_n through the equations [102]

$$Z_{n+1} = \mathbf{A}Z_n + \mathbf{K}E_n \tag{4.28}$$

$$Y_n = \mathbf{C}Z_n + E_n, \tag{4.29}$$

where the innovations $E_n = Y_n - \mathbb{E}[Y_n|Y_n^-]$ are equivalent to the innovations U_n in (4.1) and thus have covariance matrix $\mathbf{\Phi} \equiv \mathbb{E}[E_n E_n^T] = \mathbf{\Sigma}$. This representation, typically denoted as "innovation form" SS model (ISS) [102], 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 (4.1). The advantage of this representation is that it allows to form "submodels" which exclude one or more scalar processes from the observation equation (4.28) leaving the state equation (4.29) unaltered. In particular, the submodel excluding the driver process Y_i has observation equation:

$$Y_{jk,n} = \mathbf{C}_{(jk)}Z_n + E_{jk,n}, \tag{4.30}$$

where the subscript (jk) denotes the selection of the rows with indices j and k in a matrix. As demonstrated in [119], the submodel (4.28-4.30) is not in ISS form, but can be converted into ISS by solving a Discrete Algebric Riccati equation (DARE). The prediction error variances needed for the determination of the GC measures can be computed from the identification of the full model (4.1) through OLS regression (SS-GC) [17, 102], LASSO regression by solving the problem (4.22) (SS-LASSO) [8] or by using ANN through the solution of the problem (4.27) (SS-ANNs). Given that $E_{j|ijk,n} = U_{j,n}$, the error variance of the full regression can be obtained as the j^{th} diagonal element of the error covariance matrix $\lambda_{j|ijk} = \Sigma(j,j)$. On the other hand, from the parameters of the full model $(\mathbf{A}_1, \ldots, \mathbf{A}_p, \mathbf{\Sigma})$ the partial variance $\lambda_{j|jk}$ can be retrieved by exploiting the theory of State-Space (SS) models by extracting the first diagonal element of the covariance matrix of the innovations $E_{jk,n}$.

GC-mBTS based on SS models (SS-mBTS)

For the identification of a State-Space model it is necessary retrieve the AR coefficients matrix $(\mathbf{A})$ and the corresponding residuals covariance matrix $(\mathbf{\Sigma})$ as defined in (4.1). In this case the, $M \times Mp$ matrix of coefficients is defined as matrix of zeros and it is iteratively filled with those coefficients estimated from the mBTS algorithm by solving the problem (4.7) for each couple (i,j). In this way we obtain a sparse matrix of coefficients in which the elements different from zero express the lagged variables which are useful for the prediction of the present state of the target $Y_{j,n}$. The residuals are simply obtained as product between the matrix of lagged variables $(\mathbf{Y}_{n-k})$ and the resulting sparse matrix of AR coefficients. Then the SS representation of the sub-model as reported in (4.28)-(4.29) is applied for the computation of the error variance $\lambda_{j|jk}$.

Partial Conditioning GC based on SS models (SS-PCGC)

Starting from Y_c and from the same n_d processes selected for the computation of PCGC, the idea here is to evaluate $F^C_{i\to j}$ with an SS approach. In this case the full model is described by (4.18) and the full error variance is $\lambda_{j|ijc} = \mathbb{E}[e^2_{j|ijc,n}]$. The error variance of the restricted regression is obtained considering an SS model with state and observation equations as follows:

$$Z_{ijc,n+1} = \mathbf{A}Z_{ijc,n} + \mathbf{K}E_{ijc,n} \tag{4.31}$$

$$Y_{jc,n} = \mathbf{C}_{(jc)}Z_{ijc,n} + E_{jc,n} \tag{4.32}$$

Also in this case, by solving a DARE equation is it possible to bring the model (4.31)-(4.32) into an ISS form, so that the covariance matrix of the innovations $E_{jc,n}$ includes the desired $\lambda_{j|jc}$ as the first diagonal element.

4.2.4 Testing the significance of GC values

Since the multivariate GC is a measure of the information transferred directly (i.e., not through indirect paths) from the driver to the target process, the assessment of its statistical significance is useful establish the existence of a directed link between the two network nodes generating the driver and target dynamics. In this work, the significance of DR and SS-GC, DR-mBTS, DR and SS-PCGC was tested generating surrogate driver and target series which share the same power spectrum of the original time series but are otherwise uncorrelated [120]. Specifically, 100 sets of surrogate time series were generated, the GC estimated for each pair of processes was compared with a threshold set at the 95^{th} percentile of its distribution on the surrogates, and was deemed at statistically significant if it exceeded the threshold. For what concern DR and SS-LASSO, DR and SS-ANN and SS-mBTS was assessed by exploiting the sparseness of the estimated AR matrix for the full model as already explained in [8].

4.3 Simulation experiments

This section reports two different simulation studies designed to evaluate the performances of the different methodologies for the GC estimation with the proposed extensions in the context of SS modelling. In the first simulation study we study the accuracy of GC estimators in a VAR process specifically configured to introduce redundancy between different processes [106]. In the second simulation study the performances of the ten different methodologies analyzed were tested on a four-dimensional VAR process previously used in different studies for the comparison of different techniques for inferring directed interactions [107, 201].

4.3.1 Design of simulation study I

Simulated multivariate time series (M=16) were generated as realizations of a VAR(3) fed by Gaussian noise with variance equal to 0.1. The simulated networks have a tree structure as depicted in Figure 4.1 and previously used in another study [106]. In this network non-zero AR parameters were set assigning randomly the lag in the range (1-3) and the coefficient value in the interval [0.6,

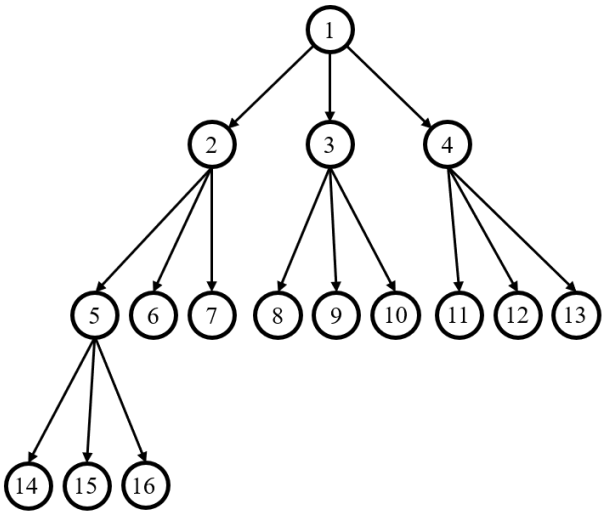

Figure 4.1: Graphical representation of the directed rooted tree of 16 nodes

0.7]. Under these constraints, 50 realization of the VAR(3) process were generated with different values of the parameter K in the range (1, 1.5, 2, 3, 5, 10, 20), so that the length of the simulated time series was $N = 48$ when K=1 and $N = 960$ when K=30. For each value of K, GC between each pair of processes was computed by all the 10 methodologies. Then, a measure of bias for the GC estimates obtained were assessed for the connections with non-zero values as explained in the following section.

4.3.2 Design of simulation study II

Simulated multivariate time series (M=4) were generated as realizations of the following VAR(5) process reported in Figure 4.2 [107]:

$$Y_{1,n} = 0.8Y_{1,n-1} + 0.65Y_{2,n-4} + U_{1,n} \tag{4.33a}$$

$$Y_{2,n} = 0.6Y_{2,n-1} + 0.6Y_{4,n-5} + U_{2,n} \tag{4.33b}$$

$$Y_{3,n} = 0.5Y_{3,n-3} - 0.6Y_{1,n-1} + 0.4Y_{2,n-4} + U_{3,n} \tag{4.33c}$$

$$Y_{4,n} = 1.2Y_{4,n-1} - 0.7Y_{4,n-2} + U_{4,n} \tag{4.33d}$$

In Equations (4.33) $\mathbf{U} = [U_1, \ldots, U_4]$ is a vector of zero-mean uncorrelated white noises with unit variance (i.e. with covariance $\mathbf{\Sigma} \equiv \mathbf{I}$). With these settings, 100 realizations of the processes were generated under different values of K parameter which was varied in the range (1, 1.5, 2, 3, 5, 10, 20) where K=1 corresponded to N=20 and K=20 to N=400. For each realization and for each value of K, GC was computed with all the methodologies introduced.

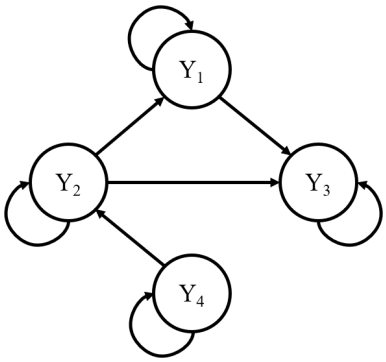

Figure 4.2: Graphical representation of the four-variate VAR (Vector Autoregressive) process according to the Equations 4.33. Networks nodes represent the four simulated processes which arrows depicting the imposed causal relationships. Self-loops depict influences from the past to the present sample of a single process.

4.3.3 Performance Evaluation

The performance of all the methodologies explored were assessed both in terms of the accuracy in estimating the strength of the GC links through a measure of bias. For each pair of network nodes represented by the processes Y_j and Y_i, the theoretical GC value ($F_{i\to j}$) was compared with the corresponding estimated value $\widehat{F}_{i\to j}$, using the Mean Absolute Percentage Error (MAPE) [59] if the theoretical link is non-null:

$$MAPE = \sum_{n=1}^{N_c} \left| \frac{F_{i\to j} - \widehat{F}_{i\to j}}{F_{i\to j}} \right| \cdot 100 \tag{4.34}$$

where N_c represents the total number of connections different from zero in the theoretical network structure. Finally, the distribution of MAPE were assessed across all the realizations and presented separately for each of the ten methodologies analyzed. The ability to detect the absence or presence of a link in a network based on the statistical significance of the GC was tested comparing the two adjacency matrices representative of the estimated and theoretical network structures. As explained in the previous Sections, this can be seen as a binary classification task where the existence (class 1) or absence (class 0) of each estimated connection is assessed using surrogate data or looking for zero/non-zero estimated links and compared with the underlying structure. Performances were assessed through the computation of false negative rate (FNR), measuring the number of non-null links classified as null; false positive rate (FPR), measuring the number of null links classified as non-nulla and Area Under Curve (AUC) which represents a trade-off between FNR and FPR [8, 48]. For each realization one single value of each parameter was computed and a distribution was then obtained for each methodologies analyzed (50 values for the simulation study I and 100 for the simulation study II). As additional performance parameter, the total computation time required for completing the two simulation studies was calculated using an implementation of the methodologies in MATLAB environment on a PC with a four cores Intel i7 (CPU clock speed 3.40 GHz), 8 GB DDR4 RAM.

4.3.4 Statistical Analysis

The methodologies were firstly divided in two different groups of five on the basis of whether or not are embedded in the SS framework. For each simulation study, were performed four different repeated measures two-way ANOVA tests, one for each performance parameters (MAPE, FNR, FPR, AUC) to evaluate the effects of different values of K (varied in the range [1, 1.5, 2, 3, 5, 10, 20]) and different methodologies for estimating GC, on performance parameters (based on state-space (SS-) models or on double regression (DR-)). In order to test the differences between sub-levels of ANOVA factors were then performed Tukey's post-hoc tests.

4.4 Results

4.4.1 Results of simulation study I

The results of the two-way repeated measures ANOVAs, expressed in terms of F-values and computed separately on all the performance parameters considering K and TYPE (the method used) as main factors, are reported in Table 4.1 for double regression strategy, and Table 4.2 for state-space representation.

The two-ways ANOVAs reveal a strong statistical influence of the main factors K and TYPE and of their interaction on all the performance parameters analyzed. It is important to note that the level K=1 was not considered for the SS approach because for the computation based on OLS (SS-GC) DARE equation did not converge.

Figure 4.3 reports the distribution of the MAPE according to the interaction between the factor K and TYPE for both strategies based on double regression (panel a) and based on SS models (panel b). The comparison between the two approach of GC computation (DR and SS) shows very different trends. It is immediately evident how the use of SS models increases drastically the accuracy in the estimation of the link strength. In fact, irrespective of identification method chosen, the use of SS approach reduces the bias in any case even when the value of K ratio is very close to 1. In all cases the performances in estimating the link strength improve with the increase of the number of data samples available (K value far from 1).

Figure 4.4 reports the distribution of the parameters FNR, FPR and AUC according to the interaction $K \times TYPE$. The analysis of the rate of false negatives (Figure 4.4 a-b) shows that the number of links classified as negatives increases with the decreasing of amount of data available (K decreasing from 30 to 1) especially for the GC computation through OLS identification (DR-GC and SS-GC). Average value of FNR for LASSO and ANNs seems to remain stable independently from the approach used (double regression or State space model). The analysis of false positives (panels

Factors	**DoF**	**MAPE**	**FNR**	**FPR**	**AUC**
TYPE	(4, 196)	368.7***	1265.4***	3860.9***	1811***
K	(6, 294)	673.4***	1195.2***	180.15***	1174.5***
TYPE $\times$ **K**	(24, 1176)	275.6***	638.41***	78.63***	260.4***

Table 4.1: 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 $TYPE$ (method used, i.e. DR-GC, DR-LASSO, DR-mBTS, DR-PCGC, DR-ANN) on the performance parameters of GC estimation (MAPE) and of network reconstruction (FNR, FPR, AUC).***, $p < 10^{-5}$.

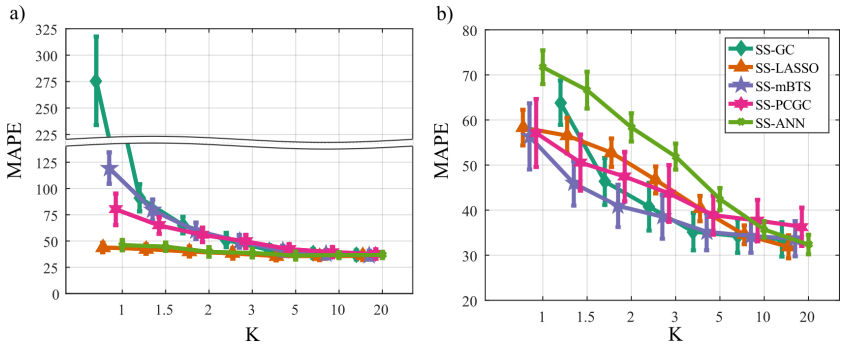

Figure 4.3: Distribution of MAPE parameter computed for the non-null links for the five methodologies based on double regression approach (panel a) and for those based on SS models (panel b). In both cases, was considered the interaction factor $K \times TYPE$ expressed as mean value and 95% confidence interval of the parameter computed across 50 realization of simulation study I.

c-d) shows very different trends with average values of FPR that remain quite stable if GC value is computed through double regression strategy (panel c) or with a state space representation of VAR model (panel d). This is not the case for ANN that show very high value of FPR, independently from K ratio, when used in combination with the double regression strategy. When K is equal to 1, while LASSO shows a reduction of 10% when SS representation is used, mBTS shows an opposite trends with FPR value around 20% with no statistical differences between LASSO and mBTS in the SS case as highlighted by post-hoc test (panel d).

The overall performances are displayed in the panels e and f with the average trends of AUC parameter. For all the cases analyzed, the use of SS representation guarantees an improvement in the detection of the true causality structure with the ANN that shows the lowest value of AUC even when the amount of data samples are not scarce (K=20 panel e) when combined with DR strategy. When SS representation is used the situation becomes opposite with ANNs that shows the highest values of AUC especially when the number samples is very scarce (K=1). The results highlights that especially when the number of data samples decreases the classical approach based on double regression and OLS identification performs a random reconstruction of the network structure ($AUC \sim 50\%$). The situation deteriorates passing to the SS representation in which it is not possible the computation of GC due to the non-convergence of DARE equation. However in a condition far from the above mentioned one, all the methodologies show comparable performance.

4.4.2 Result of simulation study II

The results of the two-way repeated measures ANOVAs, expressed in terms of F-values and computed separately on all the performance parameters considering K and TYPE (the method

Factors	**DoF**	**MAPE**	**FNR**	**FPR**	**AUC**
TYPE	(4, 196)	73.34***	333.4***	595.4***	3011.6***
K	(6, 294)	360.8***	349.6***	195.1***	3772.2***
TYPE $\times$ **K**	(24, 1176)	249.4***	292.5***	155.7***	2314.6***

Table 4.2: 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 $TYPE$ (method used, i.e. SS-GC, SS-LASSO, SS-mBTS, SS-PCGC, SS-ANN) on the performance parameters of GC estimation (MAPE) and of network reconstruction (FNR, FPR, AUC).***, $p < 10^{-5}$.

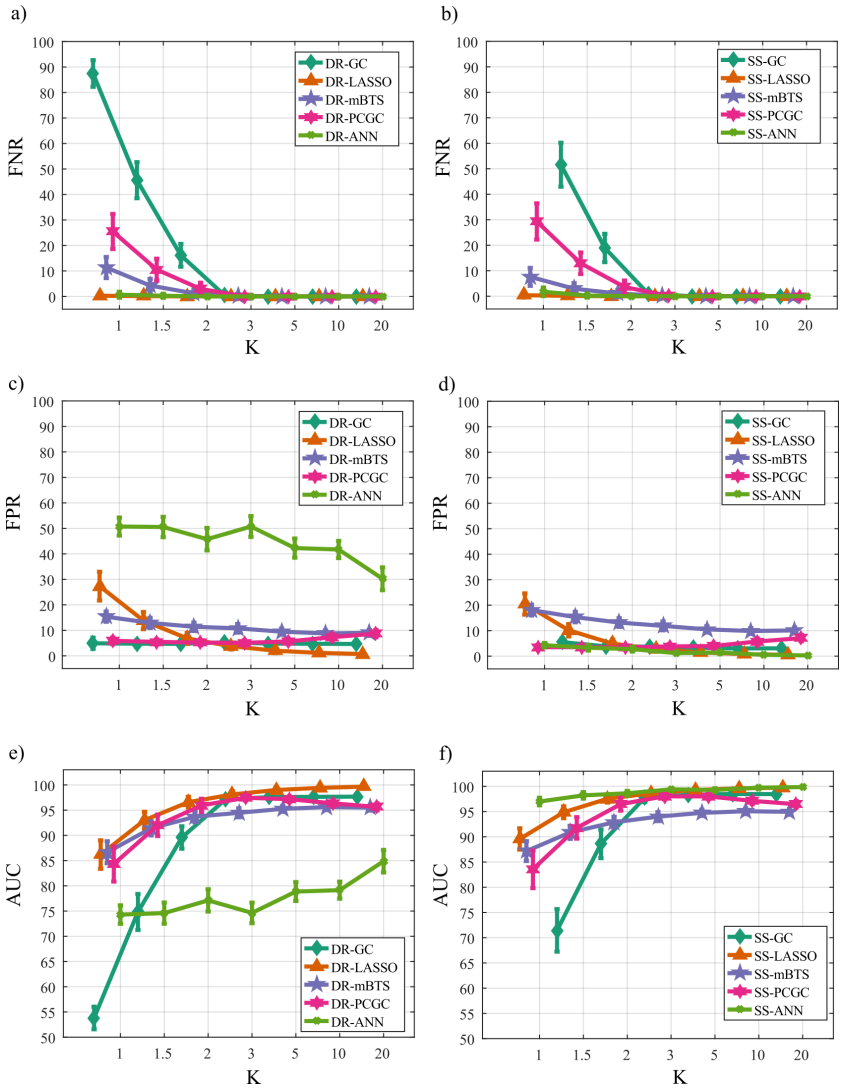

Figure 4.4: Distribution of percentage values for FNR, FPR and AUC parameters computed for the five methodologies based on double regression approach (panels a-c-e) and for the same but those based on SS models (panels b-d-f). In both cases, was considered the interaction factor $K \times TYPE$ expressed as mean value and 95% confidence interval of the parameter computed across 50 realization of simulation study I.

used) as main factors, are reported in Table 4.3 for double regression approach, and Table 4.4 for state-space representation. The two-ways ANOVAs reveal a strong statistical influence of the main factors K and TYPE and of their interaction on all the performance parameters analyzed. It is important to note that the level K=1 also in this case was not considered for the SS approach because for the computation based on OLS (SS-GC) DARE equation did non converge.

Figure 4.5 reports the distribution of the MAPE according to the interaction between the factor K and TYPE for both approaches based on double regression (panel a) and based on SS models (panel b). The analysis shows very different trends in the case of using DR or SS strategy: In the case of partial conditioning criterion (PCGC) there is a strong reduction of MAPE and the associated variance if SS is used independently from the number of data samples available. This is also true when OLS identification method is used (green trends). Also for mBTS only when K is less from 10 there is a strong reduction of MAPE when state-space strategy is used. The trends for ANNs remain similar if the identification problem is solved with a double regression (DR-ANN) or

Factors	DoF	MAPE	FNR	FPR	AUC
TYPE	(4, 396)	108.8***	629.4***	475.6***	322.6***
K	(6, 594)	216.9***	681.9***	35.8***	686.7***
TYPE × K	(24, 2376)	24.8***	119.9***	72.8***	35.7***

Table 4.3: 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 $TYPE$ (method used, i.e. DR-GC, DR-LASSO, DR-mBTS, DR-PCGC, DR-ANN) on the performance parameters of GC estimation (MAPE) and of network reconstruction (FNR, FPR, AUC).***, $p < 10^{-5}$.

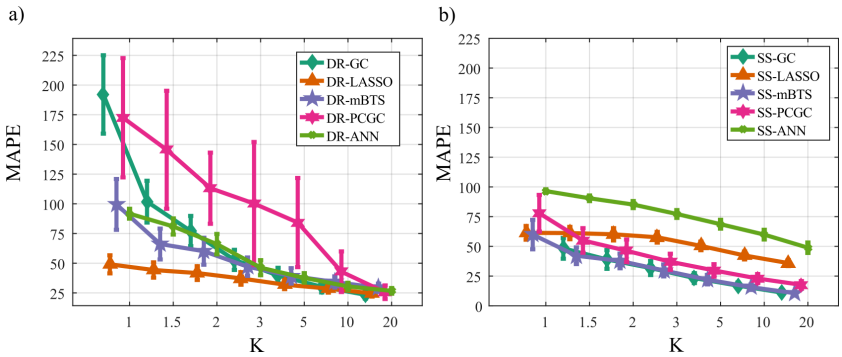

Figure 4.5: Distribution of MAPE parameter computed for the non-null links for the five methodologies based on double regression approach (panel a) and for those based on SS models (panel b). In both cases, was considered the interaction factor $K \times TYPE$ expressed as mean value and 95% confidence interval of the parameter computed across 100 realization of simulation study II.

with a single one (SS-ANN) followed by SS identification.

Figure 4.6 reports the distribution of the parameters FNR, FPR and AUC according to the interaction $K \times TYPE$. The analysis of Figure 4.6 a-b shows that the number of links classified as negatives increases with the decreasing of amount of data available (K decreasing from 30 to 1) especially for GC computation through OLS solution (DR-GC and SS-GC). The trends can be divided in two different groups: the first one which includes partial conditioning (PCGC), OLS solution (GC) and Artificial neural networks (ANN); the second one including mBTS and LASSO regression. In the latter case, the value of FNR seems to be unaffected from the strategy used (double regression or state space representation). In the first case instead, there is a reduction of $\sim 20\%$ when K is equal to 1 for ANN and PCGC passing from double regression to the approach composed of full regression and state space models. Even if less evident, in the same condition of data samples available (K=1) there is a reduction in FNR of SS-mBTS respect to DR-mBTS.

The analysis of false positives (panels c-d) shows opposite trends compared to that already described for FNR. In fact, independently from the data sample available FPR is very close to zero

Factors	DoF	MAPE	FNR	FPR	AUC
TYPE	(4, 196)	624.1***	484.5***	480.5***	246.9***
K	(6, 294)	349.1***	396.9***	132.6***	1454.2***
TYPE × K	(24, 1176)	77.4***	179.5***	93.4***	262.9***

Table 4.4: 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 $TYPE$ (method used, i.e. SS-GC, SS-LASSO, SS-mBTS, SS-PCGC, SS-ANN) on the performance parameters of GC estimation (MAPE) and of network reconstruction (FNR, FPR, AUC).***, $p < 10^{-5}$.

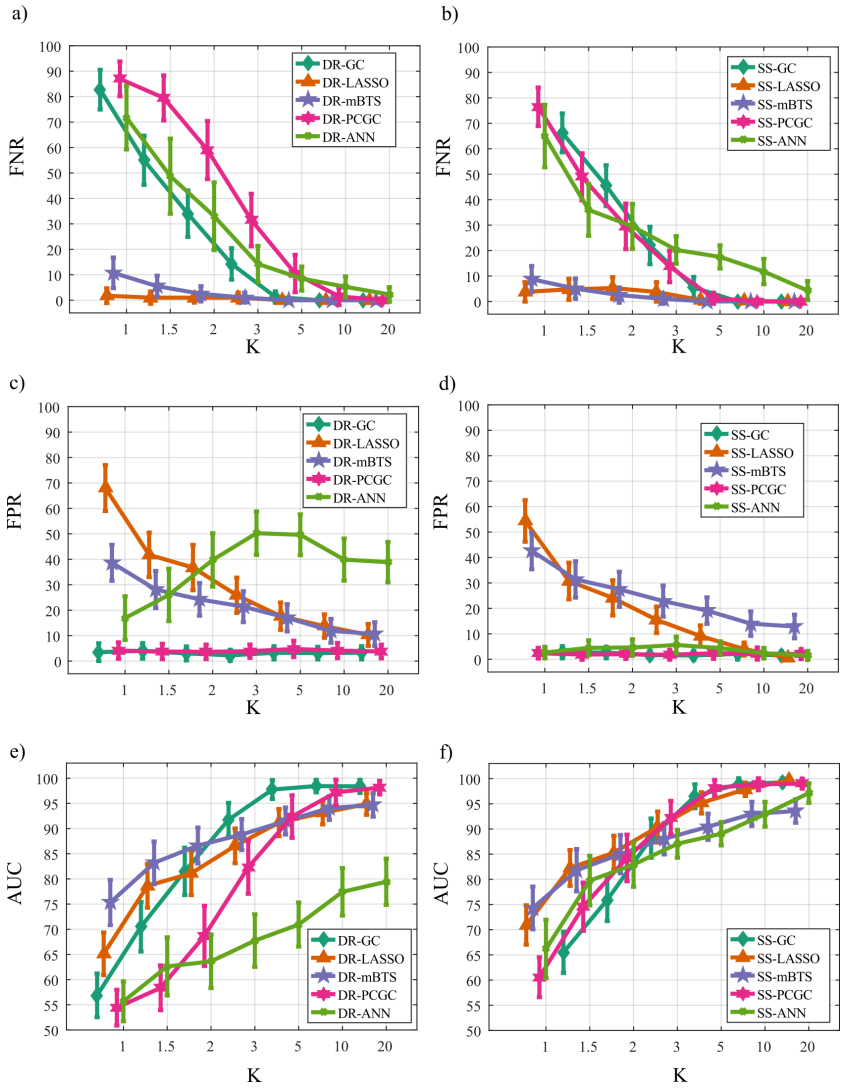

Figure 4.6: Distribution of percentage values for FNR, FPR and AUC parameters computed for the five methodologies based on double regression approach (panels a-c-e) and for the same but those based on SS models (panels b-d-f). In both cases, was considered the interaction factor $K \times TYPE$ expressed as mean value and 95% confidence interval of the parameter computed across 50 realization of simulation study II.

for GC based on OLS (DR-GC, SS-GC) and for GC based on partial conditioning (DR-PCGC, SS-PCGC). In the opposite direction there is the group formed by LASSO and mBTS which show very high values of FPR especially when K value is less than 10, with LASSO showing greater values of FPR respect to mBTS (panel c, DR-LASSO and DR-mBTS). With the use of state-space instead, the two methods do not show any statistical difference when $K \leq 3$. A particular case is that depicted by ANNs which, when used to solve double regression, shows an high unstable trend with FPR value increasing with the number of data samples. However, by introducing the SS representation, FPR assumes a stable trend very close to zero (panel d - light green line).

The overall performances are displayed in the panels e and f with the average trends of AUC parameter. For all the cases analyzed, the use of SS strategy guarantees an improvement in the detection of the true causality structure with the ANNs that if used to solve a double regression problem show very low values of AUC if compared with all the other approaches. In a condition of data paucity (K=1) mBTS shows the best performance in reconstructing the network structure

when double regression approach is exploited. On the other hand there are no statistically significant differences between LASSO and mBTS when the SS representation is introduced as highlighted in the panel f (violet and orange lines). For conditions where the number of data is large enough ($K \geq 5$), all the methods can be used and the combination with SS models guarantee the highest values of AUC parameter for OLS solution (SS-GC), LASSO regression (SS-LASSO) and partial conditioning (SS-PCGC).

Table 4.5 reports the total computation time required for the computation of the two simulation studies with the five different methodologies implemented for the solution of a double regression analysis or for a single one combined with the state-space models. The analysis of the Table 4.5 highlights a huge heterogeneity in the computation time required for performing the simulation studies. Except the methods that are based on a "training" phase (i.e those based on LASSO and ANN), using state-space modeling requires a considerable amount of additional computation time that can be three order of magnitude greater that the solution of a double regression problem. The method based on OLS is the least time consuming if combined with the double regression strategy and represents the most time consuming if combined with State space modelling. In the case of LASSO the difference between using a DR or SS strategy is not so different with computation time required that is quite stable and in the same order of magnitude. This is not the case for ANN which is faster when used in combination with SS strategy in respect to the DR.

4.5 Discussion

In this study, state-space strategy for the computation of GC with different methodologies available in the literature has been introduced. These measures have been extensively tested and their robustness was investigated on artificial data from multivariate systems with different distinguish characteristics such as density of connected node, links strength and the presence or not of the "curse of dimensionality" problem. In the first simulation study was used a dynamical system with directed tree structure in which redundant information is present [106]. In the second one, was used a smaller system highly connected, with a predefined causality structure previously used in different studies [107, 184].

From the results here obtained it is clear how the use of state-space representation of a VAR

Methods	**Computation Time**
DR – GC	231
DR – LASSO	$1.6 \cdot 10^4$
DR – mBTS	$5.3 \cdot 10^3$
DR – PCGC	480
DR – ANN	$2.6 \cdot 10^5$
SS – GC	$2.6 \cdot 10^4$
SS – LASSO	$1.9 \cdot 10^4$
SS – mBTS	$5.4 \cdot 10^3$
SS – PCGC	$9.9 \cdot 10^3$
SS – ANN	$9.2 \cdot 10^3$

Table 4.5: Total computation time (in seconds) required for each method to complete simulation studies performed

model can overcome the well known problem related with the solution of a double regression problem in the Granger causality estimation [139, 151]. In particular the results here obtained confirmed those obtained in different works [17, 102, 103, 139], for which it has been pointed out how it could be necessary to perform the identification procedure of the restricted model with a VAR ideally of infinite order. As a result, both GC estimation in time and frequency domain exhibits a strong bias effect. Our results, document that it is possible to extend the SS representation for different GC methodologies and how this has a strong positive effect on the entire estimation process independently from the method under analysis.

For what concerns the "curse of dimensionality" effects, in the literature is argued how the ratio between the number of data samples available and the regression coefficients to be estimated should be at least equal to 10 in order to overcome "curse of dimensionality" problems [9, 24, 48]. If this condition is not fulfilled, it has been experienced a severe decrease in accuracy due to the well known problem of multi-collinearity between regressors that, in the context of GC analysis, are represented by the lagged variables of the entire dynamical system. The results here obtained show how the performance in estimating Granger causality, in terms of MAPE and AUC parameters, worsen with the decrease of data samples available for the estimation process and even in another context, we can speculate about the possibility that these results could be driven by the multi-collinearity effects during the VAR model identification necessary for the GC estimation [72, 127]. In this conditions of strong data paucity, methodologies based on variable selection techniques or based on the possibility to make a sub-selection of the legged variables (LASSO regression, ANNs, mBTS and PC) seem to perform better.

LASSO regression and mBTS show the best performance in reconstructing the network structure if compared with all the methodologies tested. This is especially true in conditions of strong data paucity ($K < 3$) where, for LASSO regression multi-collinearity is reduced through variable selection and for mBTS is reduced by selection only specific lagged variables that brings the value of K to increase. A completely different issue is the error in estimating the link strength measured by MAPE parameter. In fact, for LASSO and ANN always appears to be very high and this behavior could be related with the shrinkage that controls the sparsity of the network with the side effect of a greater bias introduced in the estimation process [46, 68, 114].

A remark on the use of Artificial Neural Networks for Granger causality estimation should be made. The results of the two simulation studies highlighted the impossibility to use in an effective way this methodology if combined with a double regression strategy. However, our implementation in a combined strategy with a state space model, show outstanding performance if compared to those obtained with the classical approach based on double regression and OLS estimator. To the best of our knowledge, in the current literature there are no works which explore this aspect and for this reason it is only possible to hypothesize that this behavior could be related with a different training process experienced by the neural network when the information about the past of the driver is removed from the system.

Lastly, we analyzed the computation time required to complete all the simulation studies for each different GC estimation method combined with the strategy of double regression and with that based on state-space modelling. It is possible to see that the strategy including the double regression requires less time if compared with the counterpart based on SS models, except for the case of ANNs which requires more time in the DR strategy. This is due to the fact that in DR-ANN

case it is necessary to train $M + 1$ neural networks, one for the full model and one for each single process in the dynamical system considered. In the SS case, the least time consuming methods is mBTS which as here implemented creates as an outcome a sparse AR, like that obtained using LASSO regression, that does not requires any statistical assessment since the sparsity of estimated GC network is representative of the sparseness of estimated AR matrix. On the other hand, the most time consuming method is that based on OLS mainly because during the assessment of the estimated network through the surrogate analysis, it is necessary to identify one hundred different VAR-SS models which is a time consuming process as already pointed out in our recent works [8, 24].

4.6 Conclusion

This work documented that the combined use of any methods for estimating Granger causality, with SS strategy allows a very high accurate reconstruction of the causality network. In particular, we showed how it is possible overcoming the problems related with the "curse of dimensionality" by using methodologies based on conditioning or variables selection. The results clearly showed that an high value of accuracy of the estimation process, even in conditions of strong data paucity, has to be supported by high computational resources. Moreover, the variability resulting from the analysis of dynamical systems, with very different characteristics, suggests that could be necessary to perform the analysis with several methods with the aim to isolate the "real" interactions from those due to the method used thus leading to erroneous conclusions and misinterpretation.

Conclusion

The research activity here described was carried out with the primary purpose of introducing and testing different methodologies for the study of "information processing" in multiple interacting dynamical systems and giving new insights in the study of complex dynamical systems across multiple domains.

In Chapter 1 of this thesis work I evaluated the usefulness of penalized regression techniques for the identification of linear vector autoregressive models. The simulation study performed showed how LASSO and Elastic-Net can accomplish the identification procedure with very high accuracy especially in conditions of strong data paucity. Furthermore, the application on real EEG data evidenced consistent patterns strictly related with the motor imagery physiology, with different autoregressive parameters that could be used for a classification task in order to distinguish between rest and task conditions. These findings suggested that when classical OLS estimator cannot be used LASSO might represents the most suitable method even with limited computational power available. Since all the linear causality estimators are based on the identification of VAR models, penalized regression techniques could be useful for the computation of coupling measures both in time and frequency domains.

In Chapter 2 LASSO regression was tested for the computation of different parametric measures of information transfer in networked coupled stochastic processes. When compared with the traditional OLS estimator, LASSO showed highly accuracy both in the estimation of information transferred between coupled processes and in retrieving the topological structure of the underlying network. On the other hand, in favourable conditions of data size the results are fully overlapped with classical OLS confirming the appropriateness of embedding LASSO into the existent framework for the analysis of information dynamics. In the network physiology context, the main findings regard the detection of significant information transfer within the brain-body sub networks and reflect well-defined physiological mechanisms, such as the respiratory and heart rate effects on the pulse arrival time.

The results obtained in Chapter 3 documented that artificial neural networks ca be used in combination with state-space models for the identification of linear parametric models, allowing computationally reliable and accurate estimation of Granger causality in its conditional and unconditional forms. In particular, it has been shown how this combined approach leads to overcome the decrease in accuracy reported for traditional least squares when it needs to be performed in unfavourable conditions of data availability. Artificial neural networks have been demonstrated very useful to assess the statistical significance of Granger causality estimates, favouring the reconstruction of the network topology underlying the observed dataset without the need to employ time-consuming asymptotic or empirical procedures for significance assessment. The implementation of the proposed

approach for the study of physiological networks and coupled electronic oscillators documented its usefulness in practical applications, supported by the observation of interactions patterns similar to those found in previous studies where the datasets were first studied in terms of bivariate Granger causality. All the findings suggested that Artificial neural networks are able to detect the strongest interactions providing output patterns of information dynamics which are more straightforward and easy to interpret than those obtained with OLS.

In the last Chapter I performed an extensive comparative analysis between ten different methods for GC estimation, some of which firstly developed in this thesis work. The results of the simulations studies performed, documented that the combined use of any of the methods for Granger causality estimation with state-space strategy allows a very high accurate estimation of the link strength and provides an accurate reconstruction of the causality network. The variability of the obtained results for different coupled systems, suggested that when dealing with a Granger causality estimation problem it might be useful to compare the networks obtained with different methodologies to improve the interpretation of the results.

In conclusion, in this thesis work I tested different methodologies derived from convex optimization theory, for the identification of linear models that can be used for the analysis of the information flow among different stochastic processes of a complex system. Furthermore, I demonstrated how these new tools can be used for the dissection of the information processed in a network of multiple interacting dynamical systems. The tools here introduced have also been tested in very heterogeneous applicative contexts, such as the new field of Network Physiology, the study of brain connectivity from scalp electroencephalographic signals and, lastly, the study of the information transferred between different non-linear chaotic oscillators, as a tool to study the recently introduced concept of remote synchronization.

Next steps of this research may include the non-linear extension of penalized regression techniques and Granger causality based on artificial neural networks in order to have new tools for the study of complex non-linear dynamics in different real word systems. Furthermore, given the absence of an information decomposition tool in the hypothesis of non-stationarity of time series analysed, adaptive penalized regression techniques may be used for the study of time-varying information processing.

Bibliography

[1] T. Bossomaier, L. Barnett, M. Harré, and J. Lizier, "An introduction to transfer entropy: Information flow in complex systems. cham, switzerland: Springer international publishing; 2016."

[2] F. Hassanibesheli, N. Boers, and J. Kurths, "Reconstructing complex system dynamics from time series: a method comparison," *New Journal of Physics*, vol. 22, no. 7, p. 073053, 2020.

[3] K. McGoff, S. Mukherjee, N. Pillai, *et al.*, "Statistical inference for dynamical systems: A review," *Statistics Surveys*, vol. 9, pp. 209–252, 2015.

[4] L. Astolfi, F. Cincotti, D. Mattia, M. G. Marciani, L. A. Baccala, F. D. V. Fallani, S. Salinari, M. Ursino, M. Zavaglia, and F. Babiloni, "Assessing cortical functional connectivity by partial directed coherence: simulations and application to real data," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 9, pp. 1802–1812, 2006.

[5] L. Astolfi, F. Cincotti, D. Mattia, M. G. Marciani, L. A. Baccala, F. de Vico Fallani, S. Salinari, M. Ursino, M. Zavaglia, L. Ding, *et al.*, "Comparison of different cortical connectivity estimators for high-resolution eeg recordings," *Human brain mapping*, vol. 28, no. 2, pp. 143–157, 2007.

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

[7] L. Barnett and A. K. Seth, "The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference," *Journal of neuroscience methods*, vol. 223, pp. 50–68, 2014.

[8] Y. Antonacci, L. Astolfi, G. Nollo, and L. Faes, "Information transfer in linear multivariate processes assessed through penalized regression techniques: Validation and application to physiological networks," *Entropy*, vol. 22, no. 7, p. 732, 2020.

[9] A. Schlögl, "A comparison of multivariate autoregressive estimators," *Signal processing*, vol. 86, no. 9, pp. 2426–2429, 2006.

[10] J. Liu, S. Ji, J. Ye, *et al.*, "Slep: Sparse learning with efficient projections," *Arizona State University*, vol. 6, no. 491, p. 7, 2009.

[11] M. Billinger, C. Brunner, and G. R. Müller-Putz, "Single-trial connectivity estimation for classification of motor imagery data," *Journal of neural engineering*, vol. 10, no. 4, p. 046006, 2013.

[12] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[13] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[14] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "A framework for the local information dynamics of distributed computation in complex systems," in *Guided self-organization: inception*, pp. 115–158, Springer, 2014.

[15] M. Prokopenko, F. Boschetti, and A. J. Ryan, "An information-theoretic primer on complexity, self-organization, and emergence," *Complexity*, vol. 15, no. 1, pp. 11–28, 2009.

[16] A. Porta, V. Bari, B. De Maria, A. C. Takahashi, S. Guzzetti, R. Colombo, A. M. Catai, F. Raimondi, and L. Faes, "Quantifying net synergy/redundancy of spontaneous variability regulation via predictability and transfer entropy decomposition frameworks," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2628–2638, 2017.

[17] V. Solo, "State-space analysis of granger-geweke causality measures with application to fmri," *Neural computation*, vol. 28, no. 5, pp. 914–949, 2016.

[18] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, "Network physiology reveals relations between network topology and physiological function," *Nature communications*, vol. 3, no. 1, pp. 1–9, 2012.

[19] N. Sciaraffa, J. Liu, P. Aricò, G. D. Flumeri, B. Inguscio, G. Borghini, and F. Babiloni, "Multivariate model for cooperation: bridging social physiological compliance and hyperscanning," *Social Cognitive and Affective Neuroscience*, 2020.

[20] Y. Antonacci, J. Toppi, D. Mattia, A. Pietrabissa, and L. Astolfi, "Estimation of brain connectivity through artificial neural networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 636–639, IEEE, 2019.

[21] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 477–485, 2009.

[22] L. Minati, P. Chiesa, D. Tabarelli, L. D'Incerti, and J. Jovicich, "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*, vol. 25, no. 3, p. 033107, 2015.

[23] L. Minati, "Experimental synchronization of chaos in a large ring of mutually coupled single-transistor oscillators: Phase, amplitude, and clustering effects," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 24, no. 4, p. 043108, 2014.

[24] Y. Antonacci, L. Astolfi, and L. Faes, "Testing different methodologies for granger causality estimation: A simulation study," *EURASIP 2020*, 2020.

[25] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[26] K. J. Friston, "Functional and effective connectivity: a review," *Brain connectivity*, vol. 1, no. 1, pp. 13–36, 2011.

[27] P. L. Nunez and B. A. Cutillo, *Neocortical dynamics and human EEG rhythms*. Oxford University Press, USA, 1995.

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

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

[30] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[31] L. A. Baccalá, K. Sameshima, G. Ballester, A. Do Valle, and C. Timo-Iaria, "Studying the interaction between brain structures via directed coherence and granger causality," *Applied signal processing*, vol. 5, no. 1, p. 40, 1998.

[32] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.

[33] G. Plomp, C. Quairiaux, C. M. Michel, and L. Astolfi, "The physiological plausibility of time-varying granger-causal modeling: normalization and weighting by spectral power," *NeuroImage*, vol. 97, pp. 206–216, 2014.

[34] J. Toppi, D. Mattia, M. Risetti, R. Formisano, F. Babiloni, and L. Astolfi, "Testing the significance of connectivity networks: Comparison of different assessing procedures," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 12, pp. 2461–2473, 2016.

[35] M. H. Ahmad, R. Adnan, and N. Adnan, "A comparative study on some methods for handling multicollinearity problems," *Matematika*, vol. 22, pp. 109–119, 2006.

[36] A. Schlögl and G. Supp, "Analyzing event-related eeg data with multivariate autoregressive parameters," *Progress in brain research*, vol. 159, pp. 135–147, 2006.

[37] Y. Antonacci, J. Toppi, S. Caschera, A. Anzolin, D. Mattia, and L. Astolfi, "Estimating brain connectivity when few data points are available: Perspectives and limitations," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4351–4354, IEEE, 2017.

[38] I. Rish and G. Grabarnik, *Sparse modeling: theory, algorithms, and applications*. CRC press, 2014.

[39] S. Haufe, K.-R. Müller, G. Nolte, and N. Krämer, "Sparse causal discovery in multivariate time series," in *Causality: Objectives and Assessment*, pp. 97–106, 2010.

[40] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 66–75, 2007.

[41] P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 969–981, 2005.

[42] T. R. Mullen, C. A. Kothe, Y. M. Chi, A. Ojeda, T. Kerth, S. Makeig, T.-P. Jung, and G. Cauwenberghs, "Real-time neuroimaging and cognitive monitoring using wearable dry eeg," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 11, pp. 2553–2567, 2015.

[43] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[44] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[45] H. Zou and T. Hastie, "Regression shrinkage and selection via the elastic net, with applications to microarrays," *JR Stat Soc Ser B*, vol. 67, pp. 301–20, 2003.

[46] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[47] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[48] Y. Antonacci, J. Toppi, D. Mattia, A. Pietrabissa, and L. Astolfi, "Single-trial connectivity estimation through the least absolute shrinkage and selection operator," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6422–6425, IEEE, 2019.

[49] C. A. Porro, M. P. Francescato, V. Cettolo, M. E. Diamond, P. Baraldi, C. Zuiani, M. Bazzocchi, and P. E. Di Prampero, "Primary motor and sensory cortex activation during motor performance and motor imagery: a functional magnetic resonance imaging study," *Journal of Neuroscience*, vol. 16, no. 23, pp. 7688–7698, 1996.

[50] F. Pichiorri, F. D. V. Fallani, F. Cincotti, F. Babiloni, M. Molinari, S. Kleih, C. Neuper, A. Kübler, and D. Mattia, "Sensorimotor rhythm-based brain–computer interface training: the impact on motor cortical responsiveness," *Journal of neural engineering*, vol. 8, no. 2, p. 025020, 2011.

[51] J. Decety, D. Perani, M. Jeannerod, V. Bettinardi, B. Tadary, R. Woods, J. C. Mazziotta, and F. Fazio, "Mapping motor representations with positron emission tomography," *Nature*, vol. 371, no. 6498, pp. 600–602, 1994.

[52] D. Rathee, H. Cecotti, and G. Prasad, "Single-trial effective brain connectivity patterns enhance discriminability of mental imagery tasks," *Journal of neural engineering*, vol. 14, no. 5, p. 056005, 2017.

[53] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[54] O. P. Burdakov, C. Kanzow, and A. Schwartz, "Mathematical programs with cardinality constraints: Reformulation by complementarity-type conditions and a regularization method," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 397–425, 2016.

[55] R. J. Tibshirani, J. Taylor, *et al.*, "Degrees of freedom in lasso problems," *The Annals of Statistics*, vol. 40, no. 2, pp. 1198–1232, 2012.

[56] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[57] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[58] A. Anzolin and L. Astolfi, "Statistical causality in the eeg for the study of cognitive functions in healthy and pathological brains," *Sapienza University of Rome*, 2018.

[59] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, 2016.

[60] J. Toppi, N. Sciaraffa, Y. Antonacci, A. Anzolin, S. Caschera, M. Petti, D. Mattia, and L. Astolfi, "Measuring the agreement between brain connectivity networks," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 68–71, IEEE, 2016.

[61] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller, "A fully automated correction method of eog artifacts in eeg recordings," *Clinical neurophysiology*, vol. 118, no. 1, pp. 98–104, 2007.

[62] L. Faes, S. Erla, and G. Nollo, "Measuring connectivity in linear multivariate processes: definitions, interpretation, and practical analysis," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[63] F. Pichiorri, G. Morone, M. Petti, J. Toppi, I. Pisotta, M. Molinari, S. Paolucci, M. Inghilleri, L. Astolfi, F. Cincotti, *et al.*, "Brain–computer interface boosts motor imagery practice during stroke recovery," *Annals of neurology*, vol. 77, no. 5, pp. 851–865, 2015.

[64] S. Arlot, A. Celisse, *et al.*, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[65] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to pattern recognition: a matlab approach*. Academic Press, 2010.

[66] N. J. Guliyev and V. E. Ismailov, "A single hidden layer feedforward network with only one neuron in the hidden layer can approximate any univariate function," *Neural computation*, vol. 28, no. 7, pp. 1289–1304, 2016.

[67] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[68] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[69] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[70] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.

[71] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, 2009.

[72] M. Irfan, M. Javed, and M. A. Raza, "Comparison of shrinkage regression methods for remedy of multicollinearity problem," *Middle-East Journal of Scientific Research*, vol. 14, no. 4, pp. 570–579, 2013.

[73] D. Vidaurre, C. Bielza, and P. Larrañaga, "Classification of neural signals from sparse autoregressive features," *Neurocomputing*, vol. 111, pp. 21–26, 2013.

[74] M. Hamedi, S.-H. Salleh, and A. M. Noor, "Electroencephalographic motor imagery brain connectivity analysis for bci: a review," *Neural computation*, vol. 28, no. 6, pp. 999–1041, 2016.

[75] N.-J. Huan and R. Palaniappan, "Neural network classification of autoregressive features from electroencephalogram signals for brain–computer interface design," *Journal of neural engineering*, vol. 1, no. 3, p. 142, 2004.

[76] C. Grefkes, S. B. Eickhoff, D. A. Nowak, M. Dafotakis, and G. R. Fink, "Dynamic intra-and interhemispheric interactions during unilateral and bilateral hand movements assessed with fmri and dcm," *Neuroimage*, vol. 41, no. 4, pp. 1382–1394, 2008.

[77] L. Faes, A. Porta, and G. Nollo, "Information decomposition in bivariate systems: theory and application to cardiorespiratory dynamics," *Entropy*, vol. 17, no. 1, pp. 277–303, 2015.

[78] M. Zanetti, L. Faes, G. Nollo, M. De Cecco, R. Pernice, L. Maule, M. Pertile, and A. Fornaser, "Information dynamics of the brain, cardiovascular and respiratory network during different levels of mental stress," *Entropy*, vol. 21, no. 3, p. 275, 2019.

[79] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[80] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," *Progress in neurobiology*, vol. 77, no. 1-2, pp. 1–37, 2005.

[81] S. Schulz, F.-C. Adochiei, I.-R. Edu, R. Schroeder, H. Costin, K.-J. Bär, and A. Voss, "Cardiovascular and cardiorespiratory coupling analyses: a review," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1997, p. 20120191, 2013.

[82] L. Faes, G. Nollo, F. Jurysta, and D. Marinazzo, "Information dynamics of brain–heart physiological networks during sleep," *New Journal of Physics*, vol. 16, no. 10, p. 105005, 2014.

[83] R. P. Bartsch, K. K. Liu, A. Bashan, and P. C. Ivanov, "Network physiology: how organ systems dynamically interact," *PloS one*, vol. 10, no. 11, 2015.

[84] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Local measures of information storage in complex distributed computation," *Information Sciences*, vol. 208, pp. 39–54, 2012.

[85] L. Faes, M. Javorka, and G. Nollo, "Information-theoretic assessment of cardiovascular variability during postural and mental stress," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pp. 67–70, Springer, 2016.

[86] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[87] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Information modification and particle collisions in distributed computation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 3, p. 037109, 2010.

[88] E. Schneidman, W. Bialek, and M. J. Berry, "Synergy, redundancy, and independence in population codes," *Journal of Neuroscience*, vol. 23, no. 37, pp. 11539–11553, 2003.

[89] A. B. Barrett, "Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems," *Physical Review E*, vol. 91, no. 5, p. 052802, 2015.

[90] F. Lombardi, J. W. Wang, X. Zhang, and P. C. Ivanov, "Power-law correlations and coupling of active and quiet states underlie a class of complex systems with self-organization at criticality," in *EPJ Web of Conferences*, vol. 230, p. 00005, EPJ Web of Conferences, 2020.

[91] F. Lombardi, M. Gómez-Extremera, P. Bernaola-Galván, R. Vetrivelan, C. B. Saper, T. E. Scammell, and P. C. Ivanov, "Critical dynamics and coupling in bursts of cortical rhythms indicate non-homeostatic mechanism for sleep-stage transitions and dual role of vlpo neurons in both sleep and wake," *Journal of Neuroscience*, vol. 40, no. 1, pp. 171–190, 2020.

[92] A. Porta, V. Bari, B. De Maria, and M. Baumert, "A network physiology approach to the assessment of the link between sinoatrial and ventricular cardiac controls," *Physiological Measurement*, vol. 38, no. 7, p. 1472, 2017.

[93] J. Krohova, L. Faes, B. Czippelova, Z. Turianikova, N. Mazgutova, R. Pernice, A. Busacca, D. Marinazzo, S. Stramaglia, and M. Javorka, "Multiscale information decomposition dissects

control mechanisms of heart rate variability at rest and during physiological stress," *Entropy*, vol. 21, no. 5, p. 526, 2019.

[94] D. Widjaja, A. Montalto, E. Vlemincx, D. Marinazzo, S. Van Huffel, and L. Faes, "Cardiorespiratory information dynamics during mental arithmetic and sustained attention," *PLoS One*, vol. 10, no. 6, 2015.

[95] M. Zanetti, T. Mizumoto, L. Faes, A. Fornaser, M. De Cecco, L. Maule, M. Valente, and G. Nollo, "Multilevel assessment of mental stress via network physiology paradigm using consumer wearable devices," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2019.

[96] G. Valenza, A. Greco, C. Gentili, A. Lanata, L. Sebastiani, D. Menicucci, A. Gemignani, and E. Scilingo, "Combining electroencephalographic activity and instantaneous heart rate for assessing brain–heart dynamics during visual emotional elicitation in healthy subjects," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2067, p. 20150176, 2016.

[97] A. Greco, L. Faes, V. Catrambone, R. Barbieri, E. P. Scilingo, and G. Valenza, "Lateralization of directional brain-heart information transfer during visual emotional elicitation," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 317, no. 1, pp. R25–R38, 2019.

[98] M. Wibral, J. Lizier, S. Vögler, V. Priesemann, and R. Galuske, "Local active information storage as a tool to understand distributed neural information processing," *Frontiers in neuroinformatics*, vol. 8, p. 1, 2014.

[99] L. Barnett, J. T. Lizier, M. Harré, A. K. Seth, and T. Bossomaier, "Information flow in a kinetic ising model peaks in the disordered phase," *Physical review letters*, vol. 111, no. 17, p. 177203, 2013.

[100] T. Dimpfl and F. J. Peter, "Using transfer entropy to measure information flows between financial markets," *Studies in Nonlinear Dynamics & Econometrics*, vol. 17, no. 1, pp. 85–102, 2013.

[101] S. Stramaglia, J. M. Cortes, and D. Marinazzo, "Synergy and redundancy in the granger causal analysis of dynamical networks," *New Journal of Physics*, vol. 16, no. 10, p. 105003, 2014.

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

[103] L. Faes, D. Marinazzo, and S. Stramaglia, "Multiscale information decomposition: Exact computation for multivariate gaussian processes," *Entropy*, vol. 19, no. 8, p. 408, 2017.

[104] Y. Antonacci, J. Toppi, S. Caschera, A. Anzolin, D. Mattia, and L. Astolfi, "Estimating brain connectivity when few data points are available: Perspectives and limitations," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4351–4354, 2017.

[105] Y. Antonacci, J. Toppi, D. Mattia, A. Pietrabissa, and L. Astolfi, "Single-trial connectivity estimation through the least absolute shrinkage and selection operator," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6422–6425, 2019.

[106] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Causal information approach to partial conditioning in multivariate data sets," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[107] E. Siggiridou and D. Kugiumtzis, "Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model," *IEEE Transactions on Signal Processing*, vol. 64, no. 7, pp. 1759–1773, 2015.

[108] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[109] Y. Antonacci, J. Toppi, D. Mattia, A. Pietrabissa, and L. Astolfi, "Estimation of brain connectivity through artificial neural networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 636–639, 2019.

[110] S. Smeekes and E. Wijler, "Macroeconomic forecasting using penalized regression methods," *International journal of forecasting*, vol. 34, no. 3, pp. 408–430, 2018.

[111] R. Pernice, M. Zanetti, G. Nollo, M. De Cecco, A. Busacca, and L. Faes, "Mutual information analysis of brain-body interactions during different levels of mental stress," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6176–6179, IEEE, 2019.

[112] H. Lütkepohl, *Introduction to multiple time series analysis*. Springer Science & Business Media, 2013.

[113] H. Zou, T. Hastie, R. Tibshirani, *et al.*, "On the "degrees of freedom" of the lasso," *The Annals of Statistics*, vol. 35, no. 5, pp. 2173–2192, 2007.

[114] X. Sun, *The Lasso and its implementation for neural networks*. PhD thesis, National Library of Canada= Bibliothèque nationale du Canada, 2000.

[115] L. Faes, A. Porta, G. Nollo, and M. Javorka, "Information decomposition in multivariate systems: definitions, implementation and application to cardiovascular networks," *Entropy*, vol. 19, no. 1, p. 5, 2017.

[116] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

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

[118] L. Faes, A. Porta, G. Nollo, and M. Javorka, "Information decomposition in multivariate systems: definitions, implementation and application to cardiovascular networks," *Entropy*, vol. 19, no. 1, p. 5, 2017.

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

[120] T. Schreiber and A. Schmitz, "Improved surrogate data for nonlinearity tests," *Physical review letters*, vol. 77, no. 4, p. 635, 1996.

[121] A. Porta, L. Faes, G. Nollo, V. Bari, A. Marchi, B. De Maria, A. C. Takahashi, and A. M. Catai, "Conditional self-entropy and conditional joint transfer entropy in heart period variability during graded postural challenge," *PLoS One*, vol. 10, no. 7, 2015.

[122] J. Toppi, N. Sciaraffa, Y. Antonacci, A. Anzolin, S. Caschera, M. Petti, D. Mattia, and L. Astolfi, "Measuring the agreement between brain connectivity networks," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 68–71, 2016.

[123] A. Porta, G. D'addio, S. Guzzetti, D. Lucini, and M. Pagani, "Testing the presence of non stationarities in short heart rate variability series," in *Computers in Cardiology, 2004*, pp. 645–648, IEEE, 2004.

[124] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[125] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[126] J. T. Lizier, N. Bertschinger, J. Jost, and M. Wibral, "Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work," 2018.

[127] S. Silvey, "Multicollinearity and imprecise estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 31, no. 3, pp. 539–552, 1969.

[128] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[129] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[130] S. M. Abd Elrahman and A. Abraham, "A review of class imbalance problem," *Journal of Network and Innovative Computing*, vol. 1, no. 2013, pp. 332–340, 2013.

[131] D. Chetverikov, Z. Liao, and V. Chernozhukov, "On cross-validated lasso in high dimensions," tech. rep., Working Paper, UCLA, 2019.

[132] S. Schulz, J. Haueisen, K.-J. Bär, and A. Voss, "Multivariate assessment of the central-cardiorespiratory network structure in neuropathological disease," *Physiological measurement*, vol. 39, no. 7, p. 074004, 2018.

[133] A. Lin, K. K. Liu, R. P. Bartsch, and P. C. Ivanov, "Dynamic network interactions among distinct brain rhythms as a hallmark of physiologic state and function," *Communications Biology*, vol. 3, no. 1, pp. 1–11, 2020.

[134] N. T. Kuipers, C. L. Sauder, J. R. Carter, and C. A. Ray, "Neurovascular responses to mental stress in the supine and upright postures," *Journal of Applied Physiology*, vol. 104, no. 4, pp. 1129–1136, 2008.

[135] G. G. Berntson, J. T. Cacioppo, and K. S. Quigley, "Respiratory sinus arrhythmia: autonomic origins, physiological mechanisms, and psychophysiological implications," *Psychophysiology*, vol. 30, no. 2, pp. 183–196, 1993.

[136] C. Schäfer, M. G. Rosenblum, J. Kurths, and H.-H. Abel, "Heartbeat synchronized with ventilation," *nature*, vol. 392, no. 6673, pp. 239–240, 1998.

[137] M. J. Drinnan, J. Allen, and A. Murray, "Relation between heart rate and pulse transit time during paced respiration," *Physiological measurement*, vol. 22, no. 3, p. 425, 2001.

[138] S. I. Vrieze, "Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic).," *Psychological methods*, vol. 17, no. 2, p. 228, 2012.

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

[140] T. B. Kuo, C.-Y. Chen, Y.-C. Hsu, and C. C. Yang, "Eeg beta power and heart rate variability describe the association between cortical and autonomic arousals across sleep," *Autonomic Neuroscience*, vol. 194, pp. 32–37, 2016.

[141] Y. Kubota, W. Sato, M. Toichi, T. Murai, T. Okada, A. Hayashi, and A. Sengoku, "Frontal midline theta rhythm is correlated with cardiac autonomic activities during the performance of an attention demanding meditation procedure," *Cognitive brain research*, vol. 11, no. 2, pp. 281–287, 2001.

[142] A. Behzadnia, M. Ghoshuni, and S. Chermahini, "Eeg activities and the sustained attention performance," *Neurophysiology*, vol. 49, no. 3, pp. 226–233, 2017.

[143] A. B. Tort, S. Ponsel, J. Jessberger, Y. Yanovsky, J. Brankačk, and A. Draguhn, "Parallel detection of theta and respiration-coupled oscillations throughout the mouse brain," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[144] R. Pernice, M. Javorka, J. Krohova, B. Czippelova, Z. Turianikova, A. Busacca, L. Faes, *et al.*, "Comparison of short-term heart rate variability indexes evaluated through electrocardiographic and continuous blood pressure monitoring," *Medical & biological engineering & computing*, vol. 57, no. 6, pp. 1247–1263, 2019.

[145] A. Silvani, G. Calandra-Buonaura, R. A. Dampney, and P. Cortelli, "Brain–heart interactions: physiology and clinical implications," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2067, p. 20150181, 2016.

[146] F. Jurysta, J.-P. Lanquart, V. Sputaels, M. Dumont, P.-F. Migeotte, S. Leistedt, P. Linkowski, and P. Van De Borne, "The impact of chronic primary insomnia on the heart rate–eeg variability link," *Clinical neurophysiology*, vol. 120, no. 6, pp. 1054–1060, 2009.

[147] F. Jurysta, J.-P. Lanquart, P. Van De Borne, P.-F. Migeotte, M. Dumont, J.-P. Degaute, and P. Linkowski, "The link between cardiac autonomic activity and sleep delta power is altered in men with sleep apnea-hypopnea syndrome," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 291, no. 4, pp. R1165–R1171, 2006.

[148] N. Wiener, "The theory of prediction," *Modern mathematics for engineers*, 1956.

[149] A. Porta and L. Faes, "Wiener–granger causality in network physiology with applications to cardiovascular control and neuroscience," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 282–309, 2015.

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

[151] L. Barnett, A. B. Barrett, and A. K. Seth, "Misunderstandings regarding the application of granger causality in neuroscience," *Proceedings of the National Academy of Sciences*, p. 201714497, 2018.

[152] S. M. Kay, *Modern spectral estimation: theory and application*. Pearson Education India, 1988.

[153] K. Hornik, M. Stinchcombe, H. White, *et al.*, "Multilayer feedforward networks are universal approximators.," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[154] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[155] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[156] M. F. Pagnotta, G. Plomp, and D. Pascucci, "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)*, pp. 611–615, IEEE, 2019.

[157] N. Talebi, A. M. Nasrabadi, and I. Mohammad-Rezazadeh, "Estimation of effective connectivity using multi-layer perceptron artificial neural network," *Cognitive Neurodynamics*, vol. 12, no. 1, pp. 21–42, 2018.

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

[159] A. Attanasio and U. Triacca, "Detecting human influence on climate using neural networks based granger causality," *Theoretical and Applied Climatology*, vol. 103, no. 1-2, pp. 103–107, 2011.

[160] A. Duggento, M. Guerrisi, and N. Toschi, "Echo state network models for nonlinear granger causality," *bioRxiv*, p. 651679, 2019.

[161] L. V. Gambuzza, A. Cardillo, A. Fiasconaro, L. Fortuna, J. Gómez-Gardenes, and M. Frasca, "Analysis of remote synchronization in complex networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 23, no. 4, p. 043103, 2013.

[162] L. Minati, "Remote synchronization of amplitudes across an experimental ring of non-linear oscillators," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 12, p. 123107, 2015.

[163] L. Minati, L. Faes, M. Frasca, P. Oświecimka, and S. Drożdż, "Apparent remote synchronization of amplitudes: A demodulation and interference effect," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 6, p. 063124, 2018.

[164] K. Sun, S.-H. Huang, D. S.-H. Wong, and S.-S. Jang, "Design and application of a variable selection method for multilayer perceptron neural network with lasso," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 6, pp. 1386–1396, 2016.

[165] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American statistical association*, vol. 77, no. 378, pp. 304–313, 1982.

[166] L. Faes, D. Marinazzo, and S. Stramaglia, "Multiscale information decomposition: Exact computation for multivariate gaussian processes," *Entropy*, vol. 19, no. 8, p. 408, 2017.

[167] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[168] Y. Tran, R. Thuraisingham, N. Wijesuriya, H. Nguyen, and A. Craig, "Detecting neural changes during stress and fatigue effectively: a comparison of spectral analysis and sample entropy," in *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pp. 350–353, IEEE, 2007.

[169] L. J. Trejo, K. Knuth, R. Prado, R. Rosipal, K. Kubitz, R. Kochavi, B. Matthews, and Y. Zhang, "Eeg-based estimation of mental fatigue: convergent evidence for a three-state model," in *International Conference on Foundations of Augmented Cognition*, pp. 201–211, Springer, 2007.

[170] T. Takahashi, "Complexity of spontaneous brain activity in mental disorders," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 45, pp. 258–266, 2013.

[171] C. J. Stam, "Nonlinear dynamical analysis of eeg and meg: review of an emerging field," *Clinical neurophysiology*, vol. 116, no. 10, pp. 2266–2301, 2005.

[172] L. Minati, "Time series from ring oscillators." `http://www.lminati.it/listing/2015/a/time_series/figure_8/`, 2015. top_row.mat.

[173] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*, vol. 751. John Wiley & Sons, 2013.

[174] G. P. Zhang, "Avoiding pitfalls in neural network research," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 1, pp. 3–16, 2006.

[175] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," in *Advances in Neural Information Processing Systems*, pp. 1731–1741, 2017.

[176] Y. Li, C. Wei, and T. Ma, "Towards explaining the regularization effect of initial large learning rate in training neural networks," in *Advances in Neural Information Processing Systems*, pp. 11674–11685, 2019.

[177] Y. Antonacci, L. Astolfi, A. Busacca, R. Pernice, G. Nollo, and L. Faes, "Model-based transfer entropy analysis of brain-body interactions with penalized regression techniques," in *2020 11th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*, pp. 1–2, IEEE, 2020.

[178] S. Scardapane and D. Wang, "Randomness in neural networks: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 2, p. e1200, 2017.

[179] L. Faes, D. Marinazzo, F. Jurysta, and G. Nollo, "Linear and non-linear brain–heart and brain–brain interactions during sleep," *Physiological measurement*, vol. 36, no. 4, p. 683, 2015.

[180] M. Javorka, J. Krohova, B. Czippelova, Z. Turianikova, Z. Lazarova, K. Javorka, and L. Faes, "Basic cardiovascular variability signals: mutual directed interactions explored in the information domain," *Physiological Measurement*, vol. 38, no. 5, p. 877, 2017.

[181] M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente, "Measuring information-transfer delays," *PloS one*, vol. 8, no. 2, p. e55809, 2013.

[182] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," *Journal of Neuroscience*, vol. 35, no. 8, pp. 3293–3297, 2015.

[183] T. Scagliarini, L. Faes, D. Marinazzo, S. Stramaglia, and R. N. Mantegna, "Synergistic information transfer in the global system of financial markets," *Entropy*, vol. 22, no. 9, p. 1000, 2020.

[184] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Financial networks based on granger causality: A case study," *Physica A: Statistical Mechanics and its Applications*, vol. 482, pp. 65–73, 2017.

[185] D. Chorozoglou, D. Kugiumtzis, and E. Papadimitriou, "Testing the structure of earthquake networks from multivariate time series of successive main shocks in greece," *Physica A: Statistical Mechanics and its Applications*, vol. 499, pp. 28–39, 2018.

[186] L. Angelini, M. De Tommaso, D. Marinazzo, L. Nitti, M. Pellicoro, and S. Stramaglia, "Redundant variables and granger causality," *Physical Review E*, vol. 81, no. 3, p. 037201, 2010.

[187] M. Paluš, V. Albrecht, and I. Dvořák, "Information theoretic test for nonlinearity in time series," *Physics Letters A*, vol. 175, no. 3-4, pp. 203–209, 1993.

[188] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[189] Y. Yang and L. Wu, "Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling," *Journal of Statistical Planning and Inference*, vol. 174, pp. 52–67, 2016.

[190] I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," *Physical Review E*, vol. 82, no. 1, p. 016207, 2010.

[191] L. Faes, G. Nollo, and A. Porta, "Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique," *Physical Review E*, vol. 83, no. 5, p. 051112, 2011.

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

[193] G. Nolte, A. Ziehe, N. Krämer, F. Popescu, and K.-R. Müller, "Comparison of granger causality and phase slope index," in *Causality: Objectives and Assessment*, pp. 267–276, 2010.

[194] M.-H. Wu, R. E. Frye, and G. Zouridakis, "A comparison of multivariate causality based measures of effective connectivity," *Computers in biology and medicine*, vol. 41, no. 12, pp. 1132–1141, 2011.

[195] E. Florin, J. Gross, J. Pfeifer, G. R. Fink, and L. Timmermann, "Reliability of multivariate causality measures for neural data," *Journal of neuroscience methods*, vol. 198, no. 2, pp. 344–358, 2011.

[196] A. Fasoula, Y. Attal, and D. Schwartz, "Comparative performance evaluation of data-driven causality measures applied to brain networks," *Journal of neuroscience methods*, vol. 215, no. 2, pp. 170–189, 2013.

[197] K. J. Blinowska, "Review of the methods of determination of directed connectivity from multichannel data," *Medical & biological engineering & computing*, vol. 49, no. 5, pp. 521–529, 2011.

[198] E. Olejarczyk, L. Marzetti, V. Pizzella, and F. Zappasodi, "Comparison of connectivity analyses for resting state eeg data," *Journal of neural engineering*, vol. 14, no. 3, p. 036017, 2017.

[199] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[200] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[201] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, D. Klan, R. Bauer, J. Timmer, and H. Witte, "Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems," *Signal processing*, vol. 85, no. 11, pp. 2137–2160, 2005.

List of publications

International Journal Papers:

1. **Y. Antonacci**, L. Astolfi, G. Nollo and L. Faes: *"Information transfer in linear multivariate processes assessed through penalized regression techniques: Validation and application to physiological networks"*, Entropy 2020, 22(7), 732; https://doi.org/10.3390/e22070732.

2. **Y. Antonacci**, L. Minati, L. Faes, R. Pernice, G. Nollo, J. Toppi, A. Pietrabissa and L. Astolfi: *"Estimation of Granger causality through Artificial Neural Networks: applications to physiological systems and chaotic electronic oscillators"*, PeerJ Computer Science, 2020, (under review).

3. R. Pernice, **Y. Antonacci**, M. Zanetti, A. Busacca, D. Marinazzo, L. Faes and G. Nollo: *Multivariate Correlation Measures Reveal Structure and Strength of Brain-Body Physiological Networks at Rest and During Mental Stress*, Frontiers in Neuroscience, 2020 (in Press).

4. G. Mijatovic, **Y. Antonacci**, T. Loncar-Turukalo, L. Minati and L. Faes: *"An Information-Theoretic Framework to Measure the Dynamic Interactions between Neural Spike Trains"*, IEEE Transactions on Biomedical Engineering, 2020, (under review).

Papers in International Conferences indexed in ISI Web of Science:

1. **Y. Antonacci**, L. Astolfi and L. Faes: *"Testing different methodologies for Granger causality estimation: A simulation study"*, 28^{th} European Signal Processing Conference (EUSIPCO), 2020, https://www.eurasip.org/Proceedings/Eusipco/Eusipco2020/pdfs/0000940.pdf, (Oral presentation).

2. **Y. Antonacci**, L. Astolfi, A. Busacca, R. Pernice, G. Nollo and L. Faes: *"Model-Based Transfer Entropy Analysis of Brain-Body Interactions with Penalized regression techniques"*, 11^{th} Conference of the European Study Group on Cardiovascular Oscillations (ESGCO), 2020, 10.1109/ESGCO49734.2020.9158165, (Oral presentation).

3. **Y. Antonacci**, L. Faes and L. Astolfi: *"Information Dynamics Analysis: A new approach based on Sparse Identification of Linear Parametric Models"*, 42^{th} Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, 10.1109/EMBC44109.2020.9176114, (Oral Presentation).

4. **Y. Antonacci**, J. Toppi, D. Mattia, A. Pietrabissa and L. Astolfi: *"Estimation of brain connectivity through Artificial Neural Networks"*, 41^{th} Annual International Conference of the

IEEE Engineering in Medicine & Biology Society (EMBC), 2019, 10.1109/EMBC.2019.8856585, (Poster Presentation).

5. **Y. Antonacci**, J. Toppi, D. Mattia, A. Pietrabissa and L. Astolfi: *"Single-trial Connectivity Estimation through the Least Absolute Shrinkage and Selection Operator"*, 41^{th} Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2019, 10.1109/EMBC.2019.8857909, (Oral Presentation).

6. **Y. Antonacci**, J. Toppi, S. Caschera, A. Anzolin, D. Mattia and L. Astolfi: *"Estimating brain connectivity when few data points are available: Perspectives and limitations"*, 39^{th} Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2017, 10.1109/EMBC.2017.8037819, (Oral Presentation).

7. J. Toppi, N. Sciaraffa, **Y. Antonacci**, A. Anzolin, S. Caschera, M. Petti, D. Mattia and L. Astolfi: *"Measuring the agreement between brain connectivity networks"*, 38^{th} Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2016, 10.1109/EMBC.2016.7590642, (Oral Presentation).

Scientific Curriculum

Yuri Antonacci Ph.D. candidate
github.com/YuriAntonacci
yuri.antonacci@uniroma1.it

Education and Training

University of Palermo, Palermo, Italy — July 2020 – January 2022 (expected)
- Research Fellow (FIS/07), Department of Physics and Chemistry "Emilio Segrè"

University of Rome "La Sapienza", Rome, Italy — October 2016 – date
- Ph.D. Student in Bioengineering (ING-INF/06), Department of Computer, Control and Management Engineering "Antonio Ruberti"

IRCCS Fondazione Santa Lucia, Rome, Italy — October 2016 – date
- Visiting Student at laboratory of Neuroelectrical Imaging and BCI

University of Palermo, Palermo, Italy — September 2019 – December 2019
- Visiting Student at Department of Engineering

IRCCS Fondazione Santa Lucia, Rome, Italy — January 2017 – September 2017
- Fellowship at laboratory of Neuroelectrical Imaging and BCI

University of Rome "La Sapienza", Rome, Italy — June 2016–October 2016
- Scholarship at Department of Computer, Control and Management Engineering "Antonio Ruberti"

IRCCS Fondazione Santa Lucia, Rome, Italy — April 2015 – January 2016
- Internship at laboratory of Neuroelectrical Imaging and BCI

University of Rome "La Sapienza", Rome, Italy — A.Y. 2013/14–2015/16
- Master's degree in Biomedical engineering (INF-INF/06), 110/110 Cum Laude
- Advisor: Prof. Laura Astolfi

University of Rome "La Sapienza", Rome, Italy — A.Y. 2008/09–2012/13
- Bachelor's degree in Clinical engineering (INF-INF/06), 103/110
- Advisor: Prof. Domenico Caputo

Research Activity

Methodological context
- Development of methodologies for multivariate time series analysis in the time domain (prediction methods) and in the context of information theory for the description of the coupling between different dynamical systems.

Applications

- Electroencephalographic signals analysis for the study of information flow between different brain regions.
- Characterization of multi-organ physiological systems in different physiological states.

Participation in research groups

University of Rome "La Sapienza", Rome, Italy

- Bioengineering and Bioinformatic Laboratory (BiBiLab), Department of Computer and Management Engineering.
- Head: Prof. Laura Astolfi

IRCCS Fondazione Santa Lucia, Rome, Italy

- laboratory of Neuroelectrical Imaging and BCI
- Head: Donatella Mattia, M.D., Ph.D.

University of Palermo, Palermo, Italy

- Observatory of Complex Systems (OCS), Department of Physics and Chemistry "Emilio Segrè"
- Head: Prof. Rosario Nunzio Mantegna

University of Palermo, Palermo, Italy

- Laboratory of Optics and Optoelectronics (LOOX), Department of Engineering
- Responsible: Prof. Alessandro Busacca

Academic Activity

University of Rome "La Sapienza", Rome, Italy

- Integrative teaching activity for the basic course of mathematics, Faculty of business management (60 hours classroom teaching). A.Y. 2018/2019
- Seminars and tutoring for the "Industrial Neuroscience" course (Biomedical Engineering, ING-INF/06, 9CFU), A.Y. 2016/17, 2017/18, 2018/19, 2019/20
- Seminars and tutoring for the course: "Modeling of biological systems" (Biomedical Engineering, ING-INF/06, 9CFU), A.Y. 2016/17, 2017/18
- Seminars and tutoring for the course: "Methods for the analysis of biomedical signals" (Biomedical Engineering, ING-INF/06, 12CFU), A.Y. 2016/17, 2017/18
- Co-supervision of two internships and one master thesis on biomedical engineering (Dissertation October 2019)

University of Palermo, Palermo, Italy 21 December 2019

- *"Stima della connettività cerebrale e delle interazioni fisiologiche in bioelettronica e bioingegneria"*, seminar lecture

Awards and Achievements

- Grant for the project *Avvio alla Ricerca* (AR11916B88F7079E) titled *"Development of a new approach based on Information Theory and machine learning for the detection of physiological states in humans"* funded by University of Rome "La Sapienza", 2019.

– Grant for the project *Avvio alla Ricerca* (AR1181643695F5CD) titled *"Development of a toolbox based on Artificial Neural Network, for monitoring the effects on brain networks of a BCI-based rehabilitation treatment in stroke patients"* funded by University of Rome "La Sapienza", 2018.
– Grant for the project *Avvio alla Ricerca* (AR11715C82385545) titled *"Development of a Brain Computer Interface system based on brain functional connectivity for rehabilitation applications"* funded by University of Rome "La Sapienza", 2017.
– Student award received from Brain Computer Interface Society, 2018
– *"Awards for best master thesis on disability theme"* received from University of Rome "La Sapienza", A.Y. 2014/2015.

Participation in Research Projects

Principal Investigator for the following projects:

– *Avvio alla Ricerca* (AR11916B88F7079E) titled *"Development of a new approach based on Information Theory and machine learning for the detection of physiological states in humans"* funded by University of Rome "La Sapienza", 2019.
– *Avvio alla Ricerca* (AR1181643695F5CD) titled *"Development of a toolbox based on Artificial Neural Network, for monitoring the effects on brain networks of a BCI-based rehabilitation treatment in stroke patients"* funded by University of Rome "La Sapienza", 2018.
– *Avvio alla Ricerca* (AR11715C82385545) titled *"Development of a Brain Computer Interface system based on brain functional connectivity for rehabilitation applications"* funded by University of Rome "La Sapienza", 2017.

Participant in the following projects:

– PRIN (U-GOV PRJ-0167) titled *"Stochastic Forecasting in complex systems"* funded by the Italian Ministry of education. P.I: Prof. Rosario Nunzio Mantegna, 2020
– *Progetti di Ateneo* (RP11816436CDA44C) titled *"Sviluppo di algoritmi per l'analisi di potenziali evento-correlati in presenza di jitter"* funded by University of Rome "La Sapienza". P.I: Prof. Laura Astolfi, 2018
– *Progetti di Ateneo* (RM11715C82606455) titled *"EMBRACING: Estimating Multiple-Brain connectivity in Autism during Cooperative Interaction: anew tool for real-time hyperscanninG"* funded by University of Rome "La Sapienza". P.I: Prof. Laura Astolfi, 2017

Participation to Conferences and Scientific Meetings

– 28^{th} European Signal Processing Conference (EUSIPCO), was held virtually, 7 December 2020.
– 11^{th} conference of the European Study Group on Cardiovascular Oscillations (ESGCO), 15 July 2020.
– 42^{th} Annual International Conference of the IEEE, EMBS, was held virtually, 20-24 July 2020.
– 41^{th} Annual International Conference of the IEEE, EMBS, Berlin, Germany, 23-27 July 2020.
– 2019 OHBM Annual meeting, Rome, italy, 9-13 June 2019.
– *"XIX Workshop on convex optimization in finance"*, Rome, Italy, 15-23 January 2018.
– *"Winter School: Machine and Deep Learning for Neurological Diseases"*, Pavia, 3-7 December

2018

– *"First International Summer Institute on Network Physiology – ISINP"*, Como, Italy, 24-29 July 2017

Participation to Master classes

– *"Coding in Python and Machine Learning"*, Rome, italy, 20 hours, May-June 2019
– *"Frank-Wolfe variants for optimization over convex sets"*, Rome, A.Y. 2016/17
– *"Advances in Bioengineering"*, Rome, A.Y. 2016/17
– *"Technical Scientific writing"*, Rome, A.Y. 2016/17

Editorial activity

Reviewer for several international journals

– see Publons profile: https://publons.com/researcher/1578746/yuri-antonacci/

Reviewer editor:

– Frontiers in Social Neuroergonomics
– Frontiers in Autonomic Neuroscience
– Frontiers in Fractal and Network Physiology

Software

I release MatLab codes which implement algorithms for biomedical data analysis and signal processing developed during my research activity (see https://github.com/YuriAntonacci)

– 2020: **PID-LASSO** - Matlab Tool for the computation of Partial Information Decomposition and conditional Granger Causality based on LASSO parametric Identification
– 2020: **ANN-GC** - Matlab Tool for the computation of conditional and unconditititional Granger Causality based on the combination of state-space models and Artificial neural networks
– 2020: **S-MVAR** - Matlab Tool for the identification procedure of Multivariate Autoregressive models with different penalized regression techniques

Personal Skills and Competences

Mother Tongue: Italian

Other Languages: English

– Intensive course of four weeks in general English, Purley Languages School, London, UK (Level achieved: C1)
– general English course, Darcy School of Languages, Rome, Italy (Level achieved: B2)

Computer Skills

– Excellent knowledge of MatLab environment (daily use)
– good knowledge of LaTex environment (daily use)
– good knowledge of Python environment (occasional use)
– Competent with most of the programs of the Microsoft Office Suite

Technical skills

- Competent with EEG acquisition devices such as BrainAmp, gTech.
- good knowledge of software for the analysis of brain signals such as EEGLAb and BrainVision Analyzer
- Competent with technologies used for the fabrication of micro electric mechanic systems (mems) such as sputtering, reactive ion etching and photo-lithography

Rome, 31 January 2020 [1]

[1] I hereby authorize you to use my personal details contained in this document (D.lgs.n. 196/2003)